Interpretable Machine Learning for Electro-encephalography

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Sebastian Mathias Keller

aus Deutschland

2020

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Volker Roth, Dissertationsleiter

Prof. Dr. Thomas Vetter, Korreferent

Basel, den 17. November 2020

Prof. Dr. Martin Spiess, Dekan

It's no use going back to yesterday,
because I was a different person then.
— Lewis Carroll

To my parents Henriette and Jürgen
and to my brother Jan-Philipp

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Dr. Volker Roth, whose expertise was invaluable in formulating the research questions and methodology. This thesis would not have been possible without his guidance and support. His door was always open – both literally and figuratively – which is easily taken for granted when in fact it shouldn't be.

I gratefully acknowledge the funding received towards my PhD from the Swiss National Science Foundation.

I greatly appreciate the support received through the collaborative work undertaken with Prof. Dr. Peter Fuhr. He always made me feel welcome in his group and introduced me to the fascinating world of EEG. During our numerous and invaluable discussions I have learned much about clinical neurology. My thanks also go out to the support I received from Prof. Dr. Ute Gschwandtner who both motivated and challenged me through constructive criticism and many productive discussions.

I very much like to thank Prof. Dr. Thomas Vetter for agreeing to be a co-examiner of my dissertation which requires both significant time and effort.

I would like to acknowledge my colleagues Dr. Dinu Kaufmann, Aleksander Wieczorek, Dr. Menorca Chaturvedi, Dr. Antonia Meyer, Dr. Mario Wieser, Damian Murezzan, Maxim Samarin and Fabricio Arend Torres for their wonderful collaboration. Research work is collaborative work and with colleagues such as yourselves it also was a lot of fun – thank you all.

I am especially grateful to Dr. Mario Wieser – while pursuing a PhD there are many good times and a few bad ones but with you the good times were better and the bad ones bearable.

I am also very grateful to all members of the Fuhr group and the Vetter group, to the computer admins and the administrative staff of the Computer Science department.

In addition, I would like to thank my family and friends for their wise counsel and sympathetic ear. You are always there for me.

Basel, 5. Oktober 2020 S. M. K.

Abstract

While behavioral, genetic and psychological markers can provide important information about brain health, research in that area over the last decades has much focused on imaging devices such as magnetic resonance tomography (MRI) to provide non-invasive information about cognitive processes. Unfortunately, MRI based approaches, able to capture the slow changes in blood oxygenation levels, cannot capture *electrical* brain activity which plays out on a time scale up to three orders of magnitude faster. Electroencephalography (EEG), which has been available in clinical settings for over 60 years, is able to measure brain activity based on rapidly changing electrical potentials measured non-invasively on the scalp. Compared to MRI based research into neurodegeneration, EEG based research has, over the last decade, received much less interest from the machine learning community. But generally, EEG in combination with sophisticated machine learning offers great potential such that neglecting this source of information, compared to MRI or genetics, is not warranted. In collaborating with clinical experts, the ability to link any results provided by machine learning to the existing body of research is especially important as it ultimately provides an intuitive or interpretable understanding. Here, interpretable means the possibility for medical experts to translate the insights provided by a statistical model into a working hypothesis relating to brain function. To this end, we propose in our first contribution a method allowing for ultra-sparse regression which is applied on EEG data in order to identify a small subset of important diagnostic markers highlighting the main differences between healthy brains and brains affected by Parkinson's disease. Our second contribution builds on the idea that in Parkinson's disease impaired functioning of the thalamus causes changes in the complexity of the EEG waveforms. The thalamus is a small region in the center of the brain affected early in the course of the disease. Furthermore, it is believed that the thalamus functions as a pacemaker – akin to a conductor of an orchestra – such that changes in complexity are expressed and quantifiable based on EEG. We use these changes in complexity to show their association with future cognitive decline. In our third contribution we propose an extension of archetypal analysis embedded into a deep neural network. This generative version of archetypal analysis allows to learn an appropriate representation where every sample of a data set can be decomposed into a weighted sum of extreme representatives, the so-called archetypes. This opens up an interesting possibility of interpreting a data set relative to its most *extreme* representatives. In contrast, clustering algorithms describe a data set relative to its most *average* representatives. For Parkinson's disease, we show based on deep archetypal analysis, that healthy brains produce archetypes which are different from those produced by brains affected by neurodegeneration.

Contents

List of Figures

List of Tables

1 Introduction

1.1 General Motivation

Accordingst to statistics compiled by Dell EMC, healthcare institutions have seen a 878% health data growth rate since 2016 compared to 2018. Quantifying the storage needs, this translates into healthcare institutions throughout the world having to manage an average of 8.41 petabytes of data in 2018 which is almost a ninefold increase within two years. More importantly, this trend continuous unbroken, making it all the more important for adequate data analysis tools to keep up with the ever increasing demand. Naturally, data collection has increased especially in domains related to imminent health threats such as neurodegenerative diseases. While the global increase of average life expectancy is without a doubt a major social achievement, it comes at the price of confronting societies across the globe with age related diseases such as Alzheimer's disease (AD) or Parkinson's disease (PD) in ever increasing case numbers. According to Dorsey and Bloem [2018], the growth rate of PD has now surpassed that of AD. Still, with over 50 million people globally suffering from AD and around 6 million suffering from PD, the trend points toward a *slow pandemic of neurodegeneration* – a pandemic which cannot be controlled through vaccination. Given these prospects, i. e. an increasing need for better diagnostic, monitoring and treatment options in neurodegeneration and the currently unfolding data deluge, it is high time for the medical and the machine learning community to join forces in every aspect – from basic medical research and clinical research down to the daily clinical workflow. But while basic medical research is often eager to embrace new statistical methods and even drives their development by putting out new data sets, increased proximity to the patient seems to correlate with an increased reluctance to adopt new methods. While regulatory issues and personal responsibilities within the clinical setting certainly play a major role, models proposed by the machine learning community often neglect the importance of offering an intuitive or interpretable access to clinicians. An example is the publication by Biswal et al. [2019] where an electro-encephalographic recording (EEG) serves as the input of a deep learning pipeline with the diagnostic report as its output. From the machine learning point of view this is an exciting achievement, but unfortunately unlikely to find much support form the clinical community as no clinician would sign a diagnostic report without the possibility to check *why* the diagnosis is what it is. Similar examples can be found all over the intersection of machine learning and medicine – albeit not always justified. In essence, for bridging the gap between machine learning and the practice of medicine, proposed statistical models need to take into account the need of medical experts to comprehend an automated decision making process without having to become an expert of machine learning themselves. To that end, a

promising approach is to design models with *interpretability* in mind. But as this concept still defies a general formal definition, the best practical approach is working in close cooperation with experts in the field as models attractive to medical practitioners are likely those models showing a high degree of interpretability. Nevertheless, from a machine learning perspective well established methods with the potential to increase interpretability are readily available. A classical approach includes sparsity inducing penalties such as the ℓ_1-norm which has become very popular due to lasso [Tibshirani, 1996]. More recently, deep latent space models [Kingma and Welling, 2013] which identify low-dimensional representations of the data, often provide possibilities of understanding a statistical model *without* requiring a detailed understanding of the underlying statistical method.

The present work focuses on the intersection between machine learning and electrophysiology, a

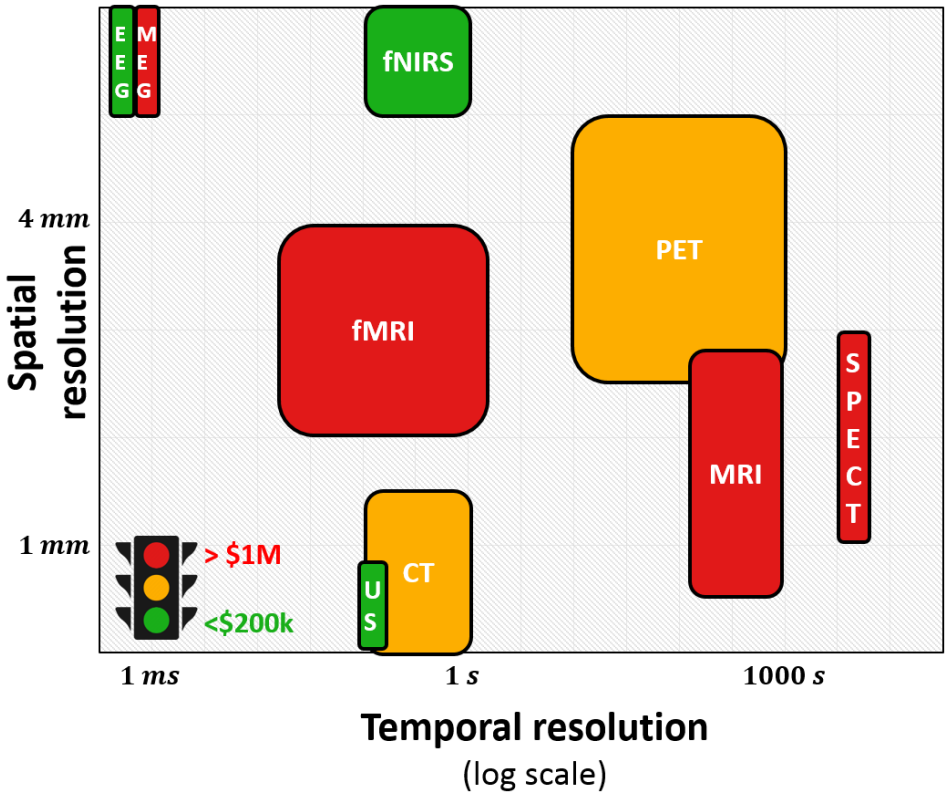


Figure 1.1.1 – Comparison of popular devices for analyzing brain structure and/or dynamics.

subfield of neurology. At the point of writing, no known drug is commercially available for slowing down or even stopping neurodegenerative processes in Alzheimer related disorders (ARD) or PD – in other word: beyond treating the symptoms there exists no standard of care to speak of. This certainly explains the heightened interest in any kind of modality able to provide – preferably noninvasively – measurements related to brain structure and dynamic brain activity. Figure 1.1.1 shows a collection of popular devices used for this purpose. Furthermore, for each device its temporal as well as spatial resolution is indicated along with a rough estimate of the cost of each device. The least expensive devices are electroencephalography (EEG), functional near infrared spectroscopy (fNIRS) and ultra sound (US) followed by computer tomography (CT) and positron emission spectroscopy (PET). Among the most expensive devices are magnetic resonance tomography (MRI), functional magnetic resonance tomography (fMRI), single photon emission computer tomography (SPECT) and

magnetoencephalography (MEG). Despite its low spatial resolution, EEG is a device that is especially attractive as it is non-invasive, not even requiring any contrast agent, inexpensive and mobile, and allows for extended periods of measurement. As EEG is a tool to study functional aspects of the nervous system it provides potentially massive amounts of data in form of multiple continuous wave forms. Considering that the established clinical practice for EEG analysis is visual inspection by a human, EEG offers much room for computer-aided analysis based on modern machine learning methods. Moreover, as EEG has been essential to the practice of neurology for over 60 years, there is a large body of research to rely on for developing automated solutions. But the long standing history of EEG makes it also necessary for new statistical models to build on pre-existing practices in order to be accepted by the medical community. This begins by possibly using band power as features for analyzing EEG, upon which the last fours decades of EEG research have been mainly focused, and mandates caution when proposing models which cannot be interpreted in the sense that a connection to previous clinical research cannot easily be established.

1.2 Interpretability and Challenges with Neurodegenerative Diseases

The whole spectrum of neurodegenerative diseases poses many challenges – from diagnosis, especially *early* diagnosis, to monitoring disease progression and potential treatment and to predict future cognitive health based on presently available brain measurements. As with many diseases, the root of this problem lies with its complexity: Both genetic predisposition as well as lifestyle may play important roles. But while neurodegenerative diseases ultimately impact the brain their origin might be elsewhere. PD for example, is currently believed to originate in the gut and might thus be present in an individual many years before inducing functional or structural changes in the brain [Kalia and Lang, 2016]. In PD, it is also the drastic heterogeneity of the disease which makes the development of biomarkers – for whatever task – a difficult undertaking with uncertain outcome but worthwhile nevertheless. In a review article of biomarker research for PD over the last two decades Yilmaz et al. [2019] conclude that "[d]espite intensive effort, biomarker research for the detection of prodromal stage, diagnosis and progression of Parkinson's disease . . . falls short of expectations." They go on by stating that "[a]lthough several biomarkers are currently available, none of them is specific enough for diagnosis, prediction of future PD or disease progression". A likely conclusion would be that successful biomarkers for PD will have to be compound biomarkers based on multiple modalities of which EEG might be one. But in more general terms – not specific to PD – the design of interpretable models poses an additional challenge in case of brain derived measurements which might be summarized in the following question: How strong is the foundation on which interpretability is built if the causal mechanisms underlying brain function remain obscured? In cardiology – although not without problems – machine learning algorithms characterizing electrocardiographic waveforms might be interpreted in light of the main function of the heart and the associated mechanical activity. But the brain has not a singular well defined function that could be considered its main function. Therefore, interpretable models of machine learning for electrophysiology might have an added layer of complication as not only black box algorithms should be avoided but because the system of interest itself – the brain – is to a large extent still a black box. This position is debated in an article with the title "Could a neuroscientist understand a microprocessor?" where the authors [Jonas and Kording, 2017] posit that if at all times the inputs and outputs of a microprocessor would be measured and made available for subsequent analysis it would still remain impossible to derive from these measurements alone a meaningful description of the hierarchy of information processing in the microprocessor.

1.3 Contributions and Outline of the Thesis

The thesis is divided into three parts. In the first part a method for non-convex regression is proposed based on the log-norm which is a parametric family of functions where the parameter γ is used to interpolate between the ℓ_1-norm and the ℓ_0-pseudonorm. This allows for the exploration of models of different complexity, trading off model bias and variance. In principle, model selection can then include the γ parameter along with the parameter λ which regulates the strength of the penalization for a given γ parameter. Effectively this makes the parameter space 2-dimensional which of course impacts on the time required for potential cross validation but on the other hand, it allows exploring a large range of sparsity patterns, approximating best subset selection for $\gamma \to +\infty$. With a variable transformation the non-convex log-penalty is mapped onto the convex ℓ_1 penalty while ensuring that the same transformation applied onto the convex objective function leads to a transformed objective where all stationary points are guaranteed to be global minimizers. As a result, standard algorithms of convex penalized regression can be applied to this non-convex problem. Following the "bet on sparsity principle", we show an application example based on spectral EEG features can provide a very sparse set of predictors highlighting important differences between a group of healthy controls and patients suffering from Parkinson's disease. With the possibility to explore sub-ℓ_1 penalties, increased sparsity directly impacts an the perceived level of interpretability by clinical experts.
The second part presents an extension of archetypal analysis [Cutler and Breiman, 1994]. While the original method describes a data set as a convex mixture of extreme representatives, where this mixture occurs in data space, the proposed extension learns an appropriate latent representation on which to perform archetypal analysis. This is especially important for image-like data where a convex mixture of images could not provide adequate results. As a solution to this problem, a deep learning approach based on an information bottleneck architecture is used to learn a latent representation allowing for additive mixing of latent representatives. As a result, the decoded images can be interpreted as mixtures of those images associated with the latent archetypes. This method of deep archetypal analysis is then applied on different data sets. Application cases include a task of sentiment analysis based on different facial expressions and a task of identifying archetypal molecules. Furthermore, EEG scalp topographies based on spectral features are analyzed in order to identify archetypes associated with cognitive decline in Parkinson's disease. This application highlights a promising approach in dealing with high dimensional time series EEG data while providing an intuitive understanding based on a highly structured latent space.
The third part presents an analysis where a measure of signal complexity of EEG waveforms, recorded at baseline, is used to predict cognitive decline over 3 years in patients with Parkinson's disease. It is shown that low complexity of EEG waveforms in the frequency range between $4-8$Hz are associated with cognitive decline over a period of 3 years. Interestingly, this association is significant only if EEG is measured in eyes open condition while eyes closed condition remains inconclusive. Considering that *different* brain networks are active depending on whether visual information is processed or not, this result is in agreement with similar research [Miraglia et al., 2016] where eyes open condition contained more information about brain health.

1.4 List of Publications

- *Cognitive decline in Parkinson's disease is associated with reduced complexity of EEG at baseline*
  SM Keller, U Gschwandtner, A Meyer, M Chaturvedi, V Roth, P Fuhr
  Accepted for publication in Brain Communications 2020.

- *Learning Extremal Representations with Deep Archetypal Analysis*
  SM Keller, M Samarin, F Arend Torres, M Wieser, V Roth
  Accepted for publication in International Journal of Computer Vision, 2020.
- *Tsallis entropy of EEG correlates with future cognitive decline in patients with Parkinson's disease: 1219*
  SM Keller, A Meyer, J Bogaarts, U Gschwandtner, P Fuhr, V Roth
  Movement Disorders 34, 2019.
- *Deep Archetypal Analysis* [**oral**, **best paper runner up**]
  SM Keller, M Samarin, M Wieser, V Roth
  German Conference on Pattern Recognition, 171-185, 2019.
- *Computational EEG in Personalized Medicine: A study in Parkinson's Disease*
  SM Keller, M Samarin, A Meyer, V Kosak, U Gschwandtner, P Fuhr, V Roth
  Machine Learning for Health, ML4H: a workshop at NeurIPS 2019.
- *Invexity Preserving Transformations for Projection Free Optimization with Sparsity Inducing Non-convex Constraints* [**oral**]
  SM Keller, D Murezzan, V Roth
  German Conference on Pattern Recognition, 682-697, 2018
- *Interfacial exchange interactions and magnetism of bilayers*
  R Yanes, E Simon, SM Keller, B Nagyfalusi, S Khmelevsky, L Szunyogh, U Nowak
  Physical Review B 96 (6), 064435, 2017.
- *Bayesian markov blanket estimation*
  D Kaufmann, S Parbhoo, A Wieczorek, SM Keller, D Adametz, V Roth
  International Conference on Artificial Intelligence and Statistics, PMLR 51:333-341, 2016.
- *Copula archetypal analysis* [**oral**]
  D Kaufmann, SM Keller, V Roth
  German Conference on Pattern Recognition, 117-128, 2015.

2 Related Work

In this chapter, related work on sparse regression under convex constraints is discussed, highlighting important models and algorithms as they relate to the proposed method for ultra-sparse regression under non-convex constraints. Furthermore, a short summary of important concepts of information theory are summarized which are the basis for understanding the information bottleneck principle and the deep information bottleneck. The variational autoencoder framework is also discussed as it shares computational and architectural similarities with the deep information bottleneck. The chapter closes with a brief introduction to electroencephalography, EEG for short, as electrophysiology will be the domain of various applications.

2.1 Statistical Learning With Sparsity

Machine learning models are mathematical models containing statistical assumptions about the data-generating process. If only a small fraction of the model parameters are *non-zero* such models are referred to as a *sparse* models. In statistics, one of the most prominent sparse models is the Lasso Tibshirani [1996]. In deep learning, inducing sparsity has helped design more compact architectures resulting in more efficient computation at test time Alvarez and Salzmann [2016].
While the initial motivation for turning to sparsity was improved *interpretability*, both statistical and computational efficiency have been recognized as possible benefits emerging from the paradigm of sparse modeling.

2.1.1 Regularization for Sparsity

Regularization is the process by which certain explanations, i.e. models, are favored relative to others. A very general preference to introduce to the model selection process relates to the philisophy of "William of Ockham"[1] and is known as *Occam's razor* or *lex parsimoniae*. In the context of model selection, Occam's razor states that one should *not use more parameters than necessary*, which indicates a trade-off between the complexity of a model and its explanatory power. It also gives rise to the notion of parsimonious models, i. e. models with optimal parsimony or models with just the right amount of predictors needed to explain the data well. Candidate parsimonious models are found by penalyzing the more complex models and then sorting potential models from least overfit on a

[1] English Franciscan friar William of Ockham (c. 1287–1347), a scholastic philosopher and theologian

test set to greatest. Finally, models with lowest overfitting score are usually the best candidates for models with optimal parsimony. In practice, sparsity regularizers are often used for model selection in order to identify the optimal *complexity-vs-predictability* trade-off. Especially regularization based on the ℓ_1 norm seems to be ubiquitous throughout many fields of mathematics and engineering.

Linear Regression Models and Least Squares

Let $y \in \mathbb{R}^n$ be a response vector and $X \in \mathbb{R}^{n \times p}$ be a matrix of predictors. The problem of *linear regression* is commonly written as

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\}, \tag{2.1}$$

where β_0 is an intercept term. The unknown parameters or coefficients β_j are typically estimated based on a set of training data $(x_1, y_1) \ldots (x_N, y_N)$. For solving problems of the form given in eq. 2.1, *least squares* is the most widely used estimation method, in which the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ are found by minimizing the residual sum of squares (RSS):

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2, \tag{2.2}$$

with $f(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j$. As long as the observations (x_i, y_i) used for training, represent independent random draws from their population, minimizing the residual sum of squares remains a reasonable approach.

Subset selection

Often the least squares estimate $\hat{\beta}$ in eq. 2.1 has *not* the highest prediction accuracy possible. This is due to least squares estimates usually exhibiting a *low bias/large variance* characteristic. But this also implies that prediction accuracy may be improved by trading-off some of the bias in order to reduce the variance of the predicted values. In statistics, this is known as the bias-variance trade-off. Practically, it means shrinking or setting some of the coefficients β_i to zero. As a result the overall prediction accuracy may improve.
Subset selection methods retain only a subset $p_{sub} < p$ of the total number p of variables. Least squares estimation is then applied only on the variables remaining in the subset in order to estimate the coefficients of $\hat{\beta}$. Effectively, this means setting $p - p_{sub}$ coefficients β_i to zero.

Best subset selection For low-dimensional problems a brute-force strategy can be applied where all possible subsets are evaluated. Although strategies have been proposed to perform best subset selection without having to try *all* possible subsets Furnival and Wilson [1974], such an approach remains limited in its ability to scale to very high dimensional problems.

Forward stepwise selection This approach refrains from examining *all* possible subsets of meaningful predictors and rather builds-up a "good" solution in a successive manner, starting from an empty set. Consequently, stepwise methods cannot guarantee optimality of the inferred solution under

any criterion function, but in practice they tend to provide useful results. While stepwise methods are not a brute-force strategy, they are *greedy* in the sense that they produce a *nested* sequence of models.
While different variants of forward stepwise selection exist, they generally start by adding an intercept term $\bar{y}$ to the empty active set. In subsequent steps, predictors are sequentially added if they improve the overall model fit. After having added a new predictor to the model, forward stepwise usually adjusts the current model by checking whether variables already in the active set need to be removed in light of the new predictor that has just been added. With this strategy of building up a final model, forward stepwise algorithms are in fact performing a combination of backward elimination and forward selection. Advantages over subset selection methods include (i) a reduced computational burden as well as (ii) lower variance, implying potentially improved prediction accuracy. Especially for a large number of predictors where computational cost would prohibit the use of subset selection methods, stepwise approaches present a viable alternative.

Incremental Forward Stagewise FS_ϵ

Unlike forward stepwise selection, forward stagewise adds variables to the model *without* adjusting the variables that have already entered the model in previous steps. It can therefore be considered an even more constrained strategy compared to forward stepwise selection. Forward stagewise also starts with an empty set and then proceeds to update the variable most correlated with the current residual. The update strategy of forward stagewise can be described as "slow fitting" as updates are performed in small increments of ϵ. The update procedure is continued till none of the variables have correlation with the residuals. With step counter k, and initialization $r^0 = \mathbf{y}$ and $\beta^0 = \mathbf{0}$, FS_ϵ is computed as follows:

Compute $j_k \in \underset{j\in\{1,\dots,p\}}{\operatorname{argmax}} \left|(r^k)^T \mathbf{X}_j\right|$ and update:

1. $\beta_{j_k}^{k+1} \leftarrow \beta_{j_k}^{k} + \epsilon \cdot sgn\left[(r^k)^T \mathbf{X}_{jk}\right]$
2. $r^{k+1} \leftarrow r^k - \epsilon \cdot sgn\left[(r^k)^T \mathbf{X}_{jk}\right]\mathbf{X}_{jk}$

where $\beta_{j_k}^k$ is the j_k^{th} coordinate of β^k. With the following sparsity properties [Freund et al., 2013]:

$$||\beta^k||_1 \leq k \cdot \epsilon \quad \text{and} \quad ||\beta^k||_0 \leq k$$

FS_ϵ is attractive from a statistical point of view due to its ability to provide regularized solutions.

Breiman's nonnegative garotte

While any form of subset selection effectively sets some of the coefficients β_i to zero by excluding them from the set of predictors, the nonnegative garotte *both* shrinks and zeroes coefficients to improve on the least squares estimate based on the full set of predictors. The nonnegative garotte estimator proposed in Breiman [1995] is optimizing the following objective:

$$\hat{\beta} = \min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} c_j x_{ij} \hat{\beta}_j^{LS} \right)^2 + \lambda \sum_{j=1}^{p} c_j \right\} \qquad (\lambda > 0) \tag{2.3}$$

This estimator takes as input the least squares solution $\hat{\beta}^{LS}$ based on the full set of predictors and chooses $c_1, \ldots, c_p$ with $c_j \geq 0$ for all $j \in \{1, \ldots, p\}$ in order to scale the least squares estimate. As some of the c_j might be zero, the garotte is implicitely selecting a subset of relevant predictors. While the results achievable with the garotte are simpler equations that often show better predictive accuracy (unless a large portion of the *true* predictors are non-negligible), it is not defined for $p > N$ as it relies on the least squares estimate $\hat{\beta}^{LS}$.

Tibshirani's lasso

The *least absolute shrinkage and selection operator* proposed by Tibshirani [1996], or "lasso" for short, is an estimator directly motivated by the nonnegative garotte [cite canada]. But in contrast to the garotte, which is defined only in case of $N < p$, the lasso estimator is also defined for $N > p$. In principle, this was made possible by removing the dependence on the least squares estimator and by introducing an absolute value constraint to the regression problem:

$$\hat{\beta}^{lasso} = \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (\lambda > 0) \tag{2.4}$$

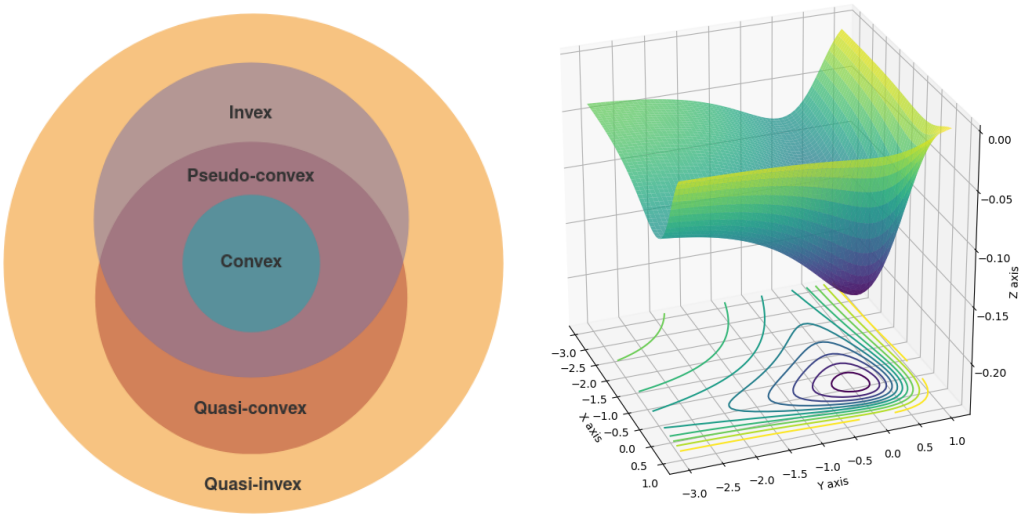
In the years since its inception, the lasso estimator has become *the* alternative to subset regression for obtaining a sparse or parsimonious model. Mathematically, the ℓ_1-penalized least squares problem in equation 2.4 is a convex problem, and together with the sparsity of the final solution this can be leveraged to greatly improve computational efficiency.

2.1.2 Invexity

Over time, a large class of optimization algorithms for solving convex optimization objectives over convex feasible regions has been established. However, while these assumptions often lead to a convenient treatment of the problem, many mathematical formulations of practical problems exist, where these requirements are not fulfilled. For such problems, finding common characteristics with convex problems would often help to establish theoretical results or develop algorithms. The strategy to generalize the definition of convexity while keeping – if possible – properties of interest, has lead to an important generalization of convex functions, establishing the notion of *invexity*. Invex functions have the property that all stationary points are global minimizers. The relation of convex and invex functions, as well as other extensions of convexity, are shown in figure 2.1.1a. Formally, invexity is defined as follows [Mishra and Giorgi, 2008]:

Definition 1. *Assume $X \subseteq \mathbb{R}^n$ is an open set. The differentiable function $f : X \to \mathbb{R}$ is invex if there exists a vector function $\eta : X \times X \to \mathbb{R}^n$ such that $f(x) - f(y) \geq \eta(x, y)^T \nabla f(y)$, $\forall x, y \in X$.*

The particular case of a (differentiable) convex function is obtained from definition 1 by choosing $\eta(x, y) = x - y$. Invex functions and quasi-convex functions are, according to figure 2.1.1a, two classes with only *partial* overlap. An example of an invex function which is *not* quasi-convex is shown in figure 2.1.1b: Clearly, the levelsets of that function are not convex but every stationary point is a global minimizer, making it an invex function.



(a) Relation between properties

(b) Invex but not quasi-convex

Figure 2.1.1 – Invexity is an extension of convexity where all stationary points are required to be global minimizers.

2.2 Concepts from Information Theory

Information theory traditionally revolves around three core questions: (i) what is information, (ii) how much information does a signal contain and (iii) how much information can be reliably transmitted over a channel? With regard to machine learning, especially questions (ii) and (iii) establish a strong link between information theory and core tasks of machine learning. In principle, this relation can be understood through the lens of "source coding" and "channel coding" [Duchi, 2016].

Source coding Source coding refers to question (ii) and can be seen as a *data compression* problem where information provided by a source S is first compressed, and then decompressed with the goal to recover the original information. In machine learning terms, this problem is similar to observing a sequence of data points $X_1,\dots,X_n$, distributed according to an unknown distribution $p(\mathcal{X})$. A core task of machine is then to construct a model that *efficiently* encodes that data, i. e. estimating an empirical version $\hat{p}$ of the unknown distribution p. In doing so, it is generally hoped to gain insights into the data generating mechanism.

Channel coding Channel coding refers to question (iii) and describes the data transmission problem of information theory. It is similar to source coding except that in between the compression and decompression step, a channel is introduced which provides an additional source of noise. As this channel is distorting the compressed information to be reconstructed during the decompression step, channel coding essentially studies redundancies necessary in order to ensure the recovery of the original information content. Through the lens of machine learning, an unknown compressor f transforms a sequence of data points $X_1,\dots,X_n$ resulting in noisy observations $f(X_1),\dots,f(X_n)$. These observations are transmitted via a channel $p(Y|f(X))$ to the decompressor, such that the original information has to be reconstructed from a sequence $X_1,\dots,X_n$). The goal of machine learning – when performing

estimation or inference – is usually estimating $f(\cdot)$ or any other aspects of the probability distribution of the source S. An important difference compared to information theory is that in machine learning the unknown compression function f is not a matter of choice but a given.

2.2.1 Information and Entropy

The central concept of information theory is *entropy*, which is closely linked to *information*. In the following, we give an axiomatic definition of information [Effenberger, 2013]. The concept of entropy of a random variable will then be introduced as the expected amount of information contained in a realization of that random variable.

Information Given a probability space (Ω, Σ, P), i. e. a space consisting of (i) a sample space Ω which contains all possible outcomes, (ii) an event space, i. e. a set of events Σ and (iii) a probability measure P assigning each event $\sigma \in \Sigma$ a probability between 0 and 1. How can the information content of an event contained in that space be defined? The following four axioms will lead to a natural definition of the information content h of an event. Here, natural refers to the fact that a *unique* mapping between the probability of the occurrence of an event and the non-negative real numbers will emerge, which will define the information content.

Axioms:

- h is non-negative: $h : \Sigma \to \mathbb{R}^+$
- h is sub-additive: For any two messages $\omega_1, \omega_2 \in \Sigma$ we have $h(\omega_1 \cap \omega_2) \leq h(\omega_1) + h(\omega_2)$, where equality holds if and only if ω_1 and ω_2 are independent
- h is continuous and monotonic with respect to the probability measure P
- Events with probability 1 are not informative: $h(\omega) = 0$ for $\omega \in \Sigma$ with $P(\omega) = 1$

For a mapping $h(\cdot)$ to fulfill these requirements simultaneously, the *only* choice is the logarithm, which leads to the following definition of information:

Definition 2. *Let (Ω, Σ, P) be a probability space. Then the information h of an event $\sigma \in \Sigma$ is defined as*

$$h(\sigma) := h(P(\sigma)) = -\log_b(P(\sigma)),$$

where b denotes the basis of the logarithm.

By choosing the basis of the logarithm to be $b = 2$ or $b = e$, the unit of h if fixed, leading to the unit of information of "bit" or "nat", respectively.

Entropy Generally, entropy is the expected information content. For a discrete random variable the entropy is defined as follows:

Definition 3. *Let X be a random variable on some probability space (Ω, Σ, P) with values in the integer or the real numbers. Then its entropy $H(X)$ is defined as the expected amount of information of X,*

$$H(X) := \mathbb{E}[h(X)].$$

Assuming a random variable X which takes on integer values only, the entropy $H(X)$ given in definition 3 can be evaluated to:

$$H(X) = \Sigma_{x \in \mathbb{Z}} P(X = x) h(P(X = x)) = -\Sigma_{x \in \mathbb{Z}} P(X = x) \log(P(X = x)) \tag{2.5}$$

For a real-valued, continuous random variable X, the so–called *differential entropy* is obtained, which is given as:

$$H(X) = \int_{\mathbb{R}} P(X = x) h(P(X = x))\, dx \tag{2.6}$$

Possible interpretations of entropy There are three general perspectives concerning the interpretation of entropy. (i) Entropy measures the *average* amount of information one expects to obtain from a given random variable X, if realized. (ii) Entropy is the average information *missing* if realizations of the random variable X are unknown. (iii) Entropy quantifies the average reduction of *uncertainty* about the possible values of a random variable X having observed one or more realizations.

Joint and conditional entropy Extending the definition of entropy given in definition 3 to two or more variables leads to the so–called *joint entropy* which quantifies the expected uncertainty in a joint distribution of random variables. Alternatively, joint entropy can be interpreted as quantifying the *expected information* of that joint distribution of random variables, as described in the previous paragraph.

Definition 4. *Let X and Y be discrete random variables on some probability spaces. Then the joint entropy of X and Y is given by*

$$H(X, Y) = -\mathbb{E}_{X,Y}[\log P(x, y)] = -\Sigma_{X,Y} P(x, y) \log P(x, y),$$

where $P_{X,Y}$ denotes the joint probability distribution of X and Y and the sum runs over all possible values x and y of X and Y, respectively.

The definition of entropy can also be extended to a conditional form $H(X|Y)$, where the *conditional entropy* of two random variables X and Y quantifies the expected uncertainty (or the expected information, depending on interpretation), remaining in a random variable X under the condition that Y was observed. By observing Y, the expected uncertainty in X might be reduced:

Definition 5. *Let X and Y be discrete random variables on some probability spaces. Then the* conditional entropy *of X given Y is given by*

$$H(X|Y) = -\mathbb{E}_{X,Y}[\log P(x|y)] = -\Sigma_{X,Y} P(x,y) \log P(x|y),$$

where $P_{X,Y}$ denotes the joint probability distribution of X and Y.

2.2.2 Quantifying shared information

Information can be shared between two (or more) random variables. Mutual information is an entropy-based measure quantifying the mutual dependence of random variables. Stated differently, mutual information measure how far two (or more) random variables are from being independent. In the following, the point-wise mutual information i is introduced. An expression for the mutual information of two random variables is then obtained as the expected value of the point-wise mutual information of all realizations.

Point-wise mutual information An expression for the shared information content of two events can be obtained based on the axioms and the definition of information given at the beginning of section 2.2.1. The definition of *point-wise mutual information i* is as follows:

Definition 6. *Let x and y be two events of a probability space (Ω, Σ, P). Then their point-wise mutual information i is given as:*

$$\begin{aligned} i(x;y) &:= -\log\left(\frac{P(x,y)}{P(x)P(y)}\right) && (2.7)\\ &= -\log\left(\frac{P(x|y)}{P(x)}\right) && (2.8)\\ &= -\log\left(\frac{P(y|x)}{P(y)}\right) && (2.9) \end{aligned}$$

where the sums are taken over all possible values x of X and y of Y.

Mutual information The expectation value of the point-wise mutual information of two random variables is the mutual information and quantifies the amount of shared information between the two variables:

Definition 7. *Let X and Y be two discrete random variables. Then the mutual information $I(X;Y)$ is given as the expected point-wise mutual information,*

$$\begin{aligned} I(X;Y) &:= \mathbb{E}_{X,Y}[i(x;y)] && (2.10)\\ &= \Sigma_y \Sigma_x P(x,y)\, i(x;y) && (2.11)\\ &= -\Sigma_y \Sigma_x P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) && (2.12) \end{aligned}$$

where the sums are taken over all possible values x of X and y of Y.

The extension to the continuous case in order to obtain the *differential mutual information* is straightforward. Generally, mutual information is both symmetric and non-negative but cannot be interpreted as a metric as it does not fulfill the triangle inequality. It is interpreted as the information (i. e. entropy) shared by the two variables. As $I(X;X) = H(X)$, entropy can be interpreted as "self-information".

Multi-information and conditional mutual information Mutual information can naturally be extended to cases of more than two random variables using conditional entropies. Mutual information of the multivariate case is often referred to as *multi-information*. For three random variables X_1, X_2, X_3 the multi-information is given by

$$I(X_1; X_2; X_3) := I(X_1; X_2) - I(X_1; X_2|X_3). \tag{2.13}$$

The last term in the above expression is the so–called "conditional mutual information" of X_1 and X_2 given X_3 and is defined as follows:

$$I(X_1; X_2|X_3) := \mathbb{E}_{X_3}[I(X_1; X_2)|X_3] \tag{2.14}$$

A possible interpretation of $I(X_1; X_2|X_3)$ is as a quantification of the average common information shared by X_1 and X_2 that is *not* contained in X_3. While the extension of mutual information to more than two variables is easily possible, the multivariate case might not be as straightforward to interpret: Mutual information $I(X;Y)$ is a non-negative quantity, but multi-information can also take on negative values.

2.2.3 Kullback-Leibler divergence

Given two probability distributions on the same base space Ω, the Kullback-Leibler divergence (KL-divergence) [Kullback and Leibler, 1951] is a measure for how one probability distribution is different from the second, defined as the reference probability distribution. The KL-divergence is also called *relative entropy*:

Definition 8. *Let P and Q be two discrete probability distributions over the same base space Ω. Then the KL-divergence of P and Q is given by*

$$D_{KL}(P||Q) := \Sigma_{\omega\in\Omega} P(\omega) \log \frac{P(\omega)}{Q(\omega)}$$

Properties of the KL-divergence: (i) non-negative $D_{KL}(P||Q) \geq 0$; non-symmetric $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

As the KL-divergence is non-symmetric and does not fulfill the triangle inequality, it does not constitute a measure in the mathematical sense. Definition 8 can also be written as $\mathbb{E}_P[\log P - \log Q]$ which implies an interpretation as "expected distance of P from Q", measured in terms of the information content. Alternatively, $D_{KL}(P||Q)$ is the average number of extra bits needed to code samples from P

using a code book based on Q. Expressing the KL-divergence in terms of entropies allows for the following equality:

$$D_{KL}(P||Q) = -\mathbb{E}_P[\log q(x)] + \mathbb{E}_P[\log p(x)] = H^{cross}(P,Q) - H(P) \tag{2.15}$$

Herein, $H^{cross}(P,Q)$ is the so–called *corss entropy* of P and Q, defined as follow:

$$H^{cross}(P,Q) = -\mathbb{E}_P[\log Q] \tag{2.16}$$

Based on cross entropy, a closed form of the KL-divergence can be obtained for many families of probability distributions: Given two normal distributions $P \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Q \sim \mathcal{N}(\mu_2, \sigma_2^2)$, the analytic form of the KL-divergence is as follows:

$$D_{KL}(P||Q) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right) \tag{2.17}$$

2.3 Variational autoencoder

Given a data set of high dimensional inputs $\{X_1, X_2, \ldots\}$ the task in generative modeling usually implies learning the distribution $P(\mathbf{X})$. But after successfully approximating $P(\mathbf{X})$ the output of such a model would be a probability assignment for each input datum, which might limit the range of useful applications of such a model. Arguably, a more interesting model would allow the sampling of new data which follows the learned distribution $P(\mathbf{X})$. Variational autoencoders [Kingma and Welling, 2013, Rezende et al., 2014] attempt to solve this problem by explicitly modeling $P(X|z;\theta)$, where z is a latent space variable and θ contains the parameters of the model. Assuming a distribution of $\mathbf{z} \sim P(\mathbf{z})$ from which can easily be sampled, the data distribution $P(\mathbf{X})$ can be written as follows:

$$P(\mathbf{X}) = \int P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z})\,d\mathbf{z} \tag{2.18}$$

The goal then becomes finding the parameters θ which maximize $P(\mathbf{X})$, where the approximation of $P(\mathbf{X})$ is performed based on samples of $\mathbf{z}$, such that:

$$P(\mathbf{X}) \approx \frac{1}{n}\sum_{i=0}^{n} P(\mathbf{X}|\mathbf{z_i}) \tag{2.19}$$

The problem with the approximation of $P(\mathbf{X})$ described in equation 2.19 is that a maximum likelihood method would require a large amount of samples and furthermore, as most $P(\mathbf{X}|\mathbf{z}) \approx 0$, such an approach would not be computationally efficient. The problem that the variational autoencoder (VAE) has to solve, is to learn a distribution $Q(\mathbf{z})$ where $\mathbf{z} \sim Q(\mathbf{z})$ generates $P(\mathbf{X}|\mathbf{z}) \gg 0$.

Assuming that such a distribution $Q(\mathbf{z})$ can be learned, the overall goal still remains the approximation of $P(\mathbf{X})$ in equation 2.19. But while calculating

$$P(\mathbf{X}) = \mathbb{E}_{\mathbf{z}\sim P(\mathbf{z})}P(\mathbf{X}|\mathbf{z}) \tag{2.20}$$

is still unpractical, with the distribution $Q(\mathbf{z})$ generating $P(\mathbf{X}|\mathbf{z}) \gg 0$, the following problem would be more efficient to compute:

$$P(\mathbf{X}) = \mathbb{E}_{\mathbf{z}\sim Q(\mathbf{z})} P(\mathbf{X}|\mathbf{z}). \tag{2.21}$$

This leads to the question how $\mathbb{E}_{\mathbf{z}\sim P(\mathbf{z})}$ and $\mathbb{E}_{\mathbf{z}\sim Q(\mathbf{z})}$ are related? The answer is provided by the following relationship, which is derived in [Kingma and Welling, 2013]:

$$\log P(\mathbf{X}) - D_{KL}[Q(\mathbf{z})||P(\mathbf{z}|\mathbf{X})] = \mathbb{E}_{\mathbf{z}\sim Q}[\log P(\mathbf{X}|\mathbf{z})] - D_{KL}[Q(\mathbf{z})||P(\mathbf{z})], \tag{2.22}$$

where $D_{KL}[Q(\mathbf{z})||P(\mathbf{z}|\mathbf{X})] = \mathbb{E}_{\mathbf{z}\sim Q}[\log Q(\mathbf{z}) - \log P(\mathbf{z}|\mathbf{X})]$ is the Kullback-Leibler divergence introduced in definition 8. This relation provides a path to solving the problem of maximizing $P(\mathbf{X})$ with respect to the model parameters θ. As the KL-divergence between two random variables is always non-negative, it follows from the above relation that $\log P(\mathbf{X}) > \log P(\mathbf{X}) - D_{KL}[Q(\mathbf{z})||P(\mathbf{z}|\mathbf{X})]$. The solution is thus to maximize the lower bound in order to get an estimate of $P(\mathbf{X})$. This leaves open the question how to obtain $Q(\mathbf{z})$ in the right side of equation 2.22? But instead of modeling $Q(\mathbf{z})$, we will think of this distribution as conditioned on $\mathbf{X}$, such that a neural network will in fact learn the distribution $Q(\mathbf{z}|\mathbf{X})$. Assuming $Q(\mathbf{z}|\mathbf{X})$ to be spherical Gaussian, i.e. $Q(\mathbf{z}|\mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}; 0, \mathbf{I})$, the neural network will output a mean $\boldsymbol{\mu}$, and a diagonal covariance matrix. The distribution $Q(\mathbf{z}|\mathbf{X})$ is the so–called encoder, as it encodes the input datum X into its latent representation $\mathbf{z}$.

The distribution left to learn is $P(\mathbf{X}|\mathbf{z})$, shown on the right side of equation 2.22, which will also be modeled by a neural network. Let $f(\mathbf{z})$ be the output of that network, and assume $P(\mathbf{X}|\mathbf{z})$ to be i.i.d. Gaussian. Then the datum $\mathbf{X}$ is given as $\mathbf{X} = f(\mathbf{z}) + \eta$, where $\eta \sim \mathcal{N}(0, \mathbf{I})$. This leads to a simple ℓ_2-training loss of $||\mathbf{X} - f(\mathbf{z})||^2$. The distribution $P(\mathbf{X}|\mathbf{z})$ is the so–called decoder, as it decodes the latent representation z into a reconstruction of the original input $\tilde{\mathbf{X}}$.
It was already shown in equation 2.17, that under certain circumstances, the KL-divergence can have closed form solutions. By choosing the prior distribution $P(\mathbf{z})$, shown in the right side of equation 2.22, to be $\mathcal{N}(0, \mathbf{I})$, the expression $D_{KL}[Q(\mathbf{z}|\mathbf{X})||P(\mathbf{z})]$ has in fact a closed form solution. Together with the relation $\mathbb{E}_{\mathbf{z}\sim Q(\mathbf{z}|\mathbf{X})} \log P(\mathbf{X}|\mathbf{z}) \propto ||\mathbf{X} - f(\mathbf{z})||^2$ this leads to the following loss function for the variational autoencoder:

$$\text{Loss}^{\text{VAE}} = ||\mathbf{X} - f(\mathbf{z})||^2 - \lambda \cdot D_{KL}[Q(\mathbf{z})||P(\mathbf{z})], \tag{2.23}$$

where the first term refers to the loss between $\mathbf{X}$ and its reconstruction $f(\mathbf{z}) = \tilde{\mathbf{X}}$ while the second term acts as a regularizer whose influence can be moderated via λ.

Note: While training the decoder network simply involves standard backpropagation, training the encoder is more intricate, as it is not obvious how to apply gradient descent through the latent samples $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. The problem lies with the sampling of the latent $\mathbf{z}$, which is stochastic in nature. A solution known as the "reparametrization trick" is presented in [Kingma and Welling, 2013]. Essentially, it allows for the encoder term $D_{KL}[Q(\mathbf{z})||P(\mathbf{z})] = \mathbb{E}_{\mathbf{z}\sim Q}[\log Q(\mathbf{z}) - \log(p(\mathbf{z}))]$ to express the gradient of the expectation as the expectation of a gradient.

2.4 The information bottleneck principle

The information bottleneck principle proposed by Tishby et al. [2000c] is a variational principle and designed to identify the *relevant information* in a signal $x \in X$. In that context, relevant information is defined as being the information contained within X that provides information about another signal $y \in Y$. For example, identifying the person shown on a facial image or simply identifying the gender of the person shown on that same image certainly involves different combinations of facial features. Generally, the goal of the information bottleneck (IB) is to find a short code for the signal X that preserves the maximum information about Y. The method itself is an information-theoretic approach, solving the following optimization problem:

$$\min_{P(T|X)} I(X;T) - \beta I(T;Y), \tag{2.24}$$

where X, Y and T are random vectors. The result of the optimization is the vector T which maximally preserves information about Y while simultaneously compressing X. The positive parameter β balances the trade-off between compression of X and preservation of Y: a high mutual information $I(X;T)$ corresponds to a low compression while a high value of $I(T;Y)$ indicates more relevant information about Y is preserved within T. As T, the compressed representation of X, is a function of X it is independent of Y given X, i. e. $T \perp\!\!\!\perp Y|X$. Consequently, the three variables can be written as the Markov chain $T - X - Y$. This implies that T cannot contain more information about Y than the original data X. The formulation of the IB given in equation 2.24 is general and does not depend on the type of the X, Y distribution. In the following, special cases of the IB assuming Gaussian random variables will be introduced along with an extension of the IB principle to deep neural networks.

2.4.1 Gaussian information bottleneck

In general, the IB problem in equation 2.24 cannot be solved analytically. But assuming X and Y to be joint multivariate Gaussian variables, Chechik et al. [2005] has shown that the problem becomes analytically tractable. From the assumption that

$$(X,Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}\right), \tag{2.25}$$

it follows that the solution T of equation 2.24 is also Gaussian distributed. The compressed representation T can be written as $T = AX + e$, where $e \sim \mathcal{N}(0, \Sigma_e)$ is independent of X. This implies that $T \sim \mathcal{N}(0, A\Sigma_X A^\top + \Sigma_e)$ and leads to the following optimization problem of the Gaussian information bottleneck:

$$\min_{A,\Sigma_e} I(X; AX + e) - \beta I(AX + e; Y) \tag{2.26}$$

It can be shown that the Gaussian information bottleneck problem has in fact an analytical solution: for a fixed β, equation 2.26 is minimized for $\Sigma_e = I$ while A admits to an analytic expression given in [Chechik et al., 2005].

2.4.2 Sparse Gaussian information bottleneck

Sparsity of the compression variable T in the Gaussian IB in equation 2.26 can be promoted – without imposing any norm penalty – by requiring the matrix A to have diagonal form, i. e. $A = \text{diag}(a_1, \ldots, a_n)$. Following [Rey et al., 2014], the sparsity requirement allows to rewrite equation 2.26 as a minimization problem over a diagonal matrix D, where $d_{ii} > 0$ and $D = A^\top A = \text{diag}(a_1^2, \ldots, a_n^2)$. The objective of the sparse information bottleneck is as follows:

$$\min_{A^\top A = \text{diag}(a_1^2, \ldots, a_n^2)} I(X; AX + e) - \beta I(AX + e; Y) \tag{2.27}$$

where $e \sim \mathcal{N}(0, I)$ is independent of X.

2.4.3 Deep information bottleneck

The *deep information bottleneck* is a variational approximation to the original information bottleneck of [Tishby et al., 2000c]. It was proposed by [Alemi et al., 2016] and allows to parameterize the information bottleneck model in equation 2.24 using a neural network. This leads to the following optimization problem

$$\min_{\phi, \theta} I_\phi(X; T) - \lambda I_{\phi,\theta}(T; Y), \tag{2.28}$$

where a parametric form of the conditionals $P_\phi(T|X)$ and $P_\theta(Y|T)$ is assumed and λ controls the degree of compression. This problem admits to a similar structure as the objective of the variational autoencoder in equation 2.23. Consequently, this problem can be solved using a similar network architecture, given appropriate expressions for the mutual information terms $I(X; T)$ and $I(T; Y)$. Following [Wieczorek and Roth, 2020], these take the form:

$$\begin{aligned} I(X; T) &= \mathbb{E}_{P(X)} D_{KL}(P(T|X) || P(T)) && (2.29) \\ I(T; Y) &= \mathbb{E}_{P(X,Y)} \mathbb{E}_{P(T|X,Y)} \log P(Y|T) + H(Y) && (2.30) \\ &= \mathbb{E}_{P(X,Y)} \mathbb{E}_{P(T|X)} \log P(Y|T) + H(Y) && (2.31) \end{aligned}$$

Similar to the variational autoencoder, the conditional in $P(Y|T)$ is obtained by sampling from the latent representation T.

2.5 Electroencephalography (EEG)

The electric time-dependent potential, measured non-invasively on the scalp, is a robust correlate of dynamic neocortical function. On average, a single electrode captures signals generated by tissue masses containing between roughly 100 million and 1 billion neurons [Nunez et al., 2006]. By homogeneously covering the scalp with electrodes, a spatially resolved picture of brain function can be provided. As a clinical tool, scalp EEG is used in monitoring and treating illnesses such as brain tumors, strokes, epilepsies, infectious diseases, mental retardation, severe head injury, drug overdose, sleep and metabolic disorders, and ultimately brain death.

2.5.1 The origin of human scalp EEG

Figure 2.5.1A shows the three primary divisions of the human brain, according to [Nunez et al., 2006]: (i) brainstem, (ii) cerebellum and (iii) cerebrum:

- **brainstem:** structure through which nerve fibers relay signals between spinal cord and higher brain centers (bidirectional)
- **thalamus:** relay station and important integrating center for all sensory input to the cortex, except for smell
- **cerebellum:** fine control of muscle movements in addition to a potential role in cognition

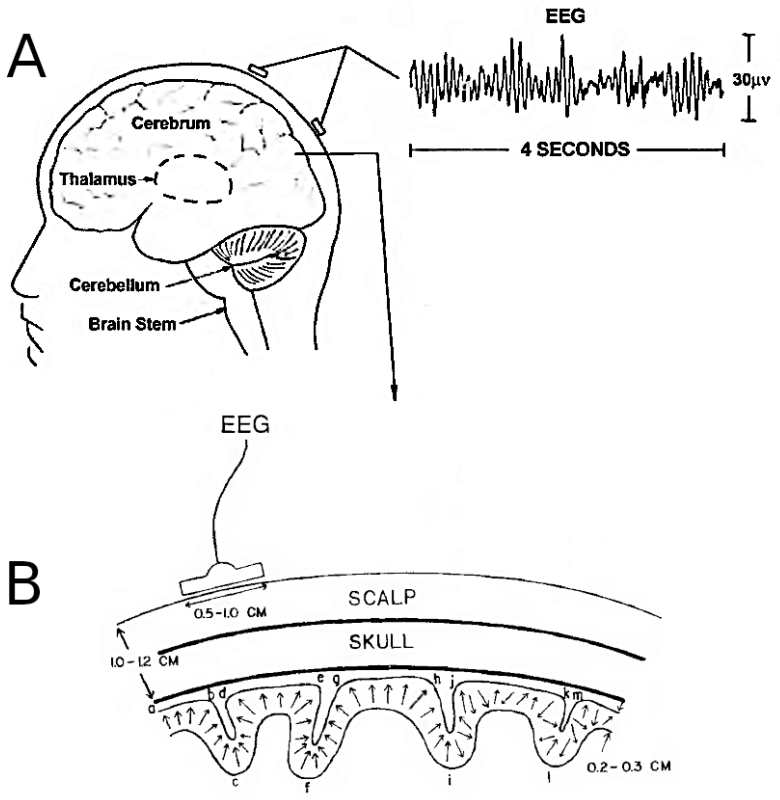


Figure 2.5.1 – (A) The three primary divisions of the human brain are brainstem, cerebellum and cerebrum. Scalp EEG originates mainly in the outer portion of the cerebrum, known as the cerebral cortex. (B) Section of cerebral cortex. Assemblies of pyramidal cells, which are the cortical generators of EEG, are shown as a layer of directed dipoles. (Source: Nunez et al. [2006])

The thalamus is embedded into the cerebrum, the largest of the three primary divisions of the human brain, which divides almost equally into two halves. The *cerebral cortex*, with a surface area between 1600 to $4000cm^2$, is the structure which likely generates most of the electric potentials measured by EEG. It is a layer of varying thickness, between 2 to 5 mm, representing the outer portion of the cerebrum. In humans, the cortex (also known as neocortex) contains about 10^{10} densely

inter-connected neurons, with the number of connections per neuron estimated between 10^4 to 10^5. Colloquially, the cells of the cortex are referred to as *gray matter* – although they turn gray only post-mortem when stained by anatomists. Just below the cortical layer begins the white matter which is composed of nerve fibers, also known as axons. In humans, white matter volume outweighs that of gray matter. Different regions of the cortex are connected through white matter fibers or corticocortical axons. Only a small percentage of white matter fibers connect the thalamus to the cortex. These are known as thalamocortical axons. In lower mammals, thalamocortical fibers are much more numerous compared to humans. Nevertheless, changes in human scalp EEG can be observed as thalamocortical connections are impeded, e. g. Parkinson's disease is associated with a loss of dopaminergic neurons in the substantia nigra, a region in the midbrain, impacting the functioning of basal ganglia-thalamocortical circuits.
Pyramidal neurons, shown in figure 2.5.2, have a pyramidal shaped cell body and are named accordingly. They form the most numerous excitatory cell type in mammalian cortical structures. Pyramidal cells have two distinct dendritic trees – the basal dendrites emerge from the base and the apical dendrites from the apex of the pyramidal cell body. Dendrides *receive* signals from other neurons and relay them to the cell body, while axons relay signals originating from the cell body to dendrides of other neurons. As pyramidal neurons are oriented parallel to each other they give the cortex a columnar structure. Importantly, this highly ordered arrangement potentially allows the summation of individual neuronal potentials such that a summed positive effect could reach the recordable range of a few microvolts. However, this also requires pyramidal cells to discharge *synchronously*. It was estimated in [Nunez et al., 2006] that 60,000,000 pyramidal neurons must be synchronously active in order to produce scalp potentials that can be recorded with non-invasive EEG. These cortical generators of EEG can be thought of as dipole layers with source strength varying as a function of cortical location as shown in figure 2.5.1B.

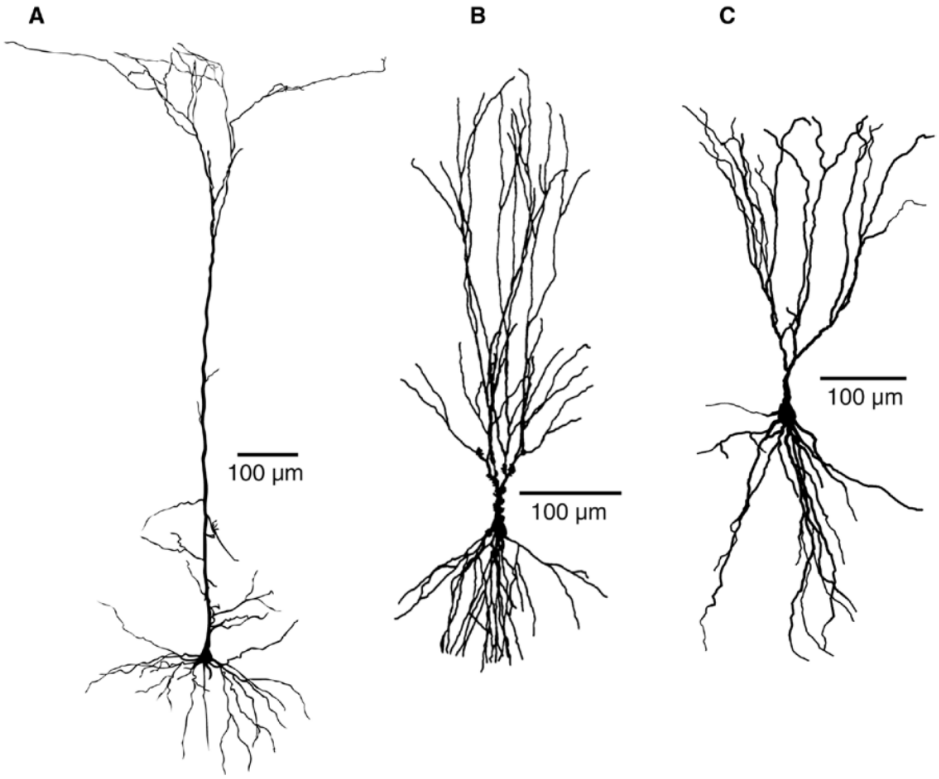


Figure 2.5.2 – Typical pyramidal neurons in different brain regions. (A) Pyramidal neuron in the rat somatosensory cortex. (B) Pyramidal neuron in the rat hippocampus. (C) Pyramidal neuron in the primary olfactory cortex of the mouse. The overall morphology is comparable in humans. (Source: Bekkers [2011])

3 Ultra-sparse Model Identification and Learning with Invexity

Machine learning methods are increasingly introduced as the future paradigm of data analysis even in areas characterized by high complexity and low fault tolerance, such as precision medicine, drug development and autonomous driving. But with low fault tolerance and the potentially high cost of individual errors (e. g. misdiagnosis or loss of life), the ability to understand – from a human perspective – the decision-making process of a trained machine learning model becomes an important requirement. The goal of *understanding* these decisions is often related to more tangible outcomes, e. g. ensuring safety, avoiding unethical decision-making or increasing fairness. The common basis for reaching these goals, and ultimately the enabler of understandable decision-making in machine learning, is *interpretability*. According to Biran and Cotton [2017], "interpretability is the degree to which a human can understand the cause of a decision". Typically, interpretability in machine learning is thought of as being a continuum where a higher degree of interpretability correlates with a better human understanding of why a model makes certain decisions. Unfortunately, a formal framework for evaluating or benchmarking machine learning models does not exist, mainly due to a lack of consensus regarding a formal definition of interpretability [Doshi-Velez and Kim, 2017]. However, the need for interpretable machine learning, especially when equating interpretability with accountability, remains undisputed. This is especially true for clinical decision-making, where possible "high stakes" scenarios are abundant and, as a consequence, the deployment of machine learning encounters considerable resistance. Interpretability might help overcome that resistance by offering a possibility for clinicians to interrogate, understand, debug and even improve machine learning models [Ahmad et al., 2018]. The qualitative definition of interpretability adopted here, refers explicitly to "the degree to which *a human* can understand the cause of a decision". The reference to a "human-in-the-loop" when defining interpretability implies a general assumption about the class of systems for which a decision-making process can be made interpretable: Social sciences have shown that explanations preferred by humans are "contrastive" [Lipton, 1990], i. e. humans tend not to ask why a certain decision was made, but rather why the decision made was taken *instead* of any other possible decision. This contrastive way of human thinking can provide satisfying explanations, and therefore interpretability, only in cases where potential underlying causes influencing the outcome are few. Otherwise, the sheer number of potential contrasts would quickly make it too confusing for humans to extract explanations that would also be considered interpretable. Consequently, humans "need to hope that the world is not as complex as it might be" [Hastie et al., 2015] in order to find interpretable explanations or, in more statistical terms, that the underlying data generating process is *sparse*. Constraining machine learning models to produce sparse solutions has proven to be a very successful approach in recent decades [Tibshirani, 1996, Friedman et al., 2001, Hastie et al., 2015]. The basic idea of these constraints is to

promote the estimation of model parameters having at most $q \ll p$ nonzero coefficients, where p is the dimensionality of the data. However, sparsity promoting constraints have also found successful application in low-rank matrix completion [Candès and Tao, 2010], structured sparsity [Jenatton et al., 2011] or sparse PCA [Mattei et al., 2016].
In this chapter, sparsity promoting constraints are explored in order to recover the underlying signal in a given data set. We focus on linear regression with a total of N observations of an outcome variable $y_{i=1\ldots N}$ and p associated predictor variables $x_i = (x_{i1}, \ldots x_{ip})^T$. In sparse linear regression our goal is to predict the outcome variable $\mathbf{Y}$ with high accuracy while using only a subset of the p predictor variables. In the following model of linear regression with noise term e_i,

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \tag{3.1}$$

providing a sparse estimate of $\boldsymbol{\beta}$ means identifying those dimensions j of the predictor variables $\mathbf{X}$ which can be ignored, i. e. set to zero, for the task of accurately predicting $\mathbf{Y}$.
Generally, contrastive explanations allow humans to ignore the *complete* explanation for a given decision and to concentrate on the differences between two explanations leading to two different decisions, i. e. two different contrasts. For example, a physician might want to know why a certain medication showed success in one patient but failed in another. An explanation highlighting the most prominent differences, e. g. the non-responding patient had a certain combination of genes making the drug less effective, would certainly be easy for a human to understand. With only a small number of nonzero predictors and only linear relations between predictor and outcome variables, promoting sparsity in linear regression models provides a setting in which formulating contrastive explanations is facilitated. Therefore, linear models, in combination with sparsity constraints, promote interpretability.

The problem solved by lasso [Tibshirani, 1996] is the following:

$$\min_{\beta} \ \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_{p=1} \tag{3.2}$$

The lasso penalty, i. e. the ℓ_1-norm, promotes sparse solutions by setting coefficients β_i to be exactly zero or shrinking them towards zero. As such, lasso performs both variable selection and regularization. Nevertheless, it is often thought of as a *convex surrogate* for best-subset selection,

$$\min_{\beta} \ \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_{p=0}, \tag{3.3}$$

where the ℓ_0-pseudonorm penalizes any non-zero coefficients of $\boldsymbol{\beta}$. However, the surrogate ℓ_1-norm can be sub-optimal for model selection as it both shrinks and selects, potentially leading to overly dense models [Zhang, 2010, Friedman, 2012]. In turn, this implies that improvements are possible compared to lasso, such that the same or even better prediction accuracies might be reached while simultaneously increasing the degree of sparsity of the estimated solution. Given the link between sparsity and interpretability, i. e. sparser solutions are potentially better interpretable, going below the ℓ_1-norm holds much promise. On the other hand, the optimization problem in equation 3.2 becomes non-convex when replacing the $\ell_{p=1}$-norm with a "sub ℓ_1" penalty. In that case, a major disadvantage is that computationally attractive algorithms such as forward stagewise and Frank Wolfe, both popular gradient based projection free optimization algorithms, cannot be used anymore due to their reliance on convex constraints.
In this chapter we propose a method to extend the applicability of these algorithms to problems of

the form

$$\min_x f(x) \quad s.t. \quad g(x) \leq \kappa, \tag{3.4}$$

where $\kappa \in \mathbb{R}$, $x \in \mathbb{R}^p$, $f(x)$ is a differentiable invex objective function and $g(x)$ is an arbitrary, typically non-convex constraint. The forward stagewise algorithm is especially of interest due to its close relation to lasso: it can be shown that incremental forward stagewise regression solves the monotone lasso problem, a version of the lasso that enforces monotonicity [Hastie et al., 2007]. In this context, monotonicity refers to the coefficient paths which are constrained to be monotone non-decreasing. According to [Hastie et al., 2007], "[t]hese monotone paths are exactly equivalent to the paths of the forward-stagewise algorithm". But comparing the original lasso problem [Tibshirani, 1996] to forward stagewise regression it turns out that these problems optimize different objectives: forward stagewise tries to minimize the arc-length of the solution path while lasso optimizes the cost function at each point of the solution path. Consequently forward stagewise produces smoother solution paths compared to lasso, while retaining most of its properties (incremental forward stagewise solutions converge to *monotone* lasso solutions as the increment or step size ϵ tends towards zero [Hastie et al., 2007]). A generalized version, which works for generic convex problems, is described in [Tibshirani, 2015]. Frank-Wolfe was introduced in [Frank and Wolfe, 1956]. But contrary to forward stagewise, Frank Wolfe is a point-estimator and will therefore *not* construct a solution path. Both forward stagewise and Frank-Wolfe linearise the target function at each step and need a constraint for which the linearized problem is easily solved in order to be efficient. This is usually only the case for convex constraints. The concept of invexity, which is a generalization of convexity and ensures that all local optima are also global optima, was introduced in [Ben-Israel and Mond, 1986] – with a minor correction provided by Giorgi [1995] – and described in detail in [Mishra and Giorgi, 2008]. Overall, invexity is a concept not very well known to the machine learning community, although it is occasionally applied in the domain of optimization, e.g. [Dinuzzo et al., 2011, Li et al., 2014].

In the following, we focus on problems of the form of equation 3.4. We will provide a theorem which defines a class of monotone component-wise transformations $x_i = h(z_i)$. Applied to the non-convex constraint $g(x)$, these transformations produce a *convex* constraint $G(z) = g(h(z))$. Assuming invexity of the original function $f(x)$ as in equation 3.4, that same transformation $h(\cdot)$ produces a transformed objective $F(z) = f(h(z))$ which is also invex. As a consequence, for algorithms relying on a non-zero gradient ∇F to produce new update steps, invexity ensures that these algorithms will move forward as long as a descent direction exists. Here, we specifically focus on constraints $g(x)$ which can be made quasi-convex in a new variable $z \in \mathbb{R}^p$ by way of coordinate-wise transformations $x_j = h(z_i)$, $j = 1, \ldots, p$. This subset of problems is still relatively large. For instance, many problems in *sparse regression* with "sub ℓ_1"-penalties fall into this class. The main advantage of having quasi-convex constraints in the new variables is that certain optimization techniques, which rely on a convex feasible region, can now be applied. In particular, we focus on projection-free algorithms of the *forward stagewise* type or on the highly related class of Frank Wolfe algorithms. These algorithms have properties which make them interesting in light of practical applications. For instance, forward stagewise methods are closely related to well-studied boosting algorithms, they are conceptually simple and computationally efficient, and they allow a dense sampling of the whole *solution path* – i. e. the set of all solutions for a sequence of increasing constraint values – without any additional computational costs. Frank-Wolfe algorithms, on the other hand, are well studied from a theoretical point of view, and (local) convergence guaranties for problems involving non-convex functions f are available. In case the full solution path is of interest, and arguably in most practical applications this is the case (e. g. performing model selection for the constraint value κ), the use of variable

transformations $t : \mathbb{R}^p \to \mathbb{R}^p,\ x \mapsto z$ might induce local extrema in the minimization problem in eq. (3.4). This in turn would pose a problem for gradient-based optimization strategies. We will show that this problem can be circumvented for transformations $h^{-1} : \mathbb{R}^p \to \mathbb{R}^p,\ x \mapsto z$ that fulfill certain criteria – basically invertability and smoothness of both h an h^{-1}. Such transformations keep the function $F(z) = (f \circ h)(z) = f(h(z))$ invex in the new variable z. This means that gradient-based methods evaluating $\nabla_z F(z)$ cannot get stuck in local minima. And assuming a sufficiently relaxed constraint value, the constructed solution path will indeed end at the unconstrained solution $\min_z F(z)$, provided that the minimum exists. Despite the fact that $F(z)$ has this invexity property, we cannot guarantee *joint invexity* of $F(z)$ and $G(z) = g(h(z))$, which would be necessary to prove that the solution path connects pointwise-optimal solutions. Panels (a) and (b) of Figure (3.0.1) show a Gaussian function $f(x)$ (solid lines) where the positions of the respective minima differ only slightly. The problem is minimizing the function $f(x)$ over the depicted non-convex feasible region $g(x)$ (dotted lines). The result of the variable transformation, the functions $F(z)$ and $G(z)$, are shown in panels (c) and (d) where the constraint has now become convex. Minimization is performed in the z-space where one observes that the solution paths have changed considerably for only minor changes in the location of the minimum. In general, it would not be realistic to expect a guarantee of reconstructing the optimal solution path for a non-convex optimization problem. Nevertheless from a practical point of view the correctness of both the starting point and the end point of the solution path is of considerable value as all solution paths will eventually converge towards the same end point as the constraint is relaxed. A proof is provided in the appendix A.1.1. Furthermore any local extrema or saddlepoints of $F(z)$ introduced at the border of the constraint region can always be escaped as κ is increased and no "re-starts" will ever become necessary. This too is a direct consequence of the invexity of $F(z)$. The solution path in the original variable is obtained by transforming back into x-space.

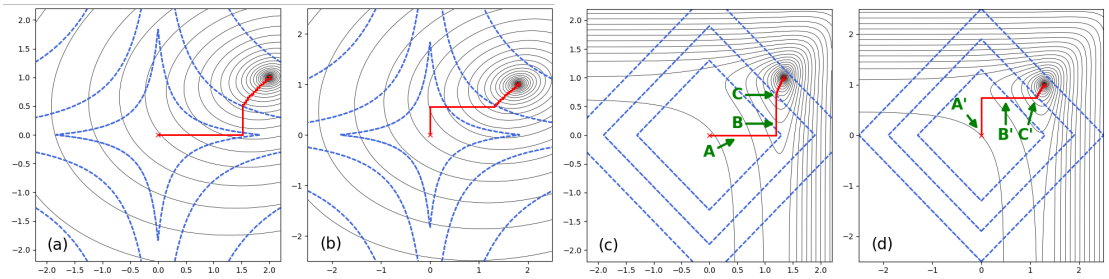


Figure 3.0.1 – The minimum of the Gaussian function $f(x)$ in panel (a) is located at $(2.0/1.0)$ and at $(1.8/1.0)$ in panel (b). After transforming $f(x)$ and the feasible region $g(x)$ into the new variable z, $F(z) = (f \circ h)(z) = f(h(z))$ has become invex whereas $G(z) = (g \circ h)(z) = g(h(z))$ is now convex (panels c,d). Using the forward stagewise algorithm which requires a convex feasible region, optimization is performed in z-space. Due to the differently located minima, the solution path for panels (a,c) is constructed by including first the dimension x_1 (A) into the active set followed by x_2 (B) once the correlation of x_1 with the residual has become small enough. In panels (b,d) the reverse sequence is observed: first x_2 (A') is included followed by x_1 (B'). From points C, C' onwards the weights of both dimensions is increased in alternating fashion till the minimum is reached for a sufficiently relaxed constraint value κ. The difference in solution paths reflects the non-convex nature of the optimization problem and illustrates that solution paths in such a setting can in general not be guaranteed to connect pointwise-optimal solutions.

	Constraint $g(x)$	Transformation $x_j = h(z_j)$	Transformed Constraint $g(z)$
Log Penalty	$\sum_j \log(\gamma x_j + 1)$	$x_j = \frac{1}{\gamma}(\exp(z_j) - 1)$	$\sum_j z_j$
Log-Group Lasso	$\sum_i w_i \log(\gamma \|\|x_{I_i}\|\|_\infty + 1)$	$x_j = \frac{1}{\gamma}(\exp(z_j) - 1)$	$\sum_i w_i \|\|z_{I_i}\|\|_\infty$
Inverse Tangent Penalty	$\sum_j \operatorname{atan}(\frac{1+2\gamma x_j}{\sqrt{3}}) - \frac{\pi}{6}$	$x_j = \frac{\sqrt{3}\,\tan(z_j + \frac{\pi}{6}) - 1}{2\gamma}$	$\sum_j z_j$
Rational Polynomials	$\sum_j \frac{x_j}{1+\gamma x_j/2}$	$x_j = \frac{2z_j}{\gamma z_j - 2}$	$\sum_j z_j$

Table 3.1 – Overview of non-convex constraints, element-wise transformations and convex transformed constraints. It is assumed that $x_j \geq 0 \ \forall j$ and $z_j \geq 0 \ \forall j$ for the log and group log penalties, $\frac{\pi}{3} \geq z_j \geq 0 \ \forall j$ for the inverse tangent penalty and $\frac{2}{\gamma} \geq z_j \geq 0 \ \forall j$ for the rational polynomials penalty.

3.1 Transformations ensuring invexity of the objective

Invexity, as defined in definition 9, is an extension of convexity, i. e. every convex function is also an invex function. To recover the usual definition of a differentiable convex function one has to set $\eta(z, z') = z - z'$ in definition 9.

Definition 9. *Let $Z \subseteq \mathbb{R}^p$ be an open set. The differentiable function $F : Z \to \mathbb{R}$ is invex if there exists a vector function $\eta : Z \times Z \to \mathbb{R}^p$ such that $F(z) - F(z') \geq \eta(z, z')^T \nabla_z F(z'), \ \forall z, z' \in Z$.*

An alternative definition of invexity is given by Ben-Israel and Mond [1986] in the following theorem:

Theorem 1. *F is invex if and only if every stationary point is a global minimum.*

The proof can be found in [Ben-Israel and Mond, 1986]. We now define a class of transformations $h(\cdot)$ under which the invexity property of function f is preserved.

Theorem 2. *Let $X, Z \subseteq \mathbb{R}^p$ be open sets, and let $f : X \to \mathbb{R}$ be invex and differentiable. Let h be a differentiable bijective function $h : Z \to X, z \mapsto x$ with differentiable inverse h^{-1}. Then $F(z) = (f \circ h)(z) = f(h(z))$ is invex on Z.*

Proof. Invexity of $f = F \circ h^{-1}$ and the chain rule imply

$$\begin{aligned}(F \circ h^{-1})(x) - (F \circ h^{-1})(y) &\geq \eta(x, y)^T \nabla (F \circ h^{-1})(y) \\ &= \eta(x, y)^T \nabla F(h^{-1}(y)) \nabla h^{-1}(y).\end{aligned}$$

If $\nabla_z F(z^\star) = 0$, there exists an $y \in X$ s.t. $h(z^\star) = y$ and $h^{-1}(y) = z^\star$. It follows that $(F \circ h^{-1})(x) \geq F(z^\star) \ \forall x \in X$. Since h is one-to-one, $F(z) \geq F(z^\star) \ \forall z \in Z$. Hence, every stationary point of F yields a global minimum on Z, so F is invex on Z. □

Note: As the proposed optimization method with non-convex constraints relies on every stationary point of the unconstrained objective function being a guaranteed global minimum, the class of quasiconvex functions is in general not permissible here. Furthermore the class of invex functions and the class

of quasiconvex functions have only partial overlap, e.g. $f_1(x) = x^3, x \in \mathbb{R}$, is quasiconvex, but not invex, since $x = 0$ is a stationary point which is not a minimum point whereas $f_2(x, y) = x^3 + x - 10y^3 - y$ is invex as no stationary points exist, but not quasiconvex: following Definition 9 and choosing $z' = (0, 0)$, $x = 2$ and $y = 1$, yields $f(x, y) - f(z') < 0$ but $(x - y)\nabla f(z') > 0$. On the other hand, the class of pseudoconvex functions presents an unnecessary restriction as every pseudoconvex function is invex whereas the reverse is not true [Giorgi, 2008].

A particular sub-class of bijective functions $h(\cdot)$ on $Z \subseteq \mathbb{R}^p$ consists of strictly monotone increasing functions that are defined in a coordinate-wise manner, i.e. $h_j = h(z_j)$, $j = 1, .., p$ and map 0 onto itself[1], i.e. $h(0) = 0$. We will restrict ourselves to this type of coordinate-wise transformations for the remainder of this chapter.

Application Example I: Logarithmic Constraints for Sparse Regression.

In the context of regression, one possibility to ensure interpretability is to enforce sparsity of the coefficients. We now discuss transformations $h(\cdot)$ for families of constraint functions that are frequently used in the context of sparsity (for a list of example functions, see Table 3.1). One interesting class of non-convex constraints uses the concavity of the logarithm. These logarithmic constraints arise naturally as a means to interpolate between the ℓ_0-pseudonorm and the ℓ_1-norm:

$$g(x; \gamma) = \sum_{j=1}^{p} \log(\gamma |x_j| + 1). \tag{3.5}$$

Used as a regularizer, such a constraint will likely increase the sparsity of the solution of a linear regression problem even in comparison to lasso. The major problem for the application of our method to a regression setting is the domain of the x_j. Due to the required monotonicity of the constraint function $g(\cdot)$ (see theorem 2), it is not possible to use penalties containing the absolute value function $|\cdot|$ on domains other than $\mathbb{R}_{\geq 0}$. This prevents the direct applicability of the method to regression and other settings where negative values can naturally occur. For regression, this problem can be circumvented by doubling the number of predictors as described in the *monotone lasso* [Hastie et al., 2007]. There x is replaced by $x_j^+ = \frac{1}{2}(|x_j| + x_j)$ and $x_j^- = \frac{1}{2}(|x_j| - x_j)$ which implies that $x_j^+ \geq 0, x_j^- \geq 0 \ \forall \ j$. The problem is thus redefined as $f(x = x^+ - x^-)$ s.t. $g(|x| = x^+ + x^-)$. This leads to a regression problem which is entirely defined on $\mathbb{R}_{\geq 0}$. Applied to least squares regression with a log-constraint, we obtain

$$\min_x \quad \sum_{i=1}^{n} (b_i - [\sum_{j=1}^{p} a_{ij} x_j^+ - \sum_{j=1}^{p} a_{ij} x_j^-])^2 \tag{3.6}$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \log(\gamma(x_j^+ + x_j^-) + 1) \leq \kappa \text{ and } x_j^+ \geq 0,\ x_j^- \geq 0 \forall\, j = 1, \ldots, p. \tag{3.7}$$

If one analyses the KKT conditions of this problem, it can be seen that $x_j^+ > 0$ implies $x_j^- = 0$. A detailed proof for general $f(x)$ and $g(x)$ can be found in the appendix A.1.2. This allows us to write

[1] This is a crucial property in the context of sparse regression, as only then the sparsity patterns in x- and z-space are identical.

$g(x^+ + x^-) = g(\tilde{x})$ with $\tilde{x} = [x^+, x^-]$:

$$\sum_{j=1}^{p} \log(\gamma(x_j^+ + x_j^-) + 1) = \sum_{j=1}^{2p} \log(\gamma \tilde{x} + 1) \leq \kappa. \tag{3.8}$$

By substituting $\tilde{x}_j = h(\tilde{z}_j) = \frac{1}{\gamma}[\exp(\tilde{z}_j) - 1]$ in eq. (3.8) this expression is transformed to the convex lasso constraint on $\mathbb{R}^{2p}_{\geq 0}$, i.e. $G(\tilde{z}; \gamma) = \sum_{j=1}^{2p} \tilde{z}_j$, while the loss function in eq. (3.6) goes from convex to being invex. Theorem 2 requires the existence of the gradient of $F(z)$ at all points, which is not the case for $F(z = 0)$ if we consider only the closed set $\mathbb{R}^{2p}_{\geq 0}$. However, we can always fulfill this requirement by simply enlarging the domain of h^{-1} to include the neighbourhood of 0 (this requires f to be differentiable at 0, but we already stated this condition previously). This is always possible, since $h^{-1}(\tilde{x}) = \log(\gamma \tilde{x} + 1)$ is well-defined and differentiable at $x = 0$. Note that this argument would not be valid for ℓ_p-pseudonorms with $0 < p < 1$, since these functions are not differentiable at zero.

Application Example II: Sparse Information Bottleneck.

The information bottleneck was introduced by Tishby et al. [2000a]. Its goal is to compress a random variable X into a variable T, such that the mutual information $I(X;T)$ is minimized while the mutual information with a target variable Y, i.e. $I(Y;T)$, is maximized. In the *Gaussian IB*, one considers jointly Gaussian random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. It follows that the optimal compression T is a noisy projection, $T = AX + \eta$ with independent standard normal noise $\eta \sim N(0, I)$ [Chechik et al., 2005]. The original formulation only allowed for discrete variables, but [Chechik et al., 2005] extended the method to Gaussian variables while [Rey and Roth, 2012] proposed an extension to variables with a joint Gaussian copula. The problem we want to improve upon, with the proposed method, is the sparse meta-gaussian information bottleneck discussed in [Rey et al., 2014], which also introduces sparsity to the compressed variable T. In this variant of the Gaussian IB, the actual optimization problem can be formulated as follows

$$\min_a \quad \log|P_{X|Y} D_a + I| - \log|P_X D_a + I| \quad s.t. \quad \log|P_X D_a + I| \leq \kappa, \tag{3.9}$$

where P_X is the correlation matrix of X, and $P_{X|Y}$ denotes the conditional correlation of X, given Y. Note that for the case $P_X = I$, the penalty in equation 3.9 reduces to

$$g_\gamma(x) = \sum_{j=1}^{p} \log(\gamma x_j + 1) \tag{3.10}$$

on the non-negative reals, which in turn allows for an element-wise transformation $x_j = h(z_j; \gamma) = \frac{1}{\gamma}(\exp(z_j) - 1)$. In general, the transformed constraint has a more complicated form with curved level sets, for which we propose a solution in section 3.2.1. A visual representation of the level-sets is shown in figure A.1.1 of the appendix.

Group-sparse constraints.

The group lasso method is a generalization of the lasso to allow for sparsity on the level of grouped variables [Yuan and Lin, 2006]. One frequently used version of the group lasso uses a $\ell_{1,\infty}$ "block norm"

$g(x) = \sum_j \|y_j\|_\infty$, where a vector y_j contains a group of variables. The constraint families in equations 3.5, 3.14, 3.15 are all of the form $g(x) = \sum_j h^{-1}(x_j)$ and can be extended to group-sparse versions that are "below" the group-lasso block-norm, if we substitute x_j by $\|y_j\|_\infty$. Since the infinity-norm of the group is defined as the maximum within the group, and since we assume that we operate on the non-negative reals, for a group y_j containing l variables we have $\|y_j\|_\infty = \max\{y_{j1}, \ldots, y_{jl}\}$. By using a strictly monotone increasing element-wise transformation $y_{ji} = h(z_{ji})$ we arrive at

$$\begin{aligned} h^{-1}(\|y_j\|_\infty) &= h^{-1}(\max\{h(z_{j1}), \ldots, h(z_{jl})\}) && (3.11)\\ &= h^{-1}(h(\max\{z_{j1}, \ldots, z_{jl}\})) && (3.12)\\ &= \max\{z_{j1}, \ldots, z_{jl}\}, && (3.13) \end{aligned}$$

which again is simply the non-negative version of the group-lasso constraint.

Other suitable non-convex constraints

In applications like image denoising, several authors, e. g. [Lanza et al., 2015], have proposed to use sparsity penalties that either involve the inverse tangent function or rational polynomials.

$$g_\gamma(x) = \sum_{j=1}^{p} \operatorname{atan}\left(\frac{1+2\gamma|x|}{\sqrt{3}}\right). \tag{3.14}$$

$$g_\gamma(x) = \sum_{j=1}^{p} \frac{|x|}{1+\gamma|x|/2}. \tag{3.15}$$

Together with the augmentation trick in equation 3.7, and similar to the log-penalties discussed above, both versions can be transformed to lasso constraints with variable transformations $h(\cdot)$ whose inverses are differentiable at zero.

3.2 Algorithms

Before discussing the algorithms, we make the following remarks.

1. The algorithms we are presenting require a convex constraint G. We consider non-convex constraint functions g which are transformed into convex functions $G = g(h(\cdot))$ by a mapping $h(\cdot)$ that fulfills the requirements of theorem 2. For an initially invex function f, invexity of the objective function $F = f(h(\cdot))$ is preserved under that same mapping h.

2. The algorithms presented here will perform update steps as long as there is a non-zero gradient. The existence of a continuation criterion for the presented algorithms is guaranteed by the invexity of the objective function: As the algorithms only consider ∇F, update steps are performed as long as the constraint remains active.

3.2.1 Forward Stagewise

General forward stagewise procedures

For two convex functions $f(x)$ and $g(x)$ the general form of the forward-stagewise method is described in [Tibshirani, 2015]. Here we assume g to be *non-convex*. Applying the doubling of predictors $x \mapsto \tilde{x} = [x^+, x^-]$ described in "Application Example I" and substituting $\tilde{x}_j = h(z_j)$ in f, we obtain the invex function $F(z) = (f \circ h)(z)$ and the convex function $G(z) = g(h(z))$. It is now possible to use projection-free optimization methods like forward stagewise to find the minimum of $F(z)$ constrained by $G(z)$. The general forward stagewise procedure is:

Initialize $z^{(0)} = 0$. Repeat while $G(z) < \kappa$:

$$L^{(j)} = \nabla F(z)\big|_{z=z^{(j)}}, \tag{3.16}$$

$$\Delta_\beta = \underset{\beta}{\operatorname{argmin}}\, \beta^t L^{(j)} \quad s.t. \quad G(\beta) \le \epsilon \text{ and } \beta \ge 0, \tag{3.17}$$

$$z^{(j+1)} = z^{(j)} + \Delta_\beta. \tag{3.18}$$

In some cases, the increment Δ_β can be found analytically. Tibshirani [2015] provides a general discussion of penalty functions that have this property. For the lasso constraint on the non-negative reals, i. e. $G(z;\gamma) = g(h(z;\gamma)) = \sum_j z_j$, the increment has the form:

Initialize $z^{(0)} = 0$. Repeat while $G(z) < \kappa$ and $L_i^{(j)} < 0$ for any i:

$$L^{(j)} = \nabla F(z)\big|_{z=z^{(j)}}, \tag{3.19}$$

$$i = \underset{i}{\operatorname{argmin}}\, L_i^{(j)}, \tag{3.20}$$

$$z^{(j+1)} = z^{(j)} + \epsilon \cdot e_i, \tag{3.21}$$

where e_i is the unit vector for dimension i. Note that log-constraints for sparse "sub-lasso" regression in equation 3.5, as well as inverse tangent penalties and rational polynomials given in table 3.1, can be easily transformed to this non-negative lasso setting.

Forward stagewise with first order Taylor approximation

In some cases, however, the transformed constraint $G(z)$ is convex but has a more complicated structure compared to the ℓ_1-region. For this case we propose the *Forward stagewise with first order Taylor approximation*. With $\epsilon \ll 0$ we are allowed to replace $G(\beta)$ in Eq. 3.17 with its Taylor expansion around zero. If the transformation $h(\cdot)$ has succeeded in mapping the non-convex constraint onto a convex constraint *close* to the ℓ_1 shape, truncating the series after the linear term is often sufficient to obtain a good approximation. For instance in case of the "Sparse Information Bottleneck", using only the linear term, the increment Δ_β is approximated as follows $\log|PD + I| \approx trace(PD) = trace(D)$. The latter identity follows from P being a correlation matrix with ones on the diagonal. In practice, this approximation leads to virtually indistinguishable results when compared to the numerically computed "true" solution of Δ_β, but at a considerable computational speed-up. We demonstrate this in the

experiment section, figure 3.3.5.

3.2.2 Frank-Wolfe algorithm

Frank-Wolfe was introduced in Frank and Wolfe [1956]. Recently it has increased in popularity due to its ability to efficiently solve a wide range of problems, as reviewed in [Jaggi, 2013]. Contrary to forward stagewise, Frank Wolfe is a point-estimator and will *not* construct a solution path. Both forward stagewise and Frank-Wolfe linearise the objective function at each step and need a constraint for which the linearised problem is easily solved in order to be efficient. This is usually only the case for convex constraints. Formally, Frank-Wolfe algorithms use simple modifications of the general forward-stagewise update steps. In particular, the computation of the increment in equation 3.17 is modified to include the *whole* constraint region $\{\beta : G(\beta) \leq \kappa\}$:

$$\Delta_\beta = \underset{\beta}{\operatorname{argmin}}\, \beta^t L^{(j)} \quad s.t. \quad G(\beta) \leq \kappa \text{ and } \beta \geq 0 \tag{3.22}$$

Further, the update in equation 3.18 has a slightly modified form using convex mixtures of the old value and the new increment Δ_β

$$z^{(j+1)} = (1-t)z^{(j)} + t\Delta_\beta. \tag{3.23}$$

Due to this high formal similarity, the use of variable transformations $h(\cdot)$ has the same implications for algorithms of the Frank-Wolfe type as it has for forward stagewise algorithms. However, it should be noted that Frank-Wolfe algorithms cannot be directly used to compute a solution path. A potential "path"-variant is discussed in [Tibshirani, 2015], but compared to forward-stagewise, this variant has a dramatically increased computational cost. On the other hand, Frank-Wolfe methods have some local convergence guarantees which are not available in this form for forward-stagewise methods. The choice between the two algorithms therefore depends on the actual requirements. Important to note is that all guarantees provided by forward stagewise and Frank Wolfe [Jaggi, 2013] apply only to the transformed problem where the constraint is convex. These guarantees have no meaning in the original problem, as neither forward stagewise nor Frank Wolfe are applicable in a non-convex setting. Nevertheless, the progress, i. e. the decrease in the loss function along the solution path, will be identical for the original and the transformed problem as shown in equation 3.24. For an optimization problem of the form of equation 3.4, the minimization is performed in z-space and the following holds:

$$\begin{aligned} F(z) - F(z') &= (f \circ h)(z) - (f \circ h)(z') \\ &= f(h(z)) - f(h(z')) \\ &= f(x) - f(x'). \end{aligned} \tag{3.24}$$

This means that for an algorithm constructing a series of intermediate solutions the difference between two such solutions will be the same in both spaces.

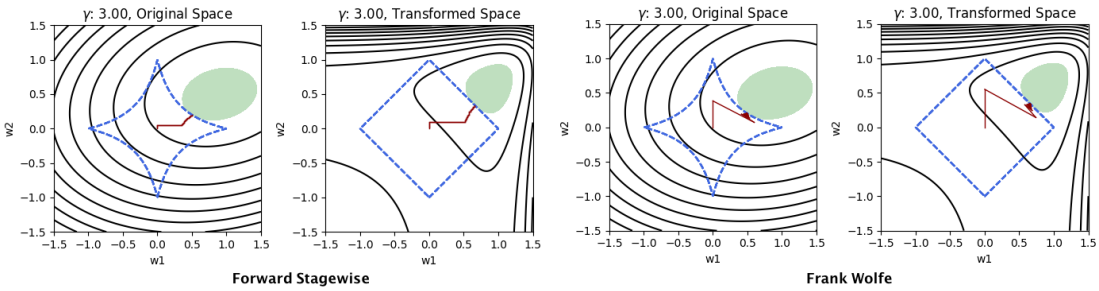


Figure 3.3.1 – Plots of two dimensional forward stagewise solution path (left) and Frank Wolfe optimization path (right). The red paths depict the solution or optimization paths, the blue dashed lines denote the boundaries of the constrained regions and the green surfaces show the area for which all values of the loss function are smaller than the solution found by the algorithm. Left to Right: Panels 2 and 4 show the transformed loss over the ℓ_1 norm, while panels 1 and 3 depict the least squares loss with the log-penalty. Plots are based on $\gamma = 5.0$ and $\kappa = 1$.

3.3 Experiments

3.3.1 Topographic plots of two dimensional solution paths

Figure 3.3.1 shows the optimization path generated by Frank Wolfe for a two dimensional problem and the solution path produced by forward stagewise for that same problem. The penalty used is given in equation 3.5. As can be seen, the boundary of the constrained region has a non-convex shape in the original problem, while the region of the transformed problem has a ℓ_1 shape. For forward stagewise, the sparsity of the path is expressed by its course parallel to the coordinate axis: first, the w_2 dimension is included into the model, followed by the w_1 dimension. All points on the path are intermediate solutions corresponding to different values of κ. Intermediate points of the Frank Wolfe algorithm, on the other hand, do in general *not* correspond to a specific value of κ.

3.3.2 Solution paths in dependence of γ

In Figure 3.3.2, we compare the solution paths generated by forward stagewise for different values of γ, based on the log penalty given in equation 3.5. The data set used to generate these plots consists of $n = 100$ samples, $p = 300$ predictors and the coefficients $x = [5,5,5,5,5,0,...,0,-1,-1,-1,-1,-1]$. Thus, there are 10 non-zero coefficients. The correlation matrix Σ is generated by $\Sigma_{ij} = min(i,j) * \frac{1}{p}$, which means the negative coefficients are highly correlated while the positive ones are uncorrelated. We add Gaussian noise with $\sigma^2 = 50$ to the response values. The top panels show the size of the coefficients in dependence of the training loss while the bottom panels show the test loss as a function of the training loss. The red vertical line depicts the minimum loss on the test set. The two leftmost panels use the log-penalty with a small γ value of 0.001. This corresponds approximately to the ℓ_1 regularization, and the coefficient paths look like a typical lasso path. The right and centre panels also use the log-penalty, but with a γ value of 0.5 respectively 3.0. We see that the coefficient paths with higher γ values are generally sparser. The test error for the best model as well as the number of active

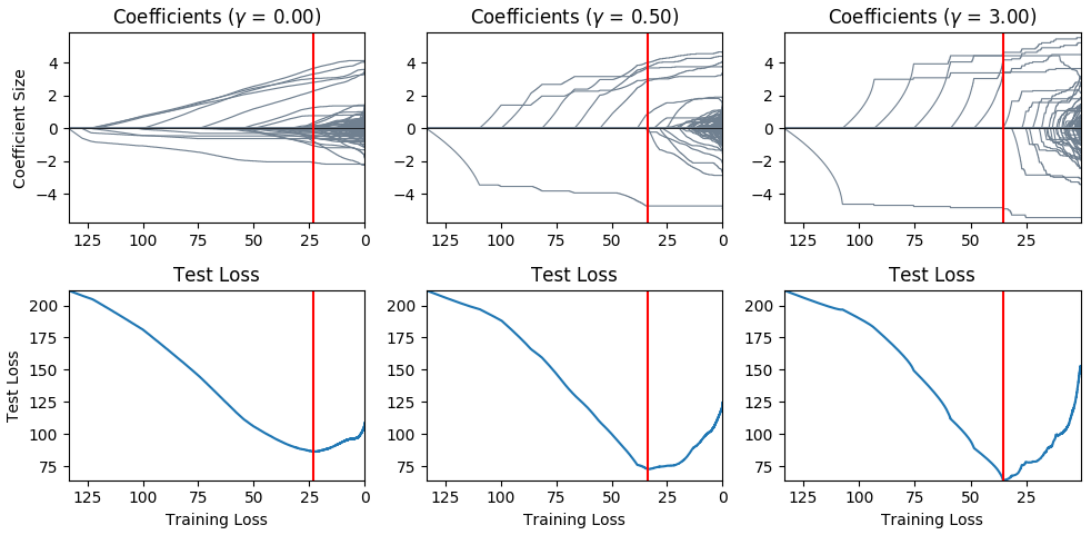


Figure 3.3.2 – From left to right: Solution paths for log penalty with $\gamma = 0.001$, $\gamma = 1$ and $\gamma = 5$. Top panels depict the size of coefficients, bottom panels the error on the test set. On the y axis, the training loss is given. Each point on the y axis corresponds to a valid model. The panel on the left side approximates the lasso problem. The center and right panels show increased sparsity compared to the lasso solutions.

coefficients decrease as γ increases. Generally, a higher γ value implies that new coefficients enter the model at a later stage and consequently, already selected coefficients will have higher absolute values as they would have had using a smaller γ.

3.3.3 Monotone increasing solution paths for forward stagewise

Hastie et al. [2007] show that the path optimized by forward stagewise differs from a solution path computed by the lasso in case of highly correlated predictors. We reproduce their experiment and show that similar observations can be made for non-convex penalties. For comparison, we use the sparsenet package by Mazumder et al. [2011], available in the R-repository. Sparsenet also constructs a path for non-convex penalties, although in their case they use the MC+ penalty proposed in [Zhang, 2010]. The experimental setup is as follows: The data consists of 60 samples with 1000 dimension. The dimensions are divided into 20 groups of the same size. Samples are drawn from a multivariate Gaussian, where the correlation between each member of a group is $\rho = 0.95$, while members of different groups remain uncorrelated. For each group there is a non-zero coefficient in the solution vector. Each coefficient is drawn from a standard Gaussian. Gaussian noise is added to the output variable with a standard deviation of $\sigma = 6$. We plot the obtained solution paths in figure 3.3.3. One observes that the effect of the monotonicity of the forward stagewise path carries over to the sparse version ($\gamma = 5$), while the sparsenet coefficients fluctuate much more. This is explained by the different objectives these algorithms optimize: Forward stagewise optimizes the arc-length of the paths, and therefore produces a much smoother appearance, i. e. these paths do not change drastically between subsequent solutions. Lasso, on the other hand, optimizes the cost function at each point of the solution path. In addition, sparse forward stagewise is more efficient in computing the solution path compared to forward stagewise, when a similarly dense solution path is requested (15 seconds for

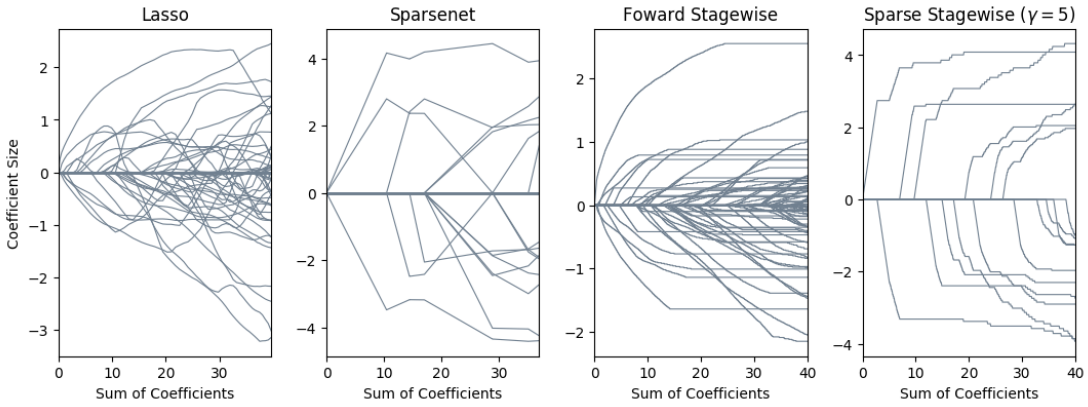


Figure 3.3.3 – Solution paths for Lasso, Sparsenet, Forward Stagewise, Sparse Forward Stagewise. As one can see, the monotonicity of the coefficient paths can be observed both in the forward stagewise and the sparse forward stagewise procedure. Both lasso and sparsenet show bigger variations in the solution paths due to the impact of correlated coefficients. This can be explained by the fact that forward stagewise minimizes the arc-length of the solution path, which adds smoothness to the coefficient paths.

sparse forward stagewise vs 45 seconds for sparsenet on a 2.9 GHz Intel Core i5).

3.3.4 Regression on artificial data

In this experiment, the goal is to assess if an increase in sparsity can help to find a better model compared to lasso regression or forward stagewise. For this purpose, a data set consisting of 50 features and 40 samples is created. In the underlying data generating process, only four coefficients are related to the target. In figure 3.3.4, the two first and the two last coefficients have a coefficient size of [-20, -10, 10, 20]. Noise with a standard deviation of $\sigma = 15$ is added to the result. All input features are correlated to each other with a correlation coefficient of 0.1. Figure 3.3.4 shows the result of this experiment. All models use cross validation to select the optimal level of sparsity. The top left panel shows lasso regression (calculation performed based on the sklearn library [Pedregosa et al., 2011]), the middle panel shows forward stagewise and the bottom panel shows *sparse* forward stagewise (with $\gamma = 5$). All models are tested on a test set and the scores on training and test set can be seen on the right. Overall, sparse forward stagewise leads to better test performance with a slightly worse training performance. If we look at the coefficients, it is apparent that sparse forward stagewise includes less spurious features as the other two models and comes closet to recovering the magnitude of the true coefficients. This result is within expectations, as the ability of a model to better approximate the ℓ_0-pseudonorm implies generally less shrinkage. This experiment can be considered as an ideal application case of the proposed method: A situation where more features than samples are available and only a small amount of correlation between input variables exists, which favours a model abtle to perform feature selection while recovering the true magnitude of the coefficients without bias.

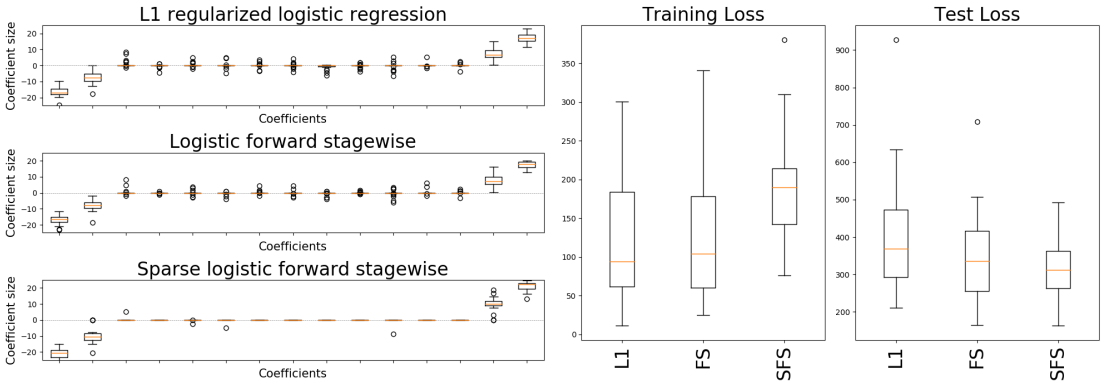


Figure 3.3.4 – Coefficient boxplots for ℓ_1 regularized logistic regression (L1), forward stagewise (FS) and sparse forward stagewise with $\gamma = 5$ (SFS). For better visualization, only 15 of the 40 coefficients are shown here (the two first and the two last, as well as eleven other). Coefficients not shown all have a median of zero with some outliers, similar to the coefficients in the center part of the 3 panels shown on the left.

3.3.5 Sparse Gaussian Information Bottleneck

We use the forward-stagewise algorithm with first order Taylor approximation introduced in section 3.2.1 for computing the solution path of the sparse Gaussian information bottleneck, i. e. we compute the evolution of the sparse compression coefficients $\mathbf{a}$ when the constraint κ is relaxed. The original algorithm proposed in [Rey et al., 2014] uses a log-Barrier method and traverses the solution path in the opposite direction: for a very large initial constraint value, this original algorithm starts at a feasible point a with strictly positive coefficients, which are then successively shrunk to zero by tightening the constraint. Typically, we are interested in sparse solutions, and this reverse traversal of the solution path is rather inefficient in practice. However, our forward-stagewise algorithm starts with the empty vector $\mathbf{a} = 0$ and successively includes new positive components when the constraint is relaxed. This conceptual difference leads to a huge difference in computational workload. On artificial data containing three "informative" features (i. e. dimensions in X which indeed have nonzero mutual information with Y), and many other noisy dimensions, our proposed forward-stagewise algorithm improves the run-time by several orders of magnitude, as shown in the left panel of figure 3.3.5. The new algorithm will introduce an error to the solution, as only the first order approximation is used, nevertheless, as one can see in the right panel of figure 3.3.5, this approximation error is negligible compared to the exact solution.

3.4 Conclusion

Our contribution is threefold: We first demonstrated how popular optimization algorithms of the forward stagewise and Frank Wolfe type can be applied to *non-convex* constraints by means of mapping the non-convex constraints onto convex ones. Assuming invexity of the initial objective function, the proposed mapping preserves this property such that the transformed objective is again invex. For gradient based optimization algorithms that require convex constraints and rely only on the gradient of the objective function to produce an update step, invexity ensures that there always exists an optimization direction as long as there is a descent direction. Secondly, we have shown that several

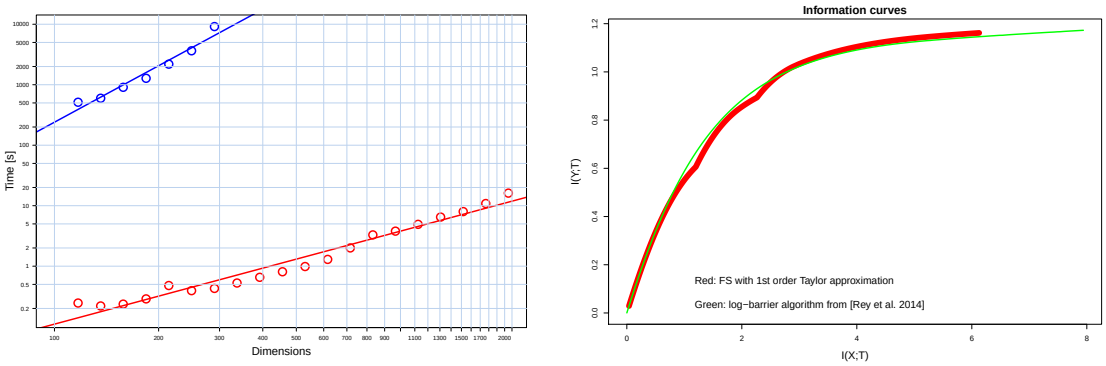


Figure 3.3.5 – (Left) Runtime experiments for the sparse meta-Gaussian information bottleneck. The data always contained 2000 samples, the dimensionality of Y was 20. There are three informative dimensions in X, and a varying number of additional noise dimensions (x-axis). Blue points/curve: algorithm proposed in [Rey et al., 2014], line is linear regression fit. Red: our proposed forward-stagewise algorithm, stopped after 10 variables have been selected. Note that this is a log-log plot. (Right) Comparison of information curves between the log-barrier method and forward stagewise with first order Taylor approximation. As one can see, forward stagewise induces a certain error into the solution, but compared to the exact solution, the error is negligible.

popular non-convex constraints can be mapped onto the ℓ_1 constraint, for which forward stagewise and Frank Wolfe are extremely efficient. Finally, in situations where non-convex penalties cannot be mapped onto the ℓ_1 region but onto a convex region close to ℓ_1, we proposed a forward stagewise approach with first order Taylor approximation. In the experiment section, we have demonstrated that in a log-constrained regression setting, the generalization performance can potentially be improved by trading-off less shrinkage for more sparsity, compared to lasso. Furthermore, we have shown that a log-constraint optimization problem, which arises naturally in the context of the sparse information bottleneck, can be solved more efficiently. This was possible by transforming the non-convex constraint in such a way that forward stagewise, a convex optimization algorithm, could be applied. Our approach was able to outperform the previous algorithm by several orders of magnitude.

4 Deep Archetypal Analysis for Interpretable Machine Learning

Previously, we used non-convex log-pseudonorms in order to perform ultra-sparse regression analysis and to identify relevant predictors X from a larger set of possible candidates. Generally, the goal of regression models is to infer the conditional probability of the target Y, given an observation x, i. e. estimating $P(Y|X=x)$. Models promoting sparsity of regression estimates are usually motivated by (i) reducing the prediction error, (ii) improving the generalization capability of the model and (iii) making the model more "interpretable". While a general consensus of how to formalize interpretability is still lacking, we think of interpretable models as those which maximize the degree to which a human can consistently explain why a model makes certain errors. While sparsity will not guarantee interpretable models, it will certainly reduce the number of non-zero predictors, and having fewer predictors is likely to increase interpretability. From the user perspective, interpretability is naturally evaluated with respect to the task or purpose of the model. However, in *unsupervised* machine learning targets Y are not available. Consequently, a learning task is often – implicitly or explicitly – defined with respect to the estimation of the joint probability density $P(X)$ over the inputs X. With respect to interpretability this begs the question how estimating $P(X)$ can provide interpretable explanations of the data? A classical examples of unsupervised learning is k-means clustering [MacQueen et al., 1967]. Generally, the goal of unsupervised clustering is to group objects into classes of similar objects, given an appropriate measure of similarity. In exploratory data analysis, clustering algorithms are often used with the expectation to recover an underlying natural grouping that might be hidden in the data. By exposing this inherent structure a possible path to interpretability is provided as, in the above defined sense, errors made by the model might be explained in a more consistent manner. K-means clustering is a special case of a Gaussian mixture model (GMM) with uniform prior weights and unit covariance. The optimized GMM will then provide the probability density of each data point x_i as $p_\theta(x_i) = \frac{1}{Z}\exp(-0.5||x_i - \mu_c||^2)$ with normalization constant Z, parameters θ and cluster means μ_c. These cluster means, also known as prototypes, derive their meaning from the assumption that a natural grouping exists in the first place, thereby justifying the use of clustering methods. If such a natural grouping does not exist, the structure of the data manifold is better explained through alternative models which respect the continuous nature of the data. Principal component analysis (PCA) [Pearson, 1901] is one such technique which aims to identify a lower-dimensional set of coordinate axes capturing the main directions of variation of the data. By reducing the initial dimensionality of the data, while identifying and discarding of directions with low variation, PCA potentially increases interpretability. For a data matrix $\mathbf{X}$, PCA identifies a rotation matrix $\mathbf{P}^T$ such that a change of basis on the data can be performed according to $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$, where $\mathbf{P}$ is chosen so that the covariance matrix of $\mathbf{Y}$ is diagonal. Dimensionality reduction is achieved by performing the

change of basis while ignoring those columns of $\mathbf{P}$ which are associated with directions of low variation. Compared to other unsupervised, linear dimensionality reduction techniques such as non-negative matrix factorization (NMF) [Lee and Seung, 1999] or independent component analysis (ICA) [Comon, 1994], the constraints governing PCA are still quite general. This has implications for the interpretability of PCA as including additional constraints, and thereby encoding prior knowledge, often helps specify an expectation with regard to possible interpretations of the data. A model closely related to PCA but governed by additional constraints was proposed by Cutler and Breiman [1994] and became known under the name of "Archetypal Analysis" (AA). Being a special type of NMF, AA aims to decompose the data matrix $\mathbf{X}$ into a row-stochastic weight matrix $\mathbf{A}$ and a matrix $\mathbf{Z}$ containing the so–called archetypes. With the additional constraint that the archetype matrix $\mathbf{Z}$ decomposes into the product of a row-stochastic weight matrix $\mathbf{B}$ and the original data matrix $\mathbf{X}$, AA approximates the data convex hull based on a polygon whose edges are the coordinates of the archetypes z_i. With the decomposition $\mathbf{X} \approx \mathbf{ABX} = \mathbf{AZ}$, every data point inside the convex polygon can be written as a weighted sum of the archetypes. Dimensionality reduction in AA is achieved by varying the number of edges, i. e. the number of archetypes, used to define the polygon which approximates the data convex hull. Archetypes are strongly connected to the kind of interpretability this model offers: As archetypes are extreme representatives of the data, inspecting the archetypes provides a sense of the variation contained within the data. Furthermore, due to the constraint that all data points x_i must decompose into a non-negative, weighted sum of the archetypes, a representation in terms of basic types is provided.

Since its conception, AA has known several advancements: In [Stone and Cutler, 1996] the authors propose an archetype model able to identify archetypes in space *and* time, named "Archetypal Analysis of spatio-temporal dynamics". A similar problem is addressed in "Moving archetypes" by Cutler and Stone [1997]. Model selection is the topic of [Prabhakaran et al., 2012], where the authors are concerned with the optimal number of archetypes needed to characterize a given data set. An extension of the original Archetypal Analysis model to non-linear kernel Archetypal Analysis is proposed by Bauckhage and Manshaei [2014], Mørup and Hansen [2012]. In [Kaufmann et al., 2015], the authors use a copula based approach to make AA independent of strictly monotone transformations of the input data. The reasoning is that such transformations should in general not influence which points are identified as archetypes. A probabilistic version of Archetypal Analysis was introduced by Seth and Eugster [2016], lifting the restriction of Archetypal Analysis to real–valued data and instead allowing other observation types such as integers, binary, and probability vectors as input. Although AA did not prevail as a commodity tool for pattern analysis, several applications have used it very successfully. In [H. P. Chan et al., 2003], AA is used to analyse galaxy spectra which are viewed as weighted superpositions of the emissions from stellar populations, nebular emissions and nuclear activity. For the human genotype data studied by Huggins et al. [2007], inferred archetypes are interpreted as representative populations for the measured genotypes. In computer vision, AA has for example been used by Bauckhage and Thurau [2009] to find archetypal images in large image collections or by Canhasi and Kononenko [2015] to perform the analogous task for large document collections. In combination with deep learning, Wynen et al. [2018] apply an archetypal style analysis to learned image representations in order to realize artistic style manipulations.

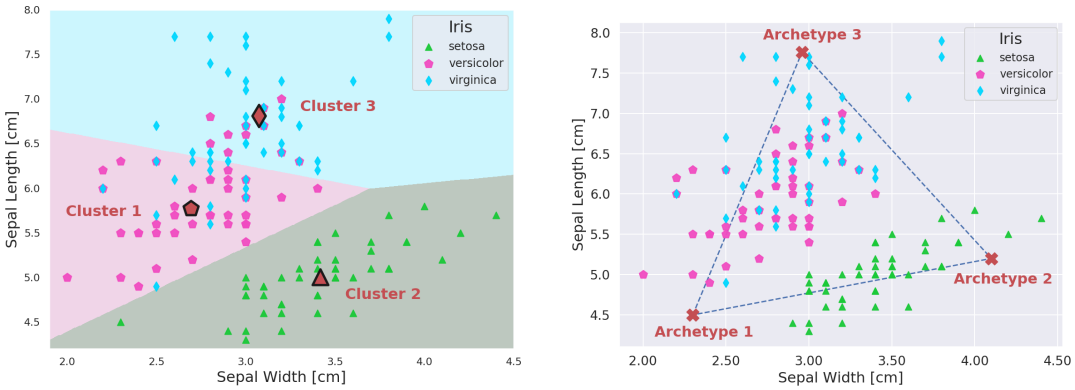
Archetypal analysis, as proposed by Cutler and Breiman [1994], has several shortcomings which we attempt to address: (i) It is a linear method and cannot integrate any additional information about the data, e.g. labels, that might be available. (ii) The feature space in which AA is performed is spanned by features that had to be selected by the user based on prior knowledge. (iii) As mixing of archetypes is performed directly on the input data, linear AA requires additivity of the input, which is a strong assumption unlikely to hold e.g. in case of image data. To address these problems,

we propose learning an appropriate *latent* feature space while simultaneously identifying suitable archetypes. We thus introduce a generative formulation of the linear archetype model, parameterized by neural networks. By introducing the distance-dependent archetype loss, the linear archetype model can be integrated into the latent space of a deep variational information bottleneck and an optimal representation, together with the archetypes, can be learned end-to-end. Moreover, the information bottleneck framework allows for a natural incorporation of arbitrarily complex side information during training. As a consequence, learned archetypes become easily interpretable as they derive their meaning directly from the included side information. Applicability of the proposed method is demonstrated by exploring archetypes of female facial expressions while using multi-rater based emotion scores of these expressions as side information. A second application illustrates the exploration of the chemical space of small organic molecules. By using different kinds of side information we demonstrate how identified archetypes, along with their interpretation, largely depend on the side information provided. The majority of the work presented in this chapter is based on [Keller et al., 2019] and [Keller et al., 2020].

Colloquially, both the words "archetype" and "prototype" describe templates or original patterns from which all later forms are developed. However, the concept of a prototype is more common in machine learning and for example encountered as cluster-centroids in classification, where a query point x is assigned to the class of the closest prototype. In an appropriate feature space such a prototype is a typical representative of its class, sharing all traits of the class members, ideally in equal proportion. By contrast, archetypes are characterized as being *extreme points* of the data, such that the complete data set can be well represented as a convex mixture of these extremes or archetypes. Archetypes thus form a polytope approximating the data convex hull. Based on the historic Iris flower data set [Anderson, 1935, Fisher, 1936], Figure 4.0.1 illustrates the different perspectives both approaches provide in exploring the data. In Figure 4.0.1a the cluster means as well as the decision boundaries in a 2-dimensional feature space are shown. The clustering was calculated using the k-Means algorithm. Each cluster mean is an *average* representative of its respective class, the aforementioned prototype. According to this clustering, the prototypical *Iris virginica* has a sepal width of 3.1cm and a sepal length of 6.8cm. On the other hand, Figure 4.0.1b shows the positions of the three archetypal Iris flowers, which represent *extreme* manifestations of the Iris species of the respective classes. The archetypal *Iris virginica* has a sepal width of 3.0cm and a sepal length of 7.8cm. All flowers within the simplex are characterized as convex mixtures of these archetypes. As flowers *outside* of that simplex will also be described as convex mixtures, the linear archetype model will approximate their location in feature space by normal projections onto the simplex' surface. With an increasing number of archetypes the approximation of the data convex hull will improve but interpretation of the individual archetypes might become more difficult. In general, a clustering approach is more natural if the existence of a cluster structure can be presumed. Otherwise, Archetypal Analysis might offer an interesting perspective for exploratory data analysis.

4.1 Exploring Data Sets Through Archetypes

Archetypal analysis (AA) was first proposed by Cutler and Breiman [1994]. It is a linear procedure where archetypes are selected by minimizing the squared error in representing each individual data point as a mixture of archetypes. Identifying the archetypes involves the minimization of a non-linear least squares loss.



(a) k-Means clustering with marked prototypes

(b) Archetypal Analysis with marked archetypes

Figure 4.0.1 – Result of a clustering procedure as well as an Archetypal Analysis, performed on the Iris data set. For clustering, the k-means algorithm was used, which is an unsupervised clustering algorithm identifying the average representatives of a data set, i.e. the cluster-centroids or prototypes. Archetypal Analysis on the other hand, seeks to identify extremes in the data set with the goal to represent individual data points as weighted mixtures of these extreme points, the so-called archetypes.

4.1.1 Archetypal Analysis

Linear AA is a form of non-negative matrix factorization where a matrix $X \in \mathbb{R}^{n \times p}$ of n data vectors is approximated as $X \approx ABX = AZ$ with $A \in \mathbb{R}^{n \times k}$, $B \in \mathbb{R}^{k \times n}$, and usually $k < \min\{n, p\}$. The so-called *archetype matrix* $Z \in \mathbb{R}^{k \times p}$ contains the k archetypes $\mathbf{z}_1, .., \mathbf{z}_j, .., \mathbf{z}_k$ with the model being subject to the following constraints:

$$a_{ij} \geq 0 \ \wedge \ \sum_{j=1}^{k} a_{ij} = 1, \quad b_{ji} \geq 0 \ \wedge \ \sum_{i=1}^{n} b_{ji} = 1 \tag{4.1}$$

Constraining the entries of A and B to be non-negative and demanding that both weight matrices are row stochastic implies a representation of the data vectors $\mathbf{x}_{i=1..n}$ as a weighted sum of the rows of Z while simultaneously representing the archetypes $\mathbf{z}_{j=1..k}$ themselves as a weighted sum of the n data vectors in X:

$$\mathbf{x}_i \approx \sum_{j=1}^{k} a_{ij} \mathbf{z}_j = \mathbf{a}_i Z, \quad \mathbf{z}_j = \sum_{i=1}^{n} b_{ji} \mathbf{x}_i = \mathbf{b}_j X \tag{4.2}$$

Due to the constraints on A and B in Eq. 4.1 both the representation of $\mathbf{x}_i$ and $\mathbf{z}_j$ in Eq. 4.2 are *convex* combinations. Therefore the archetypes approximate the data convex hull and increasing the number k of archetypes improves this approximation. The central problem of AA is finding the weight matrices A and B for a given data matrix X and a given number k of archetypes. The non-linear

optimization problem consists in minimizing the following residual sum of squares:

$$RSS(k) = \min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \mathbf{ABX}\|^2 \tag{4.3}$$

$$= \min_{a,b} \sum_{l=1}^{n} \left\| \mathbf{x}_l - \sum_{j=1}^{k} a_{ij} \sum_{i=1}^{n} b_{ji} \mathbf{x}_i \right\|^2 \tag{4.4}$$

In their original publication, Cutler and Breiman [1994] propose an alternating least squares approach for finding the archetypes: After a random initialization of the b's, Eq. 4.4 is solved for the a's. Then, given the a's, Eq. 4.4 is solved for the b's. This alternating optimization, which provably converges towards a local minimum, can be implemented using common solvers for quadratic programming. Using an active set algorithm, together with smarter initialization strategies, higher convergence rates are achieved by the archetype algorithm proposed by Bauckhage and Thurau [2009]. The *Rapid Archetypal Analysis* algorithm by Bauckhage et al. [2015] is based on a greedy Frank-Wolfe procedure and avoids, unlike the previously mentioned algorithms, the rather costly quadratic optimization routines. The example shown in Figure 4.0.1b was calculated based on our own implementation of this algorithm, which is – to our knowledge – the most efficient algorithm for solving the linear archetype problem available today.
A probabilistic formulation of linear AA is provided by Seth and Eugster [2016] where it is observed that AA follows a simplex latent variable model and normal observation model. The generative process for the observations $\mathbf{x}_i$ in the presence of k archetypes with archetype weights $\mathbf{a}_i$ is given by

$$\mathbf{a}_i \sim \mathrm{Dir}_k(\boldsymbol{\alpha}) \quad \wedge \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{a}_i Z, \epsilon^2 \mathbf{I}), \tag{4.5}$$

with uniform concentration parameters $\alpha_j = \alpha$ for all j, and weights summing up to $\|\mathbf{a}_i\|_1 = 1$. That is, the observations $\mathbf{x}_i$ are distributed according to isotropic Gaussians with means $\boldsymbol{\mu}_i = \mathbf{a}_i Z$ and variance ϵ^2.

4.1.2 A Biological Motivation for Archetypal Analysis

Conceptionally, the motivation for Archetypal Analysis is purely statistical but the method itself always implied the possibility of interpretations with a more *evolutionary flavour*. By representing an individual data point as a mixture of *pure types* or *archetypes*, a natural link to the evolutionary development of biological systems is implicitly established. The publication by Shoval et al. [2012] entitled 'Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space' made this connection explicit, providing a theoretical foundation of the 'archetype concept'. In general, evolutionary processes are multi-objective optimization problems and as such subject to unavoidable trade-offs: If multiple tasks need to be performed, no (biological) system can be optimal at all tasks at once. Examples of such trade-offs include those between longevity and fecundity in Drosophila melanogaster where long-lived flies show decreased fecundity [Djawdan et al., 1996] or predators that evolve to be fast runners but eventually have to trade-off their ability to subdue large or strong prey, e.g. cheetah versus lion [Garland, 2014]. Such evolutionary trade-offs are known to affect the range of phenotypes found in nature [Tendler et al., 2015]. In [Shoval et al., 2012] it is argued that best-trade-off phenotypes are weighted averages of archetypes while archetypes themselves are phenotypes specialized at performing a *single* task optimally. An example of an evolutionary trade-off in the space of traits (or phenospace) for different species of bats (Microchiroptera) is shown in

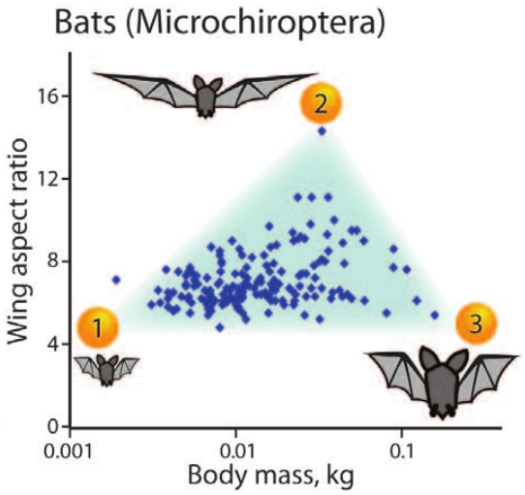


Figure 4.1.1 – Phenospace of different species of Microchiroptera. The dominant food habit of each species, and thereby the ability to procure this food source, is linked to the morphology of the animals, e.g. a higher Wing Aspect Ratio corresponds with the greater aerodynamic efficiency needed to chase high flying insects. Archetypes are extreme types, optimized to perform a single task. Proximity of a species to an archetype quantifies the level of adaptation this species has undergone with respect to the optimization objective or task. Reprinted from [Shoval et al., 2012] with permission.

Figure 4.1.1. Based on a study of bat wings by Norberg et al. [1987], each species is represented in a two-dimensional space where the axis depict Body Mass and Wing Aspect Ratio. The latter is the square of the wingspan divided by the wing area. Table 4.1 gives an account of the task the archetypes indicated in Figure 4.1.1 have evolved to performing optimally. The trade-off situation can be interpreted using Pareto optimality theory [Steuer, 1986], which was recently used in biology to study trade-offs in evolution [Schuetz et al., 2012, El Samad et al., 2005]. All phenotypes that have evolved over time lie within a restricted part of the phenospace, the so-called Pareto front, which is the set of phenotypes that cannot be improved at all tasks simultaneously. If there were a phenotype being better at all tasks than a second phenotype, then the latter would be eliminated over time by natural selection. Consequently phenotypes on the Pareto front are the best possible compromise between the different requirements or tasks.

4.2 Method

4.2.1 Deep Variational Information Bottleneck

We propose a model to generalise linear AA to the non-linear case based on the Deep Variational Information Bottleneck framework since it allows to incorporate side information Y by design and is known to be equivalent to the VAE in the case of $Y = X$, as shown in [Alemi et al., 2016]. In contrast to the data matrix X in linear AA, a non-linear transformation $f(X)$ giving rise to a latent representation $T \in \mathbb{R}^d$ of the data suitable for (non-linear) Archetypal Analysis is considered. I.e. the latent representation T takes the role of the data X in the previous treatment.

Archetype	Phenotype	Specialization
1	low aspect ratio, small body	hunting small insects near vegetation
2	high aspect ratio, medium body	hunting high flying large insects
3	low aspect ratio, large body	hunting animals near vegetation

Table 4.1 – Inferred specialization of the archetypal species of Microchiroptera indicated in Figure 4.1.1. From an evolutionary perspective, the phenotype is a consequence of the specialization, for details see [Shoval et al., 2012].

The DVIB combines the information bottleneck (IB) with the VAE approach [Tishby et al., 2000b, Kingma and Welling, 2013]. The objective of the IB method is to find a random variable T which, while compressing a given random vector X, preserves as much information about a second given random vector Y. The objective function of the IB is as follows

$$\min_{p(\mathbf{t}|\mathbf{x})} I(X;T) - \lambda I(T;Y), \tag{4.6}$$

where λ is a Lagrange multiplier and I denotes the mutual information. Assuming the IB Markov chain $T - X - Y$ and a parametric form of Eq. 4.6 with parametric conditionals $p_\phi(\mathbf{t}|\mathbf{x})$ and $p_\theta(\mathbf{y}|\mathbf{t})$, Eq. 4.6 is written as

$$\max_{\phi,\theta} -I_\phi(\mathbf{t};\mathbf{x}) + \lambda I_{\phi,\theta}(\mathbf{t};\mathbf{y}). \tag{4.7}$$

As derived in [Wieczorek et al., 2018], the two terms in Eq. 4.7 have the following forms:

$$\begin{aligned} I_\phi(\mathbf{t};\mathbf{x}) &= D_{KL}\left(p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \| p(\mathbf{t})p(\mathbf{x})\right) \\ &= \int p(\mathbf{t},\mathbf{x}) \log p_\phi(\mathbf{t}|\mathbf{x}) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} \\ &\qquad - \int p(\mathbf{x}|\mathbf{t})p(\mathbf{t}) \log p(\mathbf{t}) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} \\ &= \int p_\phi(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \log \frac{p_\phi(\mathbf{t}|\mathbf{x})}{p(\mathbf{t})} \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} \\ &= \mathbb{E}_{p(\mathbf{x})} D_{KL}\left(p_\phi(\mathbf{t}|\mathbf{x}) \| p(\mathbf{t})\right) \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} I_{\phi,\theta}(\mathbf{t};\mathbf{y}) = \; & D_{KL}\left(\left[\int p(\mathbf{t}|\mathbf{y},\mathbf{x})\,p(\mathbf{y},\mathbf{x})\,\mathrm{d}\mathbf{x}\right] \| \, p(\mathbf{t})\,p(\mathbf{y})\right) \\ = \; & \int p_\phi(\mathbf{t}|\mathbf{x},\mathbf{y})\,p(\mathbf{x},\mathbf{y}) \log \frac{p_\theta(\mathbf{y}|\mathbf{t})\,p(\mathbf{t})}{p(\mathbf{t})\,p(\mathbf{y})}\,\mathrm{d}\mathbf{t}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ = \; & \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\int p_\phi(\mathbf{t}|\mathbf{x},\mathbf{y}) \log p_\theta(\mathbf{y}|\mathbf{t})\,\mathrm{d}\mathbf{t}\right] \\ & - \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log p(\mathbf{y}) \int p_\phi(\mathbf{t}|\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{t}\right] \\ \geq \; & \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\mathbb{E}_{p_\phi(\mathbf{t}|\mathbf{x})} \log p_\theta(\mathbf{y}|\mathbf{t}) + h(Y). \end{aligned} \tag{4.9}$$

Here $h(Y) = -\mathbb{E}_{p(\mathbf{y})} \log p(\mathbf{y})$ denotes the entropy of Y in the discrete case or the differential entropy in the continuous case. The models in Eq. 4.8 and Eq. 4.9 can be viewed as the encoder and decoder, respectively. Assuming a standard prior of the form $p(\mathbf{t}) = \mathcal{N}(\mathbf{t}; 0, I)$ and a Gaussian distribution for the posterior $p_\phi(\mathbf{t}|\mathbf{x})$, the KL divergence in Eq. 4.8 becomes a KL divergence between two Gaussian distributions which can be expressed in analytical form as in [Kingma and Welling, 2013]. $I(T;X)$ can then be estimated on mini-batches of size m as

$$I_\phi(\mathbf{t};\mathbf{x}) \approx \frac{1}{m} \sum_i D_{KL}\left(p_\phi(\mathbf{t}|\mathbf{x}_i) \| \, p(\mathbf{t})\right). \tag{4.10}$$

As for the decoder, $\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\mathbb{E}_{p_\phi(\mathbf{t}|\mathbf{x})} \log p_\theta(\mathbf{y}|\mathbf{t})$ in Eq. 4.9 is estimated using the reparametrisation trick proposed by Kingma and Welling [2013], Rezende et al. [2014]:

$$\begin{aligned} I_{\phi,\theta}(\mathbf{t};\mathbf{y}) = \; & \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\mathbb{E}_{\boldsymbol{\varepsilon}\sim\mathcal{N}(0,I)} \sum_i \log p_\theta\left(\mathbf{y}_i|\mathbf{t}_i\right) \\ & + \text{const.} \end{aligned} \tag{4.11}$$

with the reparametrisation

$$\mathbf{t}_i = \boldsymbol{\mu}_i(\mathbf{x}) + \mathrm{diag}\left(\boldsymbol{\sigma}_i(\mathbf{x})\right) \boldsymbol{\varepsilon}. \tag{4.12}$$

As mentioned earlier, in the case of $Y = X$ the original VAE is retrieved [Alemi et al., 2016]. In our applications, we would like to predict not only the side information Y but also reconstruct the input X. Similar to the approach proposed in [Gomez-Bombarelli et al., 2018], we use an additional decoder branch to predict the reconstruction $\tilde{X}$. This extension requires an additional term $I_{\phi,\psi}(\mathbf{t};\tilde{\mathbf{x}})$ in the objective function Eq. 4.7 and an additional Lagrange multiplier ν. The mutual information estimate $I_{\phi,\psi}(\mathbf{t};\tilde{\mathbf{x}})$ is obtained analogously to Eq. 4.11.

4.2.2 Deep Archetypal Analysis

Deep Archetypal Analysis can then be formulated in the following way. For the sampling of $\mathbf{t}_i$ in Eq. 4.11 the probabilistic AA approach as in Eq. 4.5 can be used which leads to

$$\mathbf{t}_i \sim \mathcal{N}\left(\boldsymbol{\mu}_i(\mathbf{x}) = \mathbf{a}_i(\mathbf{x})\,Z, \boldsymbol{\sigma}_i^2(\mathbf{x})\mathbf{I}\right), \tag{4.13}$$

where the mean $\boldsymbol{\mu}_i$ given through $\mathbf{a}_i$ and variance $\boldsymbol{\sigma}_i^2$ are non-linear transformations of the data point $\mathbf{x}_i$ learned by the encoder. We note that the means $\boldsymbol{\mu}_i$ are convex combinations of weight vectors $\mathbf{a}_i$ and the archetypes $\mathbf{z}_{j=1..k}$ which in return are considered to be convex combinations of the

means $\boldsymbol{\mu}_{i=1..m}$ and weight vectors $\mathbf{b}_j$.[1] By learning weight matrices $A \in \mathbb{R}^{m\times k}$ and $B \in \mathbb{R}^{k\times m}$ which are subject to the constraints formulated in Eq. 4.1 and parameterised by ϕ, a non-linear transformation of data X is learned which drives the structure of the latent space to form archetypes whose convex combination yield the transformed data points. A major difference to linear AA is that for *deep AA* we cannot identify the positions of the archetypes $\mathbf{z}_j$ as there is no absolute frame of reference in latent space. We thus position k archetypes at the vertex points of a $(k-1)$-simplex and collect these *fixed* coordinates in the matrix Z^{fixed}. These requirements lead to an additional distance-dependent archetype loss of

$$\ell_{\text{AT}} = ||Z^{\text{fixed}} - BAZ^{\text{fixed}}||_2^2 = ||Z^{\text{fixed}} - Z^{\text{pred}}||_2^2, \tag{4.14}$$

where $Z^{\text{pred}} = BAZ^{\text{fixed}}$ are the *predicted* archetype positions given the learned weight matrices A and B. For $Z^{\text{pred}} \approx Z^{\text{fixed}}$ the loss function ℓ_{AT} is minimized and the desired archetypal structure is achieved. The objective function of *deep AA* is then given by

$$\max_{\phi,\theta} -I_\phi(\mathbf{t};\mathbf{x}) + \lambda I_{\phi,\theta}(\mathbf{t};\mathbf{y}) + \nu I_{\phi,\psi}(\mathbf{t};\tilde{\mathbf{x}}) - \ell_{\text{AT}}. \tag{4.15}$$

A visual illustration of *deep AA* is given in Figure 4.2.1. The constraints on A and B can be guaranteed by using softmax layers and *deep AA* can be trained with a standard stochastic gradient descent technique such as Adam [Kingma and Ba, 2014]. Note that the model naturally allows to be relaxed to the VAE setting by omitting the side information term $\lambda I_{\phi,\theta}(\mathbf{t};\mathbf{y})$ in Eq. 4.15.

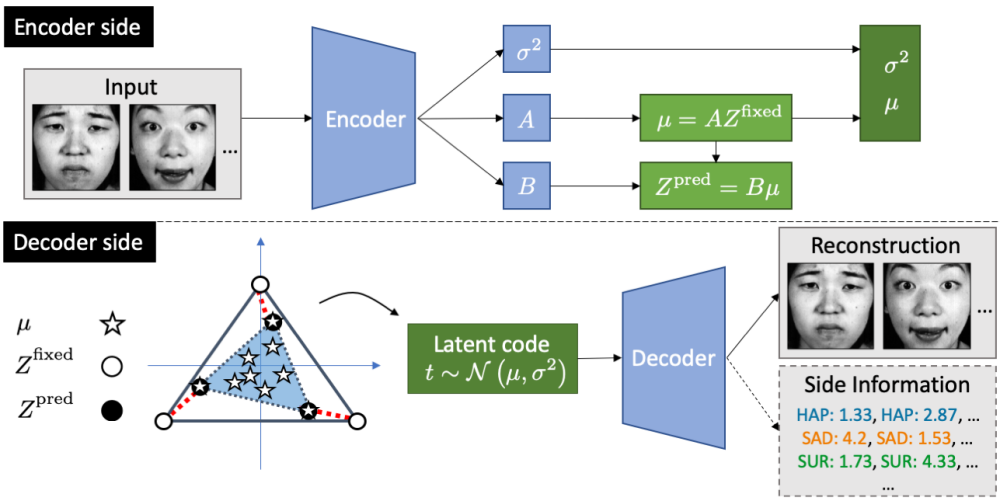


Figure 4.2.1 – Illustration of the *deep AA* model. **Encoder side**: Learning weight matrices A and B allows to compute the archetype loss ℓ_{AT} in Eq. 4.14 and sample latent variables $\mathbf{t}$ as described in Eq. 4.13. The constraints on A and B in Eq. 4.1 are enforced by using softmax layers. **Decoder side**: Z^{fixed} represent the fixed archetype positions in latent space while Z^{pred} are given by the convex hull of the transformed data point means μ during training. Minimizing ℓ_{AT} corresponds to minimizing the red-dashed (pairwise) distances. The input is reconstructed from the latent variable $\mathbf{t}$. In the presence of side information, the latent representation allows to reproduce the side information Y as well as the input X.

[1] Note that $i = 1..m$ (and not up to n), which reflects that deep neural networks usually require batch-wise training with batch size m.

4.2.3 Selecting the Number of Archetypes

In the proposed model the dimension d of the latent space and the number of archetypes k are related through the equation $k = d + 1$. The coordinates of the k archetypes coincide with the vertices of a regular d-simplex located on the unit sphere centered around the origin. Therefore, every vertex of the simplex has the same distance to the origin. Together with a spherical Gaussian prior $p(\mathbf{t}) = \mathcal{N}(\mathbf{t}; 0, I)$, this geometric construct ensures that no latent space directions is preferred over any other. Thus, in the absence of prior knowledge, this agnostic setting makes all archetypes equally important.
In principal, decoupling k and d is a valid option. But by increasing the number of archetypes k in a latent space of fixed dimension, every data set can be explained in an increasingly trivial manner. The idea of Archetypal Analysis, however, is to tolerate some noise in the generative process and to approximate the convex hull with only a limited number of vertex points. Within the framework of a variational information bottleneck, choosing $k = d + 1$ thus allows to identify the most compact latent code for a given data set. Consequently, model selection is performed by observing at which latent dimensionality the predictive mutual information, i.e. the reconstruction loss, saturates. In section 4.3.4, we demonstrate the model selection process using a held-out test set in the experiments based on the QM9 data set of small organic molecules. Additionally, in section 4.3.5, we explore an alternative prior conceptually closer to the original formulation of Archetypal Analysis.

4.2.4 The Necessity for Side Information

The goal of deep AA is to identify meaningful archetypes in latent space which will subsequently enable an informed exploration of the given data set. The *meaning* of an archetype, and thereby the associated interpretation, can be improved by providing so-called side information, i.e. information *in addition* to the input data. For non-linear latent variable models parametrized by neural networks, an interpretation of the latent space structure – depending on the data set – is often difficult, as input dimensions can be mapped to arbitrarily complex non-linear curves in latent space. In general, more non-linearity leads to more flexibility in the mapping of an input onto its latent code, which in turn leads to more ambiguity when interpreting that latent code. Supplementing the training process with additional information – which we call *side information* – can facilitate the interpretation. Consequently, the function of the side information is that of a regularizer as it restricts the class of potential mappings. If the input datum is for example an image, additional information could simply be a scalar- or vector-valued label. Using richer side information, e.g. additional images, is of course possible. In more general terms, the fundamental idea is that information about what constitutes an *archetypal* representative might not be information that is readily present in the input X but dependent on – or even defined by – the side information. Taking a data set of car images as an example, what would be an *archetypal* car? Certainly, the overall size of a car would be a good candidate, such that smaller sports cars and larger pick-ups might be identified as archetypes. But introducing the fuel consumption of each car as side information would put sports cars and pick-ups closer together in latent space, as both car types often consume above average quantities of fuel. In this way, side information guides the learning of a latent representation which is informative with respect to exactly the side information provided. Consequently, whether a data point is identified as an archetype, is not an inherent property of the data alone, but rather a function of the side information made available during training. And the selection of appropriate side information can only be linked to the questions the user of a deep AA model is interested in answering.

4.3 Experiments

4.3.1 Archetypal Analysis: Dealing With Non-linearity

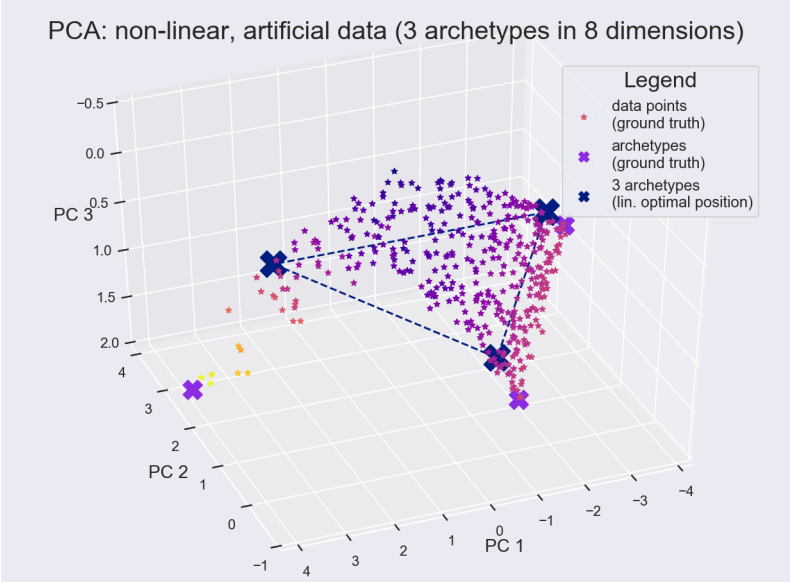
Data generation. For this experiment, data $\mathbf{X} \in \mathbb{R}^{n \times 8}$ is generated that is a convex mixture of k archetypes $\mathbf{Z} \in \mathbb{R}^{k \times 8}$ with $k \ll n$. The generative process for the datum $\mathbf{x_i}$ follows Eq. 4.5, where $\mathbf{a}_i$ is a stochastic weight vector denoting the fraction of each of the k archetypes $\mathbf{z}_j$ needed to represent the data point $\mathbf{x}_i$. A total of $n = 10000$ data points is generated, of which $k = 3$ are true archetypes. The variance is set to $\sigma^2 = 0.05$ and the linear 3-dim data manifold is embedded in a $n = 8$ dimensional space. Note that although linear and deep Archetypal Analysis is always performed on the full data set, only a fraction of that data is displayed when visualizing results.

Linear AA – non-linear data. Data is generated as described above and an additional non-linearity is introduced by applying an exponential to one dimension of $\mathbf{X}$ which results in a curved 8-dimensional data manifold. Linear Archetypal Analysis is then performed using the efficient Frank-Wolfe procedure proposed by Bauckhage et al. [2015]. For visualization, PCA is used to recover the original 3-dimensional data submanifold which is embedded in the 8-dimensional space. The first three principal components of the ground truth data are shown in Figure 4.3.1a as well as the computed archetypes (connected by dashed lines). The positions of the computed archetypes occupy optimal positions according to the optimization problem in Eq. 4.3 but due to the non-linearity in the data it is impossible to recover the three ground truth archetypes.

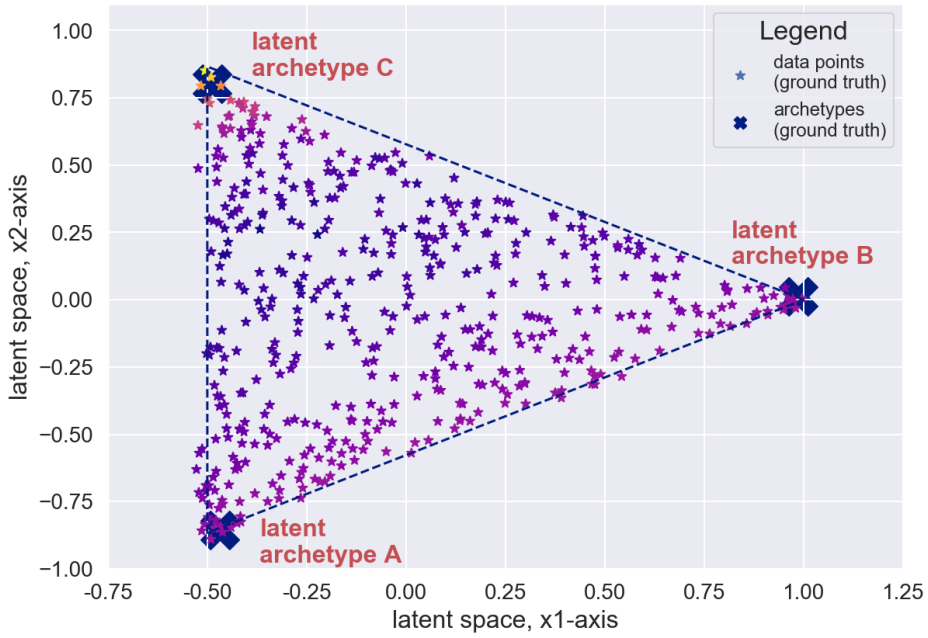
Deep AA – non-linear data. For data that has been generated as described in the previous paragraph, a strictly monotone transformation in form of an exponentiation should in general *not* change which data points are identified as archetypes. But this is clearly the case for linear AA as it is unable to recover the true archetypes *after* a non-linearity has been applied. Using that same data to train the deep AA architecture presented in Figure 4.2.1 generates the latent space structure shown in Figure 4.3.1b, where the three archetypes A, B and C have been assigned to the appropriate vertices of the latent simplex. Moreover, the sequence of color stripes shown has been correctly mapped into the latent space. Within the latent space data points are again described as convex linear combinations of the latent archetypes. Latent data points can also be reconstructed in the original data space through the learned decoder network. The network architecture used for this experiment was a simple feedforward network (2 layered encoder and decoder), training for 20 epochs with a batch size of 100 and a learning rate of 0.001.

4.3.2 Archetypes in Image-based Sentiment Analysis

The Japanese Female Facial Expression (JAFFE) database was introduced by Lyons et al. [1998] and contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral). The expressions are happiness, sadness, surprise, anger, disgust and fear. All expressions were posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects on a 5 level scale (5-high, 1-low) and each image was then assigned a 6-dimensional vector of average ratings. For the following experiments the advice of the creator of the JAFFE data set was followed to exclude *fear* images and the *fear* adjective from the ratings, as the models were not believed to be good at posing fear. All experiments based on the JAFFE data set are performed on the following



(a) Linear AA is unable to recover the true archetypes.



(b) Latent space embedding of non-linear artificial data.

Figure 4.3.1 – While linear Archetypal Analysis is in general unable to approximate the convex hull of a non-linear data set well, deep AA learns an appropriate latent representation where the ground truth archetypes can correctly be identified.

architecture[2]:

[2]The code is available via https://github.com/bmda-unibas/DeepArchetypeAnalysis

Encoder:
Input: image $\mathbf{x}$ (128×128)
→ 3×$\Big[$64 Conv. (4×4) + Max-Pool. (2×2)$\Big]$
→ Flatten + FC100
→ $\mathbf{A}$, $\mathbf{B}$, σ^2

Decoder (Image Branch):
Input: latent code $\mathbf{t}$
→ FC49
→ 3×$\Big[$64 Conv. Transpose (4×4)$\Big]$
→ Flatten + FC128×128
→ FC128×128 → 128×128 reconstruction $\tilde{\mathbf{x}}$

Decoder (Side Information Branch):
Input: latent code $\mathbf{t}$
→ FC200-5 → side information $\tilde{\mathbf{y}}$

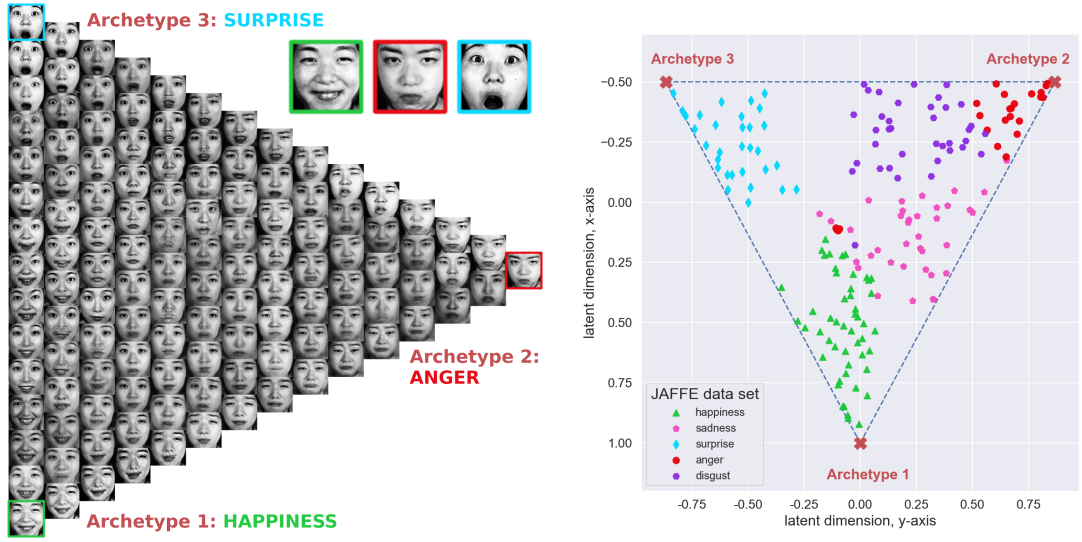
ReLU activations are used in-between layers and sigmoid activations for the image intensities. The different losses are weighted as follows: we multiplied the archetype loss by a factor of 80, the side information loss by 560, and the KL divergence by 40. In the setting where only two labels are considered, the weight for archetype loss is increased to 120. The network was trained for 5000 epochs with a mini-batch size of 50 and a learning rate of 0.0001. For training a NVIDIA TITAN X Pascal GPU was used, where a full training sessions lasted approximately 30 minutes.

JAFFE: Latent Space Structure

Emotions conveyed through facial expressions are a suitable case to demonstrate the interpretability of learned latent representation in deep AA. First, the existence of archetypes is plausible as there clearly are expressions that convey a maximum of a given emotion, i.e. a person can look extremely/maximally surprised. Second, facial expressions change continuously without having a clearly defined cluster structure. Moreover, these expressions lend themselves to being interpreted as mixtures of basic (or archetypal) emotional expressions – a perspective also enforced by the averaged ratings for each image which are essentially weight vectors with respect to the archetypal emotional expressions. Figure 4.3.2a shows the learned archetypes "happiness", "anger" and "surprise" while expressions linked to the emotion adjective "sadness" are identified as mixtures between archetype 1 (happiness) and archetype 2 (anger). Figure 4.3.2b shows the positions of the latent means where the color coding is based on the *argmax* of the emotion rating, which is a 5-dimensional vector. An analogous situation is found in case of "disgust", which, according to deep AA, is a mixture between archetype 2 (anger) and archetype 3 (surprise). Towards the center of the simplex, expressions are located which share equal weights with respect to the archetypes and thus resemble a more "neutral" facial expression.

JAFFE: Expressions As Weighted Mixtures

One advantage of deep AA compared to the plain Variational Autoencoder (VAE) is a *globally* interpretable latent structure. All latent means μ_i will be mapped inside the convex region spanned by the archetypes. And as archetypes represent extremes of the data set which are present to some



(a) Archetype latent space of the JAFFE data set. (b) Location of emotion adjectives in latent space.

Figure 4.3.2 – Deep AA with $k = 3$ archetypes identifies sadness as a mixture mostly between happiness and anger while disgust lies between the archetypes for anger and surprise.

percentage in all data points, these percentages or weights can be used to explore the latent space in an *informed* fashion. This might be especially of advantage in case of higher-dimensional latent spaces. For example, the center of the simplex will always accommodate latent representations of input data that are considered *mean* samples of the data set. Moreover, directions within the simplex have meaning in the sense that when "walking" towards or away from a given archetype, the characteristics of that archetype will either be enforce or diminished in the decoded datum associated with the actual latent position. This is shown in the Hinton plot in Figure 4.3.3 where mixture 1 is a mean sample, i.e. with equal archetype weights. Starting at this position and moving on a straight line into the direction of archetype 3 increases its influence while equally diminishing the influence of both archetypes 1 and 2. This results in mixture 2 which starts to look surprised, but not as extremely surprised as archetype 3. In the same fashion mixture 3 and 4 are the results of walking straight into the direction of archetypes 2 or 1 which results in a sad face (mixture 3) and a slightly happy facial expression (mixture 4).

JAFFE: Deep AA Versus VAE

Deep AA is designed to be a model that simultaneously learns an appropriate representation and identifies meaningful latent archetypes. This model can be compared to a plain VAE where a latent space is learned first and subsequently linear AA is performed on that space in order to approximate the latent convex hull. Figure 4.3.4a shows the interpolation in the deep AA model between two images, neither of them archetypes, from "happy" to "sad". Compared to Figure 4.3.4b, which shows the same interpolation in a VAE model with subsequently performed linear AA, the interpolation based on deep AA gives a markedly better visual impression. In case of deep AA, this is explained by the fact that all data points are mapped into the simplex which ensures a relatively dense distribution of the latent means. On the other hand, the latent space of the VAE model has no hard geometrical restrictions and thus the distribution of the latent representatives will be less dense or even "patchy",

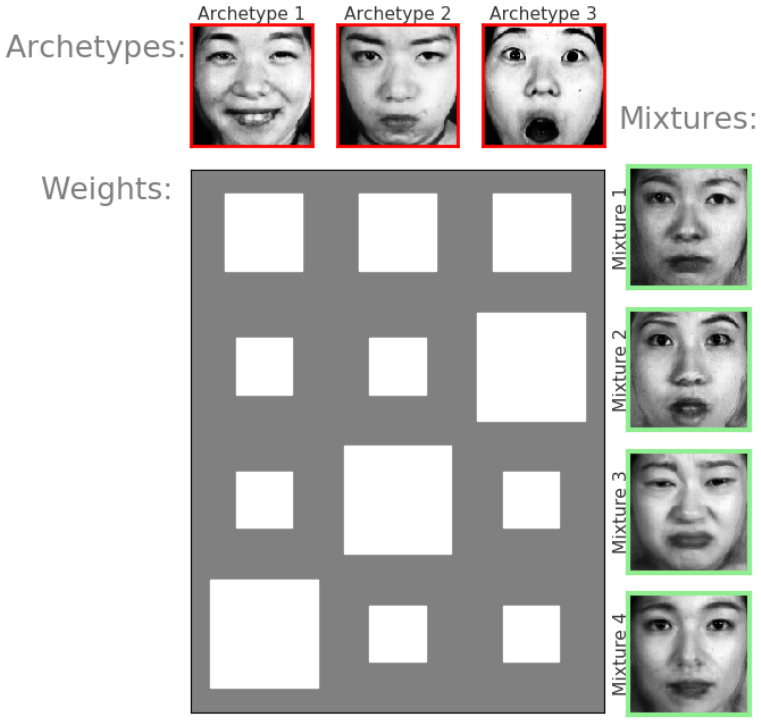


Figure 4.3.3 – Knowing the archetypes allows for an informed exploration of the latent space by *not* directly sampling latent space coordinates but by specifying a desired mixture with respect to the known archetypes.

i.e. with larger empty areas in latent space. Especially with small data sets such as JAFFE, of which less than 200 images are used, interpolation quality might be strongly affected by the unboundedness of the latent space of VAE models.

4.3.3 Stability of Inferred Archetypes: Bootstrapping Experiment

In order to demonstrate the stability of the inferred archetypes with respect to their interpretation, we evaluate our method on 40 distinct sets of bootstrap samples of the JAFFE data set. The general setup is identical to section 4.3.2. The weights of the archetype and side information loss are 1e2 and 2e2, respectively, while the weight of the reconstruction loss is set to 0.4. The weight of the KL divergence is initialized with 5e3 and then slowly decreased until it reaches the target weight of 4e1. Figures 4.5.1 and 4.5.2 show the true input images that were mapped closest to the vertices of the latent simplex when mapping the whole data set (i. e. including the bootstrap hold-out set) into the latent space at test time. The scatter plots also show the latent distribution of the whole data set. Colors indicate the argmax of the five emotion scores, which were used as side information during training. We can see that the inferred archetypes – here: the closest true input images – consist of three distinct "extreme" emotions for the majority of the runs. Importantly, some images are recognized as archetypal through multiple runs even though the training data sets had different compositions each time. Figure 4.3.6 shows the distribution of the different combinations of inferred

(a) Interpolation based on deep AA. (b) Interpolation based on VAE (with lin. AA).

Figure 4.3.4 – Interpolation between the two input images marked in red. The interpolation in the latent space of the deep AA model is qualitatively better compared to the VAE model as latent points are mapped more densely due to the simplex constraints.

archetypal emotions. We note that the predominant combination contains the emotions "surprised", "happy" and "angry".

Side Information for JAFFE. The JAFFE data set contains facial expressions posed by 10 Japanese female models. Based solely on the visual information, i.e. disregarding the emotion scores, these images could meaningfully be grouped together in a variety of ways, e.g. head shape, hair style, identity of the model posing the expressions etc. The interpretability of archetypes, in general, rests on providing side information with respect to which the learned representation shall be informative. The latent space shown in Figure 4.3.5 has been learned while providing only the emotion ratings for "sadness" and "disgust". This result illustrates how side information is shaping the structure of the learned latent representation: Comparing Figure 4.3.5 with Figure 4.3.2a, where the emotion ratings for "anger", "surprise" and "sadness" were provided as side information during training, makes clear that archetypes are not necessarily a property of the data. The final structure of the latent space is determined to a large extend by the side information and thus by the intent of the user when selection which information to provide.
While it is obvious to learn typical emotion expressions in case of JAFFE, most applications are arguably more ambiguous. In section 4.3.4, a chemical experiment is discussed, where each molecule can be described by a variety of properties. The side information introduced to the learning process will ultimately be the property the experimenter is interested in, and the learned representation will be informative with respect to that property.

4.3.4 The Chemical Universe Of Molecules

In the following section the application of deep AA to the domain of chemistry is explored. Starting with an initial set of chemical compounds, e.g. small organic molecules with cyclic cores [Visini et al.,

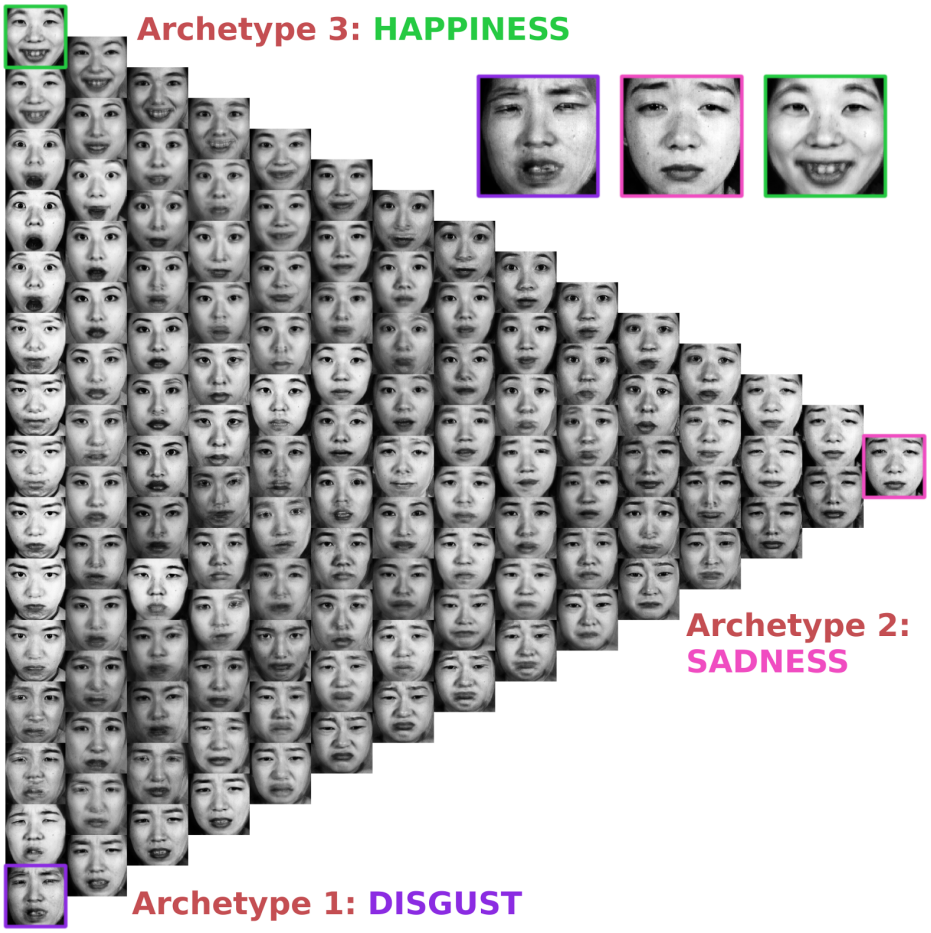


Figure 4.3.5 – Latent structure of the JAFFE data set when trained on a subset of the side information containing only the emotion ratings for "sadness" and "disgust".

2017], and iteratively applying a finite number of reactions, will eventually lead to a huge collection of molecules with extreme combinatorial complexity. But while the total number of all possible *small* organic molecules has been estimated to exceed 10^{60} [Kirkpatrick and Ellis, 2004], even this number pales in comparison to the whole chemical universe of organic chemistry. In general, the efficient exploration of chemical spaces requires methods capable of learning meaningful representations and endowing these spaces with a globally interpretable structure. Prominent examples of chemistry data sets include the family of GDB-xx data sets (generic database), e.g. GDB-13 [Blum and Reymond, 2009], which enumerates small organic molecules of up to 13 atoms, composed of the elements C, N, O, S and Cl, following simple chemical stability and synthetic feasibility rules. With more than 970 million structures, GDB-13 is the largest publicly available database of small organic molecule to date.

Exploring the Chemical Space. As discussed in section 4.1.2, Archetypal Analysis lends itself to a distinctly evolutionary interpretation. Although this is certainly a more biological perspective, the basic principle is applicable to other fields. In chemistry, the principle of *evolutive abiogenesis*

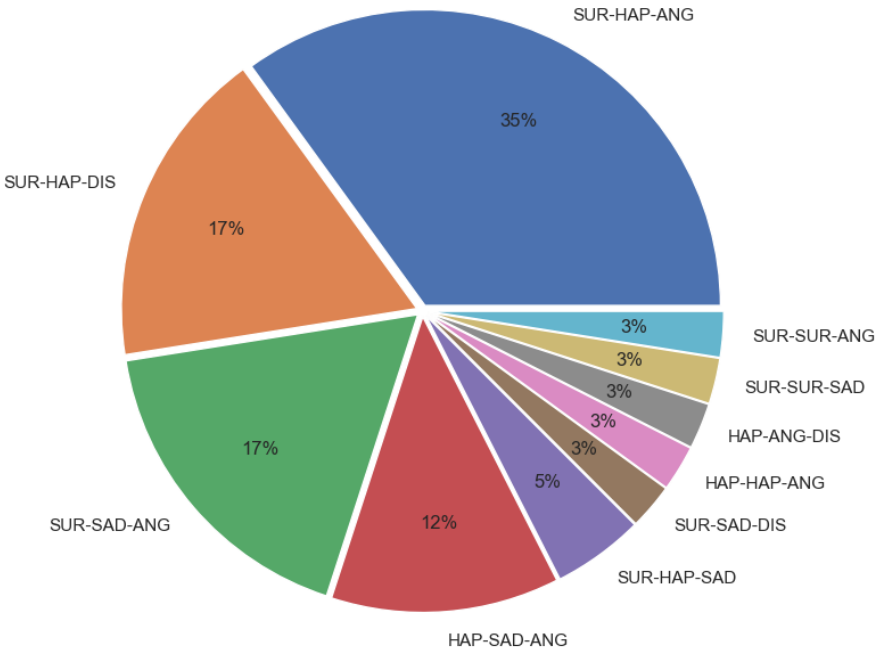


Figure 4.3.6 – Distribution of combinations of inferred archetypal emotions based on 40 bootstrap runs.

describes a process in which simple organic compounds increase in complexity [Miller, 1953]. In the following experiment a structured chemical space is learned using as side information the *heat capacity* C_v which quantifies the amount of energy (in Joule) needed to increase 1 Mol of molecules by 1 K

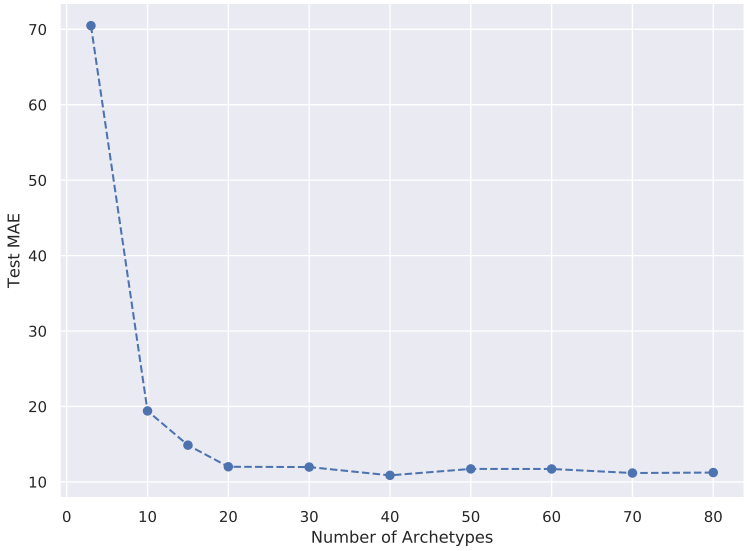


Figure 4.3.7 – Model selection on the QM9 data set: Mean absolute error (reconstruction loss) vs. number of archetypes on the test set.

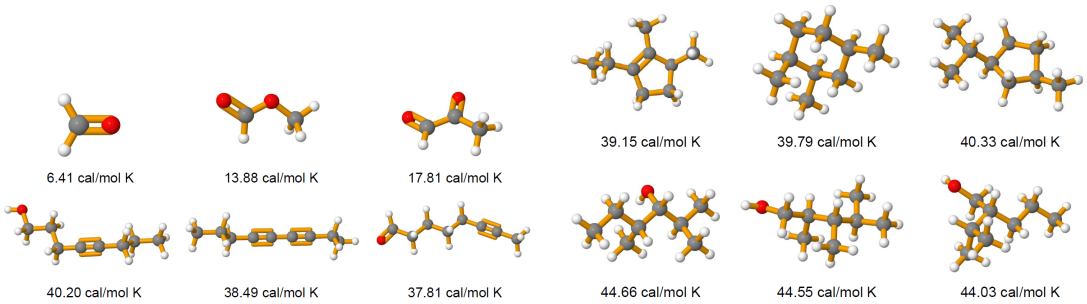


Figure 4.3.8 – Both panels illustrate a comparison between archetypal molecules, where the underlying latent representation is informative with respect to the molecular property *heat capacity*. Each row contains the three molecules of the test set that have been mapped closest to a specific vertex of the latent simplex. Panel (a) compares archetypal *linear* molecules characterized by a short chain structure versus long chained molecules. Panel (b) compares archetypal molecules with similar masses but different geometric configuration, i.e. with and without a cyclic structure.

at constant volume. A high C_v number is important e.g. in applications dealing with the storage of thermal energy [Cabeza et al., 2015]. In the following, all experiments are based on the QM9 data set [Ramakrishnan et al., 2014, Ruddigkeit et al., 2012], which contains molecular structures and properties of 134k organic molecules. Each molecule is made up of nine or less atoms, i.e. C, O, N, or F, without counting hydrogen. The QM9 data set is based on ab-initio density functional theory (DFT) calculations.

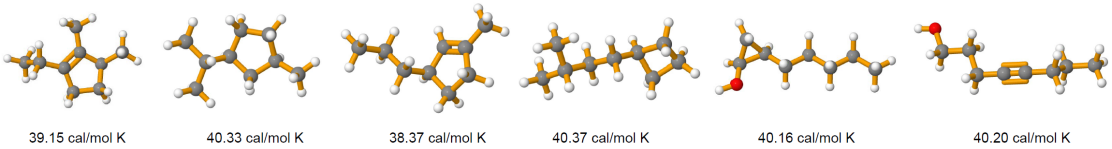


Figure 4.3.9 – Interpolation between two archetypal molecules produced by deep AA. The labels display the heat capacity of each molecule. Here, only a single example is shown but similar results can be observed for other combinations of archetypes.

Experiment Setup. A total of 204 features were extracted for every molecule using the Chemistry Development Kit [Steinbeck et al., 2003]. The neural architecture used has 3 hidden FC layers with 1024, 512 and 256 neurons, respectively, and ReLU activation functions. For all experiments, the model was trained in a *supervised* fashion by reconstructing the molecules and the side information simultaneously. In *Experiment 1*, model selection was performed by continuously increasing the number of latent dimensions. Based on the knee of the mean absolute error (MAE), the appropriate number of latent archetypes was selected. In *Experiments 2 and 3*, the number of latent dimensions was fixed to 19, corresponding to the optimal number of 20 archetypes from the model selection procedure. During training, the Lagrange multiplier λ was steadily increased by increments of 1.01 every 500

iterations. For training, the Adam optimizer [Kingma and Ba, 2014] was used, with an initial learning rate of 0.01. A learning rate decay was introduced, with an exponential decay of 0.95 every 10k iterations. The batch size was 2048 and the model was trained for a total of 350k iterations. The data set is divided in training and test set with a 90%/10% split. For visualization, the 3-dimensional molecular representations haven been created with [Jmol, 2019].

Experiment 1: Model Selection. The mean absolute error is assessed while varying the number of archetypes. The result is shown in Figure 4.3.7. Model selection is performed by observing for which number of archetypes the MAE starts to converge. The knee of this curve is used to select the optimal number of archetypes, which is 20. Obviously, if the number of archetypes is smaller, it becomes more difficult to reconstruct the data. This is explained by the fact that there exists a large number of molecules with *very similar* heat capacities but at the same time *distinctly different* geometric configurations. As a consequence, molecules with different configurations are mapped to archetypes with the similar heat capacity, making it hard to resolve the many–to–one mapping in the latent space.

Experiment 2: Archetypal Molecules. Archetypal molecules are identified along with the heat capacities associated with them. A fixed number of 20 archetypes is used for optimal exploration-exploitation trade-off, in accordance with the model selection discussed in the previous section. In chemistry, the heat capacity at constant volume is defined as $C_v = \frac{d\epsilon}{dT}\big|_{v=const}$ where ϵ denotes the energy of a molecule and T its temperature. This energy can be further decomposed into different parts, such that $\epsilon = \epsilon^{Tr} + \epsilon^{R} + \epsilon^{V} + \epsilon^{E}$. Each part is associated with a different degree of freedom of the system. Here, Tr stands for translational, R for rotational, V for vibrational and E for the electronic contributions to the total energy of the system [Atkins and de Paula, 2010, Tinoco, 2002]. With this decomposition in mind, the different archetypal molecules associated with a particular heat capacity are compared in Figure 4.3.8. In both panels of that figure, the rows correspond to the three molecules in the QM9 data set (test set) that have been mapped closest to a vertex of the latent simplex and have thus been identified as being extremes with respect to the heat capacity. Out of a total of 20 vertices, molecules in close proximity to four of them are displayed here. Panel 4.3.8a shows the configuration of six archetypal molecules. The upper three are all associated with a low heat capacity while the lower three all have a high heat capacity. This result can easily be interpreted, as the lower heat capacity can be traced back to the shorter chain length and the higher number of double bonds of these molecules, which makes them more stable and results in a lower vibrational energy V and subsequently in a lower heat capacity. The inverse is observed for the linear archetypal molecules with higher heat capacities, which show, relative to their size, a lower number of double bonds and a long linear structure. Panel 4.3.8b shows both linear (lower row) and non-linear archetypal molecules (upper row) but with similar atomic mass. Here, the non-linear molecules containing a cyclic structure in their geometry, are more stable and therefore have an overall slightly lower heat capacity compared to their linear counterparts of the same weight, shown in the second row.

Experiment 3: Interpolation Between Two Archetypal Molecules. Interpolation is performed by plotting the samples from the test set which are closest to the connecting line between the two archetypes. As a result, one can observe a smooth transition from a molecule with a ring structure to a linear chain molecule. Both the starting and the end point of this interpolation is characterized

by a similar heat capacity, such that these archetypes differ only in their geometric configuration but not with respect to their side information. As a consequence, any molecule in close proximity to that connecting line can differ only with respect to its structure, but *must* display a similarly high heat capacity. Figure 4.3.9 shows an example of such an interpolation.

Experiment 4: The Role of Side Information and the Exploration of Chemical Space. Deep AA structures latent spaces both according to the information contained in the input to the encoder as well as the side information provided. As a consequence, any molecule characterized as a *true* mixture of two or more archetypes, given a specific side information such as *heat capacity*, might suddenly be identified as archetypal should the side information change accordingly. In the following, archetypal molecules with respect to *heat capacity* as the side information are compared to archetypes obtained while providing the *band gap energy* of each molecule as the side information. In Figure 4.3.10a archetypal molecules with both the highest and the lowest heat capacities are displayed while 4.3.10b shows archetypes with highest and lowest band gap energies. The archetypes significantly differ in their structure as well as their atomic composition. For example, archetypal molecules with low heat capacity are rather small, with only few C and O atoms, while archetypal molecules with a low band gap energy are characterized by ring structures containing N and H atoms. This illustrates the essential role of side information for learning and subsequently enabling the interpretation of the latent representation.

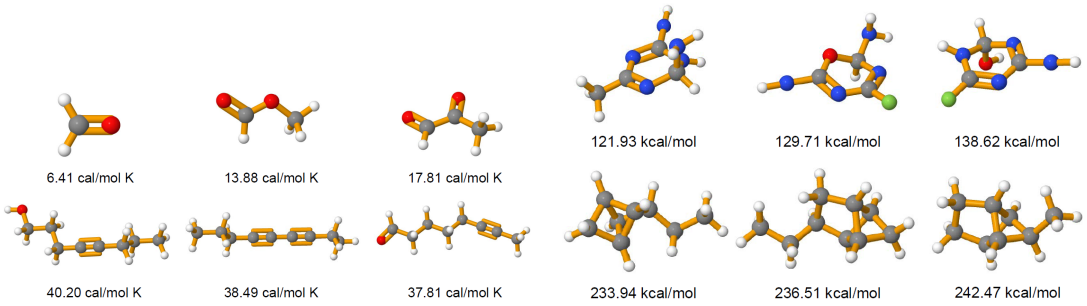


Figure 4.3.10 – Panels (a) and (b) compare archetypal molecules identified using different side information: Here, the labels correspond to the heat capacity (panel a) and the band gap energy (panel b). The rows contain the three molecules of the test set closest to the given archetype.

4.3.5 Alternative Priors For Deep Archetypal Analysis

The standard normal distribution is a common choice for the prior distribution $p(\mathbf{t})$ due to its simplicity and closed form expression for the KL divergence. However, alternative priors might influence the inferred archetypes or prove beneficial when learning the structure of the latent space. Leaving aside the wide range of well explored priors for vanilla VAEs, we explore a hierarchical prior that directly corresponds to the generative model of linear AA presented in Eq. 4.5, i.e. isotropic Gaussian noise

around a linear combination of the archetypes:

$$\mathbf{m} \sim \mathrm{Dir}_k(\boldsymbol{\alpha} = \mathbf{1}) \quad \wedge \quad \mathbf{t} \sim \mathcal{N}(\mathbf{m}Z^{\text{fixed}}, \mathbf{I}) \tag{4.16}$$

The estimation of the KL divergence given in Eq. 4.8 is based on Monte-Carlo sampling. In order to qualitatively compare the standard normal prior and the sampling Dirichlet prior, we train the respective deep AA model on the JAFFE data set with $k = 4$ archetypes, implying a 3-dimensional latent space. The architecture used is similar to the previous experiments but we additionally *learn* the variance of the decoder. The Lagrange parameters or weights in Eq. 4.15 are set to 1×10^3 for the archetype loss and to 1×10^2 for the KL divergence.
Finally, Figure 4.3.11 shows examples of the inferred archetypes for the standard normal prior (panel a) and the sampling Dirichlet prior (panel b). In conclusion, different priors do not seem to strongly affect the inferred archetypes. However, the structure of latent spaces do differ, which can be seen when projecting them onto the first two principal components as shown in Figure 4.3.12. As a reference, a uniformly filled simplex would result in a triangular shaped projection. The difference seen here is caused by large gaps in the higher-dimensional simplex when using the hierarchical prior, which we assume is mainly due to the high variance estimation of the KL divergence.
In our experience, the choice of the prior is not of primary concern for finding meaningful archetypes, as long as it encourages the latent space to be spread out inside the simplex, be that via a standard normal, a uniform or – as in this case – a hierarchical prior.

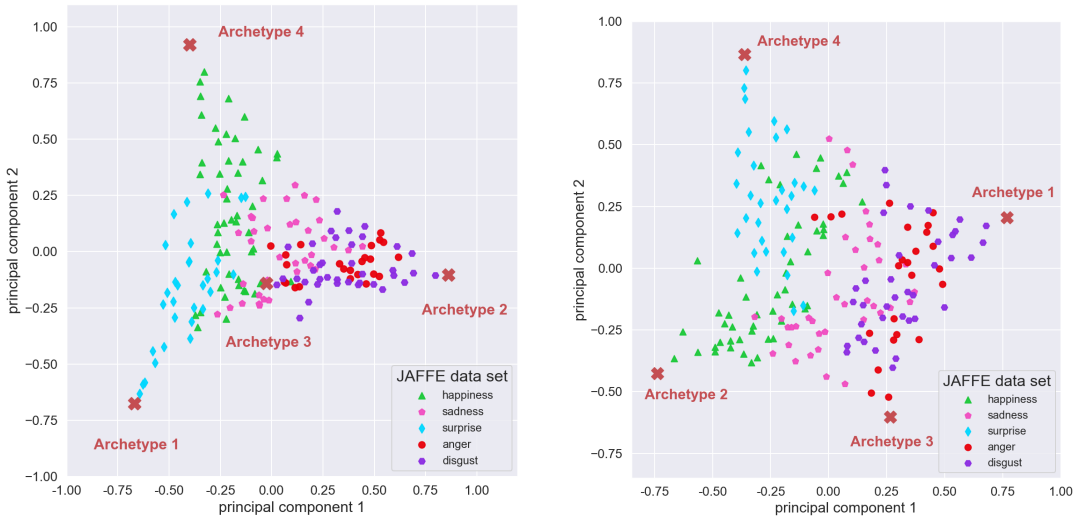
(a) Archetypes learned using the standard normal prior.

(b) Archetypes learned using the sampling Dirichlet prior.

Figure 4.3.11 – Deep AA with $k = 4$ archetypes using two different priors, which both identify similar archetypes.

4.4 Practical Considerations for using Deep AA

In the following, we provide general aspects worth considering when deciding whether the use of deep AA might be appropriate, given a specific data set.
Linear Archetypal Analysis relies on the additivity assumption, as data points are described as a weighted sum of the archetypes, with the weights constrained to be non-negative. This mixing procedure is performed directly on the input data. In deep AA, on the other hand, the mixing is performed only on the latent representation of the input data. This allows more flexibility regarding the type of input data, e.g. text, images etc., but it also relaxes the additivity assumption, as the encoder *learns* a representation on which this assumption is (approximately) valid. Nevertheless, for the interpretation of the archetypes, convex mixing should *a priori* be a justifiable assumption. In general, for data without explicit cluster structure (deep) AA poses an interesting possibility. On the ten digits in MNIST for example, a clustering might be more appropriate while deep AA could provide additional insight when applied on a single digit class. The goal of deep AA is to optimize for the most



(a) JAFFE latent space with standard normal prior. (b) JAFFE latent space with sampling Dirichlet prior.

Figure 4.3.12 – Latent spaces for the two different priors projected onto the first two principal components. The explained variances are: (a) 0.74 and (b) 0.757.

compact latent code such that interpretability of the archetypes – on a qualitative level – remains possible. Having too many archetypes would likely obfuscate the meaning of an individual archetype. Furthermore, as deep AA relies on the latent simplex as a geometrical structure, Euclidean distances need to be meaningful with respect to the dimensionality of the latent space, generally encouraging *low* dimensionality. But this is of course true for all flavors of AA. In practice, inspecting the local data density in the neighborhood of the inferred archetypes in latent space is an important post-processing step. If an archetype appears to have no latent samples close to it, it might be considered an outlier.

4.5 Conclusion

In this chapter, we introduced an extension of linear Archetypal Analysis, a technique for exploratory data analysis and interpretable machine learning. By performing Archetypal Analysis in the latent space of a deep information bottleneck, we have demonstrated that the learned representation can be structured in a way that allows it to be characterized by its most extremal or archetypal representatives. As a result, each observation in the data set can be described as a convex mixture of these extremes. Endowed with such a structure, a latent space can be explored by varying the mixture coefficients with respect to the archetypes, instead of exploring the space by uniform sampling. Furthermore, we have demonstrated the need for including side information into the process of learning latent archetypal representations. Extremeness can only be understood with respect to a given property. Therefore, providing such a property through side information is essential in order to learn interpretable latent archetypes. In contrast to the original archetype model, our method offers three advantages: First, our model learns representations in a data-driven fashion, thereby reducing the need for expert knowledge. Second, our model can learn appropriate transformations to obtain meaningful archetypes, even if non-linear relations between features exist. Third, the incorporation of side information. The application of this new method is demonstrated on a sentiment analysis task, where emotion

archetypes are identified based on female facial expressions, for which multi-rater based emotion scores are available as side information. A second application illustrates the exploration of the chemical space of small organic molecules and demonstrated how crucial side information is for interpreting the geometric configuration of these molecules.

Figure 4.5.1 – JAFFE archetypes and latent distribution trained on 20 distinct bootstrap data sets.

Figure 4.5.2 – JAFFE archetypes and latent distribution trained on 20 distinct bootstrap data sets.

5 Applications in Neurophysiology

Neurophysiology is a subfield of neuroscience whose subject is the study of functional aspects of the nervous system (as opposed to structural aspects). EEGs are recordings of large-scale *electric* signals from the nervous system, and as such belong to the branch of *electrophysiology*. In this chapter, we introduce two applications; the first illustrates the potential of clinical EEG for the prognosis of cognitive decline in Parkinson's disease, the second shows how *Deep Archetypal Analysis* can provides insights into functional changes associated with the progression of Parkinson's disease.

5.1 Preprocessing of EEG data

A standard clinical EEG is obtained by placing electrodes onto the scalp of the patient and with sampling rates of usually >250Hz the recorded signal is amplified and stored digitally. Although the goal of EEG is to accurately capture the neural activity of the subject, the recorded signal is inevitably contaminated by a range of bioelectrical and external signals not originating in the brain. The

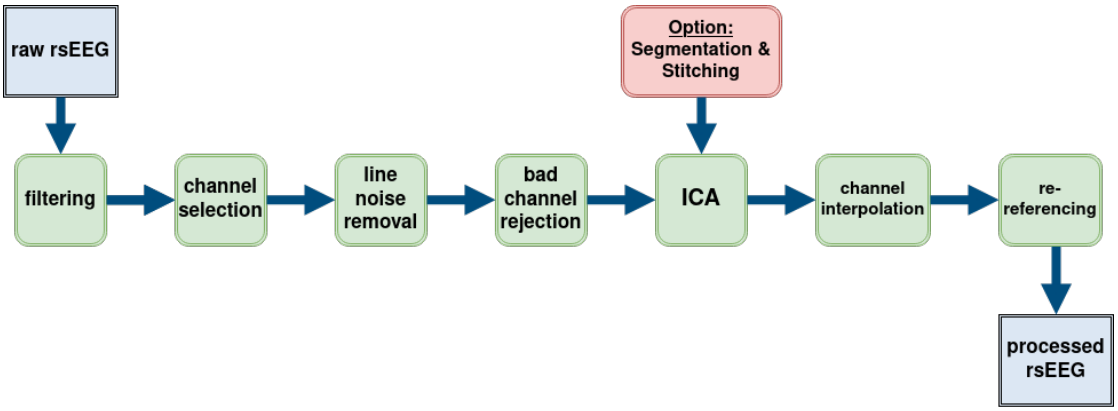


Figure 5.1.1 – A standard pre–processing pipeline for EEG recordings.

contamination of the signal of interest is caused by eye movement, cardiac activity, muscular activity, e.g. involuntary body movements, respiration and transpiration. In addition, external contaminants like power line noise, switching noise and other electronic devices are potentially captured during

the recording of the EEG. As the cortical electric activity recorded on the scalp resides in a range between $< 1\text{Hz} - 45\text{Hz}$, which overlaps with the frequency range of many of the contaminating sources, it is impossible to *unambigousley* differentiate signals of cortical origin from non–cortical signals. Nevertheless, a *clean* EEG, i.e. an EEG associated with a high degree of certainty of containing mostly signals of cortical origin, is of tremendous importance for downstream analyses. The EEG pre–processing pipeline shown in figure 5.1.1 is designed to suppress non-cortical signals and to output cleaned EEG recordings. In the following, important aspects of this pipeline are highlighted.

5.1.1 Filtering

Filtering is often the first step in any EEG pre–processing pipeline as it allows for improved visual inspection before continuing further processing of the signal. EEG recorded on the scalp reliably records neural activity up to the γ-band, i. e. up to 45Hz, and possibly beyond. Filtering is performed offline, using a finite impulse response filter (FIR filter). A high–pass and a low–pass were combined in order to obtain a pass-band filter. Figure 5.1.2 shows a raw EEG recording sampled at 19 different

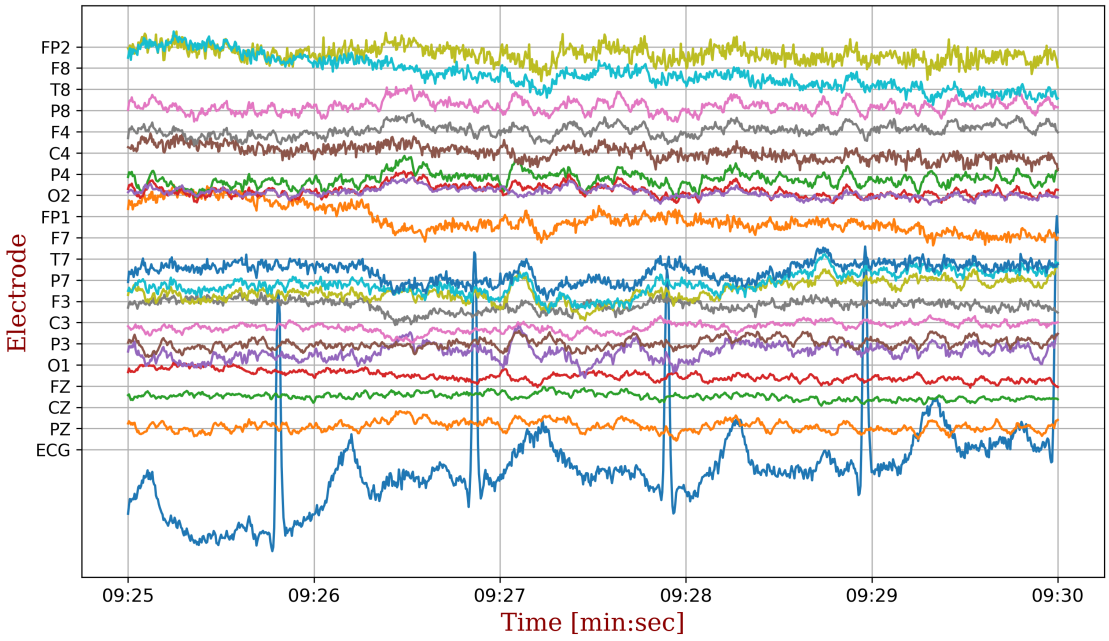


Figure 5.1.2 – Segment of a raw EEG sampled at standard 10–20 electrode locations with a sampling frequency of 250Hz. The last channel shows the patient's heart beat. Comparing the amplitude of neural and cardiac activity, one can see that brain activity signals have on average much lower amplitudes.

scalp locations covering the scalp with an average inter-electrode distance of $\approx 7\text{cm}$.
For a signal decomposed into its various sine and cosine components of different frequencies, filtering is the process by which each of these components get attenuated differently, depending on their frequency. Frequencies within the user-defined *passband* would – ideally – pass the filter unchanged while frequencies falling into the range of the *stopband* would be completely attenuated. At the cut-off frequency, the demarcation between passband and stopband, this would imply an infinitely

sharp fall-off of gain together with a perfectly flat amplitude characteristic in both the passband and the stopband. Unfortunately, these ideal filter characteristics can only be approximated, trading-off several properties like the steepness of the change of attenuation in the transition band, amplitude of the ripples in passband and stopband, and others. Figure 5.1.3 shows the frequency response

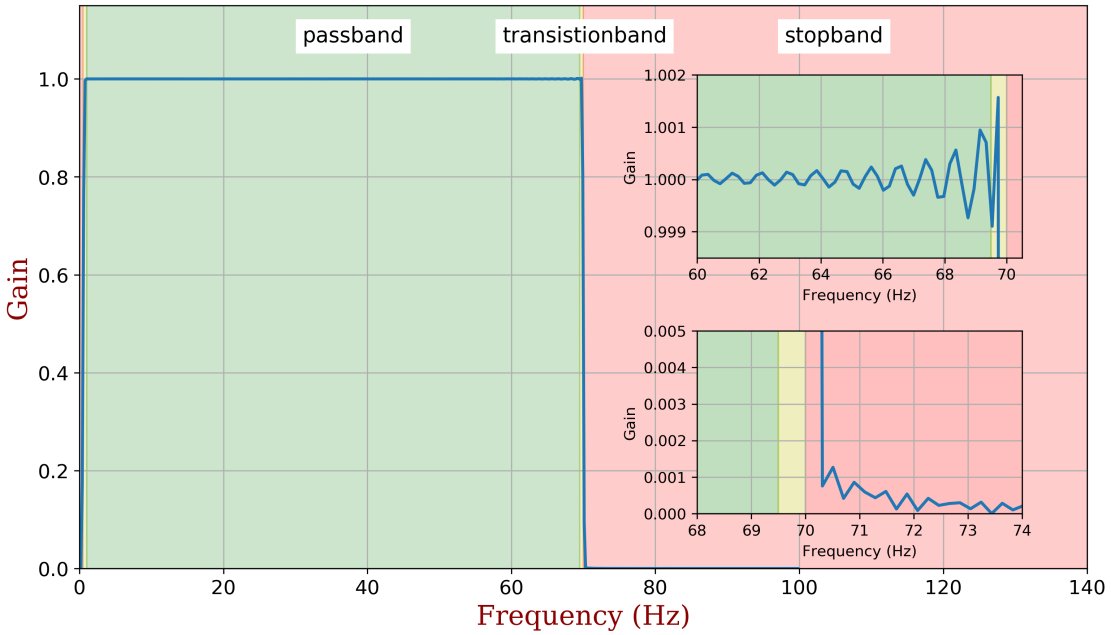


Figure 5.1.3 – Frequency response of the FIR filter used in the pre-processing pipeline. With a narrow transitionband, close-to-unit gain over the passband and strong attenuation in the stopband, ideal filter characteristics are well approximated. The inlays show the small frequency-dependent ripples around the transistionband.

of the FIR filter used for band pass filtering the raw EEG in figure 5.1.2 between 0.5Hz and 70Hz. The characteristics of an ideal filter are well approximated for the given task as the filter shows almost constant unit gain over the passband, with a quick fall-off, i. e. a narrow transistionband, while strongly attenuating the signal within the stopband. However, looking at the inlays, which show a zoomed-in view of the transition zone between pass- and transistionband (upper inlay) and between the transistionband and the stopband (lower inlay), small ripples (filter ringing) are clearly visible. Close to the transistionband the deviations from the ideal filter characteristics are maximal. Nevertheless, the frequency-dependent gain associated with these ripples is only around one-tenth of a percent, which is more than acceptable for use in the EEG pre-processing pipepline. The trade-off for FIR filter designs with low ripple effects and narrow transistionband is a large filter delay: Any signal filtered with a FIR filter experiences a *constant* shifting of its phase, independent of frequency. The closer ideal filter characteristics are approximated, the larger the delay will be. But for the subsequent analyses presented here, a large delay is of no consequence as the typical length of resting state EEG recordings is of several minutes and, in contrast to evoked potentials, there are no stimuli relative to which a neural response time would need to be measured.
Figure 5.1.4 shows the filtered EEG after applying the FIR filter of figure 5.1.3 to the raw EEG shown in figure 5.1.2. Generally, the advantage of FIR filters are (i) the constant delay over all frequencies such that the signal shape is not influenced by the phase shifts and (ii) their stability, as FIR filters are

non-recursive filters (in contrast to infinite impulse response (IIR) filters). But if large filter delays are problematic or computationally efficient filtering, e. g. for real-time applications, is necessary, then FIR filters are not recommended.

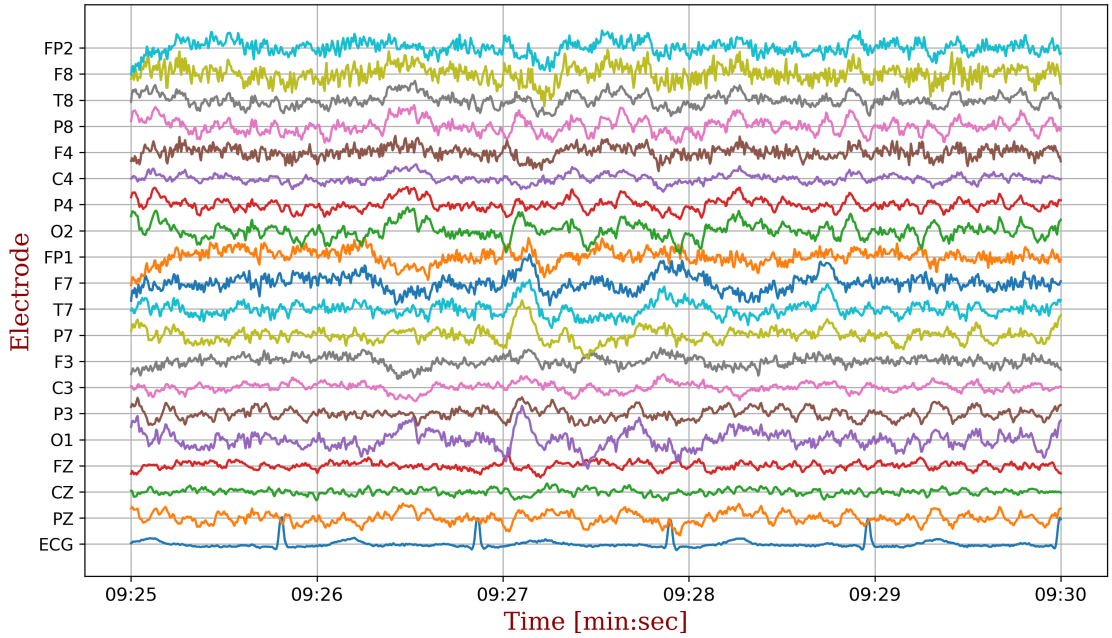
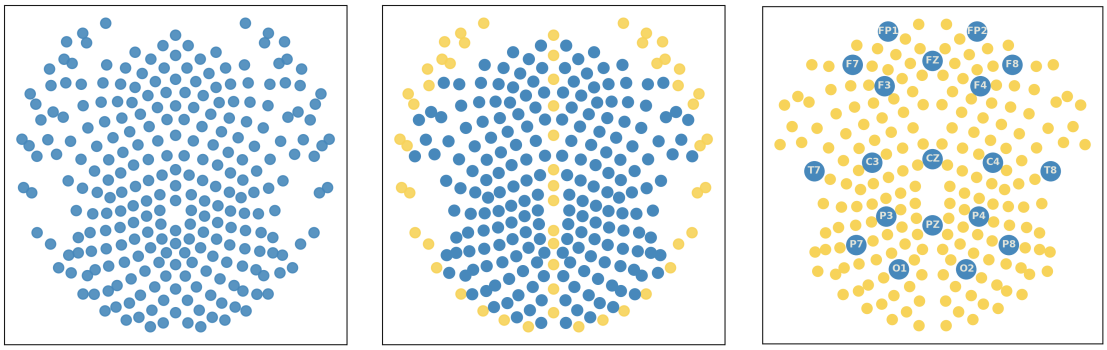


Figure 5.1.4 – Filtered EEG between 0.5Hz and 70Hz based on the FIR filter design shown in figure 5.1.3. The ECG signal has been rescaled for better visibility.

5.1.2 Channel selection

EEG can be sampled at different spatial resolutions depending on the number of electrodes available. While a dense spatial sampling of the scalp is of importance especially for source reconstruction, it generally allows for a better approximation of a reference electrode with a constant potential. This is of great importance as electric potentials are always measured with respect to a reference which has to be constant over time for EEG measurements to be reliable. Using the mean of all recording channels at each time point, the quality in approximating an inactive reference increases with spatial coverage, as the electrical potentials, integrated over the entire surface of the body, are constant. Similarly, it is assumed that the potentials, integrated over the surface of the *scalp*, fluctuate less with increasing spatial coverage.
Selecting an appropriate subset of electrodes for further analysis usually depends on the downstream task. Figure 5.1.5 shows common subsets of electrodes based on high-density EEG with 257 electrodes. The 10–20 system, a subset of electrodes commonly used in the clinical setting, is shown in panel 5.1.5c. Although high density EEG was used in the subsequent applications, the preprocessing pipeline is described based on the 19 electrodes of the 10–20 system which makes the figures less convoluted.



(a) 257 electrodes (b) 202 electrodes (c) 19 electrodes (10–20 system)

Figure 5.1.5 – Spatial sampling of the scalp with 257 electrodes implies an average inter-electrode distance of 2cm. Depending on recording quality, electrodes placed on the neck, ears and cheeks are sometimes excluded as they might show a higher portion of spurious signals, leaving up to 213 electrodes, when including the mid-line electrodes. In clinical settings, a subset of 19 electrodes is commonly used, known as the 10–20 standard.

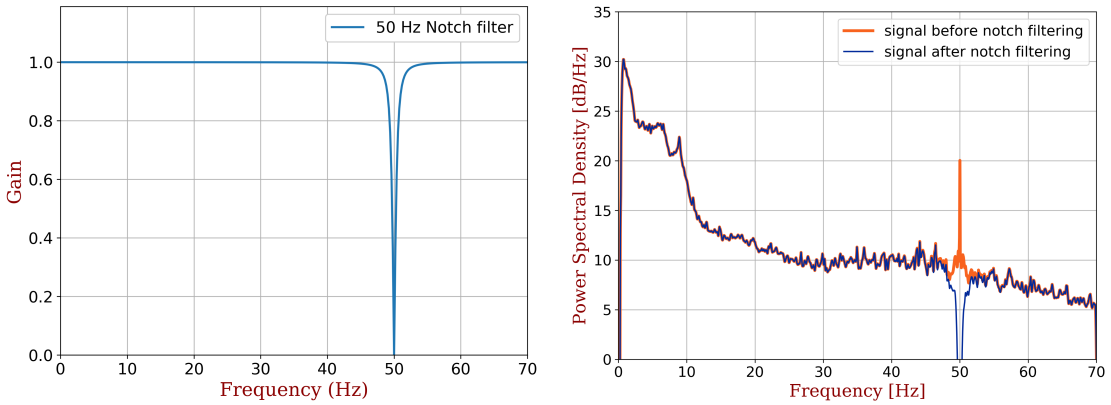
5.1.3 Line noise removal & bad channel rejection

EEG recorded with a device connected to the electric power grid is usually contaminated by power-line noise of 50Hz or 60Hz. Using a band-stop filter with a very narrow bandwidth, i. e. a high quality factor, the narrow frequency band of the line noise can be suppressed while leaving the remaining spectrum approximately unchanged. This special type of filter is known as a "notch filter". Figure 5.1.6a shows the typical magnitude response of such a filter. After applying the 50Hz notch filter to the EEG shown in figure 5.1.4 its spectrum is re-calculated and both spectra – before and after notch-filtering – are compared in figure 5.1.6b. The high q-factor ensures a narrow bandwidth which leads to distortions of the notch-filtered EEG only within a narrow band around 50Hz.

During EEG recording, it is also possible for electrodes to lose contact with the scalp due to head movement or improper attachment. Also faulty contacts in the cables connecting the electrodes to the recording device might go undetected. Especially in high-density settings with hundreds of electrodes, it is important to detect and remove these "bad" channels. Based on a range of easily computed features, such as the mean of the channel's correlation coefficients with other channels, the variance of the channel and the Hurst exponent of the channel, the quality of the individual channels is assessed. It has become standard to reject channels with a Z-score > 3 with respect to any of the aforementioned features [Nolan et al., 2010].

5.1.4 Independent Component Analysis

Electrical potential differences recorded on different electrodes positioned at different scalp locations are the result of a mixing of underlying components of activity – not all of them originating in the cortex. Common sources of electric activity beside neural generators, are eye blinks and eye movements, muscle activity, breathing, cardiac rhythm, defective electrodes and defective cables, electric line noise and unspecified external wireless signals. As only signals of cortical origin are of interest for further analysis, we would like to identify the original components of brain activity while only being able to



(a) Magnitude response (q-factor=50)

(b) Influence of the notch filter on the spectrum

Figure 5.1.6 – A narrow bandwidth notch filter is used to remove power-line noise of 50Hz. Signal distortions are kept at a minimum as can be seen by comparing the spectra before and after notch filtering.

observe a mixture of the above components. In mathematical notation, assuming linear mixing in accordance with the – also linear – Maxwell's equations [Maxwell, 1865], the problem, referred to as independent component analysis or ICA, is described as follows:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{n} \mathbf{a}_i s_i \tag{5.1}$$

with the column vector $\mathbf{x}$ denoting the scalp EEG recording, whose elements are the mixtures $x_1, \ldots, x_n$, and $\mathbf{s}$ containing the sources $s_1, \ldots, s_n$. The matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, with elements a_{ij}, is the so–called mixing matrix. Furthermore, we have assumed that we recorded n linear mixtures of n independent components.
The problem is to estimate both $\mathbf{A}$ and $\mathbf{s}$ under as general assumptions as possible. The assumption of ICA is that the components s_i are statistically independent. With an estimate of $\mathbf{A}$ obtained, its inverse $\mathbf{W}$ can be computed, from which the estimate of the underlying sources $\mathbf{s}$ follows:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{5.2}$$

As such, ICA is one method for performing *blind source separation* (BSS). One ambiguity of ICA refers to the variances of the estimated sources which cannot be determined. With $\mathbf{s}$ and $\mathbf{A}$ both being unknown, a multiplier $c \in \mathbb{R}$ could always be canceled by dividing the corresponding column $\mathbf{a}_i$ of $\mathbf{A}$ by c. Another ambiguity concerns the order of independent components or sources, which is impossible to determine. The reason for this is, that the order of the terms in the sum in 5.1 can be changed freely as both $\mathbf{A}$ and $\mathbf{s}$ are unknown. Finally, there is also ambiguity with respect to the sign of an independent component – multiplying any independent component by -1 (phase reversal) would not affect the model in 5.1.
Figure 5.1.7 shows the corresponding segment of estimated sources for the bandpass-filtered EEG shown in figure 5.1.4. ICA was performed based on the InfomaxICA proposed in [Bell and Sejnowski, 1995]. The estimated mixing matrix $\mathbf{A}$ for this EEG is shown in figure 5.1.8a while the topographical distributions of the individual independent components, projected into electrode space, are shown in figure 5.1.8b. These activation maps, through visual inspection, allow to attribute the activity

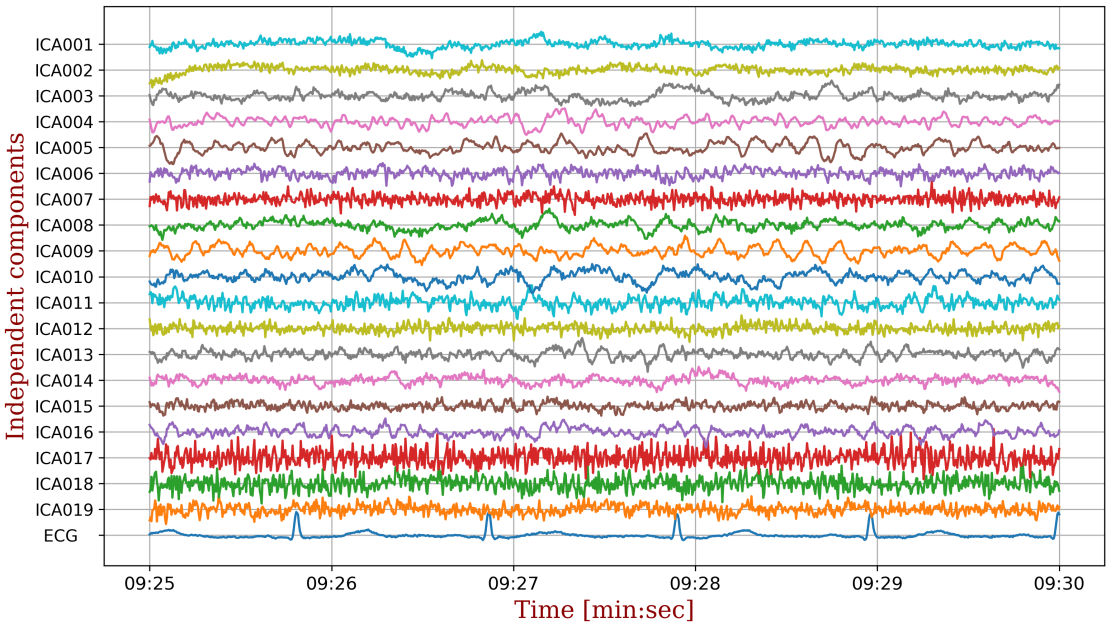
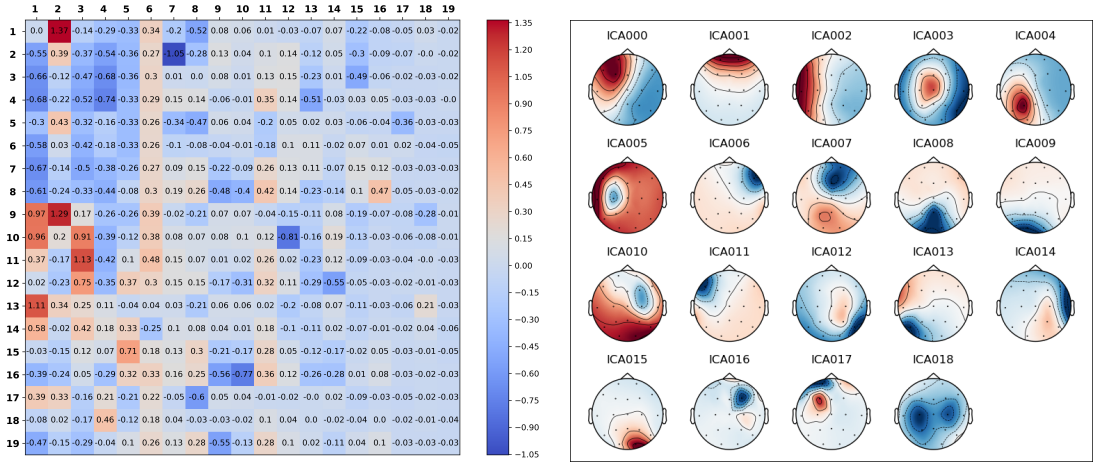


Figure 5.1.7 – Decomposition of an EEG recording into independent componenets using InfomaxICA. The segment of source time series shown here, is a decomposition of the EEG segment (in electrode space) shown in figure 5.1.4.



(a) Mixing matrix for the EEG in figure 5.1.4

(b) Topography associated with individual sources

Figure 5.1.8 – A visual inspection of activation maps reveals the likely origin of the electric activity of a given source. Removing sources of non-cortical origin and reconstructing the EEG signal in electrode space will yield a signal less affected by artifacts.

of individual sources to a likely point of origin. The topography shown in figure 5.1.8b associated with independent component "ICA001" is, for example, a typical pattern produced by eye movement. As such, removing the source time series "ICA001" and reconstructing the sensor space EEG signal would yield an EEG (almost) free of eye movement artefacts. After setting to zero the source $s_{i=1} = 0$,

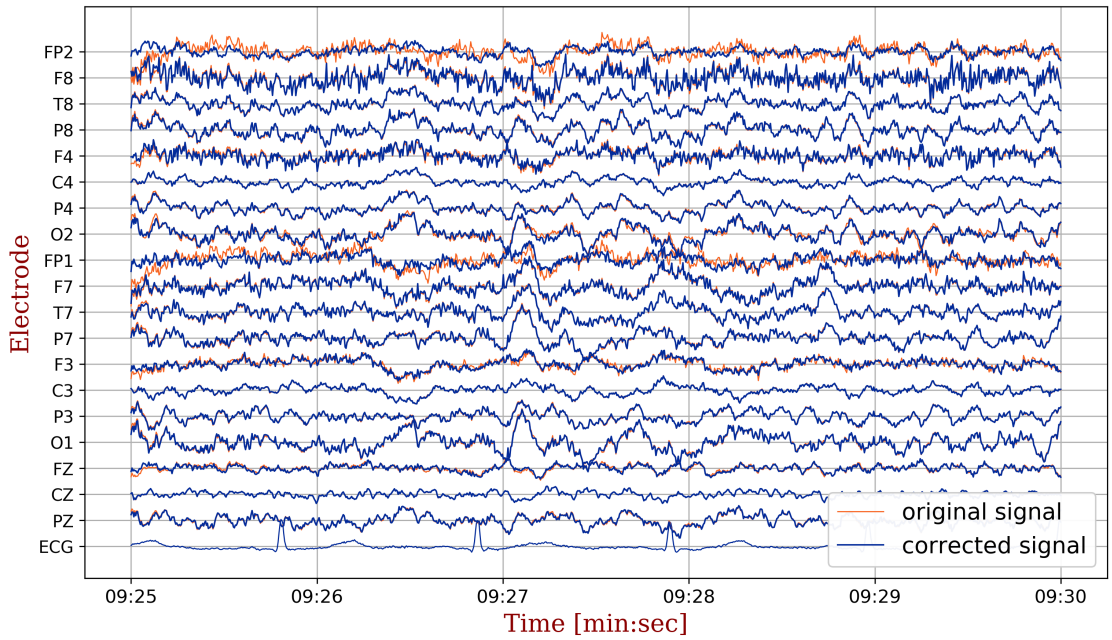


Figure 5.1.9 – After setting to zero the activity of non-cortical sources s_i of EEG in equation 5.1, a reconstruction of the signal in electrode space can be overlaid with the original signal in order to evaluate the effect of discarding certain sources.

which corresponds to component "ICA001", and additionally the sources $s_{i=15} = 0$ and $s_{i=17} = 0$, which are likely related to eye blinks, muscular activity or impedance, the signal in electrode space can be reconstructed. Figure 5.1.9 shows the original signal, with the corrected signal overlaid. The most prominent differences after removing eye movement and eye blink related sources and reconstructing the electrode time series, are visible on the frontal electrodes "FP1" and "FP2". This is exactly where one would expect eye activity related corrections to show the greatest impact.

5.1.5 Channel interpolation

One possibility to deal with channels identified as *bad*, according to section 5.1.3, is to replace them with the time course estimated based on the activity and locations of other electrodes. Commonly, such interpolation methods use weighted distance metrics such as nearest-neighbor, linear, or spline. Of course an interpolation cannot provide new information, but often interpolation is preferred over simply discarding bad channels as it makes subsequent processing more manageable. If, for example, averaging across subjects is performed as a downstream task, it is usually easier to deal with subjects when they all have an identical number of channels. Furthermore, some methods might provide more robust results if interpolated channels are provided, e. g. spatial filters such as the surface Laplacian or source reconstruction. Especially in high density settings with up to 300 electrodes, bad channels will occur regularly, making dealing with this problem quite common. The disadvantage, of course, is that interpolation reduces the rank of the data matrix which needs to be accounted for, e. g. when calculating the inverse of such a matrix [Cohen, 2014].

5.1.6 Re-referencing

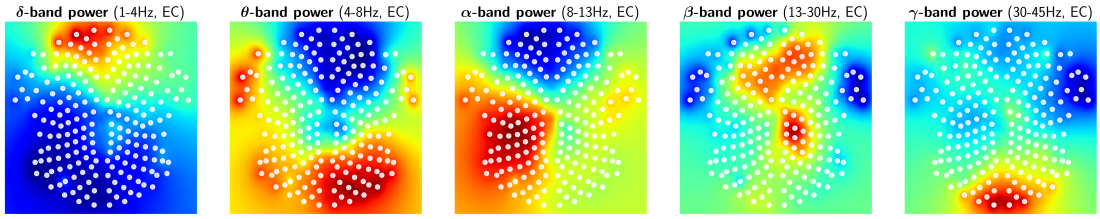
Re-referencing refers to the process of changing the reference offline after recording. Physically, *voltage* describes a *potential difference*, i. e. a voltage signal always refers to some set reference to which that difference is measured. In EEG, this reference is often an electrode with central location (e. g. FCz or Cz) or electrodes attached to the earlobes or mastoids. Generally, a reference electrode should be as little as possible influenced by brain activity. But other considerations may play a role as the choice of reference might amplify or reduce signals recorded at specific regions of the scalp, e. g. the mastoid reference tends to amplify fronto–central components. The analysis described in the following sections is based on EEG recordings that have all been re-referenced to a common average montage using the average of all "cleaned" electrodes.

5.2 Predicting cognitive decline in Parkinson's disease

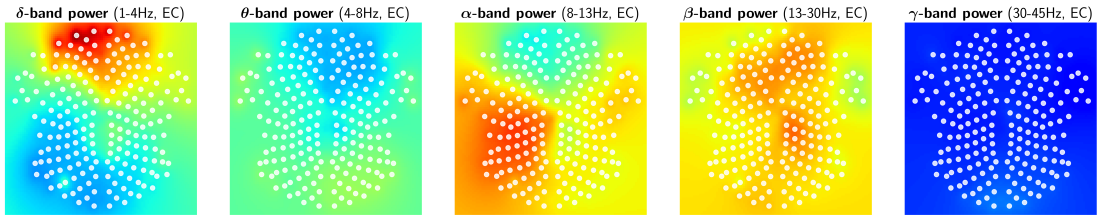
Parkinson's disease (PD) is a progressive neurodegenerative disorder which currently requires motor signs for diagnosis, but shows more widespread pathological alterations from its beginning. Compared to age-matched individuals without PD, patients have up to a six-fold increased lifetime risk of developing dementia. Although a milestone study with 20 years follow-up of initially 136 newly diagnosed patients [Hely et al., 2008] has shown that dementia was present in 83% of 20-year survivors, onset of dementia varies considerably. According to Buter et al. [2008], the 4-year prevalence of dementia was 52% while the 12-year prevalence was around 60% based on 233 patients with PD. This combination of high risk of dementia together with the varied range of individual decline onset makes the search for reliable prognostic biomarkers for assessing future cognitive deterioration in early stages of PD an important one – both for individualized counseling and treatment but also for study stratification when evaluating efficacy of pharmacological intervention. Because measures based on electroencephalography (EEG) are safe, inexpensive, and widely available, they are attractive candidate biomarkers.

5.2.1 Spectral EEG Biomarkers

Biomarkers derived from the spectral properties of EEG signals have a long standing history [Coben et al., 1983] . With the advent of affordable compute power and the publication of the Fast Fourier Transform (FFT), an efficient algorithm to compute the Discrete Fourier Transform [Cooley and Tukey, 1965], the "reformatting" of EEG data into its spectral components has lead to the identification of many candidate diagnostic, predictive, prognostic, and therapeutic biomarkers for various neurological disorders. The success of FFT in the analysis of quantitative EEG (qEEG) is due to the fact that many pathological EEG patterns translate into pathological *spectral* patterns which can be easily recognized [Matthies and Brödemann, 1981]. Because the frequency spectrum of EEG data shows decreasing power at increasing frequencies, i. e. power $\sim 1/f$, it has become standard to analyze spectral properties of EEG within five pre-defined frequency bands ranging roughly from $1-45$Hz. Higher frequencies up to 500Hz, so called "High Frequency Oscillations" (HFO), which have shown to contain important neurological information, are blocked by the skull which acts as a low-pass filter. HFO's are of great interest e. g. when localizing the generators of epileptic seizures [Gotman, 2010]. In such cases, intra-cranial electrodes are used which are placed directly onto the cortex or, in form of needles, inserted deep into the brain. For non-invasive scalp EEG, the frequency bands of interest are given in table 5.1. As an example, figure 5.2.1 shows the spatial distribution of EEG



(a) Band power normalized within each band



(b) Band power normalized to the δ-band which exhibits maximum power

Figure 5.2.1 – Intensity, spatial and temporal distribution of EEG power within the five pre-defined bands may serve as the basis for deriving spectral biomarkers. With EEG power distribution following a "$1/f$" shape, details in higher band, e. g. the γ-band, are visible only with appropriate normalization. (Red: high power, Blue: low power; segment length: 3 seconds)

power for the five standard bands. The power distribution was obtained based on a randomly selected 3-second EEG segment. The time scale at which changes in power occur is of the order of several milliseconds. Compared to fMRI, this is around three orders of magnitude faster. Most EEG based spectral biomarkers are based on group differences with respect to EEG power within specific brain regions and frequency bands [Cozac et al., 2016]. Temporal dynamic of EEG power, often requiring more sophisticated statistical models, is less common to be used in the development of spectral biomarkers. A well known model for EEG analysis which explicitly incorporates temporal dynamic is "Microstate analysis" [Lehmann et al., 1987, Lehmann, 1971]. But traditionally, temporal aspects of EEG remain under-explored.

5.2.2 Complexity based EEG Biomarkers

Beside spectral properties, the complexity of EEG signals provides another avenue in search of candidate biomarkers for various neurological conditions. The motivation is that time series, in general, can exhibit different levels of complexity while presenting (nearly) identical frequency distributions. The hope is that complexity based features might offer a second, non-redundant perspective onto EEG signals. While no single algorithmic quantification of EEG signal complexity has emerged as superior, various examples can be found where, based on different quantification schemes, complexity measures have shown great success. In [Zhang et al., 2001], the Lempel–Ziv compression algorithm [Lempel and Ziv, 1976] is used to measure the depth of anesthesia for patients undergoing vascular surgery. The analysis is based on a coarse-graining procedure of EEG signals, transforming them into a sequence of only a few distinct symbols. Subsequently, Lempel–Ziv complexity is quantified by counting the distinct patterns contained in a given sequence. In Alzheimer's disease (AD), signal complexity has

Designation	Frequency range	Amplitude
δ-band	$1-4$Hz	$>100\mu$V
θ-band	$4-8$Hz	$>100\mu$V
α-band	$8-13$Hz	$>50\mu$V
β-band	$13-30$Hz	$>30\mu$V
γ-band	$30-45$Hz	very low amplitude

Table 5.1 – EEG signals are classified according to the frequency band in which they reside. Historically, five pre-defined bands have emerged which form the bases of various biomarker studies [Jatoi and Kamel, 2017].

successfully been used to distinguish a group of early stage AD patients from a healthy control group [Houmani et al., 2015]. The complexity measure introduced and applied in this study was named "Epoch-based Entropy" and reached a cross-validated classification accuracy of 83%. More recent research is focused on combining genetic risk markers for AD with the estimated complexity of EEG based brain connectivity [Vecchio et al., 2018]. In Parkinson's disease (PD), where progression toward dementia is on average slower compared to AD, EEG based diagnostic and prognostic biomarkers have been less successful in terms of overall accuracy. Because of the generally more diverse manifestations of PD compared to AD and the additional layer of difficulty that comes with that, it is not uncommon to find computational methods that were successful in AD research also being evaluated for their benefits in PD. A promising method for early diagnosis of AD has been proposed in [Sneddon et al., 2005], where a group of 48 subjects (32 normal aging and 16 AD related disorder) has been classified with an accuracy of 92%. The complexity measure used in this study is derived from Tsallis entropy Tsallis [1988] and based on the ratio of local versus global variance estimates of the EEG time series. Based on these results, we investigated whether this complexity measure has any benefit for predicting cognitive decline in a group of PD patients.

Estimating Tsallis entropy of EEG signals

The q-entropy, also known as the Tsallis entropy (TE), was proposed by Tsallis [1988] and takes the following form:

$$S_q(p_i) = \frac{1-\sum_{i=1}^{W} p_i^q}{q-1} \tag{5.3}$$

The occurrence of event i is associated with the probability of occurrence p_i. Due to the parameter $q \in \mathbb{R}$, TE is a so-called parametric entropy where different values of q will result in different weighting schemes of the probabilities p_i. For $q \to 1$ the well-known Shannon entropy is recovered. The total number of possible events is given by $W \in \mathbb{N}$.
In this work, we focus on the TE for $q = 2$, which is approximated using an algorithmic procedure

proposed in [Sneddon, 2007]:

$$\mathrm{TE}_{q=2} = 1 - \frac{\frac{1}{N}\sum s_i^2}{\sigma^2} \tag{5.4}$$

$$s_i^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_i)^2 \tag{5.5}$$

with N: Number of bins; s_i^2: variance within the i-th bin; μ_i^2: mean of the i-th bin; σ^2: variance of entire signal.
TE is estimated by calculating the average variance within all bins s_i^2 (fast changes of the EEG signal) divided by the variance of the entire signal (slow changes of the EEG signal). The estimation procedure is as follows. For each electrode, the recorded time series is binned at local extrema, as shown in panel A of figure 5.2.2. These bins are usually different in size and contain different amounts of sample points. For example, bin 3 is wider than bin 11 and contains slightly more sample points. Then, for each bin, the within-variance s_i^2 is calculated according to equation 5.5, where x_j denotes a single sample point and μ_i is the mean of all sample points within that given bin. The total number of sample points within each bin is given as n. This operation is repeated for all N bins. The other quantity needed to estimate the entropy $\mathrm{TE}_{q=2}$ is, according to equation 5.5, the variance of the entire signal. This calculation is independent of the binning procedure.
Panel B of figure 5.2.2 shows the within-variance for each of the 12 bins. The bins are color-coded from blue to yellow for low to high within-variance, and we assigned a variance of 100% to bin 12 as it displays the highest variance of all 12 bins. In doing so, we then can give a percentage to quantify the amount of variance in each bin relative to the variance of bin 12. For example, when comparing bin 11 and bin 12, one can see that doubling the variance (bin 11: 48%, bin 12: 100%) does not imply the signal in bin 11 to have half the height of the signal in bin 12 (see panel A).
Panels D and F of figure 5.2.2 show two histograms based on the same electrode location, but measured on two different subjects. For a given signal, the variance σ of the entire time series is the width of this distribution. With the sum of the within-variance and the variance of the entire signal, an estimate of the TE for $q = 2$ can be computed according to equation 5.5. In panels C and D, the within-variation of two EEG signals with a total duration of 15 minutes is displayed. The histograms in panels D and F are based on the same signals. Both panels C and F were obtained in exactly the same fashion as panel B. The only difference is the much higher number of bins such that single bins is not visually distinguishable anymore.
At that point, this visual depiction of the variances within the different bins already qualitatively shows that there is a marked difference between both EEG recordings, especially since panel C is based on a recording from a patient suffering from PD, while panel E shows a recording based on a subject of the control group. The total variation as well as the within-variation are indicated in both panels C and E. TE can then easily be calculated, e. g. for panel C the entropy content of the signal is $1 - 1.231/1.331 = 0.075$. A similar calculation reveals that the entropy content for the signal in panel E is nearly twice as high.
In general, as entropies can take on only positive values, the quotient in equation 5.5 must never be larger than 1 to ensure positivity of the estimated entropy. This in turn implies the sum of the within-variance to be always smaller than the variance of the entire signal. From a purely mathematical perspective, it is possible to create signals that in fact would produce negative entropy estimates based on the estimator in equation 5.5. Such an occurrence is even likely if the estimator is applied directly to a raw EEG signal, usually containing a multitude of artifacts, especially if these come in

the form of severe signal distortions introduced by bad electrode contact with the scalp. However, an EEG signal pre-processed as described in section 5.1 has, in our experience, never produced a negative entropy estimate, which makes this estimator robust enough to be used in practice. Assuming an EEG signal mostly free of artifacts, the variance of the entire signal will always be larger than the mean variance over all bins. As a consequence, the quotient in equation 5.5 will always be a real positive number in the interval $(0,1)$, which in turn will restrict possible $\mathrm{TE}_{q=2}$ estimates to real positive values between 0 and 1.

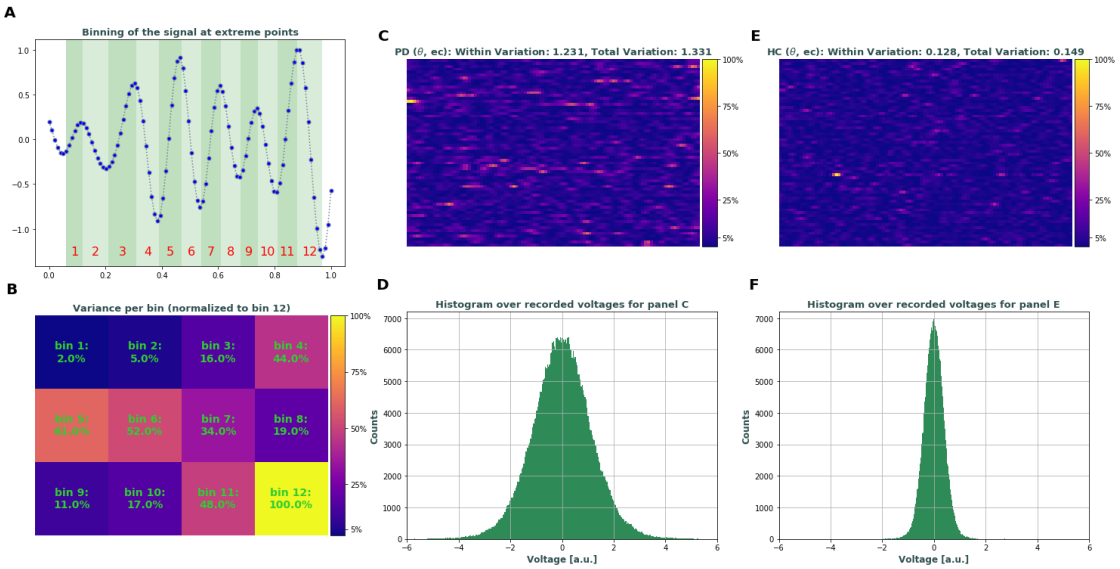


Figure 5.2.2 – (A) An EEG signal is binned at its extreme values. (B) For each of the twelve bins, the percentage in variance is indicated with respect to bin 12 exhibiting the highest variance, i.e. 100%. (C, E) For 15 minutes of EEG, the signal recorded at a single electrode was binned, and for each bin the variance was calculated. The result is displayed in the same manner as in panel (B). The signal in panel (C) exhibits a lower entropy than the signal in panel (E). (D, F) The histogram over all measured voltages for the signal in panel (C) shows a higher variance than the signal of panel (E). The signal in (C, D) is based on a patient from the PD group, while the signal in (E, F) was recorded on a patient from the healthy control (HC) group.

Tsallis entropy and relative band power provide non-redundant information

In the following analysis we demonstrate that the same neurological effect (Berger effect), captured using (i) relative band power and (ii) signal complexity quantified by $\mathrm{TE}_{q=2}$, provides two distinct perspectives. This is our first "piece of evidence" that analysis of EEG signal complexity might offer insights not contained within the spectral characteristics of EEG signals. EEG spectral band power reflects the number of neurons that discharge synchronously and is by far the most validated feature in quantitative EEG [Buzsáki and Draguhn, 2004, Klimesch, 1999]. Based on the Berger effect [Berger, 1929], which describes a significant decrease in power of α-band oscillations when cognitively normal subjects open their eyes, it can be shown that $\mathrm{TE}_{q=2}$ contains non-redundant information compared

to relative band power (rBP). For 24 healthy subjects, $TE_{q=2}$ as well as relative band power were calculated in both "eyes closed" (EC) as well as "eyes open" (EO) condition for each of the 213 electrodes in the δ, θ, α, β and γ-band. Subsequently, the 213 $TE_{q=2}$ and rBP values were grouped into 10 non-overlapping regions (frontal, temporal, parietal, central and occipital, each left and right). For each subject, the average within each region was calculated separately for the EC and the EO condition. Then the differences were computed between the region-wise averages of both conditions, i.e. EO minus EC. For each region, the mean response of the cohort, when transitioning from the EC into the EO condition, is obtained by averaging over the individual mean $TE_{q=2}$, respectively rBP values. Quantitatively, it is then observed that the magnitude of change relative to the EC condition is completely different for $TE_{q=2}$ and rBP ($p < 1e-60$). Most notably, while the rBP of the θ-band is unaffected by the transition from the EC into the EO condition, $TE_{q=2}$ of that same band increases considerably, especially in the frontal region. On the other hand, rBP of the α-band strongly decreases during EO compared to EC condition, but $TE_{q=2}$ increases. In summary, these observations show that $TE_{q=2}$, in the frequency range from $1-13$Hz, is generally higher in EO condition, while for rBP the same does not hold true. A visual comparison, including all five bands, is provided in figure 5.2.3.

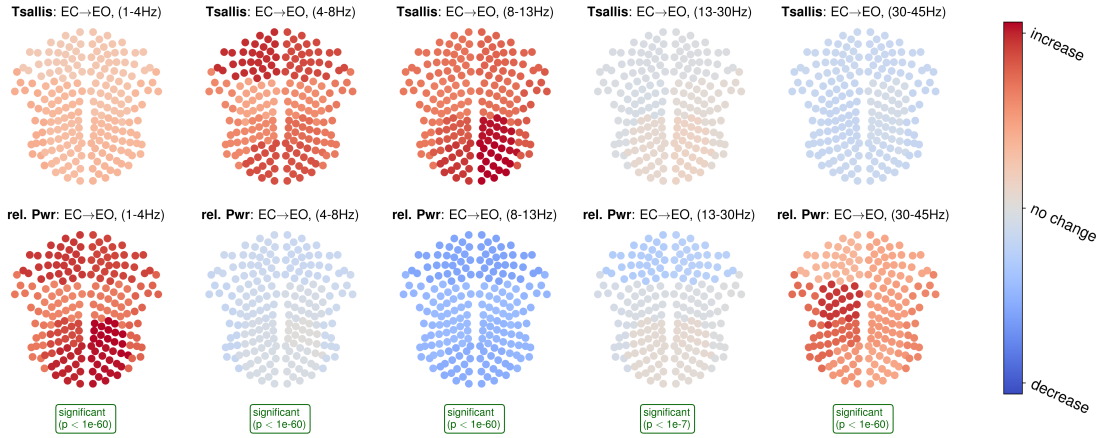


Figure 5.2.3 – Electrodes were grouped into 10 regions (frontal, temporal, parietal, central and occipital, each left and right). The upper row shows the change in TE of the EO condition relative to the EC condition for the δ, θ, α, β and γ-bands (left to right). Blue (red) stands for a lower (higher) entropy level in EO condition compared to EC condition. Color intensity corresponds to the magnitude of change in $TE_{q=2}$ relative to the EC condition. The lower row shows the change in relative band power in the identical setting as for $TE_{q=2}$ band. All changes displayed are mean values taken over the 10 individual regions and based on 24 healthy subjects. Significance values indicate that for each band a significantly different effect or reaction of the cortex is captured, depending on whether relative band power or $TE_{q=2}$ is observed. In conclusion, band power and Tsallis entropy capture different aspects of electro-physiological change and are therefore non-redundant. A multiple testing correction after Holm-Sidak, a step-down method, has been applied. Significance was tested using the Mann-Whitney U-test, a non-parametric rank test.

5.3 Spectral differences between patients with PD and healthy controls

Parkinson's disease (PD) may begin at any point during the lifetime of an individual. For many years its continued progression might go undetected before becoming clinically manifest [Kalia and Lang, 2016]. As a consequence, early and robust indicators of PD could potentially facilitate research into prevention and eventually allow for an earlier intervention in the time course of the disease. While prodromal biomarkers, i. e. markers sensitive to the disease *before* the onset of overt symptoms, might be genetic, chemical, histological or imaging based, electroencephalography based markers are of special interest: The main reasons are the non-invasive nature of routine scalp EEG and its overall low cost compared to other methods for assessing neurophysiological function. Consequently, EEG offers the possibility of population wide screenings as part of a routine check-up for the population at risk once reliable markers are validated. But given the heterogeneity of PD which can be observed in many aspects of the disease [Yilmaz et al., 2019], from pathology to clinical phenotype including disease progression, reliable biomarkers will likely be *compound markers* where EEG might provide one "piece of the puzzle".

5.3.1 Analysis

The following example demonstrates a potential application of ultra-sparse logistic regression described in chapter 3 to EEG. This application example is based on a cohort of 42 patients with PD and 24 healthy controls, both recruited from the Movement Disorders Clinic of University Hospital of Basel. The control group was matched for age, sex, and education. With a median disease duration of 2.5 years the patient cohort is in an early stage of the disease relative to the time point where symptoms, i. e. mostly motor signs, become apparent allowing for a clinical diagnosis. EEG data was recorded during eyes closed condition. After preprocessing the raw EEG data using the preprocessing pipeline described in section 5.1, relative band power was calculated for the δ-, θ-, α-, β-, and γ-band. For each band, the relative power was calculated separately for 10 regions of interest, i. e. frontal, temporal, central, parietal and occipital on each hemisphere, resulting in a total of 50 relative band power values in the range between 0 and 1. The total data set is thus comprised of $n = 66$ samples and $p = 50$ features. Spectral power was used to quantify information content of EEG due to its long standing tradition in electrophysiology [Chaturvedi et al., 2017], allowing for a qualitative validation of the result of the sparse logistic regression. By regulating the level of sparsity based on the log-norm, the number of active predictors can gradually be reduced while trading-off (some) classification accuracy. Given previous research, it is expected that not all frequency bands and scalp regions are equally important in order to differentiate healthy controls from PD patients. Consequently, it is expected that a small subset of the 50 derived relative band power features will suffice to perform this discrimination task compared to a logistic regression based on the full set of available predictors. Based on 100 random training and test set splits with a ratio of 80/20, a ℓ_2 logistic regression and two sparse logistic regressions with $\gamma = 0.0001$ and $\gamma = 2.5$ are performed. In each case, based on 5-fold cross-validation, the best model is selected. For these 3 classifiers, trained on 100 random training sets, figure 5.3.1 shows boxplots over the coefficient sizes or weights associated with each of the 50 predictors.

5.3.2 Discussion

The non-sparsity inducing ℓ_2-regularized logistic regression selects *all* predictors, as shown in the top row of figure 5.3.1. While this classifier provides an average test set accuracy of $\approx 70\%$, is shows a median coefficient size of zero for all predictors, making the identification of bands and regions of interest impossible. Nevertheless, it can be seen that the left hemispheric predictors of the θ-band show an increased negative first quartile, pointing at the possible importance of the $4-8$Hz band. The middle row shows the coefficient sizes based on sparse logistic regression with $\gamma = 0.0001$, which essentially is the ℓ_1-lasso penalty. This classifier reaches an average test set accuracy of $\approx 67\%$ revealing 8 potentially important predictors, three of which have a non-zero median over the 100 randomly split training set. Finally, the bottom row shows an even sparser logistic regression with $\gamma = 2.5$ which reveals two predictors with non-zero median, i. e. θ-power of the central left and β-power of the parietal left region: In these regions, healthy individuals tend to have lower θ-activity while having an increased β $(13-30$Hz$)$ activity compared to PD patients. The average test set accuracy for this sparsest version of logistic regression has decreased to $\approx 63\%$ while providing potentially improved possibilities of interpretation. Generally, research has shown that the classification accuracy of PD vs healthy controls based on band power might be limited. In [Chaturvedi et al., 2017], the authors shows that logistic regression based accuracies of up to median AUC $= 76\%$ are achievable by additionally including the ratio of the α_1- and θ-band $(\alpha_1 : 8-10$Hz$)$. Their analysis is based on a data set containing $N_{HC} = 41$ healthy controls and $N_{PD} = 50$ patients suffering from PD. This highlights the aforementioned need for additional information complementing EEG in order to obtain clinically relevant diagnostic biomarkers for PD.

While increased sparsity $(\gamma = 2.5)$ has led to a loss of classification accuracy, it provided an important *regional* difference between healthy controls and PD patients: Central and parietal regions on the left hemisphere. This sparse result might not only be interesting for formulating further hypotheses but might also hold practical implications. A relatively new method in the domain of EEG potentially enabling a slowing of the disease progression is Neurofeedback. Neurofeedback uses online estimations of spectral power in order to provide a feedback to the patient suffering from PD, with the goal to actively learn to shift the power spectrum from lower to higher frequencies. With daily training, it is hoped that patients learn to produce spectral characteristics of a healthy brain, thereby counteracting the degenerative process associated with PD. But as home use neurofeedback devices only have a very limited number of electrodes they cannot cover the whole scalp. It is therefor of importance to know the optimal electrode posistions, i. e. the region, based on which feedback signals should be generated in order to increase treatment outcome.

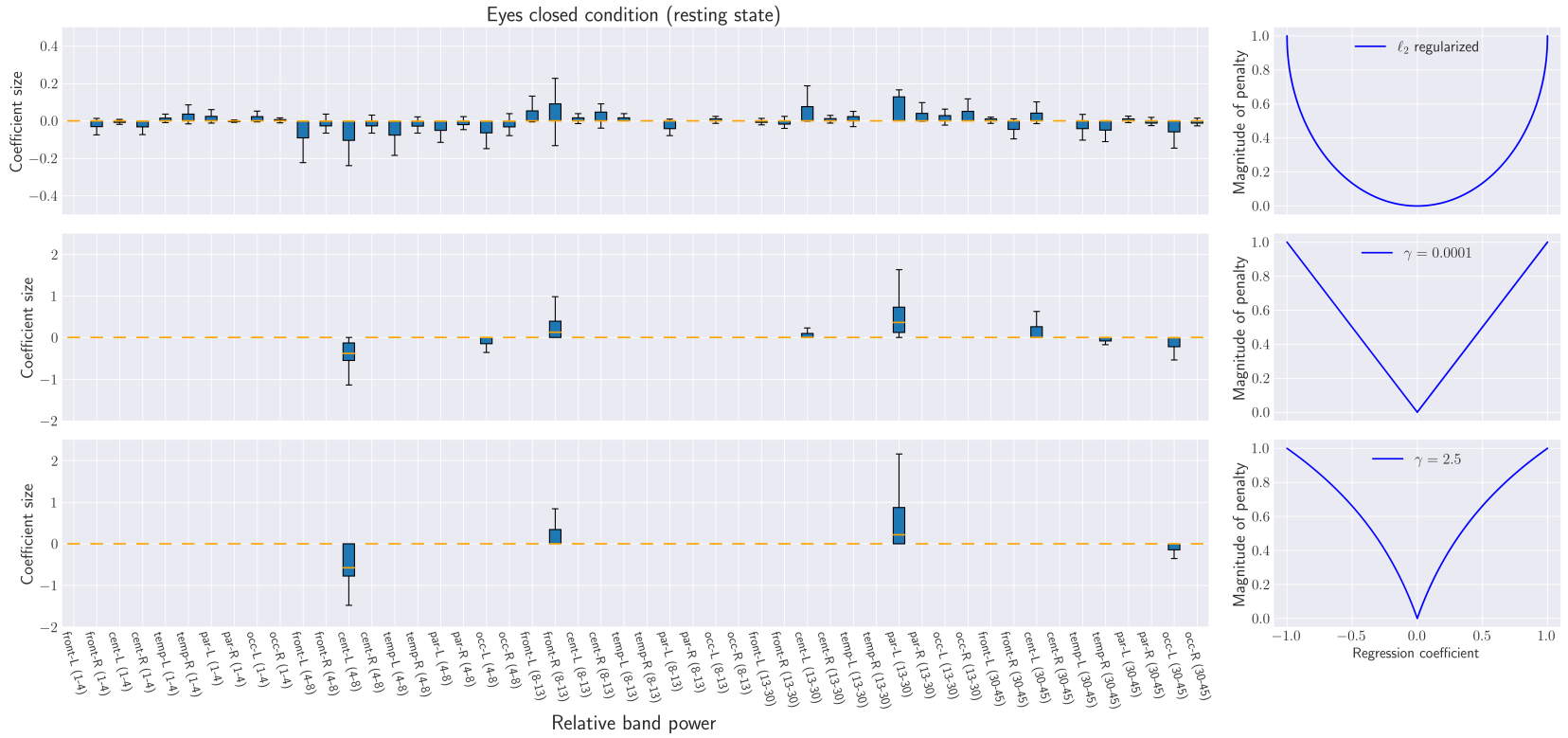


Figure 5.3.1 – Based on 100 random training and test set splits a classification of healthy controls vs patients suffering from Parkinson's disease was performed. Shown here are the boxplots over coefficient sizes. The first row shows coefficients based on ℓ_2-penalized logistic regression, the second row is a classification based on $\gamma = 0.0001$-penalized logistic regression, i. e. lasso, while the last row promotes sparsity even stronger than lasso, with $\gamma = 2.5$. Features used here are relative band power values of the five standards bands, calculated for ten distinct scalp regions leading to a total of 50 features.

5.4 Cognitive decline in Parkinson's disease is associated with reduced complexity of EEG at baseline

In a case-control study, forty-two (42) cognitively normal individuals diagnosed with PD (median age 66.5 yrs., 18 females, median education 14 yrs.) were compared with 24 healthy control subjects (HC) matched for age, sex, and education (median age 66.5 yrs., 9 females, median education 14 yrs.). Baseline EEG recordings were obtained while their eyes were open (EO) and closed (EC). Tsallis entropy (TE) of the δ, θ, α, β and γ-band was evaluated. As the θ-band showed the most pronounced differences between the PD and HC group, further analysis focused on this band. TE was then compared across groups with 16 psychological test scores at baseline, and then with follow-ups at 6-months and 3-years. Comparisons were repeated for relative band power (BP) as a predictor of cognitive decline.
Compared to healthy controls, most patients with PD showed overall lower TE at baseline. Cognitive deterioration at 3 years correlated significantly with baseline TE in EO condition ($p \leq 0.00079$), while correlation at 6 months after baseline was not significant. No significant correlation was observed between baseline TE measured in the EC condition and cognitive deterioration over 6 months and 3 years. Additional predictors taken into consideration were age, education, sex, levodopa equivalent dose (LED), disease duration and sleepiness of the patients. Age at baseline was significantly correlated with 3-year cognitive decline only in case of BP but not for TE, both measured in EO condition (TE: $p \leq 0.059$; BP: $p \leq 0.016$). Baseline sleepiness was not significant in predicting 3-year cognitive decline based on either TE or BP.
In conclusion, the lower the EEG entropy levels at baseline measured in the EO condition, the higher the probability of cognitive decline over 3 years. This makes TE a candidate for a prognostic biomarker for dementia in PD. The ability of the cortex to execute complex functions underlies cognitive health, while cognitive decline might clinically appear when compensatory capacity is exhausted.

5.4.1 Introduction

While Alzheimer's disease is the most common neurodegenerative disorder, Parkinson's disease (PD) is the fastest growing one [Dorsey and Bloem, 2018]. According to conservative estimates based on worldwide prevalence data from a 2014 meta-analysis [Pringsheim et al., 2014], the number of people suffering from PD is expected to reach 14.2 million in 2040, which would effectively double the number of cases compared to 2015. Despite being considered primarily a motor disorder, approximately 30% of patients with PD have cognitive symptoms already at initial diagnosis, and up to 80% develop cognitive symptoms at some point in their disease [Hely et al., 2008, Emre et al., 2007]. The prognosis for losing independence or life currently depends much more on neuropsychiatric and cognitive deterioration than on motor signs [Bäckström et al., 2018, Forsaa et al., 2010]. Moreover, cognition is an important aspect of quality of life for patients as well as their caregivers [Lawson et al., 2017, 2016]. Therefore, preservation and improvement of cognition in PD patients have recently become major goals for therapeutic interventions and trials. Patient care and clinical trials regarding cognition currently rely mainly on bedside assessments and psychological testing with its known difficulties, including availability and reliability, including test-retest biases. In contrast, biomarkers are objective monitors or predictors of the disease course and will improve making decisions for individual patients as well as for defining optimal populations for clinical trials for cognitive decline in PD [Dodakian et al., 2013, Cramer, 2010].
According to Stam [2014], normal cognition is characterized by electrical brain activity with an optimal

degree diversity, order, and hierarchy. Analogously, normal consciousness is characterized by an optimal entropy level of the EEG, while its reduction leads eventually to loss of consciousness and abnormal increase in incoherent thinking, such as in a psychedelic state of consciousness [Carhart-Harris et al., 2014]. Tsallis entropy (TE) of EEG, when measured during a recall task and characterized as a ratio between frontal and parietal regions, resulted in a very high accuracy for detection and treatment monitoring of mild cognitive impairment due to beginning Alzheimer's disease [Sneddon, 2007, Sneddon et al., 2005]. We therefore hypothesized a priori that TE of the EEG at baseline correlates with cognitive deterioration over a period of 3 years. Moreover, we hypothesized a posteriori that the decrease of entropy is a diffuse effect not attributable to a single location and that TE of the θ-band, measured during EO condition, is significantly more informative about future cognitive performance than the same band evaluated in EC condition. Furthermore, we hypothesized that BP as a predictor for cognitive decline, is agnostic to the recording condition of EO versus EC.

5.4.2 Material and methods

Participant demographics

The study is based on a cohort of 42 patients with PD who were recruited from the Movement Disorders Clinic of University Hospital of Basel from 2011 to 2016 by advertising in the magazine of the Swiss Parkinson's Disease Association. PD was diagnosed according to the United Kingdom Parkinson's Disease Brain Bank criteria [Gibb and Lees, 1988]. Neuropsychological assessment was carried out in all individuals when they were admitted into the study (baseline), then at 6-month and at 3-year follow-ups. Knowledge of the German language was a prerequisite for the inclusion to this study. Patients with psychiatric or organic brain disease as well as patients with complete missing data at years follow-up were excluded from the analysis. The complete consort schema is provided in figure A.2.1 of the appendix.
A group of 24 healthy controls (HC) matched for age, sex, and education was recruited from the Memory Clinic, University of Basel Center for Medicine and Aging, and from the University Hospital of Basel. The demographic characteristics of the participants are shown in table 5.2. The studies were approved by the local ethics committee (Ethikkommission beider Basel, ref. no: 135/11, 294/13, 260/09). All participants gave their written informed consent.
As all patients underwent comprehensive neuropsychological examinations, analyses showed that patients who performed all tests scored significantly higher in the MMSE (Median score: 30 vs. 28; $W = 319$; $p \leq 0.05$), and had a lower disease duration (Median score: 2 vs. 4.5; $W = 141.5$; $p \leq 0.05$) than patients with incomplete data. No other differences in demographic or disease characteristics were observed between the two groups.

Clinical neurological and neuropsychological assessments

A basic neurological examination was carried out in all individuals. All patients underwent comprehensive neuropsychological examinations. The following cognitive domains were of interest for the present study:

- Attention and psychomotor speed: Alertness (reaction time with and without sound) and Divided Attention (reaction time to visual and auditive stimulus, number of omissions) of the computerized "Test Battery of Attentional Performance" [Zimmermann and Fimm, 2007], the

	Age	Edu	Sex	MMS	UPDRS-III	LED	DisDur	KSS
HC (24)	**66.5**	**14**	**9f.**	**30**	-	-	-	-
1st Quartile	64	12	-	29	-	-	-	-
3rd Quartile	68.5	17.25	-	30	-	-	-	-
PD (42)	**66.5**	**14**	**18f.**	**29**	**14.5**	**543**	**2.5**	**3**
1st Quartile	63	12	-	28	5	305	1	2.875
3rd Quartile	72.75	16	-	30	21	1014	5	3.25

Table 5.2 – Values are presented as median values, along with values for the 25%- and 75%-quartiles. Age, education (edu), and disease duration (DisDur) are given in years. LED is given in milligrams. MMS and UPDRS-III refer to standardized psychological tests. Sleepiness (KSS) is rated according to the Karolinska Sleepiness Scale (1-extremely alert, 10-extremely sleepy).

Trail Making Test Part A [Reitan, 1958]

- Executive functions and working memory: phonemic (s-words, [Thurstone and Thurstone, 1947]) and semantic fluency (animals, [Isaacs and Kennie, 1973]), TAP Working memory (number of omissions), Digit span and Corsi Block (forward and backward) [Härting et al., 2000]
- Visuo-constructive abilities: Block Design Test [Tewes and D, 1991]

A Reliable Change Index (RCI) for each neuropsychological test was calculated for both the 6-month and 3-year follow-up after baseline. The individual RCI values were then combined into an Overall RCI for the 6-month and 3-year follow-up. As the RCI is a standardized measure, combining multiple RCI values is done through simple averaging. The Overall RCI was used as the outcome variable. The RCI for a single psychological test was calculated as the difference between the test score at either 6-month or 3-year follow-up and the test score at baseline, divided by the standard error of the difference [Jacobson and Truax, 1992]:

$$\text{RCI} = \frac{\text{follow-up} - \text{baseline}}{S_{\text{diff}}} \tag{5.6}$$

$$\begin{aligned} \text{with} \quad S_{\text{diff}} &= \sqrt{2 \cdot SE_M^2} \\ \text{and} \quad SE_M &= \text{std(baseline)} \cdot \sqrt{1 - RS} \end{aligned}$$

Here, $\mathbf{SE_M}$ is the *standard error of measurement*, **std(baseline)** is the *standard deviation of baseline* and $\mathbf{S_{diff}}$ is the *standard deviation of the errors of measurements*. The *reliability of measurement* is given by **RS**, a scalar between 0 and 1.
With approximately 5.5% of the neuropsychological data points missing, candidate values were generated based on the multiple imputation method for both predictors and missing outcomes [Little,

1992]. As a consequence, sample size was maintained and introducing potential biases, due to systematically missing data, was avoided.

EEG recording and signal processing

A total of 20 minutes of EEG was recorded at wakeful rest for each patient by using a 256-channel EEG System (Netstation 300, EGI, Inc., Eugene, OR). EEG recordings were done in the afternoons and patients were seated comfortably in a relaxing chair, instructed to open and close their eyes at regular intervals in the beginning, then closing their eyes for 15 minutes and opening them again towards the end (5 minutes). A technician present in the recording room controlled for vigilance of the patients and kept them alert. Before the EEG recording, patients were also asked to self-rate their sleepiness level from 1 to 10 by using the Karolinska Sleepiness Scale [Åkerstedt and Gillberg, 1990, Kaida et al., 2006, Miley et al., 2016]. All data were first separated into segments containing only recordings in either EO or EC condition. Subsequently, these segments were processed in an automated way by using the MATLAB based in-house software toolbox "TAPEEG" [Hatz et al., 2015], available at https://sites.google.com/site/tapeeg. EEG were filtered (Firls: 0.5–70 Hz, 50 Hz notch) at a sampling rate of 1000 Hz and an inverse Hanning window was used to stitch together shorter segments to have at least 3 minutes of cleaned EEG data. The implementation of independent component analysis ("runica") used for pre-processing was originally part of the toolbox "EEGLAB" [Delorme and Makeig, 2004]. "TAPEEG", which combines methods from "EEGLAB", "FASTER" and "Fieldtrip", was used for the entire pre-processing of EEG. As "TAPEEG" is freely available (incl. handbook/tutorial), the full pre-processing pipeline can easily be reproduced. "TAPEEG", with default settings, was used in order to detect bad channels/activations/segments. Furthermore, eye movement artifacts, traces of sleep, eye blinking, ECG and muscle artifacts were detected and removed. The average of all "good" channels was used to re-reference the EEG to a common average montage. Electrodes placed on the neck, ears, cheeks were excluded to remove spurious signals, and 213 electrodes were mapped to ten regions of interest: frontal left/right, central left/right, parietal left/right, temporal left/right, and occipital left/right. For analysis, a total of 3 minutes of EEG in EC as well as EO condition was available per patient. In 3 of 66 cases, less than 180 seconds, but more than 170 seconds of EEG data were available, which did not affect TE estimates. The 66 artifact free segments were then filtered into the following five bands: 1-4Hz (δ-band), 4-8Hz (θ-band), 8-13Hz (α-band), 13-30Hz (β-band) and 30-45Hz (γ-band). For filtering, a zero-phase band pass FIR filter with Hann window was used.

Statistical Analysis

Calculation of TE was performed as described in [Sneddon, 2007]. The algorithm was implemented in the Python (v3.5). R statistical software was used for analysis. For missing entries in test psychological data, the multiple imputation method implemented in the "MICE" R-package [Buuren and Groothuis-Oudshoorn, 2010] was used to generate a total of 20 imputed data sets. Imputation of missing psychological test values are based on all available tests from baseline, 6-month and 3-year follow-ups. Linear regression analyses were performed on each imputed data set and pooled, again using the aforementioned R-package. Given the relatively small number of patients available, the number of potential confounders was reduced by adopting a strategy proposed in (van Buuren, 2018), consisting of a stepwise forward selection, performed on each imputed data set, in order to identify significant confounders. This was followed by a majority vote over all 20 data sets in order to identify significant confounders which appeared consistently, i.e. in the majority of imputed data sets. Finally, only

confounders which appeared consistently were considered in the subsequent data analysis. Based on the "relaimpo" R-package [Grömping et al., 2006], the relative importance of TE and BP as predictors as well as the importance of the confounders was assessed. Alpha, the probability of a Type I error, was 0.05. In case of multiple testing the significance threshold was adjusted according to Holm-Sidak. Two-tailed hypothesis tests were considered throughout. If the requirements for the Welch t-test were not fulfilled, the non-parametric Mann-Whitney U-test was used. For the extreme groups, significances were not calculated. Following [Preacher et al., 2005], excluding already available data from the analysis by applying an artificial threshold, would have resulted in inflated p-values.

5.4.3 Results

TE characterized the PD and HC group in the θ-band

With 42 PD patients, each patient's EEG recorded with 213 electrodes, a total of 8,946 distinct entropy values were computed for each given band and condition. For the HC group with 24 subjects, a total of 5,112 entropy values were obtained per band and condition. For both groups, histograms were plotted. To account for imbalance regarding group size, histograms were normalized to allow for better visual. The TE histograms for each band and condition are shown in figure 5.4.1. The θ-band in "eyes open" condition shows a significant difference between healthy controls and the patient group ($p \leq 0.006$). Moreover, this band displayed the least overlap between the patient and the control group, with the EO condition showing a slightly larger separation than the EC condition. Based on this assessment, and assuming that the same pathological process underlies general Parkinsonian pathology as well as Parkinsonian cognitive decline, further investigations were focused on the θ-band. What motivated the current research is the question of whether or not the difference in signal complexity of EEG between the PD and HC group, quantified with TE, contains information about future cognitive decline.

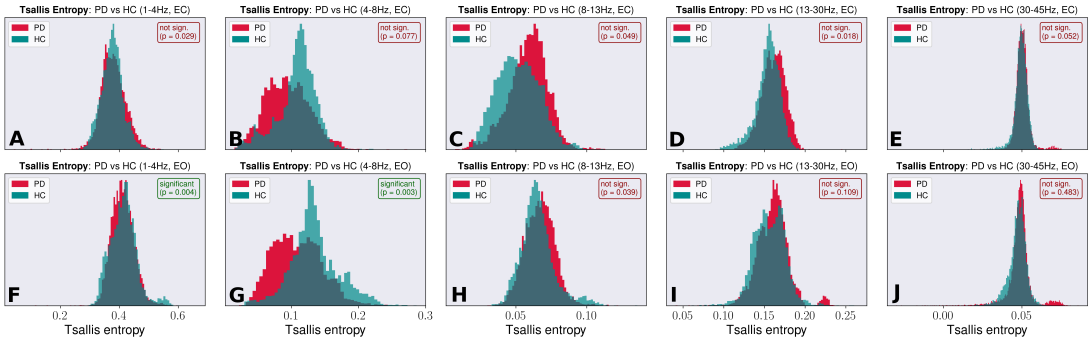


Figure 5.4.1 – Normalized Tsallis entropy histograms for the PD and the HC groups. Panels A-E show the histograms for the δ, θ, α, β and γ-band (left to right) in EC condition, while panels F-J show the histograms in the EO condition for the same bands. Given are the uncorrected p-values, based on the non-parametric Mann-Whitney U-test. Significance threshold is corrected after Holm-Sidak, a step-down method. The θ-band in "eyes open" condition shows a significant difference between healthy controls and the patient group and generally, this band displays the least overlap between both groups.

Relative band power characterized the PD and HC group in the θ-band

Histograms were also calculated for the relative band power and are shown in figure 5.4.1. Calculations were performed analogously to the calculations of the TE histograms. The same normalization procedure used in case of TE histograms was also applied to the band power histograms. Similar to the histograms of the TE, the most prominent distinction between the HC and PD groups is observed in the θ-band ($p \leq 0.001$). But while TE is reduced in PD patients compared to the HC group, the inverse is true in case of relative BP.

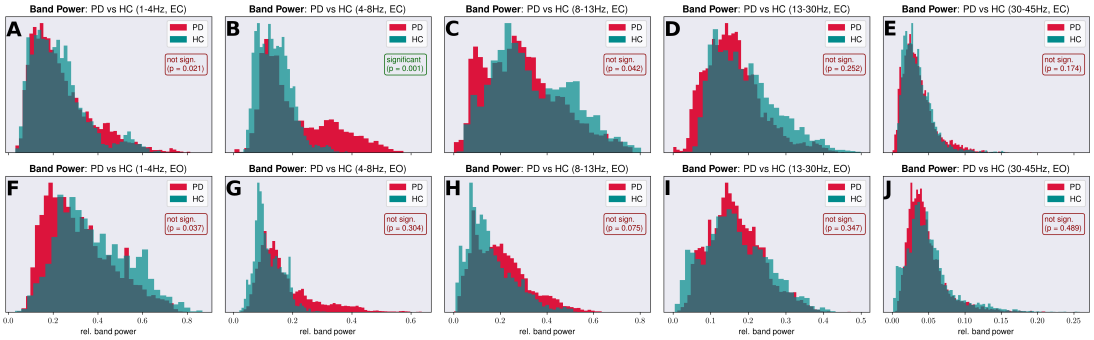


Figure 5.4.2 – Normalized relative band power histograms for the PD and the HC group. Panels A-E show the histograms for the δ, θ, α, β and γ-band (left to right) in EC condition while panels F-J show the histograms in the EO condition for the same bands. A multiple testing correction after Holm-Sidak, a step-down method, has been applied. Significance was tested using the Mann-Whitney U-test. Indicated are the non-corrected p-values while the label "not sign./significant" was given according to the corrected significance threshold.

Changes in relative BP correlated with changes in TE

The histograms for TE and rBP shown in figure 5.4.1 and 5.4.2 suggest that a change in band power is strongly correlated to a simultaneous change in TE and vice versa. As a consequence, TE might simply encode the same information as BP and not reveal new information. By assessing the degree of correlation between TE and BP it is possible to determine whether the complexity of the EEG signal and its relative power are in fact two degrees of freedom that can be regulated independently of each other. Therefore, based on all 213 electrodes, the Pearson correlation between TE and BP was calculated for each patient as well as the HC group. Figure 5.4.3 shows the result for each of the five bands, in both the EO and EC condition. Generally, the correlation between TE and BP is stronger in the PD group than HC group, which tends toward a closer-to-zero median correlation. Furthermore, correlation is stronger in the lower bands, i.e. δ ($p \leq 0.0001$) and θ ($p \leq 0.005$), with the θ-band showing the highest median correlation within the PD group. Except for the δ-band, which shows a positive correlation between TE and BP, all other bands are characterized by either a close-to-zero or an inverse relation between these two measures. On the group level, with respect to correlation, no pronounced differences between EC and EO condition exist.

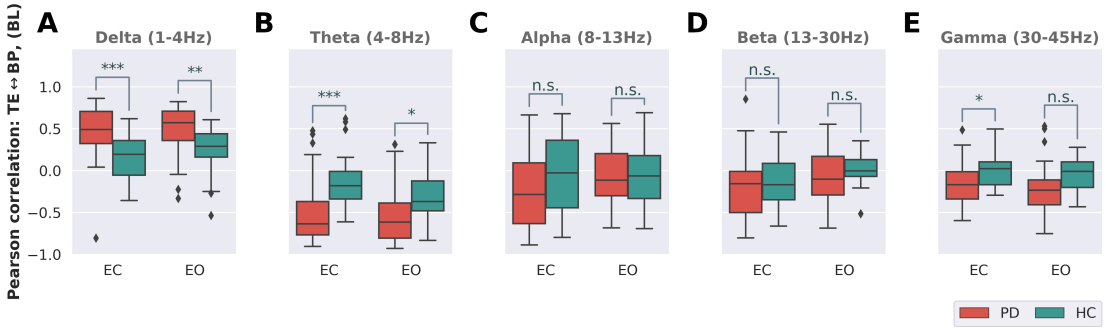


Figure 5.4.3 – Correlation between Tsallis entropy and relative band power in EC and EO condition for the patient group and the healthy controls. Except for the δ-band in panel A, all other bands (panels B-E) show a negative correlation between TE and relative band power. This relation is less pronounced for higher frequency bands. Generally, median correlation between signal power and signal complexity of EEG is stronger in the patient group, especially in the 1-8Hz range, implying that PD patients lose the ability to independently modulate power and complexity of EEG. Significance levels are based on the t-test (Holm-Sidak corrected): $p > 0.005$: n. s. – $p \leq 0.005$: * – $p \leq 0.001$: ** – $p \leq 0.0001$: ***.

In groups, TE differentiates HC, MCI and DEM but not CN

For the participants of the study, the clinical diagnoses at baseline as well as at 3-year follow-up were available. At baseline, the cohort was composed of 24 HC, 31 cognitive normal patients (PD-CN) and 11 patients suffering from mild cognitive impairment (PD-MCI). At 3-year follow-up only 24 patients were PD-CN, 10 patients were suffering from PD-MCI and 5 patients were diagnosed with Parkinson's dementia (PD-DEM). Figure 5.4.4 shows each subject in a 2-dimensional plot with the baseline TE level of the θ-band in EC and EO condition on its axes. For patients with PD, TE estimated in EC condition shows a correlation of 69.7% with TE estimated in EO condition. This is considerably lower than the 83.6% correlation between relative band power in EC and EO condition found for the same patients. The different colors of the individual markers in figure 3 indicate the cognitive status at 3-year follow-up, where mostly patients with lower baseline TE have progressed to dementia. This qualitative observation is statistically analyzed in the next section.

TE correlated with 3-year overall cognitive decline in the group of patients with PD

Here, cognitive decline is understood as an overall decline of cognitive abilities and is thus quantified based on a combination of RCI scores from multiple cognitive domains. As the RCI is designed to be a standardized score, combining multiple RCI domain scores reduces to an averaging procedure over the individual RCI values. By combining the domain-wise RCI values for "attention", "executive function", "visuo-constructive ability" and "working memory" an overall RCI for 6 months and 3 years after baseline is obtained and evaluated for the 42 PD patients, in both EC and EO condition. The evaluation is based on median TE and median relative BP of the θ-band, where the median was taken over all 213 electrodes to obtain global median values. Figure 5.4.5 shows the Overall RCI for each patient 3 years after baseline, dependent either on the patients' global baseline TE or global relative BP. Potentially significant confounders were age, education, sex, disease duration, LED and sleepiness.

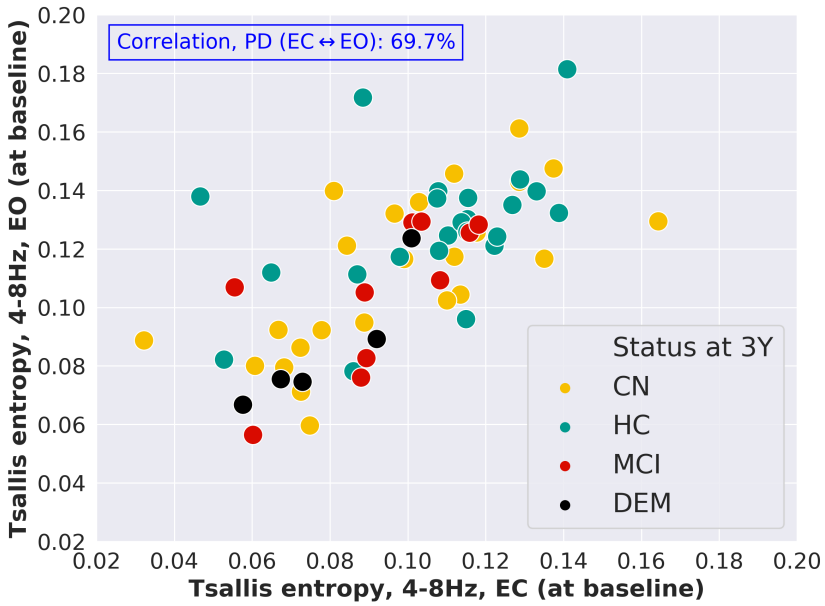


Figure 5.4.4 – TE of the θ-band at baseline in EC and EO condition, shown for each subject. The different colours encode cognitive status of each subject at 3-year follow-up. PD patients having developed overt dementia over the course of 3 years are mostly patients with lower baseline TE of the θ-band. For the group of PD patients, the relatively low Pearson correlation of 69.7% between TE in EC and TE in EO condition, points towards a sensitivity of TE with respect to this condition.

After performing the stepwise selection on all 20 imputed datasets individually, followed by a majority vote, the confounders to include in the final pooled regression, were identified (see table 5.3). As the Overall RCI for the 6-month period was almost negligible, as is shown in figure A.2.2 of the appendix, no meaningful regression analysis between either baseline TE or relative BP and the 6-month Overall RCI could be performed. The result of the pooled regression analysis for the 3-year Overall RCI are given in table 5.3. Relative θ-band power was significantly correlated with 3-year Overall RCI in both EC and EO condition ($p \leq 0.0020$ and $p \leq 0.0023$ respectively). For TE measured in EC condition, the correlation with 3-year Overall RCI was not significant ($p \leq 0.192$), whereas the correlation of TE in EO condition was highly significant ($p \leq 0.00079$). In general, the main tendency of higher TE at baseline making a 3-year decline less likely, is reversed in case of relative band power of the θ-band, where high values indicated an increased risk of 3-year cognitive decline.

TE and Age explain over 40% of the variance of the 3-year Overall RCI

For all predictors of 3-year Overall RCI listed in table 5.3, their relative importance was calculated. The results are shown in the panels C and F of figure 5.4.5. For TE in EC condition (panel A), the association with the 3-year RCI was not significant, which is reflected both by the low explained variance of TE in EC condition as well as the overall lower adjusted R-squared compared to the other settings in table 5.3. For TE in EO condition (panel B) as well as relative band power in both conditions (panels D and E), age was the second most important contributor to the variance

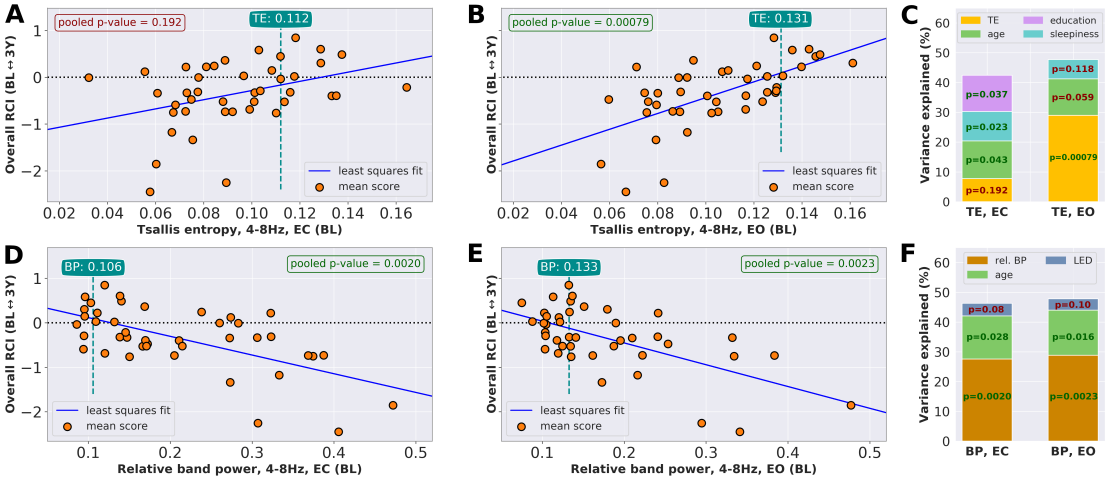


Figure 5.4.5 – Overall reliable change index (RCI) at 3-year follow-up. Panels A-C: Tsallis entropy at baseline and overall RCI are correlated significantly only in "eyes open" condition, where approx. 30% of the variance is explained by Tsallis entropy. For Tsallis entropy measured in EC condition the association is not significant. Panels D-F: For relative band power, prediction of 3-year cognitive decline is not sensitive to EC/EO condition, where both conditions explain approx. 30% of the variance. Note: The median entropy and median relative band power values of the control group in the respective condition, are indicated in the green boxes.

explained. According to Åkerstedt and Gillberg [1990], Kaida et al. [2006], the correlation between relative θ-band power and self-assessed daytime sleepiness based on the KSS, is highly significant. Based on the stepwise selection and the subsequent majority vote, both LED and sleepiness were included in the analysis. But neither predictor was below the significance level of 0.05, and only less than 10% of the variance could be explained by either LED or sleepiness.

θ-band: median Tsallis Entropy ~ compound RCI							
Time	**Condition**	**pVal: TE**	**Adj. R^2**	**pVal: Age**	**pVal: Sleepiness (KSS)**	**pVal: LED**	**pVal: Education**
3 years	EC	0.192	0.42	**0.043**	**0.023**	Not incl.	**0.037**
	EO	**0.00079**	0.48	0.059	0.118	Not incl.	Not incl.

θ-band: median rel. Band Power ~ compound RCI							
Time	**Condition**	**pVal: BP**	**Adj. R^2**	**pVal: Age**	**pVal: Sleepiness (KSS)**	**pVal: LED**	**pVal: Education**
3 years	EC	**0.0020**	0.46	**0.028**	Not incl.	0.08	Not incl.
	EO	**0.0023**	0.48	**0.016**	Not incl.	0.10	Not incl.

Table 5.3 – Pooled linear regression analysis for Overall RCI and θ-band TE or relative BP, both for EC and EO condition. The association of TE in EO condition with Overall RCI is more significant than for relative BP in either condition. Significant confounding factors for BP in both EC and EO condition was Age, while none of the confounders was significant in case of TE in EO condition. Confounders that were not included into the final pooled regression, following stepwise selection and majority vote, are marked with "not incl.".

Extreme group behaviour

Global median entropy (θ-band, EO) within the cohort of 42 patients with PD was in the range of [0.056; 0.161]. For the 10 patients showing the strongest decline (high Overall RCI values) over a period of 3 years, the median of the global entropy was in the interval [0.056; 0.105]. The 10 patients with the least decline (lowest Overall RCI values) over that same period showed global entropy values in the range of [0.071; 0.161]. Table 5.4 shows the baseline demographic of these two extreme groups. Overall, the median age of the group of strong decliners was 8 years higher than for the

	Age	Edu	Sex	MMS	UPDRS-III	LED	DisDur	KSS
PD (10) low 3Y RCI	**65.5**	**13**	**6f.**	**29**	**16.5**	**487.5**	**2**	**3**
1st Quartile	63.25	12	-	29	7.5	285	0.25	1
3rd Quartile	69.75	14	-	29	20.75	560	4.75	3
PD (10) high 3Y RCI	**73.5**	**16**	**3f.**	**28.5**	**15**	**624.5**	**2**	**3**
1st Quartile	68.25	13.75	-	28	10	337.5	2	3
3rd Quartile	79.25	17.25	-	29	30.25	1335.5	7	6.25

Table 5.4 – Demographic at baseline of the ten most and least stable patients within the cohort of 42 patients during the 3-year period from baseline. Given are the median values along with the first quartile (25%) and third quartile (75%). Stability is quantified based on patients' overall RCI value, where a low RCI is indicative of cognitive stability while a high RCI value indicates cognitive decline.

group of cognitively stable patients. Furthermore, the strong decliners had a 20% higher median LED. Otherwise both groups showed similar characteristics. Regional differences in TE of EEG in EO condition between the HC group and the two extreme groups of stable and declining patients, are shown in figure 5.4.6. The extreme group of cognitively stable patients had TE levels in the same range as the HC group across all regions. On the other hand, the group of extreme decliners showed lower TE levels, again across all regions. Regional differences were thus not present, supporting the hypothesis of decreasing TE in association with PD being a non-localized effect.

Stability of TE estimates and intra-subject variability

From a subset of patients, slightly more than 360 seconds of EEG were available after pre-processing, corresponding to approximately double the epoch length of 180 seconds used in the present analysis. For three such patients, global TE of the θ-band was re-calculated a total of 100 times in a random resampling setting, where each randomly selected epoch had a length of exactly 180 seconds. Figure 5.4.7 shows the results of the resampling procedure. Overall, the global TE estimates show very little variance over randomly selected epochs, which makes global TE, estimated with the procedure proposed in [Sneddon, 2007], a robust measure of signal complexity of EEG.

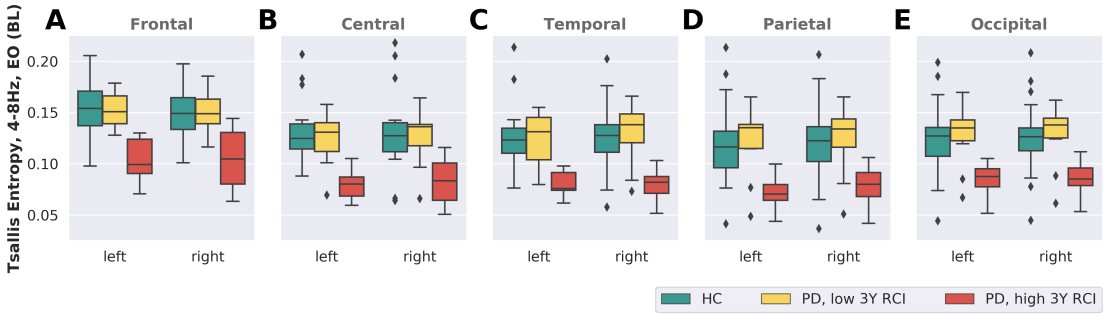


Figure 5.4.6 – Median entropy values are shown for the healthy controls (green, n=24) as well as for the extreme groups of cognitively stable individuals (yellow, n=10) and cognitive decliners (red, n=10). The most prominent difference involves a distinctly reduced TE at baseline for patients showing a strong 3-year cognitive decline. The cognitively most stable patients are indistinguishable from healthy controls based on their regional TE levels. With no apparent regional differences, decreasing TE of the θ-band during the course of PD is likely a global effect not attributable to a single location.

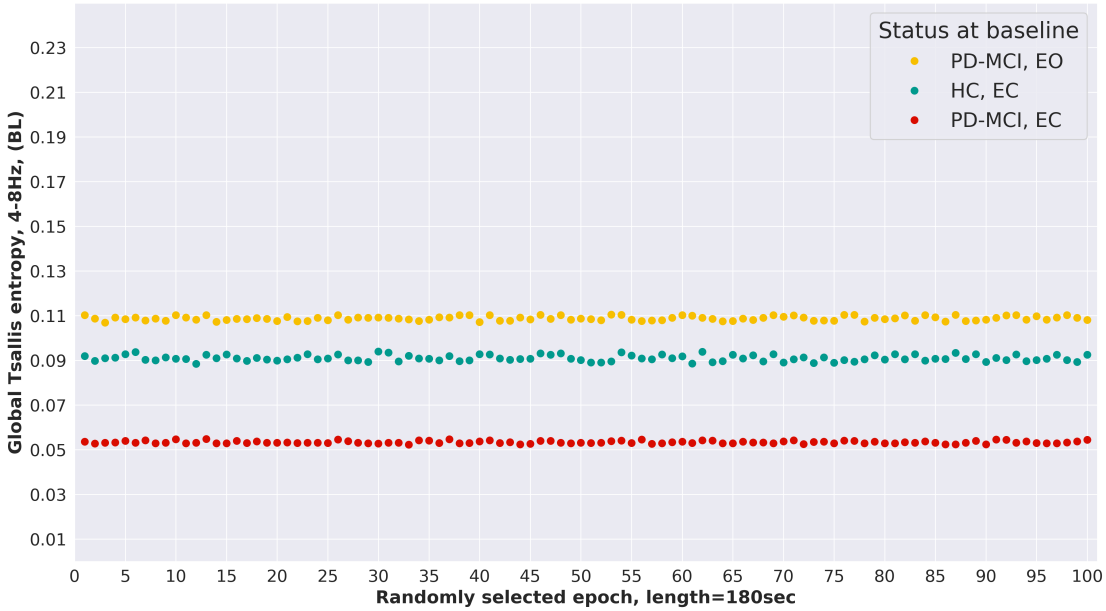


Figure 5.4.7 – For three study participants EEG with more than twice the epoch length of 180 seconds used in the present analysis was available. From these recordings, 100 epochs, each with a length of 180 seconds, were randomly sampled and the global TE of the θ-band was estimated. The results show a very low variance across TE of the different sample epochs, making global TE of the θ-band a robust feature with low intra-subject variability.

5.4.4 Discussion

Θ-band TE measured at baseline in EO condition correlates with cognitive outcome after 3 years in groups of patients with PD, independently from age, education, sex, disease duration, sleepiness and

LED. This effect is highly significant only in EO condition, while in EC condition the association with 3-year Overall RCI remains non-significant. In accordance with the a priori hypothesis, these results indicate autonomous information is contained in TE. The fact that cognitive decline can be predicted even in a cohort of patients with only a short median disease duration (2.5 yrs.) suggests sensitivity of TE already early in the course of the disease.

Daytime sleepiness is an early sign of PD [Abbott et al., 2005, Wulff et al., 2010]; moreover, daytime sleepiness in healthy people is associated with an increased θ-synchronization in EO condition [Achermann et al., 2016]. However, clinical measures of sleepiness at baseline are not significantly correlated with cognitive outcome after 3 years (see table 5.3). This result further supports the conclusion that TE at baseline predicts cognitive decline over middle and long time periods independently from any influence that daytime sleepiness might have on short-term outcome or present cognitive ability [Goldman et al., 2014]. Moreover, in our cohort of patients with relatively short disease duration and a comparatively long education, MMSE and low education are not risk factors for cognitive decline, which suggests EEG has a role to play for prognosis of cognition in PD.

This idea is supported by related findings. For instance, Klassen et al. [2011] demonstrated that background rhythm frequency and relative power in the θ-band were potential prognostic biomarkers for PD. The EEG based biomarker for changes over time in PD cognitive decline proposed by Caviness et al. [2015] is a δ-band power (2.5-4Hz) which correlated best with longitudinal neuropsychological performance changes in PD. In [Zimmermann et al., 2015], the authors conclude that global EEG slowing is a marker for overall cognitive impairment in PD.

Compared to TE, band power is influenced by a variety of unspecific factors, especially skull thickness and distance from the electrical source. The cause of these dependencies lies in the frequency-dependent attenuation by different materials like bone, galea, skin and other tissue types. Consequently, a power spectrum depends on individual anatomical and physiological features. Under the assumption that frequency-dependent attenuation only affects the amplitude of a signal, but leaves its frequency approximately unchanged, it follows that TE estimates might be less affected by an individual's anatomy. This is a consequence of the method used to estimate TE [Sneddon, 2007], as it relies only on the ratio between rapid changes (numerator) and slow changes (denominator) of the EEG signal (see equation 5.5). As a result, these entropy estimates might be less sensitive to information encoded in the amplitude of the EEG.

Moreover, as high variability of individual measurements of absolute band power usually preclude their direct inter-individual comparison, it has become standard to compare relative band power, i.e. the power in a frequency band as a percentage of the total power of the signal. While this normalization of band power is the most obvious transformation in order to facilitate inter-individual comparisons, a significant degree of variation will remain due to individual anatomical characteristics. For this reason, Klimesch [1999] suggests an individual frequency adjustment based on individual α-frequencies as an anchor point. In contrast to relative band power, comparing TE on an absolute scale is possible and does not necessarily require any adjustments.

While power-based features have a long history in EEG research, connectivity measures leverage the promising field of network neuroscience to find candidate biomarkers. M/EEG-based connectivity measures as potential biomarkers of PD progression were investigated in [Olde Dubbelink et al., 2014], while [Berendse and Stam, 2007] use M/EEG patterns of neural synchrony in PD patients to quantify the stage of the disease. While connectivity studies provide a model for detailed understanding of functional interdependency of different cortical areas and its alteration in dementia [Stam, 2014], the determination of connections, and therefore of the graph structure, may be demanding, given the large number of electrodes or sources and the different possible lengths of recording segments [Hardmeier et al., 2014]. In contrast, TE might present a shortcut by considering only possible alterations of signal

complexity as resulting from an alteration of the underlying graph structure, rather than characterizing the graph in detail.

TE of the EEG quantifies the amount of information contained in this signal and corresponds to the ratio between fast to slow oscillations [Sneddon, 2007]. Higher TE of the EEG reflects a higher complexity of the oscillatory brain activity. Oscillations that can be recorded at the surface of the head must come from synchronous discharges of more than 100 million neurons per scalp electrode [Nunez et al., 2006] working under the influence of a common pacemaker. Complexity of the recorded signal increases with the number of pacemakers, provided that the ability of the cortex to react to an increasing number of pacemakers is maintained. For normal cognition or consciousness, an optimal range of entropy of the EEG is a requirement [Carhart-Harris et al., 2014]. Abnormally increased entropy of brain activity is observed e.g. in psychedelic states, while abnormally decreased entropy is associated with reduced consciousness [Carhart-Harris et al., 2014, Zhang et al., 2001]. Decline of cognition may arguably be considered as a first step into loss of consciousness and, therefore, may also be associated with a decline of entropy.

Moreover, a slight reduction of entropy may precede a clinically evident decline of cognition, since at the very first phase of cortical dysfunction, patients recruit all available functional reserves to maintain apparently normal functioning [Peterson et al., 2015]. An example for this coping mechanism is the observation that patients with mild cognitive dysfunction "stop walking when talking" [Nieuwhof et al., 2017]. TE relates to the complexity of the cortical activity, and therefore, plausibly to the amount or complexity of information that can be processed by the cortex, which in turn is an expression of cognitive capacity. Interestingly, an artificial increase of entropy of oscillatory brain activity produced an improvement in numeracy skills in adults [Kadosh et al., 2013, 2010]. In encephalopathic and demented patients, partial or absent suppression of the EEG background activity (Berger effect) is a frequent and relatively early finding [Könönen and Partanen, 1993]. Therefore, the difference in EEG readings between HC and demented patients is greater when recorded in the EO than in the EC condition. Interestingly, TE shares a similar behavior and shows significant differences only between PD patients with and without cognitive decline over 3 years when recorded in the EO condition. Moreover, the difference of explained variance by TE measured in the EO condition, as opposed to the EC condition, points to an early deficiency of the "orienting response" [Sokolov, 1963] in the development of Parkinson's disease dementia (PD-D), which cannot be detected by BP based analysis of the EEG or neuropsychological testing before cognitive decline occurs. Loss of capacity to detect unexpected salient changes of environment may be at the base of both, the alteration of TE in EO as well as beginning cognitive decline.

The correlation of TE with cognitive outcome is observed when using global EEG, but also when the 10 regions are considered separately, as shown in figure 5.4.6. Upon visual inspection, the only distinction between the two extreme groups is a shift toward lower TE values across all regions in case of patients with a strong cognitive decline. The difference of the medians between the extreme group of cognitive stable and the group of cognitive declining patients is approximately 0.05 across all regions. Comparing the HC group and the extreme group of cognitively stable patients, no differences become apparent: judging only by their TE levels across regions and interpreting TE as a measure of cognitive health, the extreme group of cognitively stable PD patients appears as cognitively healthy as the HC group. Following Obeso et al. [2004], when the capacity of cognitively stable PD patients to compensate is exhausted, signal complexity in all regions will begin to drop, and eventually reach levels below the median TE levels of HC. The results apply to groups, and may help to define study populations for clinical trials, but cannot be applied in their present form for individual treatment decisions or counseling.

Much effort has gone into the discovery of clinically relevant biomarkers for cognitive decline in

PD [Cozac et al., 2016]. While this remains an active field of research, it is very likely that any newly emerging biomarkers will be a composite biomarker, based on more than one physiological or psychological measure. The composite prognostic biomarker for dementia in PD proposed by Liu et al. [2017], for example, is based on age at disease onset, MMSE, years of education, MDS-UPDRS III, sex, depression, and GBA mutation status. While psychological testing is affected by test-retest reliability and learning effects, and while genetic testing for PD dementia is still insufficiently validated, there is a practical advantage in having biomarkers derived from signals generally considered to be easily accessible, such as quantitative EEG.

5.4.5 Significance of the study

Defining cohorts at very high risk of cognitive decline is important in clinical trials for reaching significant results quickly with a relatively low number of patients. While the current results contribute to group characterization and, therefore, might help alone, or in combination with other parameters, to select the best groups for clinical trials, they cannot be used in their present form for individual counseling.

5.4.6 Limitations & Strengths

Limitations of the current study include unknown reliability of TE. However, for processing the EEG, we used the fully automated TAPEEG [Hatz et al., 2015] because its reliability has been demonstrated. Moreover, only 5 of the 42 PD patients developed overt dementia over the period of observation of three years, and the reliable cognitive deterioration was relatively small. Strengths of this study include a carefully matched HC group regarding age, sex and education level, as well as comprehensive neuropsychological testing of the patients with PD.

5.4.7 Conclusion

Currently, measures based on EEG for monitoring present cognition in PD, and predicting its future development, are often derived from band power or connectivity estimates. However, TE is a new measure of signal complexity that seems to be at least as sensitive, and possibly more robust against influence of age, than spectral analysis for predicting cognitive decline in groups of patients, but not individuals. Furthermore, in contrast to band power, TE is sensitive to EC/EO condition, with only the EO condition containing information with respect to cognitive decline over a period of 3 years. This might motivate new neuropsychological testing paradigms specifically designed to the EO condition.

5.5 Archetypes of Parkinson's disease

The analysis in section 5.4 has shown that the θ-band of EEG in eyes open (EO) condition is a potential biomarker for cognitive decline based both on features quantifying signal complexity or relative band power. The following experiment describes a proof-of-concept application based on these insights but applies deep archetypal analysis to EEG in order to provide a more interpretable result. The setting described in the following section allows to visualize an EEG *sequence* in latent space – instead of averaging over the whole recording to obtain a single data point, the EEG of an individual is segmented into short windows of 1 second duration with an overlap of 90% with subsequent windows in order to obtain multiple data points per individual recording.

5.5.1 Cohort description

For the following experiment, which shows the potential of deep archetypal analysis for EEG, a reduced cohort has been selected, composed of 16 patients suffering from Parkinson's disease and 8 healthy controls. Table 5.5 shows the patient demographic. From the 16 patients, half remain cognitively stable over a period of 3 years while the other half shows signs of cognitive deterioration over the same period.

	Age	Edu	Sex	MMS	UPDRS-III	LED	DisDur
HC (8)	**66.5**	**12.5**	**3f.**	**30**	-	-	-
1st Quartile	64	11.75	-	29.5	-	-	-
3rd Quartile	67	16.25	-	30	-	-	-
PD (16)	**68.5**	**13.5**	**6f.**	**29**	**14.5**	**487.5**	**2**
1st Quartile	64.75	12.75	-	28	5.75	295	0.75
3rd Quartile	72.75	16.25	-	29	20.5	858	5.25

Table 5.5 – Values are presented as median values, along with values for the 25%- and 75%-quartiles. Age, education (edu), and disease duration (DisDur) are given in years. LED is given in milligrams. MMS and UPDRS-III refer to standardized psychological tests.

5.5.2 Experiment

EEG of the patient cohort was preprocessed identically to the data used in section 5.4. For reducing the computational resources, only 20 seconds of EEG per study participant were used. While this might be sub-optimal for obtaining robust neurological insights, it is sufficient to demonstrate the feasibility of the proposed method. With a total of 24 participant, 193

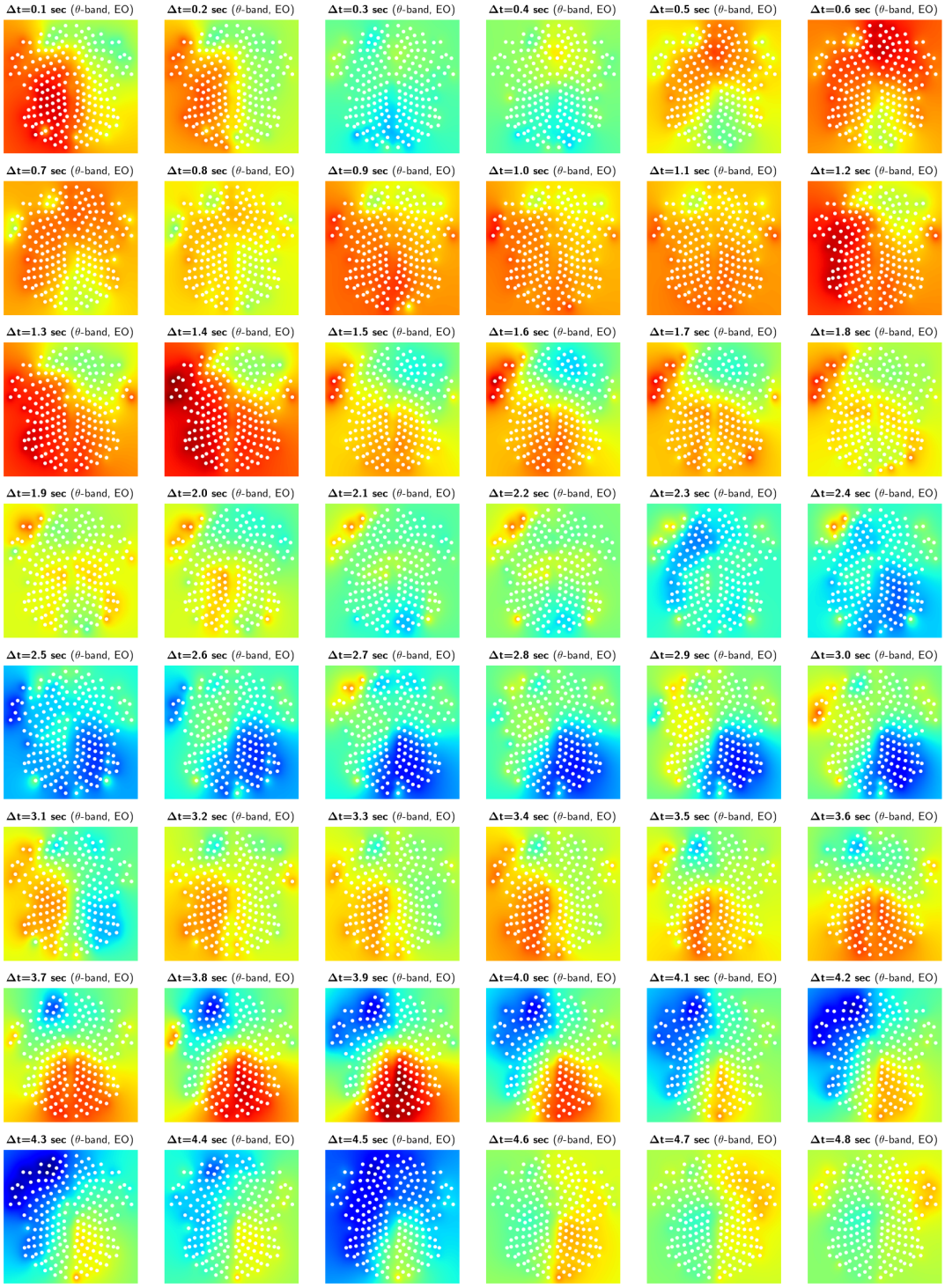


Figure 5.5.1 – Sequence of relative EEG power of the θ-band from a healthy control subject. Each topography is calculated based on a 1 second window of EEG. With a sliding factor of 0.1 seconds, the subsequent topography is calculated. The total sequence shown here is based on 4.8 seconds of EEG.

consecutive θ-power-based scalp EEG topographies were obtained using a sliding windows of 1 second with 90% overlap. After applying an inverse distance interpolation between the relative θ-power values at each electrode, this lead to a data set consisting of 4632 images of 100×100 pixel. Figure 5.5.1 shows an image sequence of 48 images for the θ-band power measured in EO condition from a healthy control. With the parameters used for the sliding window, smooth appearance of the sequence of scalp topography images is ensured: As can be seen in figure 5.5.1, most subsequent topographies are evolutions of their immediate predecessors, while sudden jump-like changes may occur where oscillatory activity switches into a different state more quickly. An example might be the first row of figure 5.5.1, topography $2 \rightarrow 3$ or the last row, again topography $2 \rightarrow 3$. Generally, this should provide an appropriate input to deep AA as smooth transitions within the data occur naturally.

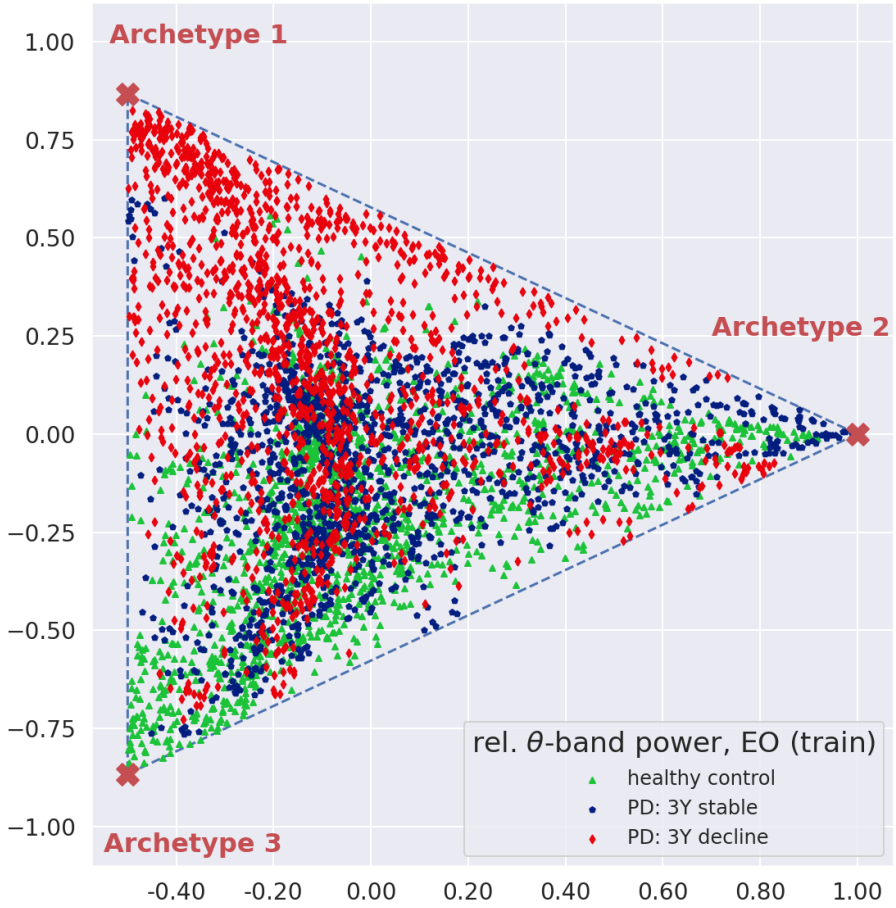


Figure 5.5.2 – Learned latent space based on scalp topographies of relative EEG power of the θ-band. Class labels are (i) healthy control, (ii) PD: 3-year stable and (iii) PD: 3-year decline. The data set was comprised of 24 subjects, each providing 193 scalp topography images.

5.5.3 Discussion

The learned archetypes are shown in figure 5.5.3. The meaning of these archetypes is interpreted based on the structure of the latent space shown in figure 5.5.2. It shows the three classes (i) healthy controls, (ii) PD: 3-year stable and (iii) PD: 3-year declining. Class separation is very low as θ-power topographies produced by neural activity within a window of 1 second are most of the time *not* representative of the general (future) health status (here: cognition). But a minority of those topographies will exhibit a spatial distribution of θ-power that a healthy brain would be more likely to produce than a brain in process of neural degeneration, and vice versa. The structure of the depicted latent space in figure 5.5.2 shows that *archetype 1* is clearly associated with spatial distributions of θ-power associated with a 3-year cognitive decline, while *archetype 3* is the extreme representative of a θ-power distribution produced by a healthy brain. On the other hand, *archetype 2* has no clear interpretation given the current labels as samples from all three classes are found close to this archetype. Interestingly, the identified archetypes allow for a qualitative comparison to the

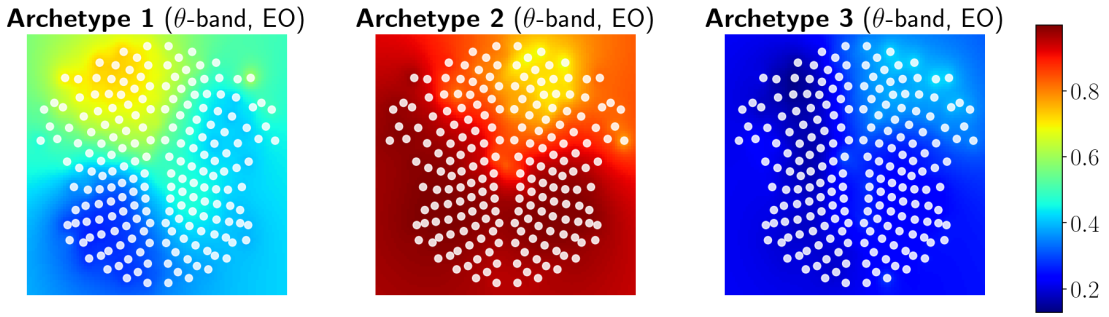


Figure 5.5.3 – Learned archetypes showing both local as well as global differences. Globally, each archetype is associated with a different average level of θ-power: (i) archetyp 1: mid power level, (ii) archetype 2: high power level and (iii) archetype 3: low power level. Locally, archetype 1 has increased frontal power while archetype 2 has increased posterior power, similar to archetype 3 but at a much higher level.

state of the art in EEG research for Parkinson's disease (PD) [Klassen et al., 2011] where a consensus has been established based on multiple independent studies stating that elevated θ-power is a marker of neurodegeneration. Inspecting archetype 3 in figure 5.5.3, it is seen that it has the overall lowest level of θ-power compared to the other two archetypes, which according to current research is a sign of a healthy brain. It thus increases confidence in the learned latent space structure that mostly samples from healthy controls are close to this archetype. More interesting, as it is unexpected, is the comparison between archetypes 1 and 2: While archetype 2 shows overall the highest level of θ-power among archetypes, PD patients with the prospect of cognitive decline produce samples accumulating around archetype 1, which shows higher levels of θ-power compared to the levels of healthy controls, but still considerably lower than the levels of archetype 2. A possible interpretation might be provided by the archetypes themselves – upon closer inspection it is seen that they are each others inverse: Archetype 1 has higher levels in the frontal regions and lower levels in the posterior regions while the situation is inverted in case of archetype 2. A possible hypothesis can be formulated stating that both the overall level of θ-power as well as the regional distribution play a role for future cognitive decline. An additional observation is provided by observing the interpolation between archetype 1 and archetype 2 as increasing proximity to archetype 2 attracts samples from all three classes while increased proximity

to archetype 1 attracts only samples from PD cognitive declining patients. Figure 5.5.4 shows samples

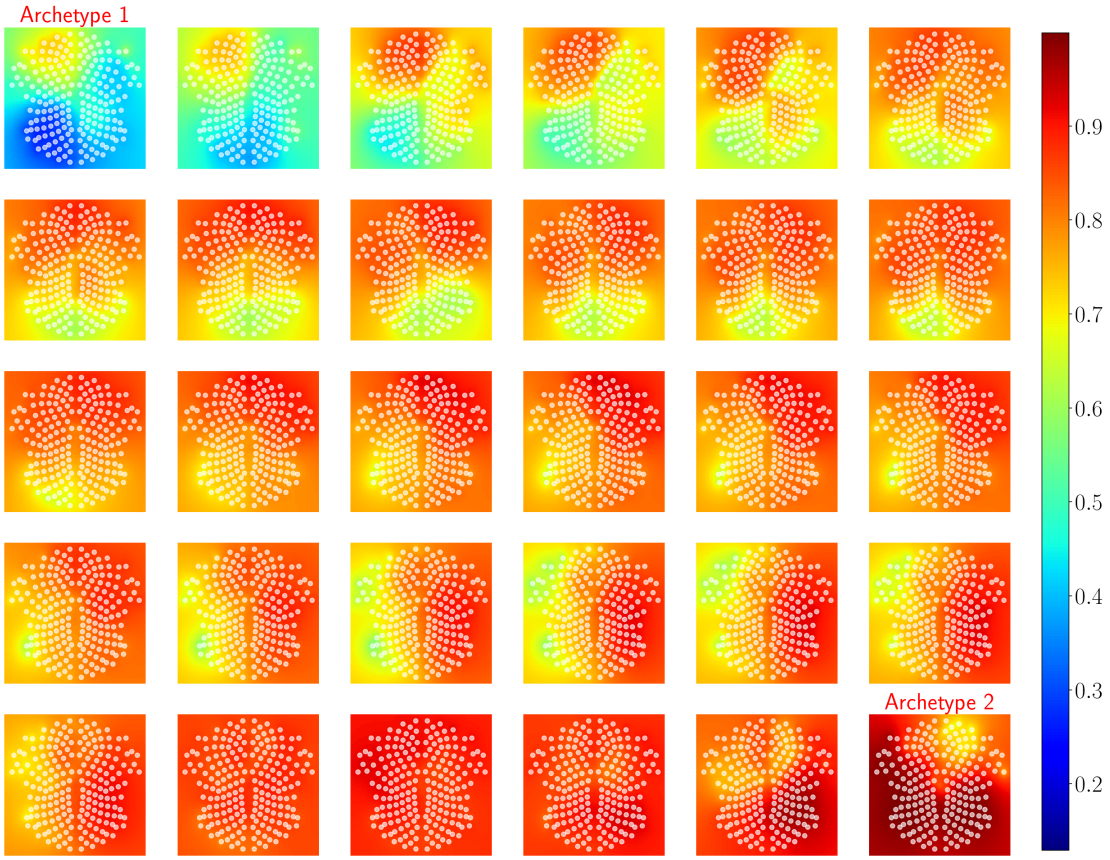


Figure 5.5.4 – Interpolation between archetype 1, which is associated with cognitively deteriorating patients, and archetype 2, which is not specific to any of the three classes.

decoded at equal distances on the connecting line between archetypes 1 and 2 in latent space. The first two rows clearly show the same distribution of θ-band power, i. e. increased in the front and reduced in the back, but at different overall power levels. Halfway into the middle row a transition is slowly beginning shifting the front/back distribution to a left/right distribution. Clearly, the fourth row now shows hemispheric imbalances in power distribution. Comparing with the latent space one could hypothesize that hemispheric imbalances (low power left, high power right) might not necessarily be a sign of deterioration compared to front/back imbalances with frontal high power and posterior low power in the θ-band.

5.5.4 Limitations & Strengths

While this proof of concept showed interesting possibilities in interpreting scalp power distributions and also opens up the possibility to analyze the dynamic behavior of EEG, the experiment in its current form is limited by the small sample size of 24 patients as well as the small number of topographies per participant of 193 images, based on only 20 seconds of EEG. Furthermore, the restriction to the θ-band might be unnecessary, as deep AA is well able to take 3-dimensional input tensors such that

further bands could be added – similar to having three channels for rgb-images, the input could be a 5 channel input, with each channel representing one of the five standard bands of EEG. Moreover, while group effect were clearly visible, the goal of biomarker research in general lies in individual counseling. Therefore, additional data for exploring *individual* latent EEG mappings would be an important next step.

6 Discussion and Outlook

A central question of this thesis was the design of new machine learning models appropriate for the analysis of clinical resting state EEG data. What makes a model *appropriate* is the degree to which the results it provides can be intuitively understood by a medical expert. This requirement of interpretability in the wider sense motivated the use of dimensionality reduction either through sparsity inducing penalties, effectively reducing the number of potential predictors, or making use of latent space models in order to provide a low dimensional representation of a data set. With neurodegenerative diseases on the rise, the clinical data used in this thesis consisted of noninvasive resting state scalp EEG recordings from 24 healthy controls and 42 patients suffering from Parkinson's disease. Identifying differences at group level at time of recording as well as prediction of future cognitive decline based on those recordings were questions of interest. But with a median disease duration of 2.5 years the cohort had not progressed much into the overt phase of the disease. Together with the large heterogeneity of Parkinson's disease, the present data set was challenging and – based on standard classification methods – expectations were moderate to identify strong group-wise differences using spectral features.

Our first contribution describes a penalized regression method. Assuming a convex objective function, the proposed method allows to interpolate between best subset selection (ℓ_0-pseudonorm) and lasso (ℓ_1-norm) as the penalty. But with all penalties "below" ℓ_1 leading to non-convex optimization problems, standard algorithms such as Forward stagewise and Frank–Wolfe cannot be used as they require a convex feasible region. The proposed solutions is a transformation of the non-convex feasible region into a convex one, such that the aforementioned algorithms become – in principle – applicable. An important requirements of any such transformation is that it when applied on the convex objective, no saddle points or extrema are induced by the transform. It can be shown that a transformation of the penalty is possible in such a manner that the transformed objective becomes *invex*. Invexity ensures that all every stationary point of the transformed objective is also a global minimum. Of course the optimization target – the combination of invex objective and convex penalty – is itself a non-convex optimization problem, but one that can be solved with standard convex optimization algorithms. In this non-convex setting the guarantee that the coefficient paths for different values of the tuning parameter still connect pointwise optimal solutions is lost. Nevertheless, the ability to produce coefficient paths is in itself an important property for enabling model selection by cross-validation. We show several application cases on artificial data and also analyze, in a setting of logistic regression, which spectral features are most important in differentiating EEG from healthy controls from those recorded on patients suffering from Parkinson's disease. This result confirmed the importance of low frequency oscillations in the $4-8$Hz range as an important marker of neurodegeneration.

Our second contribution deals with alternative features derived from EEG. Commonly, spectral features are used to quantify EEG data. The majority of research relies on quantifying the relative power in different frequency band in order to derive electroencephalographic biomarkers. But recently, the quantification of the complexity of EEG wave forms has shown promise in establishing biomarks that do not rely on spectral characteristics. We could show that complexity of EEG provides information about brain status not obtainable through spectral analysis. Using baseline EEG recordings and cognitive status measured 3 years after baseline, based on comprehensive psychological testing, complexity measured in eyes open condition was significantly correlated with changes in cognition. Interestingly, the eyes closed state was less informative regarding cognitive decline. With complexity measures potentially showing sensitivity to different brain networks, this result motivates an interesting path for future research.
Our third contribution provides a new method of analyzing EEG data based on a deep latent variable model. By translating the multivariate EEG time series into a sequence of scalp topographies based on spectral features, extreme spatial distributions of spectral power can be identified in order to obtain information about brain states associated with neurodegeneration. By extending the classical model of linear archetypal analysis to simultaneously learn an appropriate latent space along with extreme representatives of spatial power distributions, EEG data can be utilized in a more efficient manner: Instead of averaging the spectral power over the whole EEG sequence, power is estimated within a short window. Sliding this windows over the whole time series provides a sequence of short-lived brain states which form the input to the proposed deep archetypal analysis method. Applying this method on a data set consisting of healthy controls and patients suffering from Parkinson's disease, a spectral distribution was obtained which is associated with future cognitive decline. It is characterized by elevated power of the θ-band in the posterior brain region compared to the frontal region. The spectral archetype associated with a healthy brain, on the other hand, shows an inverse distribution at overall lower levels of θ-power: higher power in the front and lower power in posterior brain region.

6.1 Limitations

The major limitation – as is often the case when working on medical data – is the size of the available data set. While it is a great achievement to collect longitudinal clinical data, especially if comprehensive psychological testing is involved, data sets considers "big" by clinical standards are often perceived as small from the machine learning community. Especially the present research, which involved a very heterogeneous disease, poses statistical challenges for finding significant associations in diagnostic and predictive settings. With the long-standing goal of machine learning providing the key to a more personalized way to conduct medicine, large clinical data sets are an absolute requirement. Presently, based on the available data, only group-wise characteristics could reliably be obtained. This is certainly a step into the right direction but still a long way off from providing insights ultimately enabling individual counseling. From a statistical point of view, the design of new models is usually followed by an evaluation step in order to estimate their benefits compared to the status quo. This task would also benefit from access to larger collections of data. Nevertheless, the presented contributions have all been validated on alternative data, either artificial or natural, and provide interesting paths to future work.

6.2 Future Work

EEG in combination with machine learning, especially deep learning, is a very promising area of research. With visual inspection being the clinical standard of analyzing EEG, there is still much opportunity for providing learning based systems able to supplement in a meaningful way the daily work of clinicians. Trends in EEG recording, especially long-term recordings over several hours or days, are almost impossible to recognize based on visual inspection. The reason for this is that the largest portion of EEG that can reasonably be displayed on a single screen at any given time has a duration of 20 to 30 seconds. But even for short routine EEG recordings, which seldom last more than 30 minutes, much information remains inaccessible to the naked eye. Deep archetypal analysis revealed the potential existence of brain states associated with cognitive decline, but as such states are short lived and don't make up the majority of the time series, automated methods seem very promising. Several aspects of deep AA in combination with EEG would be worthwhile exploring: With the ability to map the EEG sequence into latent space, the dynamic of EEG is mapped onto a 2-dimensional latent path, which itself might contain interesting information, such as average path length or distribution over path lengths. Of course, deep AA does not exclusively rely on spectral power based topographies as the input data. An interesting multi-modal approach would be a combination of EEG and fMRI data which has been recorded simultaneously. Generally, focusing more on dynamical aspects of EEG based on deep learning methods such as deep AA, would be a natural aspect to include in future work. With the well known model of microstate analysis of EEG [Lehmann et al., 1987] a model for quantifying EEG dynamics has been proposed over 3 decades ago – continuing along these lines, based on more sophisticated statistical models, certainly promises interesting insights.

´

A Appendix

A.1 Ultra-sparse Model Identification and Learning with Invexity

A.1.1 Proof of consistent first coefficient selection for least squares regression

Given a least-squares regression problem $||\boldsymbol{b} - A\boldsymbol{x}||_2^2$, we can assume w.l.o.g that $z_i \geq 0 \ \forall \ i$, since we can always solve the monotone version. Let $\mathscr{A}$ be the active set and $A_{\mathscr{A}}$ be the active subset of A, i.e. the matrix containing the columns which correspond to nonzero entries $z_j > 0$. Assume $|\mathscr{A}| = 1$, $\mathscr{A} = \{k\}$. Then, at a stationary point, the inequality constraint (assuming it is active) has the form $g(\hat{z}_k) = \hat{z}_k = \kappa$. Thus,

$$\begin{aligned} f(x_k) &= \|\boldsymbol{b} - A_{\mathscr{A}} h(z_k)\|^2 \\ &= \|\boldsymbol{b} - \boldsymbol{a}_{[\cdot,k]} h(\kappa)\|^2 \\ &= \|\boldsymbol{y} - \boldsymbol{a}_{[\cdot,k]} c\|^2, \text{ where } c := h(\kappa) \\ &= \boldsymbol{b}^t\boldsymbol{b} - 2\boldsymbol{a}_{[\cdot,k]}^t \boldsymbol{b}c + \boldsymbol{a}_{[\cdot,k]}^t \boldsymbol{a}_{[\cdot,k]} c^2 \end{aligned}$$

Assume further that the inputs are standardized, i.e. $\boldsymbol{a}_{[\cdot,j]}^t \boldsymbol{a}_{[\cdot,j]} = 1$, $\forall j = 1,\ldots,2p$. Then the minimum value of $f(z_k)$ is obtained for the index k' which maximizes $\boldsymbol{a}_{[\cdot,k']}^t \boldsymbol{b}$. The latter, however, can be expressed as

$$k' = \arg\max_k -\frac{\partial f(z_k)}{\partial z_k}\Big|_{z_k=0}.$$

This follows from

$$\frac{1}{2}\nabla f(\boldsymbol{z}) = [-A^t\boldsymbol{b} + A^t A h(\boldsymbol{z})] \circ \frac{\partial h(\boldsymbol{z})}{\partial \boldsymbol{z}}$$

Note that this is exactly the first variable selected by the stagewise forward algorithm, and that the selection of the first variable does not depend on any additional parameter of $h(\cdot)$. As long as $\boldsymbol{x}$ has

only one nonzero component $x_{k'}$ and $(\nabla f(\boldsymbol{z}))_{k'} > (\nabla f(\boldsymbol{z}))_k$, $\forall k \neq k'$, it holds that

$$\begin{aligned}(\nabla f(\boldsymbol{z}))_{k'} &= \lambda \\ (\nabla f(\boldsymbol{z}))_k &= \lambda - \mu_k, \text{ with } \mu_k > 0 \, \forall k \neq k'.\end{aligned}$$

The path defined by the stagewise forward method now proceeds by increasing $z_{k'}$ until a second variable becomes active. The transition point is determined by the first μ_k to become zero, and the second variable becomes active if $(\nabla f)_k > (\nabla f)_{k'}$.

A.1.2 Proof of the implication $x_j^+ > 0 \Rightarrow x_j^- = 0$

Given augmentation functions

$$\begin{aligned}s(x^+, x^-) &= x^+ - x^- \\ t(x^+, x^-) &= x^+ + x^-\end{aligned}$$

We analyse the augmented problem

$$\begin{aligned}&\min_{(x^+, x^-)} f(s(x^+, x^-)) \\ \text{s.t.} \quad &g(t(x^+, x^-)) \leq \kappa, x_j^+ > 0, x_j^- > 0\end{aligned}$$

Replacing $g(\cdot)$ with $h^{-1}(\cdot)$, the Lagrangian of the augmented problem is:

$$\begin{aligned}\mathscr{L}(x^+, x^-) = f(s(x^+, x^-)) + \lambda h^{-1}(t(x^+, x^-)) \\ - \sum_{j=1}^{p} \mu_j^+ x_j^+ - \sum_{j=1}^{p} \mu_j^- x_j^-\end{aligned}$$

KKT conditions:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial x_j^+} = \frac{\partial f}{\partial s}\frac{\partial s}{\partial x^+} - \mu_j^+ + \lambda \frac{\partial h^{-1}}{\partial t}\frac{\partial t}{\partial x^+} &= 0 \\
\frac{\partial \mathscr{L}}{\partial x_j^-} = \frac{\partial f}{\partial s}\frac{\partial s}{\partial x^-} - \mu_j^- + \lambda \frac{\partial h^{-1}}{\partial t}\frac{\partial t}{\partial x^-} &= 0 \\
\text{complementary slackness:} \quad \mu_j^+ x_j^+ &= 0 \\
\text{complementary slackness:} \quad \mu_j^- x_j^- &= 0 \\
\text{primal feasibility:} \quad \forall j = 1,\ldots,p: \quad x_j^+ &\geq 0 \\
\text{primal feasibility:} \quad \forall j = 1,\ldots,p: \quad x_j^- &\geq 0
\end{aligned}$$

Note that $\frac{\partial s}{\partial x^+} = 1$, $\frac{\partial s}{\partial x^-} = -1$ and $\frac{\partial t}{\partial x^+} = \frac{\partial t}{\partial x^-} = 1$. It follows that $x_j^+ > 0, \lambda > 0$ implies $x_j^- = 0$:

$$\begin{aligned}
x_j^+ > 0, \lambda > 0 \quad &\Rightarrow \quad \mu_j^+ = 0 \\
&\Rightarrow \quad -\frac{\partial f}{\partial s} = \lambda \frac{\partial h^{-1}}{\partial t} > 0 \\
&\Rightarrow \quad \mu_j^- > 0 \\
&\Rightarrow \quad x_j^- = 0.
\end{aligned}$$

Likewise $x_j^- > 0, \lambda > 0$ implies $x_j^+ = 0$.
Step 2 follows from the fact that h is a strictly monotonically increasing function per definition in section 3.

A.1.3 Curved Level Sets for Information Bottleneck

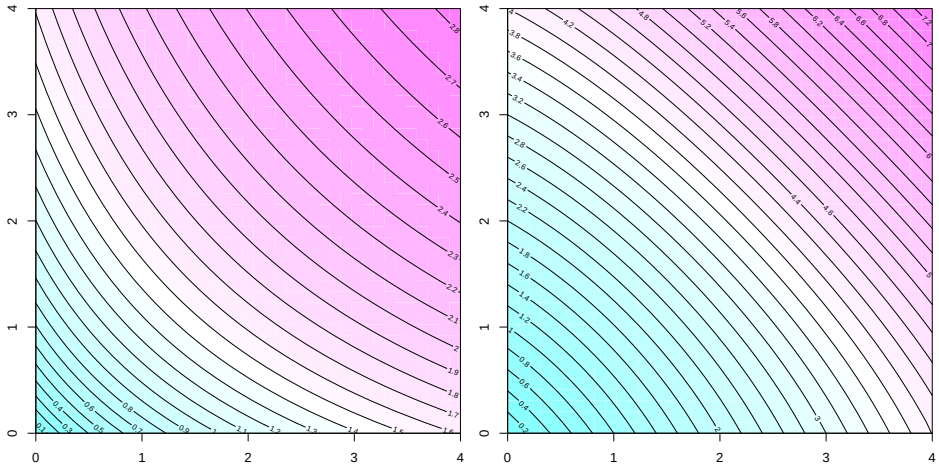


Figure A.1.1 – Contour-plot of the constraint function of the sparse IB problem in two dimensions. The correlation between the two variables is 0.7. Left: original variables lead to a concave constraint $g(x)$. Right: transformed variables result in a constraint function $g(z)$ whose sublevel sets are convex.

A.2 Reduced complexity of EEG in Parkinson's disease predicts cognitive decline

A.2.1 Consort Scheme

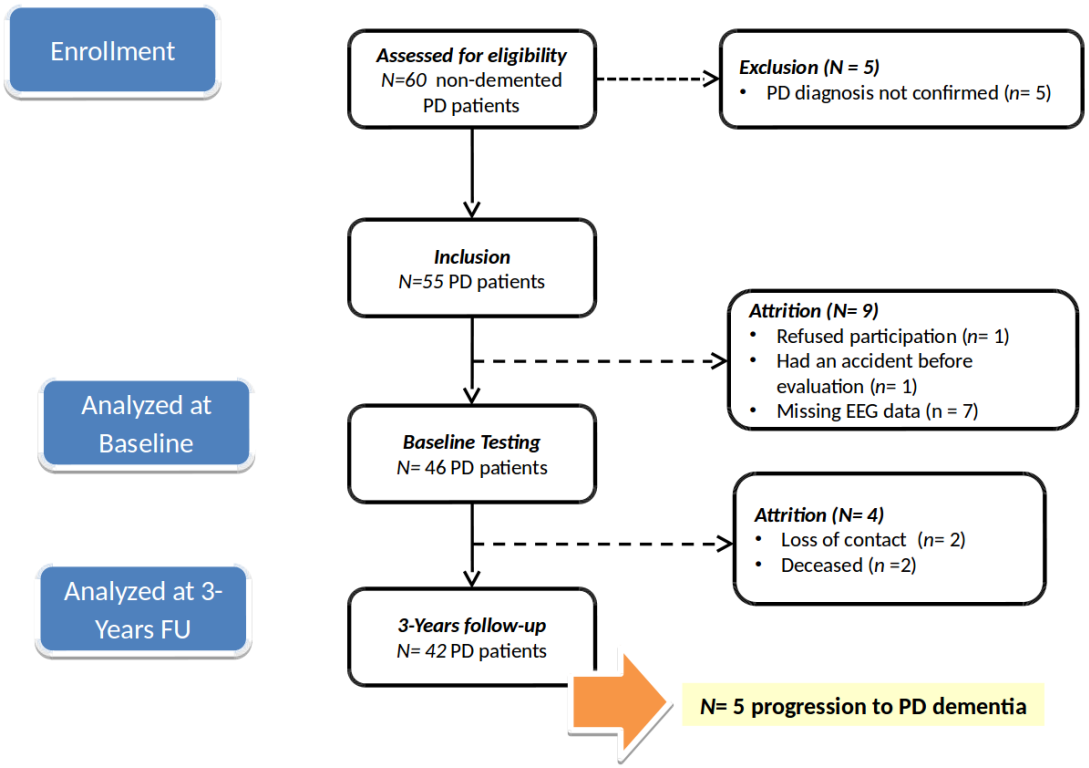


Figure A.2.1 – Consort scheme: Overview of patients who participated in the present study.

A.2.2 Overall RCI at 6-month follow-up

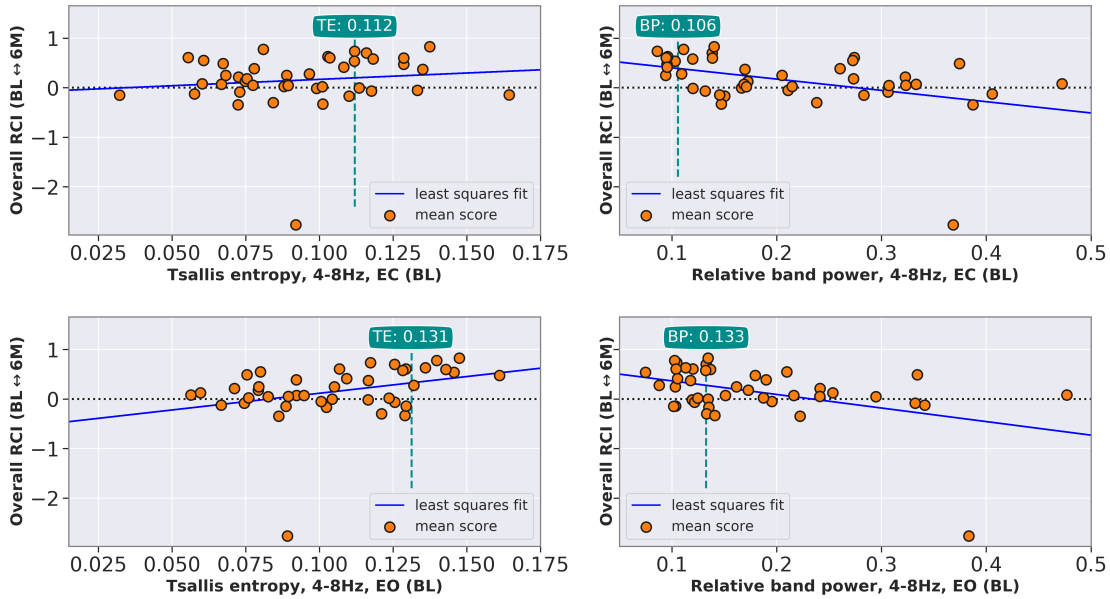


Figure A.2.2 – Overall RCI for the 6-month period after baseline. Within this period the cognitive decline is small compared to the statistical noise level. As a consequence, a quantitative analysis is not advisable. Still, the least squares fit shows identical tendencies compared to the 3-year RCI shown in figure 5.4.5, i.e. positive slope in case of TE, negative slope in case of relative band power.

Bibliography

R D Abbott, G W Ross, L R White, C M Tanner, K H Masaki, J S Nelson, J D Curb, and H Petrovitch. Excessive daytime sleepiness and subsequent development of parkinson disease. *Neurology*, 65(9): 1442–6, Nov 2005.

Peter Achermann, Thomas Rusterholz, Roland Dürr, Thomas König, and Leila Tarokh. Global field synchronization reveals rapid eye movement sleep as most synchronized brain state in the human eeg. *Royal Society open science*, 3(10):160201, 2016.

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

Torbjörn Åkerstedt and Mats Gillberg. Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52(1-2):29–37, 1990.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *CoRR*, abs/1612.00410, 2016. URL http://arxiv.org/abs/1612.00410.

Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2270–2278. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6372-learning-the-number-of-neurons-in-deep-networks.pdf.

Edgar Anderson. The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

P. Atkins and J. de Paula. *Atkins' Physical Chemistry*. OUP Oxford, 2010. ISBN 9780199543373.

David Bäckström, Gabriel Granåsen, Magdalena Eriksson Domellöf, Jan Linder, Susanna Jakobson Mo, Katrine Riklund, Henrik Zetterberg, Kaj Blennow, and Lars Forsgren. Early predictors of mortality in parkinsonism and parkinson disease: A population-based study. *Neurology*, 91(22):e2045–e2056, 2018.

C. Bauckhage and K. Manshaei. Kernel archetypal analysis for clustering web search frequency time series. In *2014 22nd International Conference on Pattern Recognition*, pages 1544–1549, Aug 2014. doi: 10.1109/ICPR.2014.274.

Christian Bauckhage and Christian Thurau. Making archetypal analysis practical. In Joachim Denzler, Gunther Notni, and Herbert Süße, editors, *Pattern Recognition*, pages 272–281. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-03798-6.

Christian Bauckhage, Kristian Kersting, Florian Hoppe, and Christian Thurau. Archetypal analysis as an autoencoder. In *Workshop New Challenges in Neural Computation 2015*, pages 8–16, 10 2015. URL https://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_03_2015.pdf.

John M Bekkers. Pyramidal neurons. *Current biology*, 21(24):R975, 2011.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

A. Ben-Israel and B. Mond. What is invexity? *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 28:1–9, 7 1986.

Henk W Berendse and Cornelis J Stam. Stage-dependent patterns of disturbed neural synchrony in parkinson's disease. *Parkinsonism & Related Disorders*, 13:S440–S445, 2007.

Hans Berger. Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. Eegtotext: Learning to write medical reports from eeg recordings. In *Machine Learning for Healthcare Conference*, pages 513–531, 2019.

Lorenz C. Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25): 8732–8733, 2009. doi: 10.1021/ja902302h. PMID: 19505099.

Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995. ISSN 00401706. URL http://www.jstor.org/stable/1269730.

TC Buter, A Van Den Hout, FE Matthews, JP Larsen, C Brayne, and D Aarsland. Dementia and survival in parkinson disease: a 12-year population study. *Neurology*, 70(13):1017–1022, 2008.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304(5679): 1926–1929, 2004.

Luisa F. Cabeza, Andrea Gutierrez, Camila Barreneche, Svetlana Ushak, Angel G. Fernandez, A. Ines Fernadez, and Mario Grageda. Lithium in thermal energy storage: A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, 42:1106 – 1112, 2015. ISSN 1364-0321.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Ercan Canhasi and Igor Kononenko. Weighted hierarchical archetypal analysis for multi-document summarization. *Computer Speech & Language*, 37, 11 2015. doi: 10.1016/j.csl.2015.11.004.

Robin Lester Carhart-Harris, Robert Leech, Peter John Hellyer, Murray Shanahan, Amanda Feilding, Enzo Tagliazucchi, Dante R Chialvo, and David Nutt. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, 8:20, 2014.

John N Caviness, Joseph G Hentz, Christine M Belden, Holly A Shill, Erika D Driver-Dunckley, Marwan N Sabbagh, Jessica J Powell, and Charles H Adler. Longitudinal eeg changes correlate with cognitive measure deterioration in parkinson's disease. *Journal of Parkinson's disease*, 5(1): 117–124, 2015.

Menorca Chaturvedi, Florian Hatz, Ute Gschwandtner, Jan G Bogaarts, Antonia Meyer, Peter Fuhr, and Volker Roth. Quantitative eeg (qeeg) measures differentiate parkinson's disease (pd) patients from healthy controls (hc). *Frontiers in aging neuroscience*, 9:3, 2017.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, 6(Jan):165–188, 2005.

Lawrence A Coben, Warren L Danziger, and Leonard Berg. Frequency analysis of the resting awake eeg in mild senile dementia of alzheimer type. *Electroencephalography and clinical neurophysiology*, 55(4):372–380, 1983.

Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

Vitalii V Cozac, Ute Gschwandtner, Florian Hatz, Martin Hardmeier, Stephan Rüegg, and Peter Fuhr. Quantitative eeg and cognitive decline in parkinson's disease. *Parkinson's Disease*, 2016, 2016.

Steven C Cramer. Stratifying patients with stroke in trials that target brain repair. *Stroke*, 41 (10_suppl_1):S114–S116, 2010.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994. doi: 10.1080/00401706.1994.10485840. URL http://digitalassets.lib.berkeley.edu/sdtr/ucb/text/379.pdf.

Adele Cutler and Emily Stone. Moving archetypes. *Physica D: Nonlinear Phenomena*, 107(1):1–16, August 1997. doi: 10.1016/s0167-2789(97)84209-1. URL https://doi.org/10.1016/s0167-2789(97)84209-1.

Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1): 9–21, 2004.

Francesco Dinuzzo, Cheng S Ong, Gianluigi Pillonetto, and Peter V Gehler. Learning output kernels with block coordinate descent. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 49–56, 2011.

Minou Djawdan, Tina T. Sugiyama, Lisa K. Schlaeger, Timothy J. Bradley, and Michael R. Rose. Metabolic aspects of the trade-off between fecundity and longevity in drosophila melanogaster. *Physiological Zoology*, 69(5):1176–1195, 1996.

Lucy Dodakian, Kelli G Sharp, Jill See, Neil S Abidi, Khoa Mai, Brett W Fling, Vu H Le, and Steven C Cramer. Targeted engagement of a dorsal premotor circuit in the treatment of post-stroke paresis. *NeuroRehabilitation*, 33(1):13–24, 2013.

E Ray Dorsey and Bastiaan R Bloem. The parkinson pandemic—a call to action. *JAMA neurology*, 75(1):9–10, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

John Duchi. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford. edu/class/stats311/Lectures/full_notes. pdf. Last visited on*, 2:23, 2016.

Felix Effenberger. A primer on information theory with applications to neuroscience. In *Computational Medicine in Data Mining and Modeling*, pages 135–192. Springer, 2013.

Hana El Samad, Mustafa Khammash, Cristian Homescu, and Linda Petzold. Optimal performance of the heat-shock gene regulatory network. *Proceedings 16th IFAC World Congress*, 16, 1 2005. URL https://engineering.ucsb.edu/~cse/Files/IFACC_HS_OPT04.pdf.

Murat Emre, Dag Aarsland, Richard Brown, David J Burn, Charles Duyckaerts, Yoshikino Mizuno, Gerald Anthony Broe, Jeffrey Cummings, Dennis W Dickson, Serge Gauthier, et al. Clinical diagnostic criteria for dementia associated with parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 22(12):1689–1707, 2007.

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

EB Forsaa, JP Larsen, T Wentzel-Larsen, and G Alves. What predicts mortality in parkinson disease?: a prospective population-based long-term study. *Neurology*, 75(14):1270–1276, 2010.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.

Robert M Freund, P Grigas, and R Mazumder. Incremental forward stagewise regression: Computational complexity and connections to lasso. In *International Workshop on advances in Regularization, Optimization, Kernel Methods and Support Vector Machines (ROKS)*, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Jerome H Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28 (3):722–738, 2012.

George M. Furnival and Robert W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4): 499–511, 1974. ISSN 00401706. URL http://www.jstor.org/stable/1267601.

T. J. Jr. Garland. Quick guides: Trade-offs. *Current Biology*, 24(2):R60–R61, 2014.

WR Gibb and AJ1033142 Lees. The relevance of the lewy body to the pathogenesis of idiopathic parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 51(6):745–752, 1988.

Giorgio Giorgi. On first order sufficient conditions for constrained optima. In *Nonlinear and Convex Analysis in Economic Theory*, pages 53–66, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

Giorgio Giorgi. On some generalizations of preinvex functions. 49, 01 2008.

Jennifer G Goldman, Glenn T Stebbins, Vania Leung, Barbara C Tilley, and Christopher G Goetz. Relationships among cognitive impairment, sleep, and fatigue in parkinson's disease using the mds-updrs. *Parkinsonism & related disorders*, 20(11):1135–1139, 2014.

Rafael Gomez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Jean Gotman. High frequency oscillations: the new eeg frontier? *Epilepsia*, 51(Suppl 1):63, 2010.

Ulrike Grömping et al. Relative importance for linear regression in r: the package relaimpo. *Journal of statistical software*, 17(1):1–27, 2006.

B H. P. Chan, Daniel Mitchell, and Lawrence Cram. Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 338, 01 2003. doi: 10.1046/j.1365-8711.2003.06099.x.

Martin Hardmeier, Florian Hatz, Habib Bousleiman, Christian Schindler, Cornelis Jan Stam, and Peter Fuhr. Reproducibility of functional connectivity and graph measures based on the phase lag index (pli) and weighted phase lag index (wpli) derived from high resolution eeg. *PloS one*, 9(10):e108648, 2014.

C Härting, HJ Markowitsch, H Neufeld, P Calabrese, K Deisinger, J Kessler, and WMS-R Wechsler Gedächtnistest. Revidierte fassung. *Hans Huber-Verlag, Bern*, 2000.

Trevor Hastie, Jonathan Taylor, Robert Tibshirani, and Guenther Walther. Forward stagewise regression and the monotone lasso. *Electron. J. Statist.*, 1:1–29, 2007. doi: 10.1214/07-EJS004.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

F Hatz, M Hardmeier, H Bousleiman, S Rüegg, C Schindler, and P Fuhr. Reliability of fully automated versus visually controlled pre-and post-processing of resting-state eeg. *Clinical Neurophysiology*, 126(2):268–274, 2015.

Mariese A Hely, Wayne GJ Reid, Michael A Adena, Glenda M Halliday, and John GL Morris. The sydney multicenter study of parkinson's disease: the inevitability of dementia at 20 years. *Movement disorders*, 23(6):837–844, 2008.

Nesma Houmani, Gérard Dreyfus, and François B Vialatte. Epoch-based entropy for early screening of alzheimer's disease. *International journal of neural systems*, 25(08):1550032, 2015.

Peter Huggins, Lior Pachter, and Bernd Sturmfels. Toward the human genotope. *Bulletin of Mathematical Biology*, 69(8):2723–2735, Nov 2007. doi: 10.1007/s11538-007-9244-7. URL https://doi.org/10.1007/s11538-007-9244-7.

Bernard Isaacs and Agnes T. Kennie. The set test as an aid to the detection of dementia in old people. *British Journal of Psychiatry*, 123(575):467–470, 1973.

Neil S Jacobson and Paula Truax. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. 1992.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

Munsif Ali Jatoi and Nidal Kamel. *Brain source localization using EEG signal analysis*. CRC Press, 2017.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011.

Jmol. Jmol: an open-source java viewer for chemical structures in 3d. 2019. URL http://www.jmol.org/.

Eric Jonas and Konrad Paul Kording. Could a neuroscientist understand a microprocessor? *PLoS computational biology*, 13(1):e1005268, 2017.

Roi Cohen Kadosh, Sonja Soskic, Teresa Iuculano, Ryota Kanai, and Vincent Walsh. Modulating neuronal activity produces specific and long-lasting changes in numerical competence. *Current Biology*, 20(22):2016–2020, 2010.

Roi Cohen Kadosh, Ann Dowker, Angela Heine, Liane Kaufmann, and Karin Kucian. Interventions for improving numerical abilities: Present and future. *Trends in neuroscience and education*, 2(2): 85–93, 2013.

Kosuke Kaida, Masaya Takahashi, Torbjörn Åkerstedt, Akinori Nakata, Yasumasa Otsuka, Takashi Haratani, and Kenji Fukasawa. Validation of the karolinska sleepiness scale against performance and eeg variables. *Clinical neurophysiology*, 117(7):1574–1581, 2006.

Lorraine V Kalia and Anthony E Lang. Parkinson disease in 2015: evolving basic, pathological and clinical concepts in pd. *Nature reviews Neurology*, 12(2):65, 2016.

Dinu Kaufmann, Sebastian Keller, and Volker Roth. Copula archetypal analysis. In Juergen Gall, Peter Gehler, and Bastian Leibe, editors, *Pattern Recognition*, pages 117–128. Springer International Publishing, 2015. ISBN 978-3-319-24947-6.

Sebastian Mathias Keller, Maxim Samarin, Mario Wieser, and Volker Roth. Deep archetypal analysis. In *German Conference on Pattern Recognition*, pages 171–185. Springer, 2019.

Sebastian Mathias Keller, Maxim Samarin, Fabricio Arend Torres, Mario Wieser, and Volker Roth. Learning extremal representations with deep archetypal analysis. *arXiv preprint arXiv:2002.00815*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. abs/1412.6980, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Peter Kirkpatrick and Clare Ellis. Chemical space. *Nature*, 432(823):1476–4687, 2004. doi: 10.1038/432823a.

BT Klassen, JG Hentz, HA Shill, E Driver-Dunckley, VGH Evidente, MN Sabbagh, CH Adler, and JN Caviness. Quantitative eeg as a predictive biomarker for parkinson disease dementia. *Neurology*, 77(2):118–124, 2011.

Wolfgang Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195, 1999.

M Könönen and JV Partanen. Blocking of eeg alpha activity during visual performance in healthy adults. a quantitative study. *Electroencephalography and clinical neurophysiology*, 87(3):164–166, 1993.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Alessandro Lanza, Serena Morigi, and Fiorella Sgallari. Convex image denoising via non-convex regularization. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 666–677. Springer, 2015.

RA Lawson, AJ Yarnall, F Johnston, GW Duncan, TK Khoo, D Collerton, JP Taylor, DJ Burn, and ICICLE-PD study group. Cognitive impairment in parkinson's disease: impact on quality of life of carers. *International Journal of Geriatric Psychiatry*, 32(12):1362–1370, 2017.

Rachael A Lawson, Alison J Yarnall, Gordon W Duncan, David P Breen, Tien K Khoo, Caroline H Williams-Gray, Roger A Barker, Daniel Collerton, John-Paul Taylor, David J Burn, et al. Cognitive decline and quality of life in incident parkinson's disease: the role of attention. *Parkinsonism & related disorders*, 27:47–53, 2016.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

D Lehmann. Multichannel topography of human alpha eeg fields. *Electroencephalography and clinical neurophysiology*, 31(5):439–449, 1971.

Dr Lehmann, H Ozaki, and I Pal. Eeg alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalography and clinical neurophysiology*, 67(3):271–288, 1987.

Abraham Lempel and Jacob Ziv. On the complexity of finite sequences. *IEEE Transactions on information theory*, 22(1):75–81, 1976.

Guocheng Li, Zheng Yan, and Jun Wang. A one-layer recurrent neural network for constrained nonsmooth invex optimization. *Neural Networks*, 50:79–89, 2014.

Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

Roderick JA Little. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

Ganqiang Liu, Joseph J Locascio, Jean-Christophe Corvol, Brendon Boot, Zhixiang Liao, Kara Page, Daly Franco, Kyle Burke, Iris E Jansen, Ana Trisini-Lipsanopoulos, et al. Prediction of cognition in parkinson's disease with a clinical–genetic score: a longitudinal analysis of nine cohorts. *The Lancet Neurology*, 16(8):620–629, 2017.

Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998. doi: 10.1109/AFGR.1998.670949. URL https://zenodo.org/record/3430156.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Pierre-Alexandre Mattei, Charles Bouveyron, and Pierre Latouche. Globally sparse probabilistic pca. In *Artificial Intelligence and Statistics*, pages 976–984, 2016.

HK Matthies and R Brödemann. Application of fast fourier transform in electroencephalography. *Biometrical Journal*, 23(8):789–794, 1981.

James Clerk Maxwell. Viii. a dynamical theory of the electromagnetic field. *Philosophical transactions of the Royal Society of London*, (155):459–512, 1865.

Rahul Mazumder, Jerome H. Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

Anna Åkerstedt Miley, Göran Kecklund, and Torbjörn Åkerstedt. Comparing two versions of the karolinska sleepiness scale (kss). *Sleep and biological rhythms*, 14(3):257–260, 2016.

Stanley L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, pages 528–529, 1953.

Francesca Miraglia, Fabrizio Vecchio, Placido Bramanti, and Paolo Maria Rossini. Eeg characteristics in "eyes-open" versus "eyes-closed" conditions: small-world network architecture in healthy aging and age-related brain degeneration. *Clinical Neurophysiology*, 127(2):1261–1268, 2016.

S.K. Mishra and G. Giorgi. *Invexity and Optimization*. Nonconvex Optimization and Its Applications. Springer Berlin Heidelberg, 2008. ISBN 9783540785620.

Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012.

Freek Nieuwhof, Bastiaan R Bloem, Miriam F Reelick, Esther Aarts, Inbal Maidan, Anat Mirelman, Jeffrey M Hausdorff, Ivan Toni, and Rick C Helmich. Impaired dual tasking in parkinson's disease is associated with reduced focusing of cortico-striatal activity. *Brain*, 140(5):1384–1398, 2017.

Hugh Nolan, Robert Whelan, and Richard B Reilly. Faster: fully automated statistical thresholding for eeg artifact rejection. *Journal of neuroscience methods*, 192(1):152–162, 2010.

Ulla M. Norberg, J. M. V. Rayner, and Michael James Lighthill. Ecological morphology and flight in bats (mammalia; chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 316(1179), 1987. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1987.0030.

Paul L Nunez, Ramesh Srinivasan, et al. *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006.

Jose A Obeso, Maria C Rodriguez-Oroz, Jose L Lanciego, and Manuel Rodriguez Diaz. How does parkinson's disease begin? the role of compensatory mechanisms. *Trends in neurosciences*, 27(3): 125–127, 2004.

Kim TE Olde Dubbelink, Arjan Hillebrand, Diederick Stoffers, Jan Berend Deijen, Jos WR Twisk, Cornelis J Stam, and Henk W Berendse. Disrupted brain network topology in parkinson's disease: a longitudinal magnetoencephalography study. *Brain*, 137(1):197–207, 2014.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

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.

Daniel S Peterson, Brett W Fling, Martina Mancini, Rajal G Cohen, John G Nutt, and Fay B Horak. Dual-task interference and brain structural connectivity in people with parkinson's disease who freeze. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(7):786–792, 2015.

Sandhya Prabhakaran, Sudhir Raman, Julia E. Vogt, and Volker Roth. Automatic model selection in archetype analysis. In Axel Pinz, Thomas Pock, Horst Bischof, and Franz Leberl, editors, *Pattern Recognition*, pages 458–467. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-32717-9.

Kristopher J Preacher, Derek D Rucker, Robert C MacCallum, and W Alan Nicewander. Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological methods*, 10(2):178, 2005.

Tamara Pringsheim, Nathalie Jette, Alexandra Frolkis, and Thomas DL Steeves. The prevalence of parkinson's disease: A systematic review and meta-analysis. *Movement disorders*, 29(13):1583–1590, 2014.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

Ralph M Reitan. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and motor skills*, 8(3):271–276, 1958.

Mélanie Rey and Volker Roth. Meta-Gaussian information bottleneck. In *Advances in Neural Information Processing Systems–NIPS 25*, 2012.

Mélanie Rey, Thomas Fuchs, and Volker Roth. Sparse meta-Gaussian information bottleneck. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 910–918, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Conference on Machine Learning*, 32(2):1278–1286, 22–24 Jun 2014.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d. URL https://pubs.acs.org/doi/10.1021/ci300415d. PMID: 23088335.

Robert Schuetz, Nicola Zamboni, Mattia Zampieri, Matthias Heinemann, and Uwe Sauer. Multidimensional optimality of microbial metabolism. *Science (New York, N.Y.)*, 336:601–4, 05 2012. doi: 10.1126/science.1216882.

Sohan Seth and Manuel J. A. Eugster. Probabilistic archetypal analysis. *Machine Learning*, 102(1):85–113, Jan 2016. doi: 10.1007/s10994-015-5498-8. URL https://doi.org/10.1007/s10994-015-5498-8.

O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon. Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science*, 336(6085): 1157–1160, 2012. doi: 10.1126/science.1217405. URL http://science.sciencemag.org/content/336/6085/1157.

Robert Sneddon. The tsallis entropy of natural information. *Physica A: Statistical Mechanics and its Applications*, 386(1):101–118, 2007.

Robert Sneddon, William Rodman Shankle, Junko Hara, Anthony Rodriquez, Donald Hoffman, and Utpal Saha. Eeg detection of early alzheimer's disease using psychophysical tasks. *Clinical EEG and neuroscience*, 36(3):141–150, 2005.

Evgeniy N Sokolov. Higher nervous functions: The orienting reflex. *Annual review of physiology*, 25 (1):545–580, 1963.

Cornelis J Stam. Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15 (10):683–695, 2014.

C. Steinbeck, Y. Q. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003.

R.E. Steuer. *Multiple Criteria Optimization: Theory, Computation and Application*. John Wiley & Sons, 1986.

Emily Stone and Adele Cutler. Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D: Nonlinear Phenomena*, 96(1-4):110–131, September 1996. doi: 10.1016/0167-2789(96)00016-4. URL https://doi.org/10.1016/0167-2789(96)00016-4.

Avichai Tendler, Avraham Mayo, and Uri Alon. Evolutionary tradeoffs, pareto optimality and the morphology of ammonite shells. *BMC Systems Biology*, 9(1), 2015. doi: 10.1186/s12918-015-0149-z. URL https://doi.org/10.1186/s12918-015-0149-z.

U Tewes and Wechsler D. Hamburg-wechsler-intelligenztest für erwachsene–revision 1991 (hawie-r). 2. korr. *Bern: Huber*, 1991.

L. L. Thurstone and T. G. Thurstone. Science research associates tests of primary mental abilities, 1947. URL https://doi.org/10.1037%2Ft02715-000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Ryan J Tibshirani. A general framework for fast stagewise algorithms. *Journal of Machine Learning Research*, 16:2543–2588, 2015.

I. Tinoco. *Physical Chemistry: Principles and Applications in Biological Sciences*. Number S. 229-313 in Physical Chemistry: Principles and Applications in Biological Sciences. Prentice Hall, 2002. ISBN 9780130959430.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000a.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000b.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000c.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

Fabrizio Vecchio, Francesca Miraglia, Francesco Iberite, Giordano Lacidogna, Valeria Guglielmi, Camillo Marra, Patrizio Pasqualetti, Francesco Danilo Tiziano, and Paolo Maria Rossini. Sustainable method for alzheimer dementia prediction in mild cognitive impairment: Electroencephalographic connectivity and graph theory combined with apolipoprotein e. *Annals of neurology*, 84(2):302–314, 2018.

Ricardo Visini, Josep Arus-Pous, Mahendra Awale, and Jean-Louis Reymond. Virtual exploration of the ring systems chemical universe. *Journal of Chemical Information and Modeling*, 57(11): 2707–2718, 2017. doi: 10.1021/acs.jcim.7b00457. URL https://doi.org/10.1021/acs.jcim.7b00457. PMID: 29019686.

Aleksander Wieczorek and Volker Roth. On the difference between the information bottleneck and the deep information bottleneck. *Entropy*, 22(2):131, 2020.

Aleksander Wieczorek, Mario Wieser, Damian Murezzan, and Volker Roth. Learning Sparse Latent Representations with the Deep Copula Information Bottleneck. *International Conference on Learning Representations (ICLR)*, 2018.

Katharina Wulff, Silvia Gatti, Joseph G Wettstein, and Russell G Foster. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, 11(8): 589–599, 2010.

Daan Wynen, Cordelia Schmid, and Julien Mairal. Unsupervised learning of artistic styles with archetypal style analysis. In *Advances in Neural Information Processing Systems*, pages 6584–6593, 2018.

Rezzak Yilmaz, Franziska Hopfner, Thilo van Eimeren, and Daniela Berg. Biomarkers of parkinson's disease: 20 years later. *Journal of Neural Transmission*, 126(7):803–813, 2019.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

X-S Zhang, Rob J Roy, and Erik W Jensen. Eeg complexity as a measure of depth of anesthesia for patients. *IEEE transactions on biomedical engineering*, 48(12):1424–1433, 2001.

P Zimmermann and B Fimm. Testbatterie zur aufmerksamkeitsprüfung (tap): Handbuch teil 2 (version 2.1). *Herzogenrath: Psychologische Testsysteme Vera Fimm*, 2007.

Ronan Zimmermann, Ute Gschwandtner, Florian Hatz, Christian Schindler, Habib Bousleiman, Shaheen Ahmed, Martin Hardmeier, Antonia Meyer, Pasquale Calabrese, and Peter Fuhr. Correlation of eeg slowing with cognitive domains in nondemented patients with parkinson's disease. *Dementia and Geriatric Cognitive Disorders*, 39(3-4):207–214, 2015.