DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
121

KRISTJAN KORJUS

Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

121

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

121

KRISTJAN KORJUS

Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (PhD) on 13th of October, 2017 by the Council of the Institute of Computer Science, University of Tartu.

Supervisors:

Prof. PhD Raul Vicente
University of Tartu
Tartu, Estonia

Opponents:

PhD Davit Bzhalava
Karolinska Institute
Department of Laboratory Medicine

PhD Ricardo Vigario
University Nova
Dept. of Physics, Faculty of Sciences and Technology

The public defense will take place on 30th of November 2017 at 4.15 pm in J. Liivi 2-405.

The publication of this dissertation was financed by Institute of Computer Science, University of Tartu.

European Union
European Social Fund

Investing in your future

ISSN 1024-4212
ISBN 978-9949-77-598-9 (print)
ISBN 978-9949-77-599-6 (pdf)

University of Tartu Press
www.tyk.ee

To the hard-working data scientists

Contents

PUBLICATIONS INCLUDED IN THIS THESIS

PUBLICATIONS INCLUDED IN THE THESIS

I J. Aru, K. Korjus, C. Murd, and T. Bachmann. Spectral signatures of the effects of caffeine and occipitally applied transcranial magnetic stimulation in a task-free experimental setup. *Journal of Caffeine Research*, 2(1):23–30, 2012.

II K. Korjus, A. Uusberg, H. Uusberg, N. Kuldkepp, K. Kreegipuu, J. Allik, R. Vicente, and J. Aru. Personality cannot be predicted from the power of resting state EEG. *Frontiers in human neuroscience*, 9:63, 2015.

III K. Korjus, M. N. Hebart, and R. Vicente. An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLOS ONE*, 11(8):e0161788, 2016.

My contribution in these articles was as follows:

I I prepared the data, performed statistical analysis, designed graphs and participated in writing the article.

II I wrote the code, performed the machine learning analysis, designed figures, wrote methods part of the article and participated in writing other parts of the article.

III I came up with the idea, performed the whole analysis, designed graphs and figures and wrote most of the article.

PUBLICATIONS NOT INCLUDED IN THE THESIS

1. A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE*, 12(4):e0172395, 2017.
2. J. Aru, K. Korjus, and E. Saar. *Matemaatika õhtuõpik*. Hea Lugu OÜ, 2014.

ABSTRACT

Data, statistical models and the predictions made by them form the backbone of data-driven scientific inquiries. In many fields, such as neuroscience and bioinformatics, data are highly complex, with many dimensions and generated by non-linear interactions. In addition, due to their inherent complexities, modern machine learning algorithms need a lot of data and parameter tuning.

The main aim of this dissertation is to illustrate and improve the data analysis process of electroencephalogram (EEG) signal recorded from the human brain. Although the current work is based on the EEG signal, most of the lessons learned and the methods developed here are applicable for problems with complex (and expensive) data.

We begin by discussing and showing the importance of data cleaning and pre-processing together with simple hypothesis testing in the context of classical statistical models.

We then continue with the application of machine learning algorithms to real-life data. As machine learning algorithms do not explicitly perform out-of-the-box statistical inference and the choice of parameters is usually complex, the focus shifts to data partitioning methods in order to perform these tasks more efficiently.

This leads to the main thesis of the dissertation: invention of a novel data partitioning approach that uses data more efficiently than current approaches in the relevant case where parameters are in need of interpretation but model weights are not.

INTRODUCTION

A job of a scientist is to come up with coherent stories that explain and make predictions about our world. These stories themselves are quite often mathematical or statistical models. In order to make good models, we need good data. As we are describing more complex phenomena, data have become more complex: there are just more of it and there are different interactions on different time-scales which make the task of meaningful scientific inquiry challenging. On the other hand, modern pattern matching algorithms, which could be summed up with a term machine learning, can reach to astronomical complexities with hundreds of millions of model weights and even higher number of computations needed to evaluate the outcome. This creates many new challenges and opportunities.

Current dissertation is a journey from data pre-processing to the interpretation of the results with an emphasis on the way we choose the model parameters and test the validity of the model. Although the current work focuses on the neural datasets, mainly the electroencephalogram (EEG) signal recorded from the scalp of human subjects, the results apply to most of the other fields with multidimensional data that are difficult or expensive to collect.

After an introduction to the required biological, technical and statistical background, three main parts of that journey are summed up in three chapters that are based on the three publications.

In Chapter 2 which is based on the publication I, we start from the basics. We understood that the incoming data is too noisy and the raw format of the data is difficult to analyze. So, we needed to clean and normalize the data, as well as to build appropriate features by decomposing the EEG time series into the frequency domain such that some properties of the signal would become more visible. In addition, we performed a simple hypothesis testing with a classical statistical inference tool called analysis of variance (ANOVA).

This study gave us interesting insights to the problems posed by neuroscientists but it also led us into troubled thinking. It might be the case that the analysis might not be accurate enough as our complex real life data does not fully compile with all the assumptions required by the data model. What is more, it is even possible that intermediate results and visualizations might have influenced our design of the final data analysis pipeline which in turn might have resulted in a circular or biased analysis. For example, recent publications are claiming that high number of studies are p-value hacked (Head et al., 2015) or they contain some circular analysis (Kriegeskorte et al., 2009).

In addition, the complex data analysis pipeline is rather specific and is probably not usable by other studies. These problems lead us to the next chapter and study.

In Chapter 3 which is based on the publication II, we have analyzed another relatively large EEG data set with machine learning algorithms. Although machine learning algorithms are very powerful, there are two major complexities related to their usage. As machine learning algorithms do not tend to explicitly work with data models (i.e. assumptions made to the data by classical statistical models), it is not possible to perform out-of-the-box statistical inference with them. Therefore, as the model and its implicit preferences are very difficult to understand, separate data is needed for choosing the best model and its parameters. In order to overcome these complications, we build an extensive system to scan a vast amount of parameters and also estimate the generalization error with an approach called nested cross-validation.

The second study left us more satisfied as we performed an extensive search over possible relationships in the data; we were (hopefully) less biased by the data and produced a bit more general pipeline which is already used by other studies. On the other side, the chosen method of nested cross-validation did not leave us any possibility to interpret the model parameters nor the model weights. As the data were scarce and signal-to-noise ratio low, the chosen pipeline was probably correct, but it left us to wonder: could we perform an efficient data partitioning such that we could still interpret the model parameters?

In Chapter 4 which is based on the publication III, we describe a novel data partitioning approach for machine learning methods. The novel approach is termed "Cross-validation and cross-testing" and it is more efficient than the current best approach for data partitioning, in the case where the interpretation of model parameters in needed but the interpretation of model weights is not needed. The possibility to interpret parameters while being more data efficient is particularly relevant in the fields where data are scarce or costly such as bioinformatics or neuroscience. In the mentioned chapter, we introduce the approach in detail, compare it to the other approaches and discuss its implications in terms of efficiency and interpretability.

Although the emphasis of this dissertation is on methods, in the process we have also shed light to some neuroscience problems. In the publication I, we are showing how caffeine and TMS impulses interact resulting in specific changes in the power spectra of EEG signals. In the publication II, we are showing with a very large sample size that it might not be possible to predict personality scores from a resting state EEG, indicating that personality traits might be more readily revealed by the brain dynamics to specific stimuli rather than by resting state dynamics. Finally, the numerical experiments that compare different data partitioning approaches for machine learning in the publication III, are carried out on different datasets from neuroscience experiments.

To give background and context to these three chapters, we introduce some common preliminaries including the main data sources analyzed (EEG signals), and a summary of machine learning algorithms, as well as common data partitioning approaches in Chapter 1. At the end of the dissertation, in the summary chapter, we also discuss some of the implications stemming from the results obtained during this dissertation. Enjoy the journey!

CHAPTER 1

PRELIMINARIES

1.1 Biological background and grand goals of science

All the PhD dissertations in the world are imagined, written, read and understood by human brains. Brain is the organ that makes our conscious existence possible. Naturally we would like to understand all the important questions about it: How to read minds? How to control machines with just your thoughts? How to cure brain diseases? What to tune in our brains in order to make us many times more intelligent? How to build artificial brains? How to transcend to immortal beings?

A good way to get closer to all the important questions is science. Although we suggest to read a book called "A History of the Brain: From Stone Age surgery to modern neuroscience" by Andrew Wickens to get a nice historical overview of the brain science (Wickens, 2014), we still mention two important milestones which are particularly important for this dissertation.

Firstly, modern brain science started with the discovery and understanding of neurons as independent and discrete building blocks of the brain. In 1906, Camillo Golgi (Golgi, 1885) and Santiago Ramon y Cajal (Cajal, 1894) shared the Nobel Price in Medicine or Physiology for their discoveries on the strucure and anatomy of the neural tissue (Golgi, 1906). In Fig. 1.1, we can see Cajal's original drawings of some neurons.

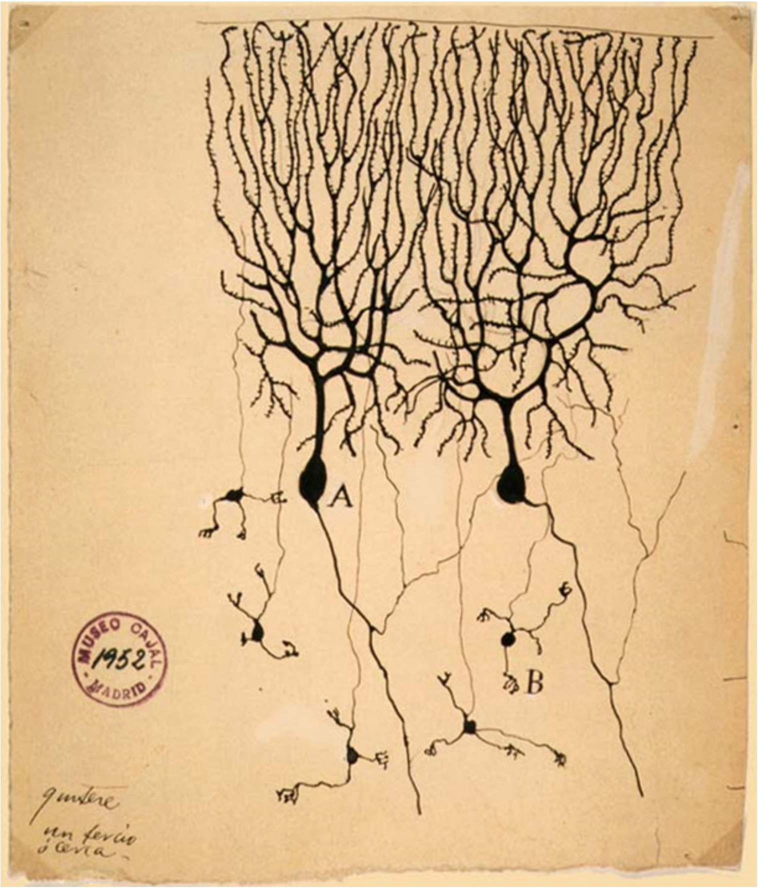

Figure 1.1: **Drawing of Purkinje cells (A) and granule cells (B)** from pigeon cerebellum by Santiago Ramón y Cajal, 1899; Instituto Cajal, Madrid, Spain

Secondly, the invention of electroencephalography (EEG), a method to record electrical activity of the brain, accelerated the study of brain even further. The first documented recording of a human EEG was performed by a German physiologist and psychiatrist Hans Berger (1873–1941) who was curious about the possibility of spontaneous telepathy (Haas, 2003). The invention of EEG has been described "as one of the most surprising, remarkable, and momentous developments in the history of clinical neurology." (Millet, 2002). In Fig. 1.2, we can see the original EEG measurements made on a human.

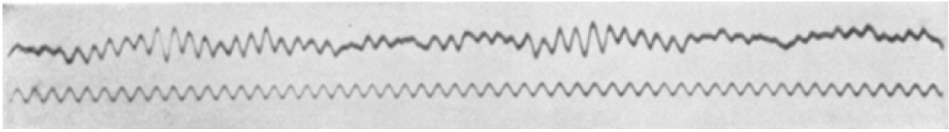

Figure 1.2: **First published Electroencephalogram of a human** (Berger, 1969). The upper trace is the real EEG signal which is the main type of signals analyzed in this dissertation. The lower trace is a sinusoidal signal drawn for comparison.

The upper trace in Fig. 1.2 is the real EEG signal. Lower trace is a periodic signal to show that the upper trace (real EEG) contains some rather rhythmic activity.

Interestingly for this dissertation, the neuro-inspired machine learning community and neuroscience community have produced a fruitful exchange which recently has fueled major improvements in both algorithmics, and understanding of information processing by the brain (Cox and Dean, 2014). As an example from the neuroscience side, remarkable activational and structural similarities between the artificial deep neural networks and early visual areas of the brains have been described (Khaligh-Razavi and Kriegeskorte, 2014), providing a common and rich framework to describe how vision occurs in artificial and biological systems.

Solving grand neuroscience questions with machine learning algorithms that have deep connections to neuroscience itself is an interesting and promising endeavor. However, as the data in this field are usually noisy, non-stationary and high-dimensional, in addition to being scarce and expensive, the deployment of machine learning on this type of data becomes a true challenge.

1.2 Technological background and challenges

To successfully understand any system in nature we need some empirical observations or a posteriori knowledge as philosophers put it. There are many ways how to collect data about the brain depending on the questions one is interested: one can cut it open and check the anatomy and structure of the brain; one can make psychological experiments or models of behavior; one can model the brain at different levels from synapses to networks of neurons. In the current dissertation, we are focusing on the methods which are recording the brain while its doing its usual job by measuring the voltage, magnetic field or change in oxygen in the blood induced by neuronal activity. Mainly we are focusing on electrical signal measured by EEG but most of the work is relevant for any type of multivariate data.

1.2.1 Electroencephalogram

Neuroimaging is a relatively new field which uses different techniques to study and measure the activity and structure of the brain. In this dissertation electroencephalogram (EEG) data are analysed (Nunez and Srinivasan, 2007).

Electroencephalography measures voltage fluctuations in the scale of μV's resulting from ionic current flows across neurons of the brain. EEG typically consist of the recording of electrical activity along the scalp with a few tens of channels being measured simultaneously.

EEG measures electricity. There are many advantages in using EEG such as a relatively low cost of equipment, very good temporal resolution, noninvasiveness etc. There are many problems associated with these measurements as well such as low spatial resolution and a tiny signal-to-noise ratio with lots of artefacts, including eye blinking, jaw movements, or sensor drifting to mention a few. For example, in the publication I and II, we needed to remove the muscle artefacts as it was a major component that contaminated the signal of true interest for our purposes (those of neuronal origin). In the first case we did it manually by going through the whole data and in the second case semi-automatically by training a system to detect the artifacts automatically.

As the signal-to-noise ratio of EEG is very low it usually takes lots of data to train more expressive models. However, data collection is usually expensive in both time and money. In this case, efficient data partitioning to extract the maximal information from the recorded data becomes an important part of the analysis pipeline to which this dissertation aims to contribute. Also, extracting relevant features or representing the data in an appropriate manner for the problem at hand is another important factor of any machine learning approach.

Time–frequency representation

Transforming data into a useful form is a common task by a data scientist. An interesting property of an EEG signal is its periodic components. Similarly to music, using Fourier analysis the analyst can decompose any signal into mixture of periodic sine waves with different frequencies (see Fig. 1.3). Importantly, for EEG recordings different frequency bands have been associated to different types of activity in the brain generated by specific mechanisms and regions.

Compromising between frequency and time resolution one can also consider shorter time-windows to proceed with the Fourier analysis. This makes it possible to visualize the power of periodic components at different moments. This way of looking at the data is called time–frequency representation (see Fig. 2.1) and it has become a very common tool for analyzing EEG signal providing an important and interpretable set of features to characterize the data (Roach and Mathalon, 2008).

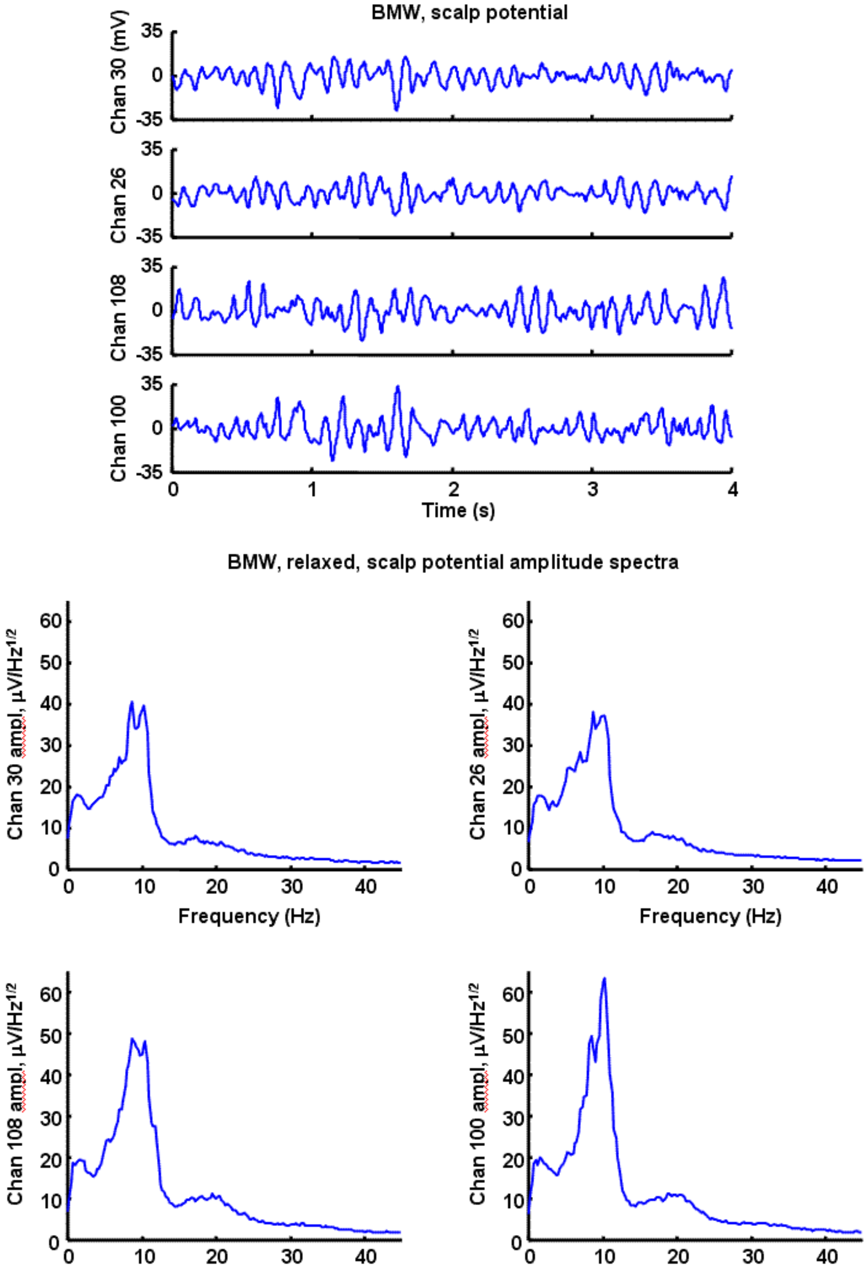

Figure 1.3: **(Upper) Alpha rhythm recorded from a healthy relaxed subject (age 25) with closed eyes.** The y-axes label *Chan* is a short form of a *channel*. This EEG was recorded at the Brain Sciences Institute in Melbourne. (Lower) The corresponding amplitude spectra based on the full five minute record reveals dominant activity in the alpha (8-13 Hz) band. (Nunez and Srinivasan, 2007)

1.3 Statistical background

1.3.1 Statistical inference

Although in this dissertation the main focus is in machine learning and in data partitioning, the whole EEG data analysis is also briefly discussed, including data cleaning, pre-processing and classical statistical inference with analysis of variance that was the topic of the publication I.

In the process of statistical inference, the analyst aims to understand properties about a population (Upton and Cook, 2014). For example, one might assume a data model, take a statistical model, fit it to the observed data and perform an hypothesis testing to reject our alternative hypothesis.

In statistical inference, data model and assumptions about the statistical model have a crucial importance. Importantly for this dissertation, if the model does not accurately reflect the nature in the best form, the conclusions maybe wrong (Breiman et al., 2001).

Next, a description of a statistical model is given. Although it is a specific method, best suited for the publication I, the underlying higher level ideas are similar for most data models and statistical inference methods.

Analysis of variance (ANOVA)

Analysis of variance (ANOVA) denotes a wide collection of statistical models designed to study the differences among group means (Girden, 1992). In particular, the simplest ANOVA method considers the ratio of the variance between groups against the variance within groups, taking into account the degrees of freedoms (related to the number of data points). This ratio produces a measure, called f-test statistic, which is used to perform a statistical significance test, called f-test.

More precisely, the procedure is as follows. Each data point can be written out as a group mean plus a deviation from it. In the equation 1.1, x denotes the data point, μ is the group mean and ϵ the deviation from the mean. Subscripts i and j denote respectively the group and the data point in the group:

$$x_{ij} = \mu_i + \epsilon_{ij}. \tag{1.1}$$

The between-group variation, denoted by SS_0, is calculated from the between-group sums of squares (see equation 1.2) where n_i's denote the group sample sizes, n total number of groups and μ the whole population mean:

$$SS_0 = \sum_{i=1}^{n} \frac{n_i(\mu - \mu_i)^2}{n_i - 1}. \tag{1.2}$$

The within-group variation, denoted by SS_1, is calculated from the within-group sums of squares (see equation 1.3). For each group, such variance is multiplied by the degrees of freedom $(n_i - 1)$:

$$SS_1 = \sum_{i=1}^{n} \frac{\sum_{j=0}^{n_i} (\mu_i - x_{ij})^2}{n_i} (n_i - 1). \tag{1.3}$$

F-ratio is then calculated as $\frac{SS_0}{SS_1}$ which is compared against an F-distribution with associated degrees of freedom to get a significance p-value and can be compared with a certain significance level α to name the difference in the group means as significant or not.

The assumptions underlying this ANOVA test are that the data is normally distributed, the variance is similar within different groups, and that the data points are independent.

In our publication I, the design of the experiment did not allow us to assume that data points are independent as we were doing many repeated measurements from the same subject which can produce some time dependent effects. Hence, we used modified scheme of repeated measures analysis of variance (rANOVA), which is a widely used tool (Gueorguieva and Krystal, 2004).

The main idea for repeated measure ANOVA is to take out the subject variability from the variability produced by the treatment and by the error terms. Such a decomposition of error terms makes it possible to still use ANOVA methods effectively even when data points are not independent.

1.3.2 Machine learning

Machine learning is a branch in the field of artificial intelligence that deals with algorithms that learn from the data. The machine learning itself consists of different type of algorithms starting from simple logistic regression and ending with deep neural network with hundreds of internal layers that learn hierarchical representations of the data (Bengio and Courville, 2016). Fig. 1.4 gives an overview of the different approaches in machine learning methods.

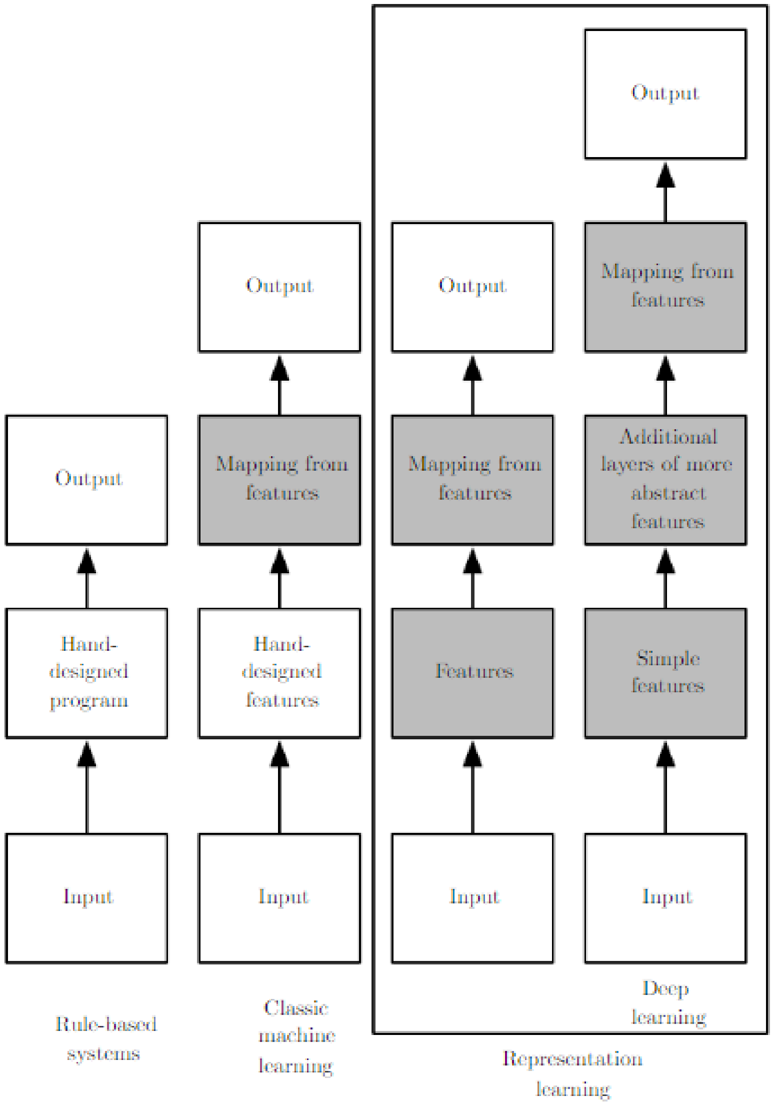

Figure 1.4: **Chart is showing different parts of AI systems.** Shaded boxes indicate components that are able to learn from data. Image taken from a book by Bengio and Courville (2016).

It can be said that machine learning consists of various algorithm components such as a optimization algorithm, a cost function, a model, and a data set (Bengio and Courville, 2016). Combining them gives a machine learning algorithm.

Another way to categories machine learning algorithms is by their output:

classification, regression, clustering, density estimation and dimensionality reduction. In this dissertation, machine learning algorithms are used for classification and in particular for binary classification. More formally, in the context of binary classification, the aim of a training phase is to learn a model f such that it would correctly classify an unseen data point x to a label $y \in \{A, B\}$: $y = f(x)$.

Importantly for this dissertation, machine learning can learn very complex representations of the data and indeed is more and more used in natural science fields with complex and noisy data such as neuroscience (Haynes and Rees, 2006) and bioinformatics (Larrañaga et al., 2006). However, many complex and non-linear machine learning algorithms need lots of data which in turn means that data efficiency is important while designing the analysis of a study.

Support vector machine

Support vector machine (SVM) is a supervised machine learning method to classify input vector into two classes (Vapnik and Lerner, 1963; Boser et al., 1992; Cortes and Vapnik, 1995; Chang and Lin, 2011). If the annotation groups are coded with $+1$ and -1 the prediction for a new data vector $x = (x_1, \ldots, x_p)$ is given by

$$\hat{y} = \mathrm{sign}(\beta_0 + \sum_{j=1}^{p} \beta_j x_j). \tag{1.4}$$

As seen in the formula 1.4, the original SVM performs a linear separation of the feature space. Naturally, the question is how to find the optimal hyperplane that separates the data. Linear SVM solves this optimization problem by finding the best linear separation between two groups by fitting a hyperplane which maximizes the gap between two groups (see Fig. 1.5).

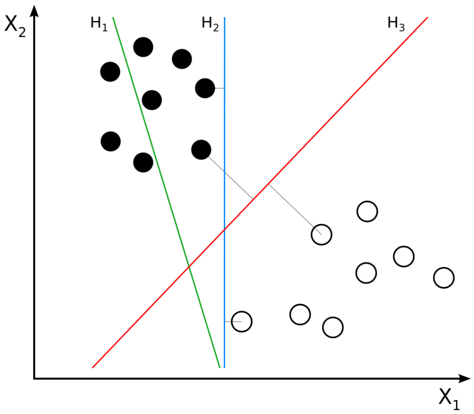

Figure 1.5: **Hyperplanes.** H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin. The goal of simple support vector machines is to find the hyperplane that maximizes certain margins with the two classes of data. (Wikipedia, 2016).

There are several extensions for more than two groups of data and also non-linear versions exist. In the publications II and III, we use linear SVMs and also non-linear SVM with a radial basis function kernel. The kernel transforms the data to a new feature space and a linear separation is performed in there.

Comparison of modern machine learning to classical statistical inference

A provocative article "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)" by L Breiman gives an extreme view of the differences between modern machine learning and classical statistical inference (Breiman et al., 2001). On the other hand, it could be said that both of the fields are using data and algorithmics to come up with explanations about our world and the difference is mainly due to terminology. I will try to give my personal balanced opinion of this issue.

In practice it could be said that there is a continuum of properties from classical statistical methods to machine learning algorithms. One of this properties is the number of model weights. On the one extreme, we have a linear model without an intercept $y = ax$, where a is the only model weight and in the other extreme we have very deep neural networks with hundreds of millions of model weight (Simonyan and Zisserman, 2014; LeCun et al., 2015). It can be said that machine learning methods usually have more model weights than classical methods.

Another continuum for any kind of analysis is the amount of pre-processing needed. For example, before deep neural network, speech to text modeling usually involved many small steps such as recognizing phonemes, words and then sentences. (Reddy, 1976; Benzeghiba et al., 2007; Anusuya and Katti, 2010). But modern methods are trained end-to-end without any feature engineering or intermediate steps (Graves et al., 2013; Hannun et al., 2014). Similar ideas were visualized in Fig. 1.4.

In addition, usually the simplicity of models makes it possible to have a stronger theoretical background about them which makes it possible to assume a data model and perform statistical inference with them without the need of a separate test set. More complex machine learning models lack this property and they do need a separate set for testing the generalization error.

To sum up, more complex models have their clear benefits but on the other hand they need more data and a separate test set. This is one of the main contributions of this dissertation: how to build an efficient data partitioning scheme such that we could still estimate the generalization error of the model and perform hypothesis testing while being able to interpret the parameters (Korjus et al., 2016).

1.3.3 Data partitioning

The goal of supervised machine learning, in particular classification, is to find a model that assigns data to separate classes as accurately as possible. To test the generality of a learned model, this model is typically applied to independent test data, and the accuracy can be reported as a measure of the quality of the classifier (Alpaydin, 2014).

Machine learning model parameters are complex and the number of pre-processing options vast. To test the different possibilities, it is quite common to divide a data set into three parts: (1) a training set and (2) a validation set and (3) a test set. Training and validation set are used many times with all the combinations of parameters to find the best combination. The test set is used to validate the generalization performance of the final classifier.

Data collection can be very expensive in biological and social sciences, and over time more data-efficient data partitioning methods have emerged (Pereira et al., 2009). The most common method is to use the training data set (1) and validation set (2) repeatedly. This method is called cross-validation.

Cross-validation

Cross-validation is a method that makes near-optimal use of the available data by repeatedly training and validating classifiers on different subsets of data, typically with a large training and a small validation set in each iteration (Molinaro et al., 2005). For example, in 10-fold cross-validation 90 percent of the data are used for training, 10 percent for validation, and in the next iteration another 10 percent of data are chosen as a validation set (see Fig. 1.6). This process is repeated ten times until all data have served as validation data once. Cross-validation is repeated with different parameter combinations, and once the best parameters have been found, the model is trained with the chosen parameters on all data that have previously been used for cross-validation and applied to the separate test set (see Fig. 4.2).

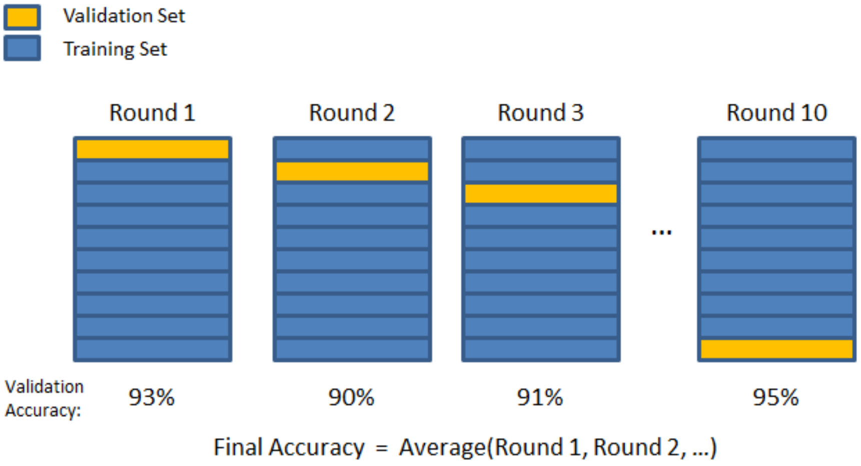

Figure 1.6: **Visualization of cross-validation.** At each round, a part of the data is left for testing (yellow) and the rest for training the model (blue). For the final estimate of the accuracy, an average of all the rounds is taken. The figure is from an article by Akyildiz (2014).

In addition to the above mentioned k-fold cross-validation, a leave-one-out cross-validation is sometimes used where the k parameter equals the number of data points. Leave-one-out cross-validation is also called exhaustive cross-validation as all the possible partitions are used.

In stratified cross-validation, the folds are generated such that they contain approximately the same proportions of labels as the original dataset. In this dissertation, all cross-validation variants are stratified.

One of the main assumptions of cross-validation states that there must be no information leakage from training sets to validation sets. With non-stationary time series such as EEG data, it can happen that data points that are close in time might introduce this information leakage.

Although cross-validation can also be seen as a method for estimating generalization error of a model (Kohavi, 1995), in the context of the thesis of this dissertation, cross-validation is used to compare different models in order to choose the best one. This maximum operator (choosing the best model parameters) makes the estimate of the generalization error biased and with a large number of different models a separate test set is needed.

In addition to the use of cross-validation with a test set, it is also used in the form of nested cross-validation and in the novel method developed in this dissertation termed Cross-validation and testing (see Chapter 4 for full details).

1.3.4 Terminology

Terminology has been a considerable reason for confusing our discussions with other people and each other. The term *parameter* has been used inconsistently in the literature, sometimes referring to the individual weights of a given model, and sometimes to the parameters that are used to optimize the learning algorithm.

Here, we use the term *parameter* to refer to a variable that is used to optimize classification performance (which may incorporate choices not directly applied to a particular classifier, including the choice of pre-processing or the choice of classifier). For parameters related to the model itself, we use the term *model weights*. The term *hyper-parameter* is used for all the choices that must be made before applying any of the approaches to the data such as choosing the size of the test set, for example.

CHAPTER 2

PREPARATION OF DATA AND STATISTICAL INFERENCE (PUBLICATION I)

This chapter is about data pre-processing and simple statistical inference, using well-known statistical methods that should be relatively easy to use and data efficient. One aim of the dissertation is to contradict this knowledge and show numerous problems with simple statistical analysis. For example, usually people believe that most of the time in data analysis is spent on preparation of the data (Zhang et al., 2003). It is also the case that usually data does not fully comply with the assumptions needed by data analysis methods.

As with two next chapters, we also introduce the neuroscientific motivation behind the study and the results. For more details, please see the publication I.

2.1 Neuroscience question

It is a well known fact that the state of the brain considerably influences information processing in the conscious and unconscious state (Barry et al., 2005; He and Raichle, 2009; Hohwy, 2009; Lee et al., 2009; Lorist and Tops, 2003; Massimini et al., 2005). And indeed, caffeine is known to affect perception, attention and psychomotor performance (Lorist and Tops, 2003; Rees et al., 1999; Ruijter et al., 2000). The general interest in this study is to understand the effect of caffeine on the human brain.

However, state dependent neural dynamics are difficult to study with typical psychophysiological methods as the tasks and task-related stimuli interact with physiological states in a manner difficult to control. One can overcome this problem by introducing the manipulation of states in a task-free manner with artificial perturbations of neural dynamics. Experiments using transcranial magnetic stimulation (TMS) simultaneously with EEG recording and in combination with controlled general brain states serve as an example of this strategy (Massimini et al., 2005; Ferrarelli et al., 2010; Massimini et al., 2010; Murd et al., 2010; Stamm et al., 2011).

To sum up, we give a subject either a placebo or caffeine bill, give a magnetic shock to the brain and record the EEG signal. Then, the EEG is represented in the time-frequency domain (see Section 1.3.1) and the effect of caffeine to the human brain state is analyzed.

2.2 Data pre-processing and statistical inference

Data were filtered and segmented around the time of the TMS stimulation. Afterwards, the channels with muscle or other artifacts were disregarded and the rest were grouped into frontal and parietal regions. In total, we had around 230 trials (mean: 233, standard deviation: 6.5) from each 8 subjects in each condition, which allows us to have the required sensitivity in the spectral analysis.

After the pre-processing, event-related spectral perturbations (ERSPs) were computed for each subject in each condition and for each electrode. This requires the calculation of the power spectrum over a sliding latency window and then averaging this power spectrum across the trials.

After considerable amount of effort, the data were ready for statistical inference analysis.

Repeated measures ANOVA (see Section 1.3.1) was run with the factors condition (caffeine and placebo), region of interest (frontal, parietal) and frequency (alpha, beta, low gamma, high gamma). The statistical significance level for hypothesis testing was fixed to $\alpha = 0.05$.

2.3 Answer to the neuroscience question

In line with previous observations that caffeine increases alertness (Clubley et al., 1979), caffeine consumption increased pre-TMS baseline gamma band power compared to placebo control in both low gamma (30-50 Hz) and high gamma (50-80 Hz) bands. Surprisingly, TMS led to decreased relative power of the post-stimulation low gamma band activity under caffeine as compared to TMS in placebo condition (see Fig. 2.1). In addition, caffeine administration was associated with a significant reduction of TMS evoked alpha power about 400 ms after TMS (see again Fig. 2.1).

We concluded that TMS related spectral perturbations are brain-state dependent and lead to different spectral signatures under different physiological conditions.

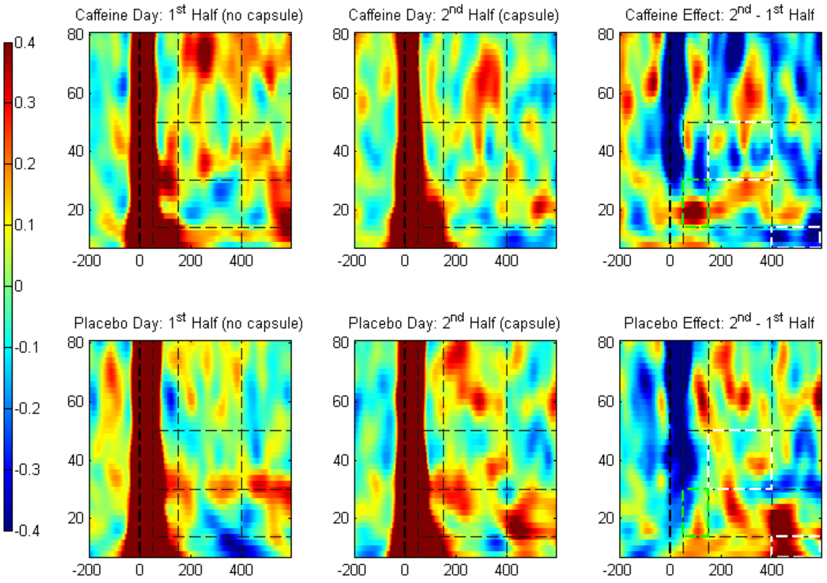

Figure 2.1: **TMS evokes stronger relative power in the placebo condition in the alpha and low gamma bands.** TMS-related spectral perturbations from frontal electrodes obtained in the caffeine and placebo conditions before capsule (left column) and after capsule (central column). The effect of the condition (2nd minus 1st half of experiment) is visualized by the right column. The vertical stripe at zero marks the onset of the TMS. The remaining vertical and horizontal stripes mark the time-frequency windows of analysis. Significant interaction between ROI and condition is marked by the green dotted line, significant main effects of condition by the white dotted lines.

2.4 Conclusion

This study was a fine introduction to the field of EEG analysis with resulted in some interesting neuroscientific conclusions. This study also lead us to troubled thinking that will be explained in the beginning of the next chapter.

CHAPTER 3

APPLICATION OF MACHINE LEARNING TO EEG (PUBLICATION II)

In the previous section, after heavy work of cleaning up the data and making everything compliant to all the assumptions of specific classical models we got a rigid pipeline that only worked for that one neuro-biological question. The aim was to do more, be more general and tackle the question in a data driven way such that the code base would be applicable to more questions and data types.

In addition, it is very likely that an important proportion of psychology studies are not reproducible (Open Science Collaboration, 2015) and it is also possible that the reason might be accidental p-value hacking (Head et al., 2015; Ioannidis, 2005). We realized ourselves that in the process of analysing the data, it is rather easy to accidentally do some steps incorrectly. For example, while setting up the whole pipeline and fixing software bugs, one usually sees some intermediate results that might influence the final results. We think that it did not happen to us but we cannot be sure either. This also led us to more exhaustive automatic search of the parameter space without looking at the data and making prior assumptions.

In the present study we asked whether it is possible to decode personality traits from resting state EEG data. EEG was recorded from a large sample of subjects ($n = 289$) who had answered questionnaires measuring personality trait scores of the 5 dimensions as well as the 10 subordinate aspects of the Big Five. According to a dominant Five Factor model, observable personality is mostly determined by five major traits – Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness (McCrae and John, 1992; McCrae and Costa, 2008).

Machine learning algorithms were used to build a classifier to predict each personality trait from power spectra of the resting state EEG data. Our results indicate that the five dimensions as well as their subordinate aspects could not be predicted from the resting state EEG data (a condition in which the subjects are not performing any particular task).

Importantly for current thesis, we used varied methods for data pre-processing (such as pooling and dimensionality reduction) and we also used a nested cross-validation which is the most common data partitioning method if the goal is to make an inference about the presence of information, where even the slightest discrimination performance indicates a statistical dependence between independent and dependent variables.

3.1 Neuroscience question

Personality can be defined as a relatively stable pattern on thinking, feeling and acting. In the current work Five Factor model personality traits are used. Their relatively high cultural universality, temporal stability, and heritability suggest that the Big Five traits may influence the brain activity enough for us to detect it (Corr, 2004; DeYoung and Gray, 2009; Kennis et al., 2013).

If traits indeed reflect individual differences in tonic brain function, then measures of baseline brain activity may provide a direct way for personality assessment. However, existing attempts to do this have generally yielded mixed results.

Most of the existing research on resting state EEG correlates of personality traits has been conducted in a hypothesis-driven way, concentrating usually on a single parameter at a time. An alternative approach would be to use data-driven techniques to first of all assess the extent to which resting state EEG signal contains information on personality and then search for relevant correlates in a more comprehensive and systematic manner. The main aim of the present study is to test such an approach. To that end we used machine learning classifiers to predict personality traits from resting state EEG signals. The classifiers were first trained using a set of data with known classes and then their performance was evaluated on data not used for the training phase.

3.2 Pre-processing and machine learning

All different methods of pre-processing, pooling, choosing the machine learning model and it's parameters are treated in this study as a parameter selection: all these decisions must be made before starting the model fitting where model weights will be estimated. In total, the system needed to choose between 648 combinations of parameters.

In the first part of the analysis, we trained statistical classifiers to map the features of the resting state EEG to personality scores of individual subjects. Given the exploratory nature of the analysis we scanned different combinations of classifier parameters and features from the EEG power spectra to find the configuration that best classified each personality trait. To avoid cherry-picking or over-fitting the results we always assessed the selected classifiers on a separate subset of subjects.

We also validated our pipeline and approach by decoding the eyes open vs eyes closed condition with an accuracy around 85.8% with $p << 0.001$. We used a binomial test against the null hypothesis that the prediction is random.

Thus, we used a nested cross-validation approach (see Section 4.1.2), which has inner and outer loops of cross validation (see Section 1.3.3). Inner 10-fold cross-validation loop is used to choose the parameters of the classifier (including different data pre-processing options, dimensionality reduction, and the choice between linear and non-linear SVM classifier). The classifier that performed the best for each personality trait in terms of misclassification rate, is used for the outer loop to estimate the misclassification rate of the selected classifier. So, the best parameters are chosen 10 times and the final misclassification rates represent the averages over these 10 partitions.

3.3 Answer to neuroscience question

We analyzed a large data set collected from 289 participants with resting state EEG recordings together with Big Five personality scores assessed with a self-report questionnaire.

The results for the binary classification of the test subjects are shown on Fig. 3.1. Although the best classification rates observed for Extraversion and Openness initially reached significance, they did not remain significant after Bonferroni or after false discovery rate correction (binomial test, corrected $p > 0.05$). This indicates that none of the personality traits were correctly classified from any of the explored combinations of resting state EEG features.

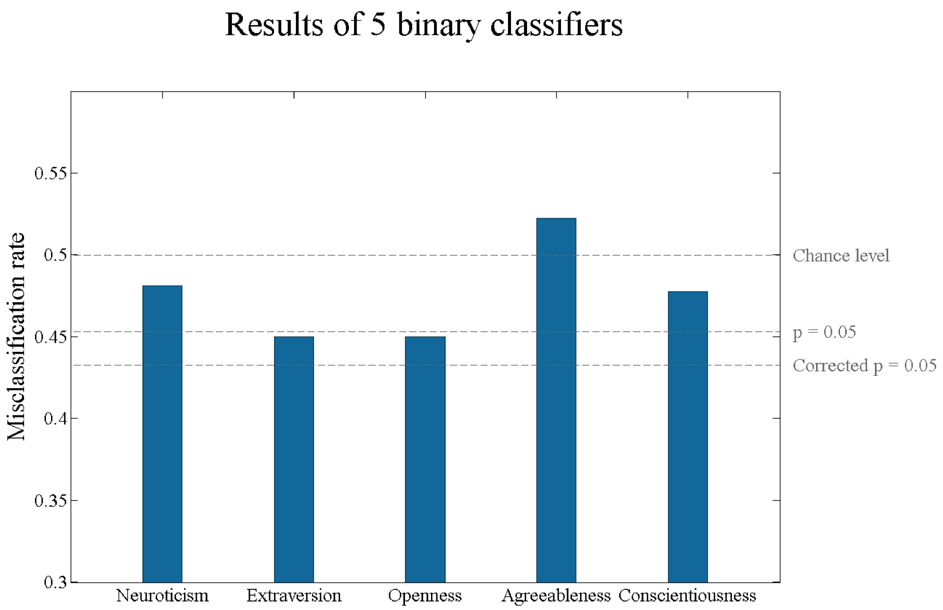

Figure 3.1: **Misclassification rates of Big Five personality traits.** The personality scores have been binarized with a median-split. None of the misclassification rates of five personality traits is statistically significant after Bonferroni or false discovery rate correction at 0.05.

Taken together, these results might indicate that personality traits might be more readily revealed by the reaction of the brain to stimuli rather than during rest state when personality traits are not classifiable from any of the explored combinations of resting state EEG features.

3.4 Conclusion

For the current study, it was relatively clear that the nested cross-validation approach is good. What is more, the whole machine learning infrastructure that was built for convincing in a negative result in the publication II, is currently actively used in various collaborations. For example, it is being used to predict the level of fatigue and attention in subjects from their EEG data.

The questions remains, what if we would have wanted to interpret the parameters which was impossible at the moment as they were different for each cross-validation fold. This study lead us to the next and main chapter of this dissertation.

CHAPTER 4

DATA PARTITIONING ALTERNATIVES FOR MACHINE LEARNING (PUBLICATION III)

As natural scientists usually want to get interpretable parameters then our analysis in the previous chapter was little worrying for us. Even with the positive outcome we would not have been able to give much explanation to the underling models that would have predicted Five Factor model personality traits. For example, we would probably not have been able to say if the predictive model is linear or non-linear. With nested cross-validation the parameters on each fold are different and it is very difficult to interpret them. On the other hand, the alternative approach where parameters are fixed with cross-validation and tested on a test set would have used data less efficiently which would have reduced the probability of finding a statistically significant result from the data. The novel efficient data partitioning approach emerged from trying to solve the conflicting goals of wanting interpretable parameters and using data more efficiently. A novel approach introduced in this chapter is the main contribution to the scientific community by this dissertation.

4.1 Common data partitioning approaches for machine learning

As we usually do not know the best possible combination of parameters in advance, supervised machine learning methods require splitting data into multiple chunks for training, validating, and finally testing classifiers (see Fig. 4.1). For finding the best parameters of a classifier, training and validation are usually carried out with cross-validation. This is followed by application of the classifier with optimized parameters to a separate test set for estimating the classifier's generalization performance. This type of data partitioning approach is the most general one.

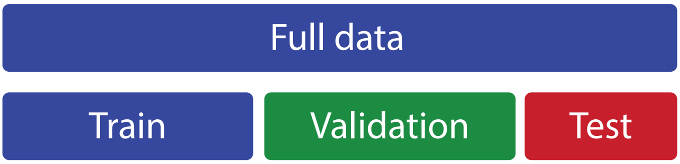

Figure 4.1: A common machine learning data partitioning approach.

4.1.1 Cross-validation and testing

Data collection can be very expensive, particularly in biological and social sciences in which for some experiments taking more than a few tens or hundreds of samples can be prohibitive. Therefore, over time more data-efficient methods for data partitioning have emerged (Pereira et al., 2009). Cross-validation, as explained in the section 1.3.3, is repeated with different parameter combinations, and once the best parameters have been found, the model is trained with the chosen parameters on all data that have previously been used for cross-validation. The generalization error is found by applying the train model to the separate test set (see Fig. 4.2). When the goal of a researcher is to build a model that generalizes well to unseen cases and that may be used in real life applications such as image recognition or text classification, this approach is probably the most generally used method for carrying out classification analyses. It makes it possible to have interpretable model parameters and also interpretable model weights.

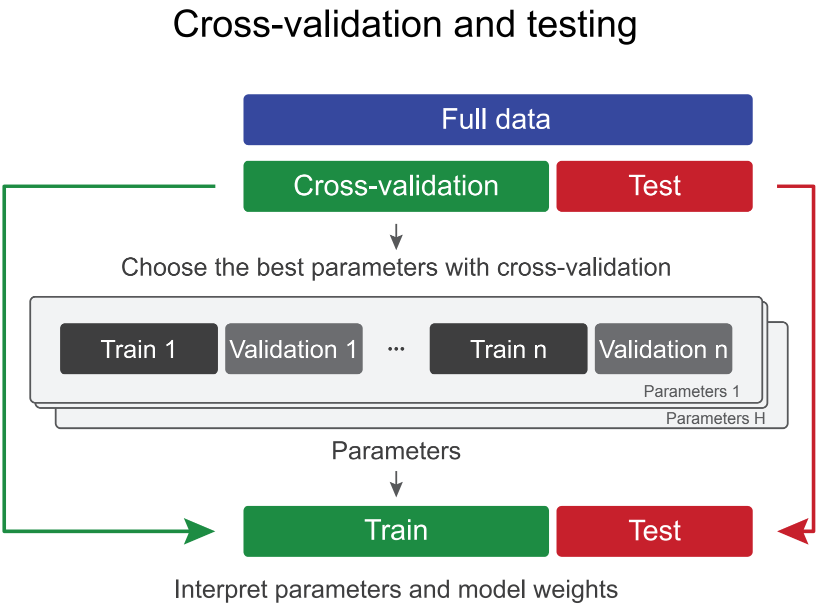
Cross-validation and testing
Full data
Cross-validation
Test
Choose the best parameters with cross-validation
Train 1
Validation 1
...
Train n
Validation n
Parameters 1
Parameters H
Parameters
Train
Test
Interpret parameters and model weights

Figure 4.2: **In the "Cross-validation and testing" approach, the data are divided into two separate sets (cross-validation set and test set) only once.** First, different models are trained and validated with cross-validation and the best set of parameters is chosen. Prediction accuracy and statistical significance of the parameters are evaluated on the test set, after training on the cross-validation set.

One difficulty of this approach is that the test set, which is used to validate classification performance, is limited in size. While cross-validation makes good use of training data, the estimation of the generalization performance of the classifier may still suffer by this limited size of the final test set. Increasing the size of the test set at the same time would come at the cost of diminishing classification performance. When data are scarce or expensive to acquire, this can become a large problem and may lead to a sub-optimal choice of classifiers and the associated parameters.

4.1.2 Nested cross-validation

One approach that has been used to overcome this difficulty is "Nested cross-validation" (Varma and Simon, 2006) (see Fig. 4.3). Here, the test set is not kept completely separate, but cross-validation is extended to incorporate all available data (outer cross-validation). In that way, all data can serve as test set once, overcoming the aforementioned trade-off. In order to still be able to optimize parameters, for a particular cross-validation iteration the training set is again divided for nested cross-validation (inner cross-validation), and once the best parameters for this iteration have been found, they are used to train a model on the current training data, which is then applied to the current test set. This approach is most useful for a researcher who does not need to build a model that generalizes well to unseen data, but who would like to describe whether there is a meaningful statistical dependence between the class labels and the given dataset, in other words whether the dataset contains information about the labels.

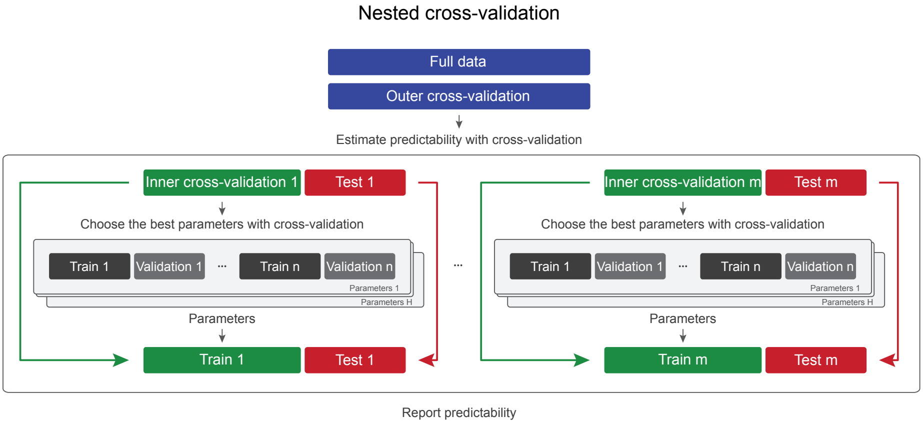

Figure 4.3: **In the "Nested cross-validation" approach, first (outer) cross-validation is performed to estimate predictability of the data.** In each iteration, data are divided into training and test sets. Before training, another (inner) cross-validation loop is used to optimize parameters. As model weights (fitted models) and parameters are different at every partition, it is not possible to report accuracy or statistical significance about a particular set of parameters or model weights.

While a Nested cross-validation procedure makes more efficient use of data, it has some drawbacks: Due to the absence of a completely separate test set it is not possible to claim that a particular model, i.e. a particular set of weights and classifier parameters, could in the future be used to classify unseen data (Hastie et al., 2011). In addition, the chosen parameters and models may vary between cross-validation iterations, making it impossible to select one set of parameters or one model as the final choice. In other words, a separate model and a separate set of parameters are chosen in each iteration and choosing any one of them would mean returning to a simple cross-validation and testing approach which would annihilate the advantage gained by nested cross-validation.

4.2 Novel data partitioning approach

There are cases in which the interpretation of parameters is desirable, even when the model is not used on unseen data. For example, for certain applications it might be useful to report that the best parameters correspond to a linear model as opposed to a non-linear one, without the need to describe the specific model weights used by the model. In another example, when using neural network as the class of machine learning algorithms, the number of layers selected during the optimization, say 3, 4 or 5 layers, may be an important choice to be communicated to other researchers and it might improve the interpretation of the results of the study.

We have invented an efficient data partitioning method to improve classification performance while keeping parameters interpretable.

4.2.1 Description of Cross-validation and cross-testing

“Cross-validation and cross-testing” starts by carrying out the common “Cross-validation and testing” (see Section 4.1.1) approach: The best parameters are chosen with cross-validation as described previously (Section 1.3.3). Once the parameters are fixed, the remaining data are used for testing the classifier. The novelty of the approach is introduced by how prediction accuracy and statistical significance are evaluated (see Fig. 4.4). Rather than keeping the test set entirely separate, a modified training set is iteratively introduced, where in each iteration the original training data plus part of the originally separate test data are used for training a classifier, and the remaining test data for testing the classification performance. We term this iterative procedure "cross-testing", because this term more accurately describes the process that is repeated than "cross-validation". Importantly, this approach maintains independence between training and test data, but makes more efficient use of test data by augmenting training data for more accurate predictions. In the next iteration, a different part of test data is added to the training data, and this process is repeated until each part of the test data has once been added to training data. The mean prediction accuracy across these different cross-validation iterations is then averaged. We refer to this novel approach as “Cross-validation and cross-testing”.

This proposed approach is expected to provide more accurate results than classical “Cross-validation and testing” by construction because it simply uses more data for model fitting, while the system for choosing the best set of parameters remains the same.

Underling assumptions of Cross-validation and cross-testing are the same as in other data partitioning methods. In particular, some care must be taken when constructing the cross-testing folds such that it would not introduce information leakage within the cross-testing set.

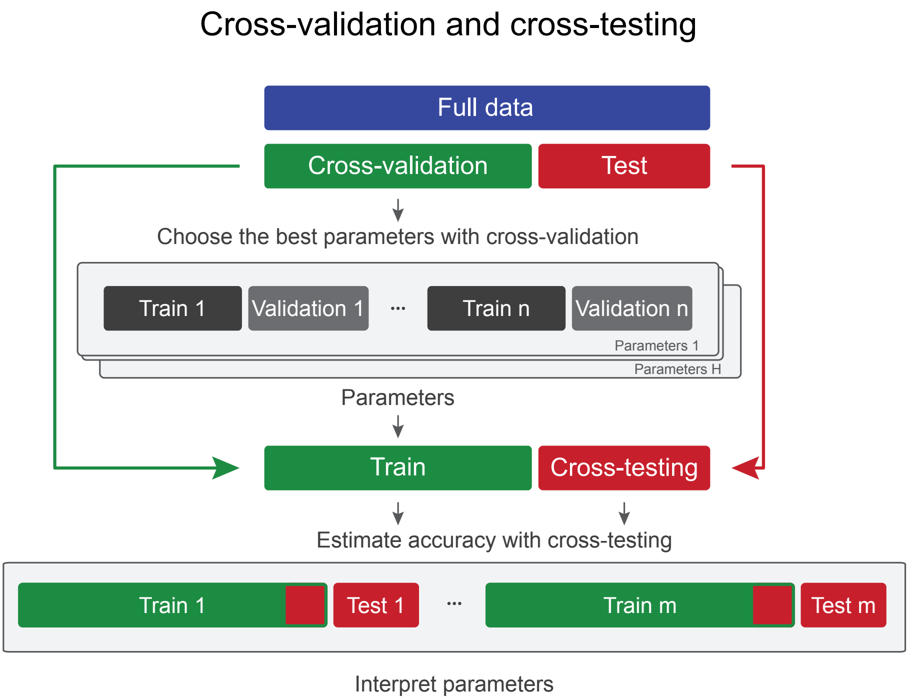

Figure 4.4: **For cross-validation and cross-testing, data are divided into two separate sets only once: a cross-validation set and a test set.** Similar to typical cross-validation, a number of iterations are carried out to choose the best parameters for the final model on the test set. Once the best combination of parameters has been chosen, the prediction accuracy and statistical significance can be evaluated on the test set with a modified cross-validation such that for each fold the original cross-validation set is repeatedly added to the training data. Due to the similarity to cross-validation, we term this approach cross-testing. While making it impossible to pick one final model on additional unseen data, the parameters that have been chosen remain interpretable.

4.2.2 Comparison to the other machine learning approaches

When a researcher is interested in publishing their model parameters, then typically the efficient and popular approach called “Cross-validation and testing” is used (Fig. 4.2). In the cross-validation set, the best parameters are chosen - usually according to highest cross-validation accuracy - and the test set is used for out-of-sample accuracy estimation.

An even more data-efficient approach for data analysis is called “Nested cross-validation” (Fig. 4.3). The approach is similar to cross-validation and testing, but the test set becomes part of an outer cross-validation loop, while parameters are optimized in inner cross-validation iterations, using only the training data of the current outer cross-validation iteration. The whole data set can be used for estimating the final accuracy and therefore has a maximum statistical power for significance analysis. However, this approach does not make it possible to publish parameters which might be sometimes desirable.

Our proposed approach can be seen as a natural extension between the two extremes described before. See Table 4.1 for a comparison of the three approaches in terms of data efficiency and parameter and model weights interpretability.

Table 4.1: **Comparison of the approaches**

Approach	Data efficiency	Possible to interpret parameters	Possible to interpret fitted model
Cross-validation and testing	Low	Yes	Yes
Proposed: Cross-validation and cross-testing	Intermediate	Yes	No
Nested cross-validation	High	No	No

Comparison of the newly proposed “Cross-validation and cross-testing”, classical “Cross-validation and testing” and “Nested cross-validation” with respect to data efficiency, and parameter and model interpretability.

4.2.3 Results

As predicted, we demonstrate the superiority of "Cross-validation and cross-testing" over "Cross-validation and testing" both in terms of accuracy and in terms of statistical sensitivity (see Fig. 4.5). This improved performance is explained by the much larger data set that is available during each testing iteration of "Cross-validation and cross-testing".

We confirmed in various numerical experiments that too large test sets quickly result in insufficient data for finding the best parameters and not enough data for fitting the best model. On the other hand, too small test sets can imply that there are not enough data for achieving statistically significant results. This trade-off for the size of the test set results in the existence of an optimal range. The proposed "Cross-validation and cross-testing" modifies the range to allow larger test set sizes because with the novel cross-testing part it is still possible to use almost all of the data for model fitting (see Fig. 4.6).

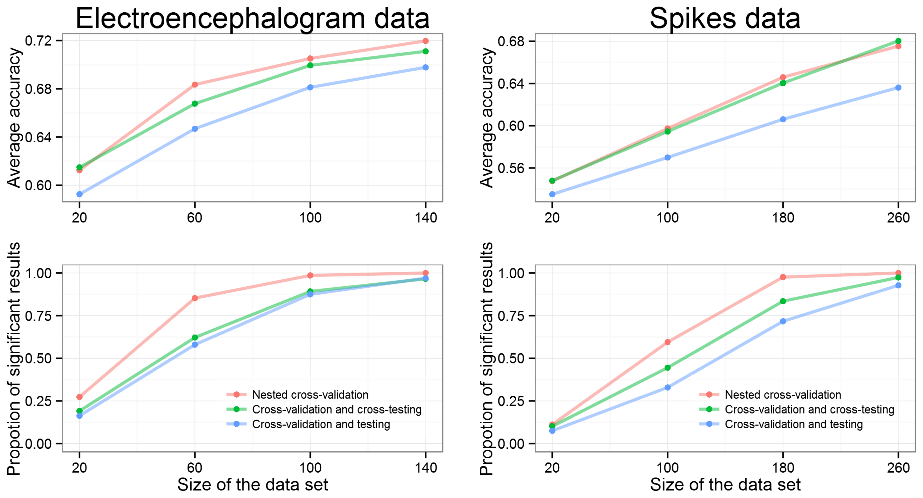

Figure 4.5: **Analysis of real data (left: EEG dataset; right: spiking dataset) with three different approaches as a function of data size with test set size fixed at 50%.** Results show the mean accuracy (upper graphs) and proportion of significant results (bottom graphs) out of the 1000 runs. More data lead to higher average accuracy and increases the proportion of significant results with the binomial test. "Nested cross-validation" outperforms other approaches while "Cross-validation and testing" gives the worst performance in terms of average accuracy and proportion of significant results.

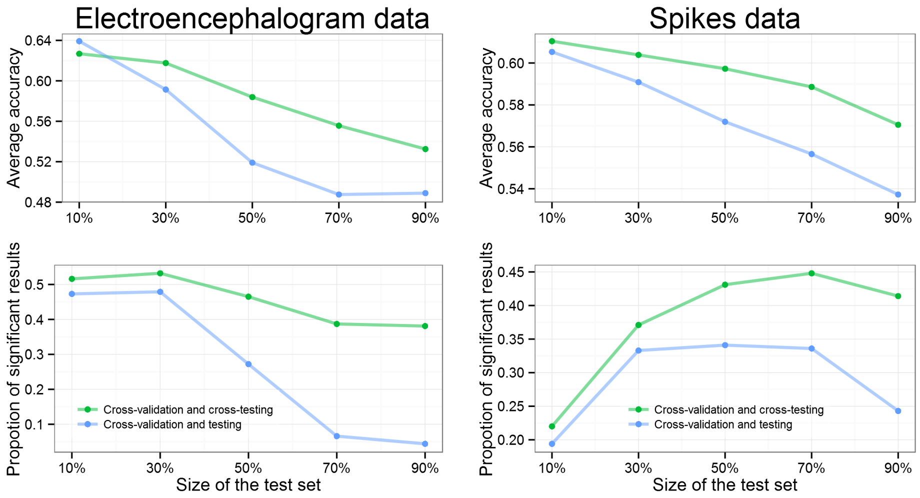

Figure 4.6: **Analysis of neuroscience data (left: EEG dataset; right: spiking dataset) with three different approaches as a function of the relative test set size.** Results show the mean accuracy (upper graphs) and proportion of significant results (bottom graphs) out of the 1000 runs. Data set size was fixed to 50 for EEG and to 100 for spikes train data set. Larger test set leads to smaller average accuracy because there is less data for choosing parameters and fitting a model. "Cross-validation and cross-testing" outperforms “cross-validation and testing” in terms of average accuracy and proportion of significant results.

4.3 Future work

It is fair to say that every contribution creates more questions than it solves. Along the way, we also noticed several lines of work which should or are already building on top of the results obtained in this dissertation. In this section, we will briefly mention possible future theoretical work related to hyper-parameters and an ongoing implementation of the main thesis of this dissertation.

4.3.1 Hyper-parameters and data

The choice of parameters for a machine learning algorithm can be somewhat solved with clever data partitioning methods but some so-called hyper-parameters must be fixed before the start of the analysis.

For example, the size of the test set is one such hyper-parameter that could be studied further. Especially because of the novel approach introduced in this dissertation and more precisely the way of splitting the problem into two sub-problems: choosing parameters and optimizing model weights. This might make it easier to come up with better heuristics for choosing the size of the test set.

In addition, this test set size might be dependent on the properties of the data. For example, in Fig. 4.5 the effect is smaller with EEG data suggesting that more efficient usage of data in the model fitting is not that important and the choice of parameters is actually the main influencer. This might influence the choice of the machine learning data partitioning method and some other hyper-parameters.

4.3.2 Implementation for common use

There are numerous programming languages and libraries for organizing the work of a data analysts such as a language R (R Core Team, 2013) and machine learning library scikit-learn in Python (Pedregosa et al., 2011). In an ongoing project, we are supervising the implementation of our novel machine learning data partitioning method in these common frameworks.

4.4 Conclusion

When using machine learning algorithms for making predictions, improving performance of a classifier can be seen as a central goal. At the same time, interpretability of models and parameters beyond the given data are in many cases desirable. Since data are often scarce or expensive to acquire, efficient use of data is another important objective. These three goals - generalization performance, interpretability, and efficient use of data - often lead to a trade-off that is resolved depending on the focus of the researcher. If the focus lies on generating a model that will generalize well to unseen data, this requires an interpretable model and interpretable model parameters. For that purpose, data can be used efficiently by the common approach known as "Cross-validation and testing". If, however, the goal is to maximize classification performance for finding a statistical dependence between class labels and data, then a better performance can be achieved by using “Nested cross-validation”, which provides a very data efficient approach by repeatedly re-using data. However, this approach does not naturally allow the interpretation of parameters and model weights.

Previously, researchers had to choose from these two extremes depending on their goal. In this study, we described a novel approach that uses cross-validation to find and fix the best combination of parameters, but importantly which then resamples test data for augmenting training data, yielding to more accurate estimation of generalization performance. We termed this new approach “Cross-validation and cross-testing” and predicted that it would outperform outperforms “Cross-validation and testing” in terms of accuracy and statistical sensitivity with simulated and empirical data sets. In particular, we tested the effects of different data sets, data set sizes and test set sizes.

While both “Cross-validation and testing” and the proposed "Cross-validation and cross-testing" make it possible to report parameters, the latter, novel approach uses data more efficiently in the final model fitting phase by reducing the test set size trade-off. This change of the trade-off occurs by reducing the detrimental effect of a larger test set size on the quality of the fitted model by augmenting the size of the training data.

It is our hope that the proposed approach can be a useful addition to the toolkit of machine learning approaches. We believe that it might be especially applicable when both data efficiency and parameter interpretations are desired.

SUMMARY

A simple conclusion of a research study, for example, that yes there is an effect of X on Y, is a result of a long and complex effort. It is amazingly difficult to choose a research topic of interest; understand current trends and problems of the related topic; invent clever methods to study it and finally carefully perform tedious experiments that should shed light on the topic of interest. But the hard work does not stop there.

Data must first be cleaned, with artefacts and outliers removed. Often a common statistical inference is enough as a next step. Choosing it correctly is difficult as complex real life data rarely complies with all the needed assumptions and it is easy to accidentally make statistical mistakes in the process. In Chapter 2, we tackled these challenges while analyzing human EEG data and performed all the step previously mentioned.

To compound the problem, very often it is not possible to use "simple" methods of statistical inference. Machine learning makes it possible to analyze more complex data while needing less pre-processing and by making fewer assumptions about the data. The apparent positive side comes with an important negative side. Machine learning methods need an approach that uses some kind of data partitioning in order to estimate the true prediction accuracy of the result.

In Chapter 3, we developed a machine learning analysis approach that uses nested cross-validation in order to perform extensive parameter search and also validate the statistical dependence between class labels and the data.

The obvious problem with the nested cross-validation approach is the lack of possibility to interpret model parameters which could lead to more insight and more efficient scientific process. The main contribution of this dissertation, as described in Chapter 4, was to categorize and analyze two most common data-partitioning schemes and invent a new one that is more efficient in the case where model parameters need interpretation but model weights do not.

The new approach is termed "Cross-validation and cross-testing". First, cross-validation is used on part of the data to determine and lock the best set of parameters. Then testing of the model on a separate test set is performed in a novel way such that on each testing cycle, part of the data is also used in a model training

phase. This gives us an improved system for using machine learning algorithms in the case where we need to interpret model parameters but not the model weights. For example, it gives us a nice possibility to be able to describe that the data has a linear relationship instead of quadratic one or that the best neural network has 5 hidden layers. The new approach is validated on different neural datasets and on a simulated dataset.

We notice a clear trend towards using machine learning to analyze neural data. This makes our contribution timely, especially in the light of how easy is to misuse statistical analysis and the need of straightforward parameter scanning with efficient use of complex models on limited data.

It would be a positive direction if a larger proportion of scientific results would be correct and reproducible and if scientists could be spending less time in data pre-processing. What is more, it would be nice if we could re-use each other's code more easily. Some of it can be achieved with modern machine learning algorithms. In natural science, description and interpretation of parameters would be a useful addition to any paper.

In this dissertation, we have invented a more efficient data partitioning approach which makes it possible to still interpret the parameters. As scanning of parameters in complex models is arguably the most time-consuming operation, sharing of them will accelerate the process of data-driven discovery in natural sciences. We hope that researchers will find it useful and will start using it and therefore improve the quality of science and also pleasure gained by doing the statistical analysis.

Bibliography

B. Akyildiz. An introduction to supervised learning via scikit learn, 2014. URL `http://bugra.github.io/work/notes/2014-11-22/an-introduction-to-supervised-learning-scikit-learn/`.

E. Alpaydin. *Introduction to machine learning*. MIT press, 2014.

M. Anusuya and S. K. Katti. Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*, 2010.

J. Aru, K. Korjus, C. Murd, and T. Bachmann. Spectral signatures of the effects of caffeine and occipitally applied transcranial magnetic stimulation in a task-free experimental setup. *Journal of Caffeine Research*, 2(1):23–30, 2012.

J. Aru, K. Korjus, and E. Saar. *Matemaatika õhtuõpik*. Hea Lugu OÜ, 2014.

R. J. Barry, J. A. Rushby, M. J. Wallace, A. R. Clarke, S. J. Johnstone, and I. Zlojutro. Caffeine effects on resting-state arousal. *Clinical Neurophysiology*, 116 (11):2693–2700, 2005.

I. G. Y. Bengio and A. Courville. Deep learning. Book in preparation for MIT Press, 2016. URL `http://www.deeplearningbook.org`.

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, et al. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007.

H. Berger. On the electroencephalogram of man. sixth report. *Electroencephalography and clinical neurophysiology*, pages Suppl–28, 1969.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

L. Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

S. R. Y. Cajal. The croonian lecture: La fine structure des centres nerveux. *Proceedings of the Royal Society of London*, 55(331-335):444–468, 1894.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

M. Clubley, C. Bye, T. Henson, A. Peck, and C. Riddington. Effects of caffeine and cyclizine alone and in combination on human performance, subjective effects and eeg activity. *British Journal of Clinical Pharmacology*, 7(2):U57–U63, 1979.

P. J. Corr. Reinforcement sensitivity theory and personality. *Neuroscience & Biobehavioral Reviews*, 28(3):317–332, 2004.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

D. D. Cox and T. Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24(18):R921–R929, 2014.

C. G. DeYoung and J. R. Gray. Personality neuroscience: Explaining individual differences in affect, behavior, and cognition. *The Cambridge handbook of personality psychology*, pages 323–346, 2009.

F. Ferrarelli, M. Massimini, S. Sarasso, A. Casali, B. A. Riedner, G. Angelini, G. Tononi, and R. A. Pearce. Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proceedings of the National Academy of Sciences*, 107(6):2681–2686, 2010.

E. Girden. *ANOVA: Repeated Measures*. Number no. 84 in ANOVA: Repeated Measures. SAGE Publications, 1992. ISBN 9780803942578.

C. Golgi. *Sulla fina anatomia degli organi centrali del sistema nervoso*. S. Calderini, 1885.

C. Golgi. The neuron doctrine: theory and facts. *Nobel lecture*, 1906.

A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

R. Gueorguieva and J. H. Krystal. Move over anova: Progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry. *Archives of general psychiatry*, 61(3):310–317, 2004.

L. F. Haas. Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(1):9–9, 2003.

A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2011.

J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.

B. J. He and M. E. Raichle. The fmri signal, slow cortical potential and consciousness. *Trends in cognitive sciences*, 13(7):302–309, 2009.

M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. The extent and consequences of p-hacking in science. *PLoS Biol*, 13(3):e1002106, 2015.

J. Hohwy. The neural correlates of consciousness: new experimental approaches needed? *Consciousness and cognition*, 18(2):428–438, 2009.

J. P. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8): e124, 2005.

M. Kennis, A. R. Rademaker, and E. Geuze. Neural correlates of personality: an integrative review. *Neuroscience & Biobehavioral Reviews*, 37(1):73–95, 2013.

S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10 (11):e1003915, 2014.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

K. Korjus, A. Uusberg, H. Uusberg, N. Kuldkepp, K. Kreegipuu, J. Allik, R. Vicente, and J. Aru. Personality cannot be predicted from the power of resting state EEG. *Frontiers in human neuroscience*, 9:63, 2015.

K. Korjus, M. N. Hebart, and R. Vicente. An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLOS ONE*, 11(8):e0161788, 2016.

N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

U. Lee, G. A. Mashour, S. Kim, G.-J. Noh, and B.-M. Choi. Propofol induction reduces the capacity for neural information integration: implications for the mechanism of consciousness and general anesthesia. *Consciousness and cognition*, 18(1):56–64, 2009.

M. M. Lorist and M. Tops. Caffeine, fatigue, and cognition. *Brain and cognition*, 53(1):82–94, 2003.

M. Massimini, F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi. Breakdown of cortical effective connectivity during sleep. *Science*, 309(5744): 2228–2232, 2005.

M. Massimini, F. Ferrarelli, M. Murphy, R. Huber, B. Riedner, S. Casarotto, and G. Tononi. Cortical reactivity and effective connectivity during rem sleep in humans. *Cognitive neuroscience*, 1(3):176–183, 2010.

R. R. McCrae and P. T. Costa. Empirical and theoretical status of the five-factor model of personality traits. *The SAGE handbook of personality theory and assessment*, 1:273–294, 2008.

R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

D. Millet. The origins of eeg. In *7th Annual Meeting of the International Society for the History of the Neurosciences (ISHN)*, 2002.

A. M. Molinaro, R. Simon, and R. M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

C. Murd, J. Aru, M. Hiio, I. Luiga, and T. Bachmann. Caffeine enhances frontal relative negativity of slow brain potentials in a task-free experimental setup. *Brain research bulletin*, 82(1):39–45, 2010.

P. L. Nunez and R. Srinivasan. Electroencephalogram. *Scholarpedia*, 2(2):1348, 2007.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

D. R. Reddy. Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531, 1976.

K. Rees, D. Allen, and M. Lader. The influences of age and caffeine on psychomotor and cognitive function. *Psychopharmacology*, 145(2):181–188, 1999.

B. J. Roach and D. H. Mathalon. Event-related eeg time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophrenia bulletin*, 34(5):907–926, 2008.

J. Ruijter, M. B. de Ruiter, J. Snel, and M. M. Lorist. The influence of caffeine on spatial-selective attention: an event-related potential study. *Clinical Neurophysiology*, 111(12):2223–2233, 2000.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

M. Stamm, J. Aru, and T. Bachmann. Right-frontal slow negative potentials evoked by occipital tms are reduced in nrem sleep. *Neuroscience letters*, 493 (3):116–121, 2011.

A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE*, 12(4):e0172395, 2017.

G. Upton and I. Cook. *A dictionary of statistics 3e*. Oxford university press, 2014.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and remote control*, 24:774–780, 1963.

S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):1, 2006.

A. P. Wickens. *A History of the Brain: From Stone Age surgery to modern neuroscience*. Psychology Press, 2014.

Wikipedia. Support vector machine — wikipedia, the free encyclopedia, 2016. URL `https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=735405069`. [Online; accessed 26-August-2016].

S. Zhang, C. Zhang, and Q. Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381, 2003.

APPENDIX

Chapter 3, publication II

The code of the whole project of the publication II can be accessed from here:

`www.github.com/kristjankorjus/PredictingPersonalityFromEEG`

Chapter 4, publication III

All datasets, source files for the figures and full code together with history of changes of the publication III can be accessed from here:

`www.github.com/kristjankorjus/machine-learning-approaches`

ACKNOWLEDGEMENTS

I am very grateful to a number of people, who have played an important role in finalising this thesis. Some of them, however, deserve special mention:

Raul Vicente, my supervisor - for moving to Estonia and giving me unfailing and extremely encouraging support.

Jaan Aru - for introducing me to the area of neuroscience and being incredibly helpful with the thesis.

Marlon Dumas - for persuading me to aim higher.

Jaak Vilo - for finding funding and believing in me.

BIIT research group - for the initial ideas and motivation.

And all the others at our Computational Neuroscience Lab: Tambet Matiisen, Ilya Kuzovkin, Ardi Tampuu, Taivo Pungas and Kristjan-Julius Laak - for the enthusiasm, knowledge, support, jokes and random meals.

Laura - for being so very supportive and carrying most of the parenting tasks with our little kid Loore.

This dissertation was supported by the Doctoral School of Information and Communication Technology (IKTDK), StudyITin.ee (IT Akadeemia) initiative and Estonian Research Council's personal research grants program (PUT438 grant) to Raul Vicente. The computationally intensive computations were performed in the University of Tartu High Performance Computing Center.

KOKKUVÕTE (SUMMARY IN ESTONIAN)

EEG ANDMETE ANALÜÜS JA ANDMEPARTITSIOONIDE ARENDAMINE MASINÕPPE ALGORITMIDELE

Empiirilise teadustöö selgrooks on andmed, statistilised mudelid ja nende põhjal tehtud ennustused. Paljudes valdkondades - nagu neuroteaduses ja bioinformaatikas - on andmed äärmiselt keerukad, mitmemõõtmelised ja tekkinud mittelinaarsete interaktsioonide poolt. Oma seesmise keerukuse tõttu vajavad kaasaegsed masinõppe algoritmid suurt hulka andmeid ja paljude paremeetrite läbikatsetamist.

Antud doktoritöö eesmärgiks on kirjeldada ja arendada inimaju signaalidest salvestatud elektroentsefalograafi (EEG) andmeanalüüsi protsessi. Vaatamata sellele, et valdav osa tööst põhineb EEG signaalidel, on enamik saadud teadmistest ja välja töötatud meetodidest rakendatavad suuremale osale keerukate (ja kulukate) andmetepõhiste valdkondade probleemidele.

Töö esimeses osas on vaatluse all andmete puhastamise ja eeltöötlemise olulisus koos klassikalistest statistilistest mudelitest lähtuvate lihtsamate hüpoteeside testimisega.

Põhifookuses on antud töö puhul masinõppe algoritmide rakendamine elulistele andmetele. Kuivõrd masinõppe algoritmid ei võimalda üldjuhul teha eksplitsiitselt lihtsaid statistilisi järeldusi, nihkub fookus efektiivsetele andmepartitsiooni meetoditele.

Doktoritöö peamiseks tulemuseks on uudne andmepartitsiooni tegemise viis, mis võimaldab kasutada andmeid oluliselt efektiivsemalt kui olemasolevad lähenemised, juhul kui parameetrid vajavad tõlgendamist, kuid mudeli kaalud mitte.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name	Kristjan Korjus
Birth	October 29th 1985 Tartu, Estonia
Citizenship	Estonian
Marital Status	Married
Languages	Estonian, English
Address	Vene 19-23, Tallinn, Estonia 10123
Contact	+37256840434 korjus@gmail.com

Education

2011–	University of Tartu, PhD student in Computer Science
2007–2011	The University of Manchester, MMATH in Mathematics (*First Class Honours*)
2005–2007	University of Bedfordshire, BA in Business Studies (discontinued)
1993–2005	Hugo Treffner's High School

Employment

2016–	Starship Technologies, Head of Computer Vision and Perception
2015–2016	Dream It Get IT Ltd, Machine Learning Engineer
2013–2015	University of Tartu, Lead Project Coordinator
2011–2013	University of Tartu, Teaching Assistant

2010–2011	KFPD GmbH, Financial Mathematician
2009–2010	Food 4 Thought Project Limited, Mathematics Tutor
2007–2008	Tark Investor OÜ, Financial Analyst

Affiliations

2017–	Tallinn University of Technology, Council of Robotics and Product Development Curriculum
2016–	Foundation for Future Technologies, Board of Directors
2013–2015	University of Tartu, Senate
2013–2015	University of Tartu, Council of Computer Science Curriculum
2012–2014	University of Tartu, Council of the Institute of Computer Science
2014	University of Tartu, Council of Strategic Planning

Publications

1. K. Korjus, M. N. Hebart, and R. Vicente. An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLOS ONE*, 11(8):e0161788, 2016.

2. A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE*, 12(4):e0172395, 2017.

3. K. Korjus, A. Uusberg, H. Uusberg, N. Kuldkepp, K. Kreegipuu, J. Allik, R. Vicente, and J. Aru. Personality cannot be predicted from the power of resting state EEG. *Frontiers in human neuroscience*, 9:63, 2015.

4. J. Aru, K. Korjus, and E. Saar. *Matemaatika õhtuõpik*. Hea Lugu OÜ, 2014.

5. J. Aru, K. Korjus, C. Murd, and T. Bachmann. Spectral signatures of the effects of caffeine and occipitally applied transcranial magnetic stimulation in a task-free experimental setup. *Journal of Caffeine Research*, 2(1):23–30, 2012.

ELULOOKIRJELDUS

Isikuandmed

Nimi	Kristjan Korjus
Sünniaeg ja -koht	20. oktoober 1985 Tartu, Eesti
Kodakondsus	Eesti
Perekonnaseis	abielus
Keelteoskus	eesti, inglise
Aadress	Vene 19-23, Tallinn, Eesti 10123
Kontaktandmed	+37256840434 korjus@gmail.com

Haridustee

2011–	Tartu Ülikool, informaatika doktorant
2007–2011	Mancesteri ülikool, MMATH matemaatikas (*First Class Honours*)
2005–2007	Bedforshire ülikool, lõpetamata BA ärinduses
1993–2005	Hugo Treffneri gümnaasium

Teenistuskäik

2016–	Starship Technologies, Masinnägemise juht
2015–2016	Dream It Get IT Ltd, Masinõppe insener
2013–2015	Tartu Ülikool, Projektijuht
2011–2013	Tartu Ülikool, Assistent
2010–2011	KFPD GmbH, Finantsmatemaatik

2009–2010	Food 4 Thought Project Limited, Matemaatika õpetaja
2007–2008	Tark Investor OÜ, Finantsanalüütik

Muud kuuluvused

2016–	Tallinna Tehnikaülikool, Tootearendus ja robootika õppekava nõukogu
2016–	Tulevikutehnoloogiate SA, Nõukogu
2013–2015	Tartu Ülikool, Senat
2013–2015	Tartu Ülikool, Arvutiteaduse õppekava programminõukogu
2012–2014	Tartu Ülikool, Arvutiteaduse instituudi nõukogu
2014	Tartu Ülikool, Strateegilise planeerimise töögrupp

Publikatsioonid

1. K. Korjus, M. N. Hebart, and R. Vicente. An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLOS ONE*, 11(8):e0161788, 2016.

2. A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE*, 12(4):e0172395, 2017.

3. K. Korjus, A. Uusberg, H. Uusberg, N. Kuldkepp, K. Kreegipuu, J. Allik, R. Vicente, and J. Aru. Personality cannot be predicted from the power of resting state EEG. *Frontiers in human neuroscience*, 9:63, 2015.

4. J. Aru, K. Korjus, and E. Saar. *Matemaatika õhtuõpik*. Hea Lugu OÜ, 2014.

5. J. Aru, K. Korjus, C. Murd, and T. Bachmann. Spectral signatures of the effects of caffeine and occipitally applied transcranial magnetic stimulation in a task-free experimental setup. *Journal of Caffeine Research*, 2(1):23–30, 2012.

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.
19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.

23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. Töö kaitsmata.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.

49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.

72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo.** Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.
89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multiparty computation. Tartu, 2015, 201 p.

95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
105. **Md Raknuzzaman.** Noncommutative Galois Extension Approach to Ternary Grassmann Algebra and Graded q-Differential Algebra. Tartu, 2016, 110 p.
106. **Alexander Liyvapuu.** Natural vibrations of elastic stepped arches with cracks. Tartu, 2016, 110 p.
107. **Julia Polikarpus.** Elastic plastic analysis and optimization of axisymmetric plates. Tartu, 2016, 114 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.
113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
115. **Tiina Kraav.** Stability of elastic stepped beams with cracks. Tartu, 2017, 126 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.

117. **Silja Veidenberg.** Lifting bounded approximation properties from Banach spaces to their dual spaces. Tartu, 2017, 112 p.
118. **Liivika Tee.** Stochastic Chain-Ladder Methods in Non-Life Insurance. Tartu, 2017, 110 p.
119. **Ülo Reimaa.** Non-unital Morita equivalence in a bicategorical setting. Tartu, 2017, 86 p.
120. **Rauni Lillemets.** Generating Systems of Sets and Sequences. Tartu, 2017, 181 p.