UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Brain Decoding for Brain Mapping

Definition, Heuristic Quantification, and Improvement of Interpretability in Group MEG Decoding

Seyed Mostafa Kia
International Doctorate School in Information
and Communication Technologies,
Università degli Studi di Trento

Advisor:
Prof. Andrea Passerini
Università degli Studi di Trento

April 2017

Abstract

In the last century, a huge multi–disciplinary scientific endeavor is devoted to answer the historical questions in understanding the brain functions. Among the statistical methods used for this purpose, brain decoding provides a tool to predict the mental state of a human subject based on the recorded brain signal. Brain decoding is widely applied in the contexts of brain–computer interfacing, medical diagnosis, and multivariate hypothesis testing on neuroimaging data. In the latest case, linear classifiers are generally employed to discriminate between experimental conditions. Then, the derived weights are visualized in the form of brain maps to further study the spatio–temporal patterns of the underlying neurophysiological activity. It is well known that the brain maps derived from weights of linear classifiers are hard to interpret because of high correlations between predictors, low signal–to–noise ratio, across–subject variability, and the high dimensionality of the neuroimaging data. Therefore, improving the interpretability of brain decoding approaches is of primary interest in many neuroimaging studies. Despite extensive studies of this type, at present, there is no formal definition for interpretability of multivariate brain maps. As a consequence, there is no quantitative measure for evaluating the interpretability of different brain decoding methods. In this thesis, as the primary contribution, we propose a theoretical definition of interpretability in linear brain decoding; we show that the interpretability of multivariate brain maps can be decomposed into their reproducibility and representativeness. As an application of the proposed definition, we exemplify a heuristic for approximating the interpretability in multivariate analysis of evoked magnetoencephalography (MEG) responses. We propose to combine the approximated interpretability and the generalization performance of the model into a new multi–objective criterion for model selection. Our results, for the simulated and real MEG *data, show that optimizing the hyper–parameters of the regularized linear classifier based on the proposed criterion results in more informative multivariate brain maps. More importantly, the presented definition provides the theoretical background for quantitative evaluation of interpretability, and hence, facilitates the development of more effective brain decoding algorithms in the future. As the secondary contribution, we present an application of multi–task joint feature learning for group–level multivariate pattern recovery in single–trial MEG decoding. The proposed method allows for recovering sparse yet consistent patterns across different subjects, and therefore enhances the interpretability of the decoding model. We evaluated the performance of the multi–task joint feature learning in terms of generalization, reproducibility, and quality of pattern recovery against traditional single–subject and pooling approaches on both simulated and real MEG datasets. Our experimental results demonstrate that the multi–task joint feature learning framework is capable of recovering meaningful patterns of varying spatio–temporally distributed brain activity across individuals while still maintaining excellent generalization performance. The presented methodology facilitates the application of brain decoding for characterizing the fine–level distinctive patterns of brain activity in group–level inference on neuroimaging data.*

Keywords

Brain Decoding; Brain Mapping; Neuroimaging; Machine Learning; Magnetoencephalography; Interpretability; Reproducibility.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my adviser Andrea Passerini for his patience, constructive feedback, and support of my research. Besides my advisor, I would like to thank the rest of my thesis committee: Lauri Parkkonen, Alexandre Gramfort, and Lorenzo Bruzzone for their insightful comments that improved this work significantly.

My special thanks also goes to Nathan Weisz for his ubiquitous motivating attitude and valuable scientific supports. I owe a debt of gratitude to Nicu Sebe and Paolo Giorgini, whose supports provided me the opportunity to continue my PhD study. I wish to thank James Haxby whose lecture on "Neural Decoding" ignited the main motivation behind my research activity.

I would like to thank my other collaborators and friends, Paolo Avesani, Emanuele Olivetti, Sandro Vega Pons, Fabian Pedregosa, and Anna Blumenthal for stimulating discussions, criticisms, and kind feedback on the content of this thesis.

Last but not the least, I would like to thank my family: my beloved wife Nastaran, and my beloved parents Masood and Mina for their sympathetic ear and spiritual supports throughout my life. I would like to dedicate this thesis to them.

Contents

List of Tables

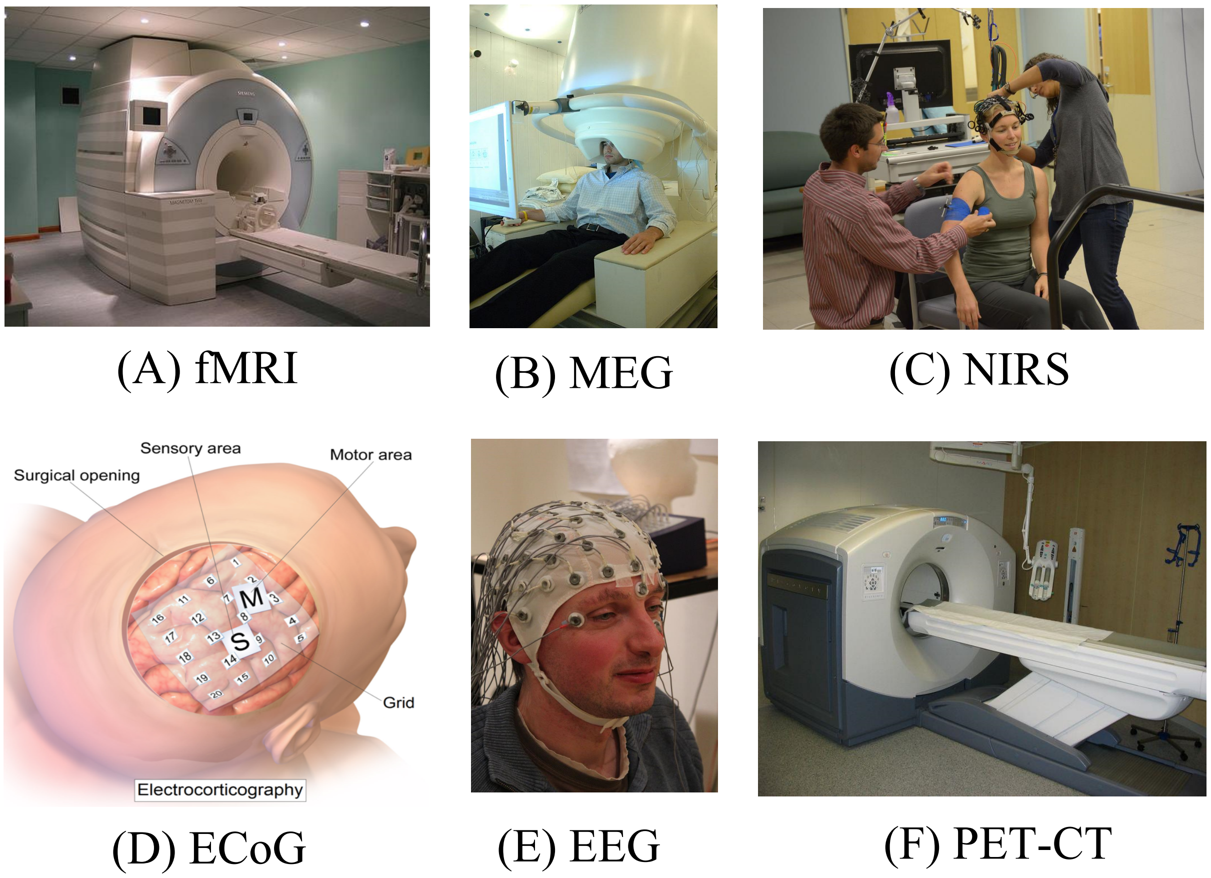
List of Figures

Chapter 1

Introduction

Understanding the nature and function of *brain* is one of the main questions that has evoked human curiosity all along the history. Ancient Greek philosophers envisaged different functions for the brain from 500 B.C.E to 200 C.E, ranging from it is being the cooling agent of body heat to the seat of a rational soul and center of sensation and understanding [42]. Nowadays, cognitive science tries to incorporate research areas that are concerned with neurophysiological and behavioral understanding of the brain, e.g., neuroscience and psychology, with variety of other research fields, such as computer science, physics, and statistics, to provide a better insight into the structure and function of the brain. As the field matures, techniques are being adopted from other areas of computational science in order to accelerate research in cognitive science.

Neuroimaging techniques (see Figure 1.1), also called brain imaging techniques, such as structural and functional Magnetic Resonance Imaging (s/fMRI) [54], Electro/Magnetoencephalography (E/MEG) [20, 37], Electrocorticography (ECoG) [103], Positron Emission Tomography (PET) [13], and Near–Infrared Spectroscopy (NIRS) [25], have become essential tools for either invasive or non–invasive imaging of the structure and function of the brain. Structural brain imaging is more concerned about the diag-


(A) fMRI

(B) MEG

(C) NIRS

(D) ECoG

(E) EEG

(F) PET-CT

Figure 1.1: Neuroimaging techniques. **(A)** Siemens MAGNETOM Trio device for structural and functional brain imaging. **(B)** CTF–275 MEG scanner for recording magnetic fields produced by electrical currents in the brain. **(C)** User preparation for a NIRS recording. **(D)** A grid of ECoG sensors implanted on sensory and motor areas. **(E)** Configuration of EEG sensors on the head for scanning electrical brain activity. **(F)** Discovery D600 PET–CT system for positron emission tomography.

nosis of large–scale brain diseases resulting from the abnormality in brain tissues [10, 179], e.g., tumors or brain injuries. On the other side, there are a variety of applications for functional brain imaging, ranging from the finer–level medical diagnosis to brain–computer interfaces and understanding brain's function.

In last three decades, the clinical application of functional brain imaging in psychiatry has impressively broadened [28, 56]. Functional brain imaging techniques are used to investigate the neural correlates of various mental disorders in order to identify biomarkers for them. These biomarkers then can be employed to investigate the effect of behavioral therapies and drug treatments. For example, resting–state functional connectivity derived from patients' fMRI are used for early identification of Alzheimer's disease and presurgical planning [169]. MEG and EEG recordings are also employed for finding the seizure onset zone in presurgical evaluation of epilepsy patients [109].

Brain–computer interface [207] (BCI) is a system that provides a real–time communication channel between the brain and an external machine. The application of neuroimaging in BCI is more focused on measuring electrical activity of brain invasively by means of intracranial implants such as ECoG [120], or non–invasively by means of EEG devices. Then an algorithm is used for online translation of the recorded brain activity to machine instructions. This technology has applications in verbal communication [50], controlling devices [209], affect recognition [1–3, 110], multi–media content retrival [59], and locomotion [200] especially for individuals with severe motor disabilities by brainstem stroke or neuro–muscular diseases such as amyotrophic lateral sclerosis.

In cognitive neuroscience [57], researchers use the recorded neuroimaging data to understand the relationship between brain activity and specific cognitive functions, i.e., to answer three key questions of *where*, *when* and *how* [1] a brain region contributes to a particular cognitive process. To do this, depending on the question of interest, an experimental protocol is designed to evoke or induce certain brain activity in human or non–human participants, while simultaneously recording neural correlates by means of functional neuroimaging devices. Then statistical analysis techniques are

[1] Here the answer to "how" question refers to finding the connection between a specific cognitive function and characteristics of the recorded neural correlates.

employed to justify the initial hypotheses about the three key questions. Here is an example of a scientific question in cognitive neuroscience [140]:

> “We here wanted to reveal whether neural excitability of the auditory cortex putatively reflected in local alpha–band power is modulated already prior to speech onset, and which brain regions may mediate such a top–down preparatory response.”

in which auditory cortex, modulation of alpha–band power, and occurrence of this modulation prior to speech onset stand for hypothesized answers to *where*, *how*, and *when* questions, respectively.

In this thesis, we are interested in the application of functional neuroimaging in understanding brain function. More specifically, we are interested in improving the interpretability of multivariate hypothesis testing approaches in order to infer more reliable, reproducible, and plausible answers to the main questions in cognitive neuroscience. Of course, the resulting methodology is also applicable to the medical diagnosis domain, but our experimental setups and discussion are more focused on the applications in confirmatory and exploratory data analysis techniques in cognitive neuroscience.

There are two schools of thought in statistical analysis for inference on neuroimaging data [32]: 1) classical statistical testing, and 2) statistical learning theory. Classical statistical testing is an in–sample generalization technique based on null–hypothesis falsification, in which, generally, a set of univariate tests, e.g., t–tests, are independently applied to each variable of interest. On the other hand, statistical learning theory is a multivariate approach that is more concerned with out–of–sample generalization. While both techniques are successfully applied for inference on neuroimaging data, they capture partially different aspects of the underlying neurophysiological activity [32].

Region–of–interest (ROI) analysis is one of the most popular methods in classical inference on neuroimaging data [71, 160]. It is typically based on the mean activity analysis, using e.g., ANOVA, on a pre–specified ROIs. The pre–specified ROIs are generally decided using prior knowledge on the studied cognitive process, and the mean activity within the ROIs are tested in different experimental conditions. Despite the popularity and simplicity of the ROI analysis method, the prerequisite for pre–selecting the ROIs limits its application especially in exploratory analysis of neuroimaging data where little is known about the brain areas involved in a cognitive function. Addressing this limitation, classical inference evolved to the new generation of exploratory whole–brain analysis such as mass–univariate hypothesis testing [64].

Mass–univariate analysis performs a large number of univariate tests on each variable, e.g., each voxel, independently. It can be employed for hypothesis testing in whole–brain exploratory analysis without the need for prior variable selection. However, it requires a procedure to handle the multiple–comparison problem (MCP) [60]. There are various methods for multiple–comparison correction based on the strong or weak control of family–wise error rate (FWER) [203,213] or false discovery rate (FDR) [16] control. Being essential for the validity of results, on the down side this correction reduces the power of statistical analysis with the increase in the number of univariate tests [64].

In statistical learning approaches, also known as brain decoding and multivariate pattern analysis (MVPA) in the literature [86, 99], a model is trained to learn the relation between the independent variables, i.e., neuroimaging data, and the dependent variables, i.e., experimental conditions. The training is performed in the framework of statistical learning theory [80]. The performance of the model is evaluated on a test set, which is different from the initial training set. If the performance is significantly above the chance level, it can be concluded that a meaningful relation exists between the recorded neural signals and the cognitive task. The statistical learning approach can possibly provide a multivariate alternative for classical univariate hypothesis testing methods. The multivariate nature of this method yields higher sensitivity to the distributed patterns of brain activities [149] and provides the possibility of capitalizing the complex interactions among the parameters of interest. Further, by employing proper validation strategies, it resolves the multiple testing problem of mass–univariate approaches [98]. In this thesis, we use *brain decoding* to refer to the application of the statistical learning theory in the neuroimaging context.

Due to the high dimensionality and limited number of samples typically associated with neuroimaging data [41, 114], linear classifiers are generally used to assess the relation between spatio–temporal brain measurements and cognitive tasks [22, 118, 157]. This assessment is performed by solving an optimization problem that minimizes a loss function by learning weights associated with each independent variable. These learned weights can then be visualized in the form of a *brain map*, in which the engagement of different brain areas in a cognitive task is illustrated. In fact, brain mapping via brain decoding can be viewed as a *pattern recovery* problem, where the goal is to recover spatio–temporal patterns of the discriminative brain activity involved in the cognitive processing of external stimuli. If successful, brain maps created by means of brain decoding can provide a comprehensive explanation regarding the nature of neural representations and brain states, and may be more informative for cognitive science than a merely decoding accuracy measure [154]. Currently, brain decoding is the gold standard in multivariate analysis of functional magnetic resonance images (fMRI) [41,86,135,149] and magnetoencephalography/electroencephalography(MEG/EEG) data [3, 34, 36, 93, 156, 167, 199]. However a number of challenges still remain, particularly regarding the interpretability of weights of classifiers, especially in group studies of neuroimaging data.

A classifier or a regression model that is trained in the statistical learning framework only answers the question of *what* is the most likely label of a given unseen sample [12]. This fact is generally known as the knowledge extraction gap [198] in the machine learning context. Thus far, much effort has been devoted to filling this gap of linear and non–linear data modeling methods in different areas such as computer vision [11], signal processing [137], chemometrics [216], bioinformatics [72], and neuroinformatics [83]. In the context of neuroimaging, this gap is generally known as the interpretation problem [88, 142, 172]. Therefore, improving the interpretability of linear brain decoding and the associated brain maps is a topic of interest in many neuroimaging studies [178]. In spite of the extensive efforts to improve the interpretability of brain decoding, there is still no formal definition for the interpretability of brain decoding. Therefore, the interpretability of different brain decoding methods is evaluated either qualitatively or indirectly by means of an intermediate property.

Group–level analyses of neuroimaging data are extremely important, as they allow for results to be generalized to new individuals. In statistical learning, an ideal group–level approach should be able to recover both structural and functional similarities and dissimilarities across different individuals. These similarities and dissimilarities generally occur at both a coarse and fine level in space and time, and can provide valuable spatio–temporal information about both the underlying macro and micro–structures of the cognitive function in question. For example, visual stimuli in general evoke a coarsely similar effect in early visual brain areas across different subjects, but the response to different types or categories of visual stimuli can differ from subject to subject at the finer level (see Ref. [87] for more examples). This across–subject functional variability makes group–

level inference on neuroimaging data challenging, particularly since there is also substantial across–subject variability in the brain structure (e.g., the different size and shape of brains) [129,164,165,180,181]. This problem is even more pronounced when one takes into account the difference in the spatio–temporal structure of noise that commonly occurs due to different external and internal sources, or manual preprocessing errors. These variations not only negatively affect the generalization performance of brain decoding, but they also make post–hoc interpretation of the derived brain maps more challenging, due to concerns about lack of reproducibility and plausibility. For these reasons, it is crucial to explore more effective decoding methods that are capable of recovering structural and functional similarities and dissimilarities in a group–level analysis of neuroimaging data.

With the aim of filling these gaps, the contribution of this thesis is two–fold:

1. A theoretical definition for the interpretability of linear brain decoding models is presented. The definition is based on cosine proximity between the estimated and true solutions of brain decoding in the parameter space. Furthermore, it is shown that the interpretability can be decomposed into the reproducibility and representativeness. As a proof of concept, a practical heuristic based on event–related fields is exemplified to quantify the interpretability of brain maps. Furthermore, the combination of interpretability and performance of brain decoding is proposed as a new Pareto optimal multi–objective criterion for model selection.

2. An application of multi–task joint feature learning [9] for accurate spatio–temporal pattern recovery at the group–level decoding of MEG data is presented. In the proposed framework, the data of each subject

is considered as a task in the multi–task learning framework, where only one decoding model is simultaneously trained over all subjects. Further, $\ell_{2,1}$ regularization [124] is employed to learn sparse patterns consistently across different subjects, i.e., to jointly learn the features across different subjects.

Regarding my first contribution, the presented definition for interpretability of linear brain decoding models provides a concrete framework for a previously abstract concept and establishes theoretical background to explain an ambiguous phenomenon in the brain decoding context. The experimental results on MEG data show that accounting for the approximated measure of interpretability has a positive effect on the human interpretation of brain decoding models. Furthermore, the proposed decomposition of the interpretability of brain maps into their reproducibility and representativeness explains the relationship between the influential cooperative factors in the interpretability of brain decoding models and highlights the possibility of indirect and partial evaluation of interpretability by measuring these effective factors. The experimental results on single–subject MEG decoding showed that adopting the new proposed criterion for optimizing the hyper–parameters of brain decoding models is an important step toward reliable visualization of learned models from neuroimaging data. Furthermore, these findings provide a step toward direct evaluation of interpretability of the currently proposed regularization strategies. Such an evaluation can highlight the advantages and disadvantages of applying different regularization strategies on different data types and facilitates the choice of appropriate regularizer for a certain application.

Regarding my second contribution, multi–task joint feature learning facilitates consistent sparse pattern recovery across individual subjects while at the same time preserving idiosyncratic structural and functional properties within each individual. By taking into account the inter–subject spatio–temporal similarities and dissimilarities of brain activity, multi–task joint feature learning provides higher interpretability for multivariate brain maps at the group–level. To my knowledge, this is the first time one uses multi–task joint feature learning in the context of group–level MEG decoding. Considering the fact that only EEG and MEG can non–invasively record brain activity at a high temporal resolution [75, 78], the proposed approach provides the possibility for recovering temporal brain dynamics within the millisecond time scale, a crucial task if we aim to understand the dynamics of human brain function [77, 79]. On the other hand, multi–task joint feature learning provides the infrastructure for combining structured regularization with stability selection in group–level multivariate analysis. While $\ell_{2,1}$ penalty combines ℓ_2 and ℓ_1 norms to enforce group sparsity, its integration with simultaneous optimization in multi–task learning also offers a variant of stability selection across a group of subjects.

The rest of this thesis is organized in the following 4 chapters:

1. In order to provide the basic background for the general audience, Chapter 2 reviews the basic concepts and terminologies that are used to develop the contributions of this thesis. To this end, the basic terminology to describe the structure and function of human brain is firstly introduced. Then the principles of brain recording and analysis using MEG data are briefly reviewed. At the end, I review the concepts behind hypothesis testing on neuroimaging data, ranging from the classical hypothesis testing to the statistical learning theory.

2. Chapter 3 presents a novel definition for the interpretability of linear brain decoding models [105, 108]. It is shown that the interpretability of multivariate brain maps can be decomposed into their reproducibility and representativeness. Then, a heuristic for approximating the interpretability in multivariate analysis of evoked MEG responses

is exemplified. Finally, I propose to combine the approximated interpretability and the generalization performance of brain decoding into a new multi–objective criterion for model selection. The results demonstrate the importance of including interpretability in the model selection for deriving more meaningful brain maps.

3. In Chapter 4, an application of multi–task joint feature learning for group–level multivariate pattern recovery in single–trial MEG decoding is proposed [106, 107]. The proposed method allows for recovering sparse yet consistent patterns across different subjects, and therefore enhances the interpretability of the decoding model in group–level analysis.

4. Finally, Chapter 5 summarizes the lessons that have been learned and states possible future directions.

Chapter 2

Background

The aim of this chapter is to provide background information about brain, magnetoencephalography (MEG), hypothesis testing, and machine learning for the readers. The basic concepts introduced in this chapter provide the formal and conceptual ingredients for understanding our contributions in the following chapters. To this end, we first introduce the basic terminology that is used to describe the brain structure. Second, we briefly describe the mechanisms and characteristics of extracranial magnetic field recording using an MEG device. Third, the principles of classic statistical hypothesis testing on the neuroimaging data are reviewed. We finalize this chapter by introducing the basic concepts in statistical learning theory.

2.1 Brain: from Neurons to the Cerebral Cortex

The *brain* is an organ contained in the skull of vertebrates and head of most invertebrate animals; brain serves as the coordinating center of the nervous system. The brain tissue is composed of two classes of cells: 1) neurons, and 2) glial cells. Glial cells are involved in structural and metabolic support. Neurons are the basic elements of the nervous system that process and transmit information via electro–chemical processes [100]. These signals are transmitted from one neuron to another via specialized inter–neuron connections called *synapses*. Synapses are key functional elements of the brain as they form modifiable communication channels between neurons [174]. This modifiability provides the possibility of changing the strength or patterns of neuro–electrical signals. This key feature provides the infrastructure for crucial brain functions such as learning and memory. The web between neurons form densely connected networks. To understand better the structural complexity of the neural networks, it is worthwhile to emphasize that the brain has around 10^{11} neurons each of which with up to $\sim 10^4$ connections.

A typical neuron is composed of a cell body or soma, dendrites, and an axon (see Figure 2.1). The electrical signals are received by the dendrites, integrated at the soma, and transmitted to the synaptic terminals via the axon. The signals that are transmitted along the axon are called *action potentials* and the received signals at dendrites are called *post–synaptic potentials*. Neurons are classified to several categories based on their structural properties. Purkinje neurons, Pyramidal neurons, Granule neurons, and Spindle neurons are examples of neuron types in the brain.

Axons are generally wrapped in a fatty insulating cover called myelin. Myelin is white, thus, the area of the brain that includes axons appears white, hence, it is known as *white matter* [see Figure 2.2(A)]. In contrast the area that contains the cell bodies of neurons and dendrites appears darker and it is called the *gray matter*. The gray matter forms the human cerebral cortex which is divided into left and right hemispheres along the sagittal plane. The types of neurons in the gray matter divide the cerebral cortex into six layers [see Figure 2.2(B)]: 1) molecular layer, 2) external granular layer, 3) external pyramidal layer, 4) internal granular layer, 5) internal pyramidal layer, and 6) polymorphic layer. The human cerebral cortex is coarsely segmented into four lobes in each hemisphere [see Figure 2.2(C)]:

1. *Occipital Lobe:* The occipital lobe contains primary visual cortex (also

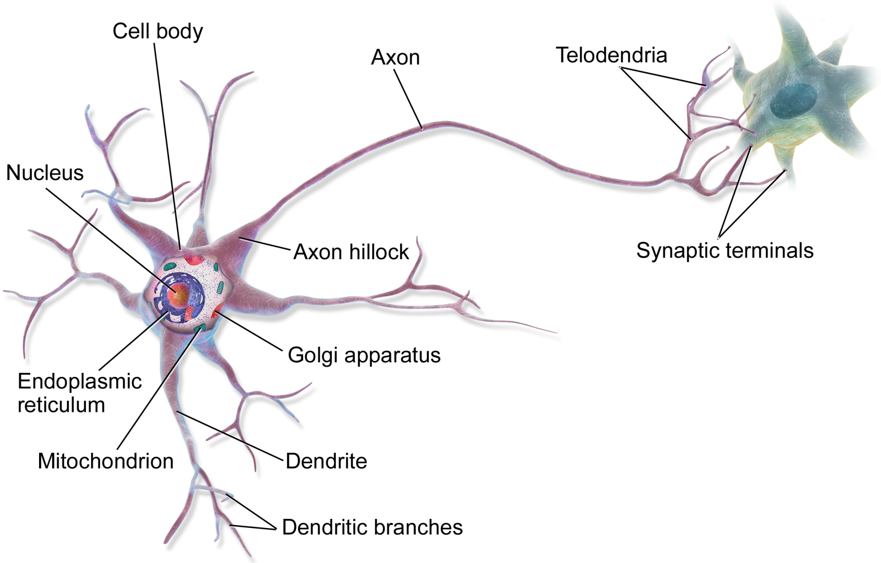


Figure 2.1: The structure of a typical neuron [206]. The electrical signals are received by the dendrites, processed at the soma, and transmitted to the synaptic terminals via the axon.

called as V1 area or striate cortex) which processes the low–level visual features such as local orientation and spatial frequency. Primary visual cortex is followed up by the ventral stream (V2 and V4 areas), and the dorsal stream (V3, and V5 areas). The ventral stream processes important information regard the identification of stimuli while the dorsal stream focuses more on the spatial aspects of motor actions in response to visual stimuli.

2. *Parietal Lobe:* The parietal lobe plays important roles in integrating sensory information, e.g., visuo–spatial processing, and language.

3. *Temporal Lobe:* The temporal lobe consists several sub–areas which are involved in associating meanings to the sensory inputs such as visual and auditory stimuli, language comprehension, and emotion processing.

4. *Frontal Lobe:* The frontal lobe is responsible for voluntary movement

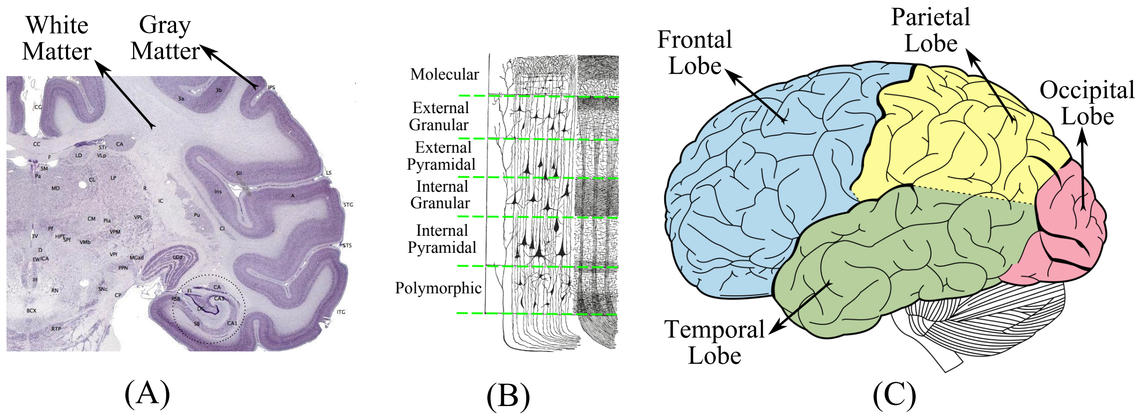


Figure 2.2: **(A)** The organization of the white and gray matter in the human brain. **(B)** The six layers of the gray matter. **(C)** The division of human cerebral cortex into occipital, parietal, temporal, and frontal lobes [204].

and performs some high–level cognitive functions such as attention, short–term memory, emotions, and planning.

2.2 Magnetoencephalography (MEG)

2.2.1 History and Mechanisms

Nowadays, neuroimaging methods that allow to explore the brain functions within the millisecond time scale provide exceptional opportunity to unveil temproal patterns of neural activity [68, 75, 77–79, 150]. Up to now, only electroencephalogram (EEG) and magnetoencephalogram (MEG) can non–invasively record neural activity at such a high temporal resolution. These methods allow for real–time tracking of brain activation sequences during sensory processing, motor planning and action, cognition, language perception and production, social interaction, and various brain disorders [73, 74, 76, 188].

According to Maxwell's equations, the post–synaptic electrical current resulting from synaptic transmission produces a magnetic field. Therefore the magnitude of the resulting magnetic field can be used as an indicator for the activation of population of neurons. The weak neuro–magnetic fields outside the human scalp were first measured by David Cohen in 1968 [37] using a copper induction coil. The weakness of the cortical magnetic fields, which are on the order of 10-10^3 femtotesla (fT), compared to the environmental noise led to the invention of superconducting quantum interference device (SQUID) [222]. Cohen used a heavy magnetically shielded room and a single SQUID detector to show that MEG can capture the brain's alpha rhythms similarly as EEG [38]. Currently, MEG devices contain around 300 SQUIDs arranged in a helmet–shaped array that cover the whole human scalp [see Figure 1.1(B)].

Measuring the magnetic fields around the scalp provides an exceptional technique to investigate the cognitive function of different brain regions especially within cortical sulci that are barely observable even with invasive intracranial brain recording techniques. The majority of magnetic field measured by SQUID are produced by the parallel pyramidal cells that are perpendicular to the cortical surface. Their electrical current flow is directed perpendicular to the cortical sheet of the gray matter. Thus magnetic fields resulting from the synchronized tangential neural activity across a population of pyramidal neurons can be sensed via SQUIDs outside the head (see Figure 2.3).

In modern MEG devices, the temporal and spatial sampling frequency is designed based on the multidimensional generalization of Nyqvist criterion to avoid any spatio–temporal aliasing [6]. The temporal and spatial sampling rate are generally $\sim$ 1000 and $\sim$ 300, respectively. The $\sim$ 300 spatial sampling rate stands for $\sim$ 300 MEG sensors which could be different from one device to another. For example CTF MEG [1] and Electa

[1] See `http://www.ctfmeg.com/`.

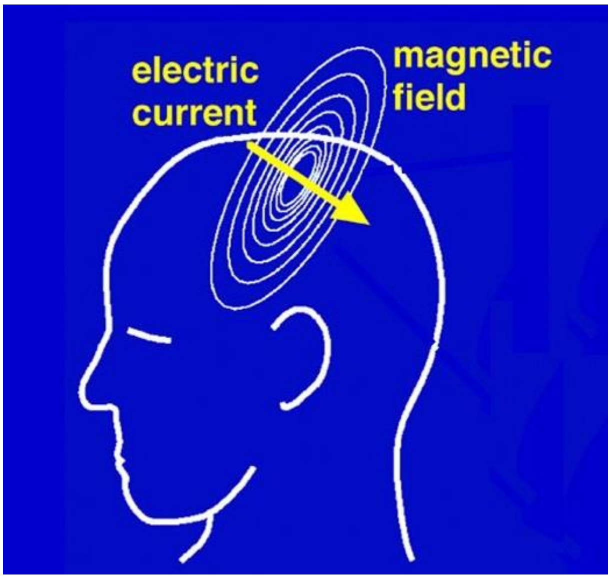


Figure 2.3: The radial magnetic fields resulting from the tangential electrical currents can be measured outside the scalp [205].

Neuromag [2] systems have 275 and 306 sensors, respectively. The MEG sensors, depending on the type of the corresponding flux transformer, i.e., a device that transforms the magnetic field to SQUID, are categorized into three main types [68]: 1) magnetometer, 2) axial gradiometer, and 3) planar gradiometer. Magnetometer sensors, with a single coil, measure only one component of the magnetic field [see Figure 2.4(A)]. Axial gradiometers consist of two vertically connected coils with opposite directions, thus, these sensors are insensitive to homogeneous fields and therefore to most of environmental noise [see Figure 2.4(B)]. Planar gradiometers consist of two twisted magnetometers placed next to each other and measure the gradient of the magnetic field in a plane roughly tangential to the head surface [see Figure 2.4(C)].

Even though the effect of environmental noise can be alleviated to some degree with astute design of flux transformers, the recorded MEG signal is often contaminated with artifacts. Eye blinks, eye movements, cardiac

[2] See `https://www.elekta.com/diagnostic-solutions/elekta-neuromag-triux.html`.

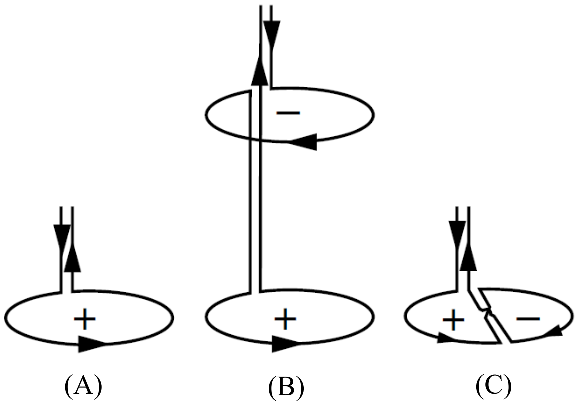


Figure 2.4: Types of flux transformers in MEG sensors [68]: (A) Magnetometer, (B) Axial gradiometer, (C) Planar gradiometer.

activity, and muscular activity are examples of biological artifacts in MEG signal. These artifacts can be partially rejected using band–pass frequency filtering or using blind–source separation methods such as independent component analysis (ICA) [94].

2.2.2 Data Analysis

Time–Domain Analysis

One of the most common methods for analyzing the EEG/MEG signals is to compute the average event–related potential/fields (ERP/ERF) [71]. ERP/ERFs are suitable for investigating the neuronal correlates of specific transient external stimuli [125]. In addition, abnormality in ERP/ERF components can be used as a clinical biomarker for diagnosing neurological diseases such as Alzheimer's [27], Parkinson's [163], and multiple sclerosis [159].

The main idea behind computing the ERP/ERF is to increase the signal–to–noise ratio (SNR). Due to the internal (such as background brain activity and other biological interference) and external (electromagnetic interference by the light sources, electricity, and peripheral devices) noise contaminations, the single trials of EEG/MEG data suffer from low SNR. One simple solution to address this problem is to compure ERP/ERF by averaging many trials in order to cancel out the random uncorrelated noise components [171]. The averaging operation is based on three main assumptions: 1) the noise components are uncorrelated with the signal of interest; 2) the signal of interest is time–locked, i.e., it has a fixed latency with respect to the stimulus onset. This type of time–locked response is also called as the *evoked response* in the literature; 3) the noise components have a zero–mean Gaussian distribution with variance of σ^2. This approach is generally known as a time–locked analysis and is available within common EEG/MEG data analysis toolboxes such as Filedtrip [153], MNE–Python [61], and EEGLAB [47].

One possible approach to interpret ERP/ERF responses is to categorize them based on their amplitude and latency [171]. ERP/ERF responses are divided into positive and negative based on the sign of their amplitudes. The $P100$, $P200$, and $P300$ are examples of well–known positive components that are evoked around 100, 200, and 300 ms after the stimulus onset, respectively. The $P100$ is typically modulated by attention in the extrastriate cortex and in response to visual stimuli [193]. The $P200$ component is involved in cognitive processes such as working memory [116] and semantic processing [53]. The $P300$ indicates higher cognitive processes and occurs in response to a variety of sensory stimuli such as visual, tactile, and auditory [161]. Due to its robustness, the $P300$ has some applications in the BCI context [158]. The $N100$ and $N170$ are examples of negative ERP/ERF components that are generally elicited in response to auditory [141] and human face [19] stimuli, respectively. Figure 2.5 illustrates schematically some well–known ERPs.

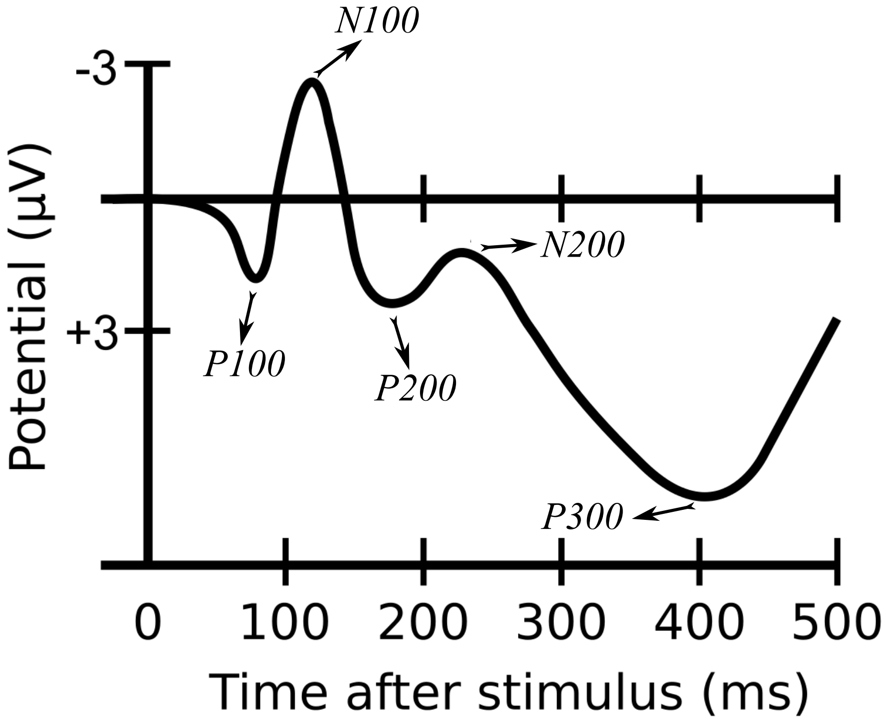


Figure 2.5: A schematic illustration of some well–known ERPs.

Time–Frequency Analysis

In computing the evoked ERP/ERF in response to external stimuli/events, one of the main assumptions is that the signal of interest is time–locked. But in fact brain responses are not always time–locked to the stimulus onset, and the timing might change slightly from one epoch to another. These jitters in time result in cancellation of positive and negative signal components when averaging the epochs. This situation might happen also in case of *induced responses*, i.e., when the response is time–locked but not phase–locked. An example for this kind of responses is Gamma oscillation in complex stimulus processing [182]. One possible approach to overcome this problem is to compute the frequency power spectrograms by transferring the signal from time domain to the time–frequency domain.

Short–time Fourier transform (SFT) and wavelet transform are two common methods for calculating time–frequency representations of EEG/MEG signals [68]. The computation is generally performed by calculating the spectral power of different frequency bands on a sliding interval of the signal. The length of intervals can be considered fixed for different frequency bands. An alternative and more effective approach is to decrease the interval length by increase in frequency. The analysis can be enhanced using the multitaper technique [136] which allows for a better control of time and frequency smoothing and reduces spectral leakage.

Source–Space Analysis

The electrical/magnetic brain activity is recorded via EEG/MEG sensors placed around the head. In sensor–space EEG/MEG data, each sensor records the electrical/magnetic activity from several sources in the brain. The goal of transferring the sensor–space data to the source–space is to estimate the source of brain activity based on the signals measured outside the head. Although the EEG/MEG data are measured simultaneously with several sensors, transforming the data to the source–space is an ill–posed problem without a unique solution. This problem is known as the *inverse problem* [68] in the context of EEG/MEG data analysis. One possible solution to derive valuable information on source distribution of brain activity is to include additional physiological information in order to put some constraints on the inverse problem. There are two main directions toward addressing the inverse problem:

1. **Parametric source models:** These approaches make some specific assumptions on the number and locations of focal sources. Generally, it is assumed that there are few active sources and their number, locations, and orientations are estimated iteratively e.g., by using standard nonlinear least–squares optimization methods [130], until the predicted electric potential or magnetic field is sufficiently close to the measured one. The equivalent current dipole model [69] and multi-

ple signal characterization [139] are two common parametric source estimation approaches.

2. **Distributed dipole models:** Unlike parametric approaches, the dipole distribution models make little assumptions on the parameters of the source model, instead they try to extract the characteristics of the data distribution in source–space in a data–driven manner. To this end, distributed dipole models assume that the sources are distributed within a volume or on a surface and then use various estimation techniques to find out the most plausible source distribution. Linear minimum–norm estimation [70] is an example of these methods.

2.3 Statistical Hypothesis Testing

The falsifiability is an indispensable principle of any scientific hypothesis [162]. The falsifiability means that before any scientific hypothesis is accepted as a theory, it must be inherently disprovable. In fact, the falsifiability provides the possibility of replacing an old theory by an enhanced one with more generalization. Statistical hypothesis testing provides a framework to measure the degree of falsifiability of a probabilistic hypothesis. In this section, we review the basic concepts behind the classical hypothesis testing approaches with focus on applications in neuroimaging.

2.3.1 Classical Hypothesis Testing

A scientific hypothesis is a proposed explanation for a general behavior of a particular phenomenon that is made based on limited observations. The validity of any scientific hypothesis is evaluated by means of statistical hypothesis testing, also known as confirmatory data analysis. Statistical hypothesis testing can be performed by adopting either a frequentist or Bayesian approach.

Frequentist Framework

In the frequentist approach, the falsifiability of a hypothesis is measured by computing the probability of erroneous inference by replicating the experiment. There are two major schools of thoughts in frequentist approach [21, 117, 119, 152]:

1. **Significance Testing (Fisher's method):** Ronald Fisher for the first time introduced the concept of significance testing in statistics [55]. The Fisher's procedure for significance testing is as follows [see Figure 2.6(A)]:

   i . Setting up the null hypothesis H_0. The aim of the experiments is to prove that the null hypothesis is false.

   ii . Choosing an appropriate test statistic T to summarize the data in real numbers.

   iii . Deriving the null distribution $p(T \mid H_0)$ analytically or by resampling.

   iv . Collecting the experimental data and calculating the test statistic in the observed data T_o.

   v . Computing the p-value $= p(T \geq T_o \mid H_0)$.

   vi . Reporting the p-value as a measure of evidence against H_0.

2. **Hypothesis Testing (Neyman–Pearson's Method):** is introduced first time in a paper by Jerzy Neyman and Egon Pearson in 1933 [145]. The Neyman–Pearson approach is applicable when the problem can be explained in the form of two disjointed hypotheses

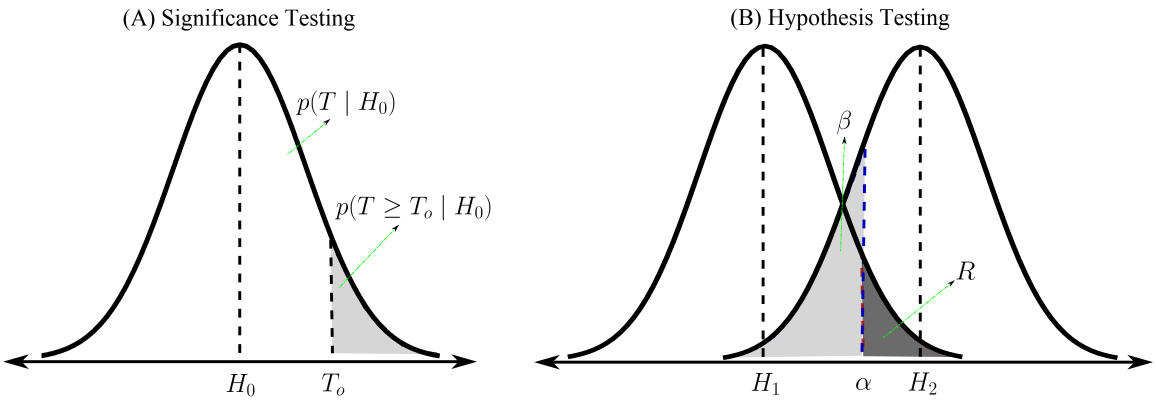


Figure 2.6: Frequentist frameworks in classical hypothesis testing: **(A)** Fisher's method for the significance testing. **(B)** Neyman–Pearson's method for the hypothesis testing.

and a meaningful cost/benefit trade–off can be set between the two. The whole procedure can be summarized as follows [see Figure 2.6(B)]:

i . Setting up two simple complementary hypotheses: the null H_1 and the alternative H_2 hypothesis. The aim of the test is to see whether we can reject H_1 in favor of H_2.

ii . Choosing an appropriate summary of the data based on a test statistic T.

iii . Deciding critical value α, so called the Type I error rate or false positive rate, and the sample size n. The α is a parameter that specifies the probability of false alarms, i.e, the probability of rejecting the null hypothesis when it is true.

iv . Computing the power of test for a given α and statistics T. The power of the test is $1 - \beta$, where β is the Type II error rate or false negative rate.

v . Computing the rejection region R on T. The rejection region is the range of values in T where the null hypothesis is rejected.

vi . Running the experiment and computing the statistic T_o on the observed data.

vii . Rejecting H_1 and accepting H_2 if $T_o \in R$, accepting H_1 and rejecting H_2 if $T_o \notin R$.

It is worthwhile to emphasize that failing to reject the H_1 in hypothesis testing must not be interpreted as the correctness of the null hypothesis, but it just shows a lack of evidence against it [147].

Bayesian Framework

Bayesian framework is an alternative for the frequentist approaches in statistical hypothesis testing [147]. In contrary to the frequentist approaches that test the data given the hypothesis, in Bayesian hypothesis testing we test the hypothesis given the data. The procedure for general Bayesian hypothesis testing for two alternative hypotheses can be summarized as follows:

1. Set up two mutually exclusive hypotheses, H_1 and H_2.
2. Run the experiment and collect the data D.
3. Use prior knowledge to specify the prior probabilities $p(H_1)$ and $p(H_2)$ where $p(H_1) + p(H_2) = 1$.
4. Specify the likelihood functions to model the data given the hypotheses: $p(D \mid H_1)$ and $p(D \mid H_2)$.
5. Compute the posterior probability of each hypothesis using the Bayes rule: $p(H_i \mid D) = \frac{p(D|H_i)p(H_i)}{\sum_{j=1}^{2} p(D|H_j)p(H_j)}$.
6. Test the hypothesis using one of the following approaches:

   i . Maximum a posteriori (MAP) approach: we accept H_1 if $p(H_1 \mid D) > p(H_2 \mid D)$ and vice versa.

   ii . Bayes factor (BF) approach: we compute the BF as $\frac{p(D|H_1)}{p(D|H_2)}$. The resulting BF can be interpreted based on Table 2.1 [95, 101].

Table 2.1: Interpretation of the Bayes factor.

Bayes Factor (BF)	**Evidence**
< 1	Negative (H_1 is rejected and H_2 is accepted)
1 to 3	Barely worth mentioning
3 to 10	Substantial (in favor of H_1)
10 to 30	Strong (in favor of H_1)
30 to 100	Very strong (in favor of H_1)
> 100	Decisive (in favor of H_1)

2.3.2 Mass–Univariate Hypothesis Testing on MEG data

The recorded MEG data represent the neural sources in space, time, and frequency domains; thus, the data contain spatio–temporal correlated structures. Therefore, an ideal approach for hypothesis testing on MEG data should consider the full range of spatio–temporal information. However, the common statistical hypothesis testing approaches on MEG data [48] fail to fully get advantage of these spatio–temporal information [64]. This fact motivates exploring new methods for statistical testing on high–dimensional data. Mass–univariate hypothesis testing is an effective approach in this direction, and it can be used to simultaneously perform a large number of univariate tests on whole spatio–temporal variables. In MEG data analysis, the mass–univariate hypothesis testing can detect the underlying neurophysiological effects with greater temporal and spatial details compared to the conventional priori–based analysis. Therefore, it is preferable to conventional analysis in exploratory studies on neuroimaging data where little is known in advance about when, where, and how an effect will occur.

Despite its effectiveness, mass–univariate hypothesis testing suffers from multiple comparisons problem (MCP). The MCP occurs in statistical hypothesis testing when a set of statistical inferences are simultaneously performed [134]. For example the MCP arises when we test concurrently a hypothesis on several data dimensions, e.g., on several MEG sensors. The MCP increases the chance of commiting the Type I error, thus, ignoring the MCP poses a threat on the reliability of multiple statistical testing [15]. Several techniques are proposed for correcting the results of multiple statistical tests. These approaches can be classified into two main categories: 1) controlling the family–wise error rate, and 2) controlling the false discovery rate.

1. **Controlling the Family–Wise Error Rate:** The family–wise error rate (FWER) is the probability of making at least one Type I error in multiple–hypothesis testing. There are several methods to strongly or weakly control the FWER such as Bonferroni correction, step–down procedure [92], step–up procedure [91], and non–parametric cluster–based permutation tests [127].

2. **Controlling the False Discovery Rate:** The false discovery rate (FDR) is defined as the expected proportion of false discoveries to all discoveries [16]. Here a discovery refers to the rejection of the null hypothesis. Controlling the FDR is less restrictive than controlling the FWER, thus, it provides more statistical power but increases the Type I error rate. So far several methods have been proposed in the literature for controlling the FDR such as controlling the FDR under dependency [18], positive FDR [177], and adaptive linear step–up procedures [17].

Being essential for validity of results, on the down side, both strong control of FWER and controlling FDR reduce the statistical power of mass–univariate analysis. One possible approach to alleviate this problem is to weakly control the FWER, which guarantees the control of FWER in case there are no experimental effects [146]. The cluster–mass test [31] is a possible method in this direction. This method was first adopted by Maris and Oostenveld [127] for non–parametric cluster–based permutation test on MEG data. The intuitive idea behind the cluster–based permutation test is that if a group of significant tests are clustered meaningfully in space, time, and frequency then the chance of committing the Type I error decreases. This method can be summarized as the following steps [64,127]:

1. Combine the MEG trials of the two experimental conditions A and B in a single dataset D.
2. Compute a random partition of D into A and B, D', by randomly permuting the trials.
3. For all the independent variables of D' in time and space (e.g., each time–bin of each sensor), compute the statistic T, e.g., t–statistic.
4. Ignore all variables with T statistic below a certain threshold. The threshold is decided based on the pre–specified α and the probability distribution of T.
5. Cluster the remaining independent variables that are adjacent in time and space.
6. Compute the cluster–level statistic T_c for each cluster, for example by summing up the statistics in each cluster.
7. Save the largest cluster level statistic as T_{max}.
8. Repeat the steps 2–7 to construct the null hypothesis of cluster–level statistics on the randomly partitioned data.
9. Perform steps 3–6 on D and save the cluster–level statistic for each cluster in T^*.
10. Use the Montecarlo method on the null hypothesis derived in step 8 to derive the p-value for each cluster obtained in step 9. The p-value

is computed by computing the proportion of T_{max} that are larger than T^*.

11. The cluster–level p-value is assigned to the all variables in that cluster. The p-value of ignored variables (not involved in any cluster) are set to 1.

In spite of its higher statistical power, the non–parametric cluster–based permutation test suffers from three main limitations: 1) since it weakly controls FWER, it is not reliable for explaining the exact spatio–temporal pattern of the underlying effect. This shortcoming makes this method more appropriate for understanding whether an effect is present in data rather than finding out exactly when and where the effect occurs [127]; 2) it is not sensitive enough to detect narrowly distributed effects in time and space [64,65]; 3) due to its univariate nature, it does not benefit from multivariate and distributed patterns across different sensors, frequency bands, and time scales. These limitations motivate exploring new approaches with higher sensitivity and specificity that enable researchers to find the exact discriminative source of neural correlates across different experimental conditions.

2.4 Statistical Learning Theory

Statistical learning theory provides an alternative for classic statistical hypothesis testing approaches. In the following text we briefly introduce the basic concepts in the statistical learning theory that are used in the rest of this thesis.

2.4.1 From Maximum a Posteriori to Risk Minimization

In the supervised statistical learning framework, the main aim is to learn a function $\Phi^* : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y}$ represent the input and output spaces, respectively. In practice, the learning process is performed on the sampled data $S = \{(\mathbf{X}, Y) \mid \mathbf{X} \subset \mathcal{X}, Y \subset \mathcal{Y}\}$ by approximating $\Phi_S : \mathbf{X} \rightarrow Y$, the so called the regression function, among a family of functions $\mathcal{H}$. Here $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^n$ are n independently and identically distributed (iid) samples drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$; based on an unknown Borel probability measure ρ and $\forall \mathbf{x} \in \mathbf{X}$ we have [43]:

$$\Phi_S(\mathbf{x}) = \int_{y \in \mathcal{Y}} y \quad d\rho(y \mid \mathbf{x}). \tag{2.1}$$

The probability measure ρ can be split into $\rho(Y \mid \mathbf{X})$ and $\rho_{\mathbf{X}}$ [43]. Unlike the marginal distribution of $\mathbf{X}$, i.e., $\rho_{\mathbf{X}}$, which is known in some cases, ρ and $\rho(Y \mid \mathbf{X})$ are unknown in advance. Therefore, the goal of learning is to estimate the predictive conditional density $\rho(Y \mid \mathbf{X})$ by training a parametric model $\rho(Y \mid \mathbf{X}, \Theta)$ where Θ denotes the parameters of the learning algorithm. In general, the parameters can be estimated by maximizing the posterior probability $\rho(\Theta \mid \mathbf{X})$ using the maximum a posteriori (MAP) estimate:

$$\hat{\Theta} = \operatorname*{argmax}_{\Theta} \rho(\Theta \mid \mathbf{X}) \propto \operatorname*{argmax}_{\Theta} \rho(\mathbf{X} \mid \Theta)\rho(\Theta). \tag{2.2}$$

The above maximization problem can be converted to the equivalent risk minimization problem by computing the negative log–likelihood:

$$\begin{aligned} \operatorname*{argmax}_{\Theta} \rho(\mathbf{X} \mid \Theta)\rho(\Theta) &= \operatorname*{argmin}_{\Theta} -\log(\rho(\mathbf{X} \mid \Theta)) - \log(\rho(\Theta)) \\ &\equiv \operatorname*{argmin}_{\Theta} \mathcal{L}(Y, \Phi(\mathbf{X})) + \lambda\Omega(\Theta) \end{aligned} \tag{2.3}$$

where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$ is the loss function, $\Omega : \mathbb{R}^p \rightarrow \mathbb{R}^+$ is the regularization term, and $\lambda \geq 0$ is a hyper–parameter that controls the amount of regularization. It is worthwhile to emphasize that in the learning paradigm presented in Eq. 3.1, we try to estimate Φ_S (and not Φ^*) on the sampled data by solving an optimization problem in $\mathcal{H}$. The *irreducible error* [81] $\varepsilon \in \mathbb{R}^n$ is the direct consequence of this approximation and provides a lower bound on the error of a model and we have:

$$\Phi_S(\mathbf{X}) = \Phi^*(\mathbf{X}) + \varepsilon. \tag{2.4}$$

The assumption on the distribution of ε drives the motivation behind the choice of the loss function $\mathcal{L}$ [211]. For example if we assume ε to have a Gaussian distribution with mean 0 and variance σ^2, we have the least squares loss function of Eq. 3.1 as

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{2} \left\| Y - \Phi(\mathbf{X}) \right\|_2^2 + \lambda \Omega(\Theta). \tag{2.5}$$

Table 2.2 summarizes some popular choices for the loss function $\mathcal{L}$.

Table 2.2: Some popular examples of the loss function.

Name	**Loss**
Least–squares loss	$\frac{1}{2} \left\| Y - \Phi(\mathbf{X}) \right\|_2^2 = \frac{1}{2} \sum_{i=1}^{n} (y_i - \Phi(\mathbf{x}_i))^2$
Logistic loss	$\sum_{i=1}^{n} \log(1 + \exp(-y_i \Phi(\mathbf{x}_i)))$
Hinge loss	$\sum_{i=1}^{n} \max(0, 1 - y_i \Phi(\mathbf{x}_i))$

2.4.2 Bias–Variance Decomposition of Error

As mentioned before, the aim of statistical learning is to find the best approximation of Φ_S among a family of functions $\mathcal{H}$, so called hypothesis space. Limiting the search space to $\mathcal{H}$ poses a restriction on finding the best match because $\mathcal{H}$ might or might not include Φ_S or even Φ^*. Thus, considering this limitation the aim of learning reduces to finding $\Phi_{\mathcal{H}} \in \mathcal{H}$, so called the target function, as the best empirical approximation of Φ_S. For example, setting $\mathcal{H}$ to a set of linear functions is a very common assumption in applying statistical learning framework on neuroimaging data. Let $\hat{\Phi} \in \mathcal{H}$ be the empirical approximation of the target function $\Phi_{\mathcal{H}}$ on the training set S where

$$\hat{\Phi} = \underset{\Phi \in \mathcal{H}}{\operatorname{argmin}} \, \mathcal{L}(Y, \Phi(\mathbf{X})). \tag{2.6}$$

Then the *expected prediction error* (EPE) associated with $\hat{\Phi}$, denoted by $\mathcal{E}_{\hat{\Phi}}$, can be computed by summing up three main contributing factors:

$$\mathcal{E}_{\hat{\Phi}} = \mathcal{E}(\Phi_{\mathcal{H}}) + \mathcal{E}_{\mathcal{H}}(\hat{\Phi}) + \varepsilon = \int_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\Phi^*(\mathbf{x}), \hat{\Phi}(\mathbf{x})) \tag{2.7}$$

where $\mathcal{E}(\Phi_{\mathcal{H}}) = \int_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\Phi_S(\mathbf{x}), \Phi_{\mathcal{H}}(\mathbf{x}))$ is generally known as the *approximation error* or the *bias* of a model. It depends strongly on the choice of the hypothesis space $\mathcal{H}$. The second term $\mathcal{E}_{\mathcal{H}}(\hat{\Phi}) = \int_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\Phi_{\mathcal{H}}(\mathbf{x}), \hat{\Phi}(\mathbf{x}))$ is known as the *sample error* or equivalently the *variance* of a model which is highly dependent on the samples in S. Fixing $\mathcal{H}$, the variance of the model decreases by increasing the number of samples n. Enlarging the hypothesis space $\mathcal{H}$ reduces the bias but has a negative effect on the variance of the model and vice versa. The relation between the sampling size and the size of the hypothesis space and their effect on the final error is typically referred as the *bias–variance trade–off* [58]. The last term $\varepsilon = \int_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\Phi^*(\mathbf{x}), \Phi_S(\mathbf{x}))$ is called *irreducible error* which provides the lower bound on the error and cannot be reduced in the learning process.

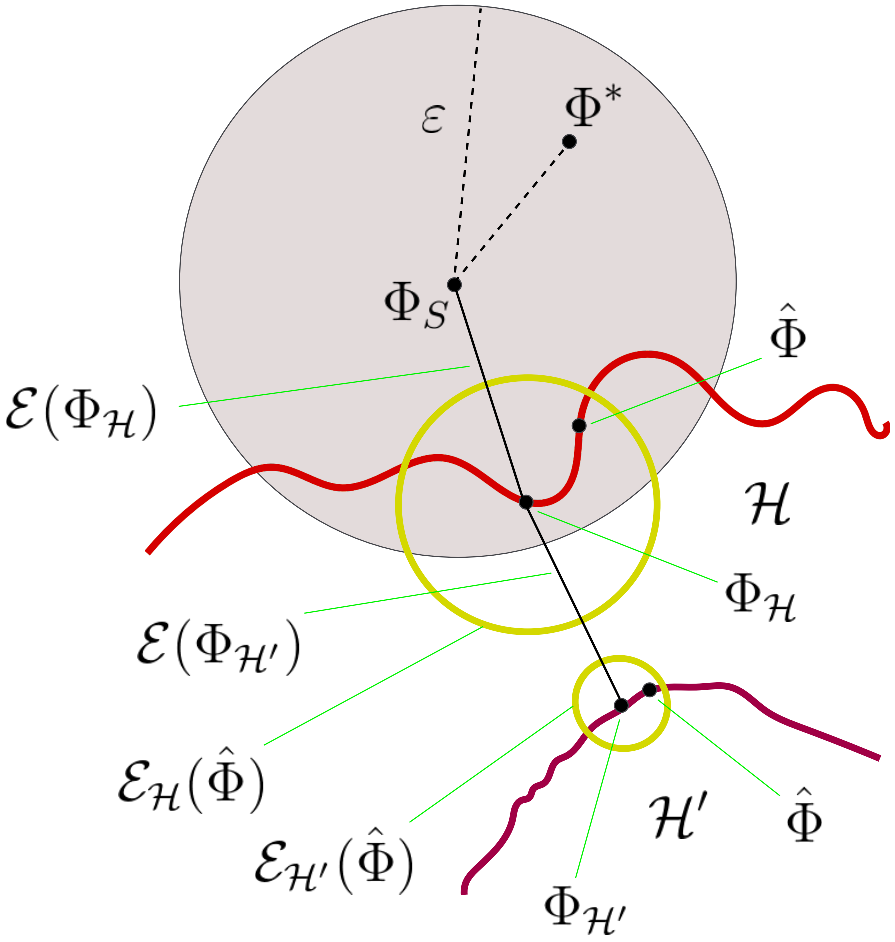

ε
Φ^*
Φ_S
$\hat{\Phi}$
$\mathcal{E}(\Phi_{\mathcal{H}})$
$\mathcal{H}$
$\Phi_{\mathcal{H}}$
$\mathcal{E}(\Phi_{\mathcal{H}'})$
$\mathcal{E}_{\mathcal{H}}(\hat{\Phi})$
$\mathcal{E}_{\mathcal{H}'}(\hat{\Phi})$
$\mathcal{H}'$
$\hat{\Phi}$
$\Phi_{\mathcal{H}'}$

Figure 2.7: The components of the error and the effect of regularization on the bias and variance of a model [81].

Figure 2.7 schematically illustrates the relation between the components of the error.

2.4.3 Regularization

The size and complexity of $\mathcal{H}$ can also be controlled by the choice of the regularization term Ω. This term, by putting prior assumptions on the distribution of parameters $\rho(\Theta)$, enforces prior knowledge into the learning process. In other words, regularization reduces the search space to $\mathcal{H}' \subset \mathcal{H}$ based on prior knowledge on the distribution of parameters. This reduction

Table 2.3: Some popular choices for Ω. Here θ_i is used to refer to the ith element of the parameter vector Θ.

Name	$\Omega(\Theta)$	Description
ℓ_2 penalty	$\sum_{i=1}^{p} \theta_i^2$	Computes squared ℓ_2-norm of the weight vectors.
ℓ_1 penalty (Lasso) [185]	$\sum_{i=1}^{p} \lvert\theta_i\rvert$	Computes ℓ_1-norm of the weight vectors.
Elastic–net [223]	$(1-\alpha)\sum_{i=1}^{p} \theta_i^2 + \alpha \sum_{i=1}^{p} \lvert\theta_i\rvert$	Combines ℓ_2 and ℓ_1 penalization using α coefficient as an extra hyper–parameter.
Group Lasso [96]	$\sum_{g \in G} \sum_{i=1}^{\lvert g\rvert} \theta_i^2$	Divides the parameters into groups G and computes the ℓ_1-norm over ℓ_2-norms of grouped parameters.
Fused Lasso [186]	$\sum_{i=1}^{p-1} \lvert\theta_{i+1} - \theta_i\rvert$	Computes the ℓ_1-norm on the difference between successive parameters.

decreases the variance of the model by the cost of increasing the bias (see Figure 2.7). As a consequence, the chance of overfitting on the training samples decreases especially when $n \ll p$. Table 2.3 summarizes some popular choices for Ω.

2.4.4 Bias–Variance Decomposition in Binary Classification

A binary classification problem is a special case of statistical learning problem where $\mathcal{Y}$ is categorical with two possible values, for example $\mathcal{Y} \in \{-1, 1\}$. Since in this case, the loss function reduces to a 0/1-loss (e.g., logistic loss or hinge loss), computing the components of EPE is different from the general regression case. One possible approach to compute the bias–variance decomposition of the error is by using the out–of–bag (OOB) technique [49, 189]. The OOB employs bootstrapping repetitions to perturb the training set and draw several training and validation sets. The perturbed data are used to compute the EPE for an estimated binary classifier Φ.

Let m be the number of perturbed training sets resulting from partitioning $S = (X, Y)$ into $S_{tr} = (X_{tr}, Y_{tr})$ and $S_{vl} = (X_{vl}, Y_{vl})$, i.e., training and validation sets. If $\hat{\Phi}^j$ is the binary classifier estimated from the jth perturbed training set, then the main prediction $\Phi^{\mu}(\mathbf{x}_i)$ for each sample in the dataset can be computed as follows:

$$\Phi^{\mu}(\mathbf{x}_i) = \begin{cases} 1 \quad if \quad \frac{1}{k_i}\sum_{j=1}^{k_i} \hat{\Phi}^j(\mathbf{x}_i) \geq \frac{1}{2} \\ 0 \qquad\quad otherwise \end{cases} \tag{2.8}$$

where k_i is the number of times that x_i is present in the test set[3]. In fact the main prediction Φ^{μ} provides an estimate of the target function $\Phi_{\mathcal{H}}$.

The computation of bias is challenging because the optimal model Φ^* is unknown. The misclassification error is one of the loss measures that satisfies a Pythagorean–type equality [184], where

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\Phi^{\mu}(\mathbf{x}_i), \Phi^*(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^{\mu}(\mathbf{x}_i)) - \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^*(\mathbf{x}_i)). \tag{2.9}$$

Because all terms of the above equation are positive, the mean loss between the main prediction and the actual labels can be considered as an upper–bound for the bias, therefore we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\Phi^{\mu}(\mathbf{x}_i), \Phi^*(\mathbf{x}_i)) \leq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^{\mu}(\mathbf{x}_i)). \tag{2.10}$$

Then, a pessimistic approximation of bias $B(\mathbf{x}_i)$ can be calculated as:

$$B(\mathbf{x}_i) = \begin{cases} 0 \quad if \quad \Phi^{\mu}(\mathbf{x}_i) = y_i \\ 1 \qquad otherwise \end{cases}. \tag{2.11}$$

Then, the unbiased and biased variances (see Ref. [49] for definitions) in each training set can be calculated by:

$$V_u^j(\mathbf{x}_i) = \begin{cases} 1 \quad if \quad B(\mathbf{x}_i) = 0 \quad and \quad \Phi^{\mu}(\mathbf{x}_i) \neq \hat{\Phi}^j(\mathbf{x}_i) \\ 0 \qquad\qquad\qquad otherwise \end{cases} \tag{2.12}$$

[3] It is expected that each sample $\mathbf{x}_i \in X$ appears (on average) $k_i \approx \frac{m}{3}$ times in the test sets.

and

$$V_b^j(\mathbf{x}_i) = \begin{cases} 1 \quad if \quad B(\mathbf{x}_i) = 1 \quad and \quad \Phi^{\mu}(\mathbf{x}_i) \neq \hat{\Phi}^j(\mathbf{x}_i) \\ 0 \qquad\qquad\qquad otherwise \end{cases} . \tag{2.13}$$

The expected prediction error of Φ can be computed as follows (ignoring the irreducible error):

$$EPE_{\Phi}(\mathbf{X}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} B(\mathbf{x}_i)}_{Bias} + \underbrace{\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}[V_u^j(\mathbf{x}_i) - V_b^j(\mathbf{x}_i)].}_{Variance} \tag{2.14}$$

2.4.5 Multi–Task Learning

Basic Concepts: Domain, Task, and Transfer Learning

The aim of this section is to provide the notation needed for the formal definition of multi–task learning. To this end, we briefly introduce basic concepts such as domain, task, and transfer learning.

In the context of statistical learning theory, a *domain* $\mathcal{D} = \{\mathcal{X}, \rho_{\mathbf{X}}\}$ is defined as a possible conjunction between an input space $\mathcal{X}$ and a marginal probability distribution $\rho_{\mathbf{X}}$. As an example in the neuroimaging context, in the multi–modal brain imaging (where several imaging techniques, e.g., fMRI and EEG, are simultaneously used) each modality represents a domain. Given a domain $\mathcal{D}$, a *task* $\mathcal{T} = \{\mathcal{Y}, \Phi\}$ is defined as a predictive function Φ from $\mathcal{D}$ to the output space $\mathcal{Y}$. For example, assume we record the brain activity when the subjects observe visual stimuli in different shapes and colors. Then, predicting the shape or the color of a particular stimulus from the recorded signal can be considered as two different tasks.

In the statistical learning theory the goal is to learn a task $\mathcal{T}$ in a certain domain $\mathcal{D}$. Assume $\mathcal{D}_S$, $\mathcal{D}_T$, $\mathcal{T}_S$, and $\mathcal{T}_T$ represent the source domain, target domain, source task, and target task, respectively. *Transfer learning* aims to benefit from the knowledge in the source domain and task in order to improve the predictive power in the target domain when $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$. Depending on the last condition, supervised transfer learning is categorized into two major branches [155]:

1. **Inductive Transfer Learning:** The necessary condition in inductive transfer learning is $\mathcal{T}_S \neq \mathcal{T}_T$, thus the relation between $\mathcal{D}_S$ and $\mathcal{D}_T$ does not matter. Further, in inductive transfer learning it is required to have some labeled data in $\mathcal{D}_T$ in order to learn Φ_T in the target domain. The goal of inductive transfer learning is to incorporate additional information in source domains and tasks in order to improve the generalization performance on the target task.

2. **Transductive Transfer Learning:** In transductive transfer learning, we have $\mathcal{T}_S = \mathcal{T}_T$ while $\mathcal{D}_S \neq \mathcal{D}_T$. Further, it is assumed that unlike the source domain there are no labeled data available in $\mathcal{D}_T$.

Multi–Task Learning

In order to solve a real–world problem, in general we need to deal with multiple related sub–problems, i.e., tasks. A trivial approach is to solve these problems independently, and ignore the useful shared information across tasks. This single–task learning (STL) approach yields sub–optimal solutions especially when few samples are available for each task. Multi–task Learning (MTL) is an inductive transfer learning approach that tries to improve the generalization performance of models by promoting information sharing across different related tasks [35]. The learning process in MTL is based on simultaneous training of several models, each of which for one task. In addition, learning multiple related tasks simultaneously effectively increases the sample size for each task. Thus, MTL is especially

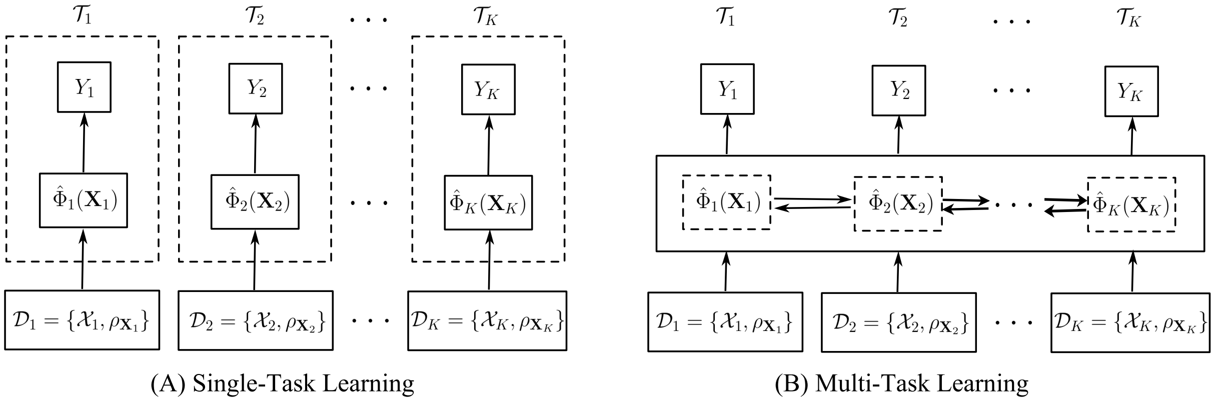

$\mathcal{T}_1$
$\mathcal{T}_2$
$\mathcal{T}_K$
Y_1
Y_2
Y_K
$\hat{\Phi}_1(\mathbf{X}_1)$
$\hat{\Phi}_2(\mathbf{X}_2)$
$\hat{\Phi}_K(\mathbf{X}_K)$
$\mathcal{D}_1 = \{\mathcal{X}_1, \rho_{\mathbf{X}_1}\}$
$\mathcal{D}_2 = \{\mathcal{X}_2, \rho_{\mathbf{X}_2}\}$
$\mathcal{D}_K = \{\mathcal{X}_K, \rho_{\mathbf{X}_K}\}$
(A) Single-Task Learning
(B) Multi-Task Learning

Figure 2.8: (A) In single–task learning the predictive functions are learned independently across subject, while (B) multi–task learning provides the possibility of sharing information across different tasks in the learning process.

advantageous over STL when there is a limited number of training samples available for each task.

Assume $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ be K corresponding pairs of domain–task. In MTL, the empirical risk minimization problem in Eg. 3.1 is reformulated as follows:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{k=1}^{K} \mathcal{L}(Y_k, \Phi(\mathbf{X}_k)) + \lambda\Omega(\Theta) \tag{2.15}$$

where $\mathbf{X}_k \in \mathbb{R}^{n_k \times p}$ represents the n_k samples in input space from domain $\mathcal{D}_k$ and under the probability distribution $\rho_{\mathbf{X}_k}$, and $Y_k \in \mathbb{R}^{n_k}$ is the output space from task $\mathcal{T}_k$. The parameters of the model $\hat{\Theta}$ are estimated by parallel minimization of the loss functions across different tasks. Unlike the common STL approach, where the predictive functions are learned independently across tasks, the parallel optimization in MTL provides the possibility for exchanging useful information among tasks. Figure 2.8 schematically illustrates this advantage of MTL over STL.

Learning Structures in Multi–Task Learning

The simultaneous learning and information sharing across tasks provides another important advantage for MTL which is the possibility of learning the structures in input or output spaces. Here the structure refers to a certain relational arrangement, e.g., correlation, between different dimensions of input spaces, i.e., features, or output spaces across the tasks. Therefore the related samples are no longer *iid* and the standard statistical learning approaches that assume independence between samples, are sub–optimal. MTL overcomes this problem as it provides the infrastructure to learn structures in input and output spaces. In formulation of the empirical risk minimization problem for MTL (Eq. 2.15), the regularization term Ω provides the possibility of learning the structures in the parameter space. Several studies investigated different regularization schemes for learning the structures in the MTL framework. Here we briefly explain three possible options in this direction:

1. **Joint Feature Learning:** Joint feature learning enables the model to capture a sparse set of features that are common across different tasks. To this end, it employs the idea of group sparsity via $\ell_{2,1}$ regularization [8,9,124,148] where

$$\Omega(\Theta) = \|\Theta\|_{2,1} = \sum_{k=1}^{K} \|\Theta_k\|_2 \, . \tag{2.16}$$

   Here $\Theta \in \mathbb{R}^{p \times K}$ is assumed to be the matrix of parameters, and Θ_k refers to its kth columns.

2. **Graph Encoding:** In this scheme it is assumed that the relationships between tasks can be encoded in the form of a graph where each task is a node and there is an edge between two nodes if two tasks are related.

It is beneficial when the existing relational structures between tasks are known in advance or they can be derived in a data–driven manner. Let E to denote the set of edges, where each edge is represented as a vector $\mathbf{e}^{(i)} \in \mathbb{R}^K$. If the ith edge connects the uth and vth nodes, then the uth and vth elements of $\mathbf{e}^{(i)}$ are set to 1 and -1, respectively. The complete graph matrix is then constructed by concatenating the edge vectors $\mathbf{R} = [\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(|E|)}] \in \mathbb{R}^{K \times |E|}$, and we have a graph–fused regularization term [121]

$$\Omega(\Theta) = \|\Theta\mathbf{R}\|_F^2 = \sum_{i=1}^{|E|} \left\|\Theta\mathbf{e}^{(i)}\right\|_2^2 \tag{2.17}$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix.

3. **Temporal Encoding:** In some applications, there are temporal structures in the feature space across different tasks. Longitudinal study on disease progression is an example of these applications [221] where a variety parameters are repeatedly measured in a period of time for a patient. In this configuration, the prediction of the value of the disease status at one time point can be considered as a task. In order to consider the temporal dependency between tasks, the regularization term should be able to encode the temporal structures in the sequence of measurements, and we have [221]

$$\Omega(\Theta) = \sum_{k=1}^{K-1} \|\Theta_k - \Theta_{k+1}\|_F^2 . \tag{2.18}$$

Chapter 3

Interpretability in Linear Brain Decoding

3.1 Introduction

Understanding the mechanisms of brain function has been a crucial topic throughout the history of science. Modern cognitive science, emerging in the 20th century, provides better insight into the functions of brain. In cognitive science, researchers usually analyze recorded brain activity and behavioral parameters to discover the answers of *where*, *when*, and *how* a brain region participates in a particular cognitive process.

To answer the key questions in cognitive science, scientists often employ mass–univariate hypothesis testing methods (see Section 2.3.2) to test scientific hypotheses on a large set of independent variables [64, 126]. Mass–univariate hypothesis testing is based on performing multiple tests, e.g., t–tests, one for each unit of the neuroimaging data, i.e., independent variables. The high spatial and temporal granularity of the univariate tests provides a fair level of interpretability. On the down side, the high dimensionality of neuroimaging data requires a large number of tests that reduces the sensitivity of these methods after multiple comparison correction [32]. Although techniques such as the non–parametric cluster–based permutation test [31, 127], by weak control of family–wise error rate, offer more sensitivity, they still experience low sensitivity to brain activity that are narrowly distributed in time and space [64, 65]. The multivariate counterpart of mass–univariate analysis, known generally as multivariate pattern analysis, have the potential to overcome these deficits. Multivariate approaches, by employing the principles behind statistical learning theory (see Section 2.4), are capable of identifying complex spatio–temporal interactions between different brain areas with higher sensitivity and specificity than univariate analysis [192], especially at the group–level [45].

Brain decoding [89] is a statistical learning approach that delivers a model to predict the mental state of a human subject based on the recorded brain signal. There are two applications for brain decoding: 1) brain–computer interfaces (BCIs) [208], and 2) multivariate hypothesis testing [32]. In the first case, a brain decoder with maximum prediction power is desired. In the second case, in addition to the prediction power, extra information on the spatio–temporal nature of a cognitive process is desired. In this study, we are interested in the second application of brain decoding that can be considered a multivariate alternative for mass–univariate hypothesis testing. Further, we mainly focus on the linear brain decoding because of its wider usage in analyzing inherently small–sample–size and high–dimensional neuroimaging data, compared to the complex [41, 114] and non–transparent [123] non–linear models.

In linear brain decoding, linear classifiers are used to assess the relation between independent variables, i.e., features, and dependent variables, i.e., cognitive tasks [22, 118, 157]. This assessment is performed by solving an optimization problem that assigns weights to each independent variable. Currently, brain decoding is the gold standard in multivariate analysis for functional magnetic resonance imaging (fMRI) [41, 86, 135, 149] and magnetoencephalogram/electroencephalogram (MEEG) studies [3, 34, 36, 93, 156, 167, 199]. It has been shown that brain decoding can be used in combination with brain encoding [143] to infer the causal relationship between stimuli and responses [202].

In *brain mapping* [112], the pre–computed quantities, e.g., univariate statistics or weights of a linear classifier, are assigned to the spatio–temporal representation of neuroimaging data in order to reveal functionally specialized brain regions which are activated by a certain cognitive task. In its multivariate form, brain mapping uses the learned parameters from brain decoding to produce brain maps, in which the engagement of different brain areas in a cognitive task is visualized. Intuitively, the interpretability of a brain decoder refers to the level of information that can be reliably derived by an expert from the resulting maps. From the cognitive neuroscience perspective, a brain map is considered *interpretable* if it enables a scientist to find answers to the three key questions: "*where*, *when*, and *how* does a brain region contribute to a cognitive function?"

3.1.1 Knowledge Extraction Gap in Brain Decoding

A classifier only tells *what* is the most likely label of a given unseen sample [12] while it provides little information regard the underlying discriminative properties. This problem is generally known as knowledge extraction gap [198] in the machine learning context. In the context of neuroimaging, the knowledge extraction gap in classification is generally known as the interpretation problem [88, 142, 172]. Therefore, improving the interpretability of linear brain decoding and associated brain maps is an important goal in the brain imaging literature [178]. There are four main reasons behind the lack of interpretability in multivariate brain mapping [7, 22, 24, 30, 82, 88, 102, 115, 118, 151, 183, 195, 197, 201]:

1. Low signal–to–noise ratio (SNR) in brain recordings [102]: Almost

all non–invasive brain imaging methods suffer from low SNR due to acquisition limitations and similarity in probability distribution of underground unrelated brain activity with the signal–of–interest. Low SNR generally reduces interpretability of the brain decoding model by decreasing its accuracy.

2. The high dimensionality of whole–scalp recordings [22, 102, 104, 195]: In brain decoding, we usually have a huge number of spatio–temporal features (on the order of 10^5) while the number of samples is limited (on the order of 10^2). This problem has two folds: 1) it causes the curse–of–dimensionality problem which affects the model accuracy, and 2) it makes the classification problem ill–posed where the number of unknown parameters is larger than the number of known data points. Although the second problem can be mitigated using regularization and sparse modeling, it still affects the interpretability of the model by decreasing parameter stability [219]. In some studies, prior knowledge is used to reduce dimensionality but unfortunately such prior knowledge is not always available. This issue supports the need for designing methods to decrease the dimensionality of the feature space without losing task–related information [30, 82, 104].

3. The high correlation between different dimensions of data [83, 195]: This problem, generally known as multicollinearity problem, reduces the stability of the model, which leads to unjustified conclusions in interpreting brain decoding models. When the feature space is highly correlated, not only the model is variable from one training run to another but also the amplitude of classifier weights is meaningless regarding the existence of the signal–of–interest.

4. Across–subject variability [102, 151]: Across–subject decoding is a meaningful process to make an inference at the group level. Unfortu-

nately, training an interpretable model across subjects is technically difficult because of variability of the underlying probability distribution of data samples from one subject to another. In practice, the interpretability of across–subject models is lower than single subject models because of decrease in both accuracy and stability of the brain decoding model.

At present, two main approaches are proposed to enhance the interpretability of multivariate brain maps: 1) introducing new metrics into the model selection procedure, and 2) introducing new hybrid penalty terms for regularization. In the following section we briefly review the current state of the art in improving the interpretability of brain decoding models.

3.1.2 State of the Art

The first approach for improving the interpretability of brain decoding concentrates on the model selection. Model selection is a procedure in which the best values for the hyper–parameters of a model are determined [118]. The selection process is generally performed by considering the generalization performance, i.e., accuracy, of a model as the decisive criterion. For example, Rasmussen et al. [166] showed that there is a trade–off between the spatial reproducibility and the prediction accuracy of a classifier; therefore, the reliability of maps cannot be assessed merely by focusing on their prediction accuracy. To utilize this finding, the authors incorporated the spatial reproducibility of brain maps in the model selection procedure. They concluded that choosing the optimal value for hyper–parameters of the model based on the combination of reproducibility and prediction metrics yields more interpretable brain decoding models. Using a similar methodology, in [14] the authors confirmed that accounting for coefficient reproducibility in the model selection procedure alleviates the coefficient instability problem in sparse brain decoding models. An analogous approach, using a different definition of spatial reproducibility, is proposed by Conroy et al. [40] where the authors illustrated that multiple models with the same classification accuracy may show completely different reproducibility level. Therefore, they proposed a model selection approach that utilizes a combination of bootstrapping and permutation testing to optimize both prediction accuracy and brain map reproducibility. They argue that optimizing hyper–parameters of the model in the accuracy–reproducibility joint space results in more interpretable decoding models. Elsewhere, Valverde and Moreno [190] experimentally showed that in the classification task optimizing just classification error rate is not enough to capture the transfer of crucial information from the input to the output of a classifier. To alleviate the problem, the authors introduced the entropy–modulated accuracy as a pessimistic estimate of the performance of a model. Furthermore to promote the interpretability of results, they introduced the normalized information transfer to avoid specialization in learning process. Beside spatial reproducibility, the stability of the classifiers [26] is another criterion that is used in combination with generalization performance to enhance the interpretability. For example, it is shown that incorporating the stability of models into cross–validation improves the interpretability of the estimated parameters [122, 215].

The second approach to improving the interpretability of brain decoding focuses on the underlying mechanism of regularization. The main idea behind this approach is two–fold: 1) customizing the regularization terms to address the ill–posed nature of brain decoding problems (where the number of samples is much less than the number of features) [138, 197], and 2) combining the structural and functional prior knowledge with the decoding process so as to enhance the neurophysiological plausibility of the models. Group Lasso [217] and total–variation penalty [186] are two effective methods using this technique [168, 212]. The first effort in this direction was made by Grosenick et al. [66]. To alleviate the multicollinearity problem in fMRI data, The authors introduced sparse penalized discriminant analysis (SPDA) for automatic selection of correlated variables. They compared SPDA with common methods like logistic regression, linear discriminant analysis (LDA), and linear support vector machine (lSVM). Their results suggest that SPDA enhances the interpretability of brain decoding models in both within and across–subject decoding scenarios. Elsewhere van Gerven et al. [192] proposed a group–wise regularization method for brain decoding on EEG data. They motivated the incorporation of prior knowledge into the regularization procedure by defining groups based on proximity of features in space, time, or frequency bands. In this way, the same sparsity profile is shared among related features. They showed the grouping strategy enhances the interpretability of the resulting models. In an MEG study, de Brecht and Yamagishi [46] presented a generalization of sparse logistic regression, called smooth sparse logistic regression (SSLR), which combines the Laplacian prior with a multivariate Gaussian prior to produce more sparse and at the same time smooth brain maps. The multivariate Gaussian prior encourages spatio–temporal smoothness and provides similar weights for neighbouring features in time and space, therefore, it selects spatio–temporally continuous groups of features. Their experiments on simulated data and real MEG data illustrated that SSLR provides more neuro–scientifically plausible brain maps. Following the idea of exploiting the data–driven extracted prior knowledge, Gramfort et al. [62] used the Total–Variation (TV) penalty to inject a spatial segmentation prior into the sparse model with ℓ_1 penalty. Their proposed method, called TV-ℓ_1, uses ℓ_1 penalization to set irrelevant features to zero and TV penalization to segment the relevant features together. On an fMRI dataset, they experimentally illustrated that their method provides better region recovery than other decoding and univariate brain–mapping strategies. They concluded that their method yields brain maps in good agreement with univariate methods like F-test while benefiting from the statistical power of multivariate methods. Grosenick et al. [67] proposed to use structural prior information, extracted from local smoothness or functional connectivity, as a graph constraint in penalization. They proposed to combine structured graph constraints with a global sparsity prior as a variation of the Graph–constrained Elastic–Net (GraphNet) for interpretable whole–brain decoding.

Recently, taking a new approach to the problem, Haufe and colleagues questioned the interpretability of weights of linear classifiers because of the contribution of noise in the decoding process [23, 83, 84]. To address this problem, they proposed a procedure to convert the linear brain decoding models into their equivalent generative models. Their experiments on the simulated and fMRI/EEG data illustrate that, whereas the direct interpretation of classifier weights may cause severe misunderstanding regarding the actual underlying effect, their proposed transformation effectively provides interpretable maps. Despite the theoretical soundness, the intricate challenge of estimating the empirical covariance matrix of the small–sample–size neuroimaging data [24] limits the practical application of this method.

3.1.3 The Gap: Formal Definition for Interpretability

In spite of the aforementioned efforts to improve the interpretability of brain decoding, there is still no formal definition for the interpretability of brain decoding in the literature. Therefore, the interpretability of different brain decoding methods are evaluated either qualitatively or indirectly (i.e., by means of an intermediate property). In qualitative evaluation, to show the superiority of one decoding method over the other (or a univariate map), the corresponding brain maps are compared visually in terms of smoothness, sparseness, and coherency using already known facts (see, for example, [195]). In the second approach, important factors in interpretability, such as spatio–temporal reproducibility, are evaluated to indirectly assess the interpretability of results (see for example Refs. [40,107,115,166]). Despite partial effectiveness, there is no general consensus regarding the quantification of these intermediate criteria. For example, in the case of spatial reproducibility, different methods such as correlation [107,166], dice score [115], or parameter variability [40,83] are used for quantifying the stability of brain maps, each of which considers different aspects of local or global reproducibility.

Although there is no formal definition for the interpretability of brain decoding models in the context of statistical learning theory, an overview of the brain decoding literature shows frequent co–occurrence of the terms interpretation, interpretable, and interpretability with the machine learning related terms such as model, classification, parameter, decoding, method, feature, and pattern. To experimentally illustrate this fact, we performed a meta–analysis on 101 papers sampled from the decoding–related studies. The AntConc [1] software was used for corpus analysis. Considering “interpretability”, “interpretable”, and “interpretation” as target words, three experiments were conducted:

1. In the first experiment, the frequency of target words were computed in the corpus. A total number of 598 hits were reported which shows on average ~ 6 hits per article. This observation confirms the pervasive usage of these terms in the machine learning and brain decoding contexts.

2. In the second experiment, the co–occurrence frequency of target words

[1] Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

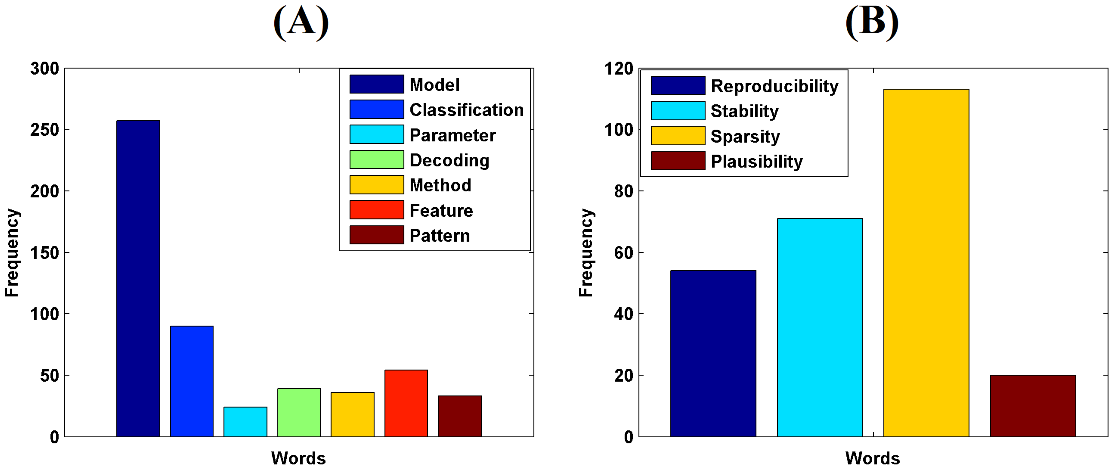


Figure 3.1: **(A)** The local co–occurrence rate of target words and machine learning related words. **(B)** The global co–occurrence rate of target words and common intuitive definitions of interpretability in brain decoding.

with machine learning related terms, such as “model”, “classification”, “parameter”, “decoding”, “method”, “feature”, and “pattern”, were counted. In order to assess the local co–occurrence, the co–occurrence window was defined from 10 words before to 10 words after the target words. Figure 3.1**(A)** summarizes the result. The local co–occurrence of target words with machine learning related terms shows the fact that they are repeatedly used to assess/explain/discuss the decoding models or their parameters. Further, the high frequency of co–occurrence with “model” illustrates the fact that talking about an “interpretable model” is very common in this context.

3. In the third experiment, the co–occurrence frequency of target words with terms “reproducibility”, “stability”, “sparsity”, and “plausibility” were counted. These terms present some commonly used intuitive explanations for interpretable models. In this case since we were interested in the global frequency of co–occurrence, the co–occurrences window is defined from 100 words before to 100 words after the tar-

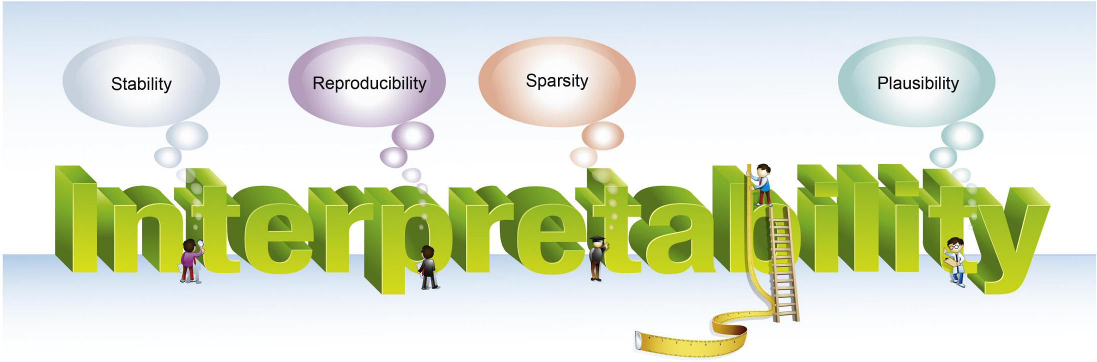


Figure 3.2: The high co–occurrence rate between the term "Interpretability" with a variety of concepts such as "Stability", "Reproducibility", "Sparsity", and "Plausibility" shows that there is no consensus over its definition and quantification.

get words. Figure 3.1**(B)** summarizes the result. The global co–occurrence of target words with these terms can be interpreted as an index on how they are connected in the literature. For example, the higher co–occurrence rate between the target words and "Sparsity" shows the fact that the more sparse models are well–accepted to be more interpretable models in the decoding studies.

3.1.4 The Contribution

With the aim of filling the aforementioned gap, our contribution is three–fold: 1) Assuming that the true solution of brain decoding is available, we present a theoretical definition of the interpretability. The presented definition is simply based on cosine proximity in the parameter space. Furthermore, we show that the interpretability can be decomposed into the reproducibility and representativeness of brain maps. 2) As a proof of the concept, we exemplify a practical heuristic based on event–related fields for quantifying the interpretability of brain maps in time–locked analysis of MEG data. 3) Finally, we propose the combination of the interpretability and the performance of the brain decoding as a new Pareto–optimal multi–objective criterion for model selection. We experimentally, on both simulated and real data, show that incorporating the interpretability into the model selection procedure provides more reproducible, more neurophysiologically plausible, and (as a result) more interpretable maps. Furthermore, in comparison with a standard univariate analysis, we show that the proposed paradigm offers more sensitivity while preserving the interpretability of results.

3.2 Materials and Methods

3.2.1 Notation and Background

Let $\mathcal{X} \in \mathbb{R}^p$ be a manifold in Euclidean space that represents the input space and $\mathcal{Y} \in \mathbb{R}$ be the output space, where $\mathcal{Y} = \Phi^*(\mathcal{X})$. Then, let $S = \{\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \mid z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ be a training set of n independently and identically distributed (i.i.d) samples drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ based on an unknown Borel probability measure ρ. In the neuroimaging context, $\mathbf{X}$ indicates the trials of brain recording, e.g., fMRI, MEG, or EEG signals, $\mathbf{Y}$ represents the experimental conditions or dependent variables, and we have $\Phi_S : \mathbf{X} \rightarrow \mathbf{Y}$ (note the difference between Φ_S and Φ^*). The goal of brain decoding is to find the function $\hat{\Phi} : \mathbf{X} \rightarrow \mathbf{Y}$ as an estimation of Φ_S. From here on, we refer to $\hat{\Phi}$ as a brain decoding model.

As is a common assumption in the neuroimaging context, we assume that the true solution of a brain decoding problem is among the family of linear functions $\mathcal{H}$ ($\Phi^* \in \mathcal{H}$). Therefore, the aim of brain decoding reduces to finding an empirical approximation of Φ_S, indicated by $\hat{\Phi}$, among all $\Phi \in \mathcal{H}$. This approximation can be obtained by estimating the predictive conditional density $\rho(\mathbf{Y} \mid \mathbf{X})$ by training a parametric model $\rho(\mathbf{Y} \mid \mathbf{X}, \Theta)$

(i.e., a likelihood function), where Θ denotes the parameters of the model. Alternatively, Θ can be estimated by solving a risk minimization problem:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \, \mathcal{L}(\mathbf{X}\Theta, \mathbf{Y}) + \lambda\Omega(\Theta) \tag{3.1}$$

where $\hat{\Theta}$ is the parameter of $\hat{\Phi}$, $\mathcal{L} : \mathbf{Y} \times \mathbf{Y} \to \mathbb{R}_0^+$ is the loss function, $\Omega : \mathbb{R}^p \to \mathbb{R}^+$ is the regularization term, and λ is a hyper–parameter that controls the amount of regularization. There are various choices for Ω, each of which reduces the hypothesis space $\mathcal{H}$ to $\mathcal{H}' \subset \mathcal{H}$ by enforcing different prior functional or structural constraints on the parameters of the linear decoding model (see, for example, [97,185,186,223]). The amount of regularization λ is generally decided using cross–validation or other data perturbation methods in the model selection procedure.

In the neuroimaging context, the estimated parameters of a linear decoding model $\hat{\Theta}$ can be used in the form of a brain map so as to visualize the discriminative neurophysiological effect. Although the magnitude of $\hat{\Theta}$ (i.e., the 2nd-norm of $\hat{\Theta}$) is affected by the dynamic range of data and the level of regularization, it has no effect on the predictive power and the interpretability of maps. On the other hand, the direction of $\hat{\Theta}$ affects the predictive power and contains information regarding the importance of and relations among predictors. This type of relational information is very useful when interpreting brain maps in which the relation between different spatio–temporal independent variables can be used to describe how different brain regions interact over time for a certain cognitive process. Therefore, we refer to the normalized parameter vector of a linear brain decoder in the unit hyper–sphere as a multivariate brain map (MBM); we denote it by $\vec{\Theta}$ where $\vec{\Theta} = \frac{\Theta}{\|\Theta\|_2}$ ($\|.\|_2$ represents the 2nd-norm of a vector).

As shown in Eq. 3.1, learning occurs using the sampled data. In other words, in the learning paradigm, we attempt to minimize the loss function with respect to Φ_S (and not Φ^*) [43]. Therefore, all of the implicit assumptions (such as linearity) regarding Φ^* might not hold on Φ_S, and vice versa. The *irreducible error* ε is the direct consequence of sampling; it sets a lower bound on the error, where we have:

$$\Phi_S(\mathbf{X}) = \Phi^*(\mathbf{X}) + \varepsilon. \tag{3.2}$$

The distribution of ε dictates the type of loss function $\mathcal{L}$ in Eq. 3.1. For example, assuming a Gaussian distribution with mean 0 and variance σ^2 for ε implies the least–squares loss function [211].

3.2.2 Interpretability of Multivariate Brain Maps: Theoretical Definition

In this section, we present a theoretical definition for the interpretability of linear brain decoding models and their associated MBMs. Consider a linearly separable brain decoding problem in an ideal scenario where $\varepsilon = 0$ and $rank(\mathbf{X}) = p$. In this case, the ideal solution of brain decoding, Φ^*, is linear and its parameters Θ^* are *unique* and neurophysiologically *plausible* [191]. The unique parameter vector Θ^* can be computed as

$$\Theta^* = \Sigma_{\mathbf{X}}^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.3}$$

where $\Sigma_{\mathbf{X}}$ represents the covariance of $\mathbf{X}$. Using Θ^* as the reference, we define the *strong–interpretability* of an MBM as follows:

Definition 1. *An MBM $\vec{\hat{\Theta}}$ associated with a linear brain decoding model $\hat{\Phi}$ is "strongly–interpretable" if and only if $\vec{\hat{\Theta}} \propto \Theta^*$.*

It can be shown that, in practice, the estimated solution of a linear brain decoding problem is not strongly–interpretable because of the inherent limitations of neuroimaging data, such as uncertainty [5] in the input and output space ($\varepsilon \neq 0$), the high dimensionality of data ($n \ll p$), and the high correlation between predictors ($rank(\mathbf{X}) < p$). With these limitations in mind, even though in practice the solution of linear brain decoding is not strongly–interpretable, one can argue that some are more interpretable than others. For example, in the case in which $\Theta^* \propto [0, 1]^T$, a linear classifier where $\vec{\hat{\Theta}} \propto [0.1, 1.2]^T$ can be considered more interpretable than a linear classifier where $\vec{\hat{\Theta}} \propto [2, 1]^T$. This issue raises the following question:

Problem 1. *Let S be a training set of n i.i.d samples drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $P(S)$ be the probability of drawing a certain S from $\mathcal{Z}$. Assume $\vec{\hat{\Theta}}$ is the MBM of a linear brain decoding model $\hat{\Phi}$ on S (estimated using Eq. 3.1 for a certain loss function $\mathcal{L}$, regularization term Ω, and hyper–parameter λ). How can we quantify the proximity of $\hat{\Phi}$ to the strongly–intrepretable solution of the brain decoding problem Φ^*?*

To answer this question, considering the uniqueness and the plausibility of Φ^* as the two main characteristics that convey its strong–interpretability, we define the interpretability as follows:

Definition 2. *Let S, $P(S)$, and $\vec{\hat{\Theta}}$ be as defined in Problem 1. Then, assume α be the angle between $\vec{\hat{\Theta}}$ and $\vec{\Theta}^*$. The "interpretability" ($0 \leq \eta_\Phi \leq 1$) of a linear brain decoding model $\hat{\Phi}$ is defined as follows:*

$$\eta_\Phi = \mathbb{E}_{P(S)}[\cos(\alpha)] \tag{3.4}$$

In practice, only a limited number of samples are available. Therefore, perturbation techniques are used to imitate the sampling procedure. Let $S^1, \ldots, S^m$ be m perturbed training sets drawn from S via a certain perturbation scheme such as jackknife, bootstrapping [51], or cross–validation [111]. Assume $\vec{\hat{\Theta}}^1, \ldots, \vec{\hat{\Theta}}^m$ are m MBMs estimated on the corresponding perturbed training sets, and α^j ($j = 1, \ldots, m$) be the angle between $\vec{\hat{\Theta}}^j$ and $\vec{\Theta}^*$. Then, the empirical version of Eq. 3.4 can be rewritten as

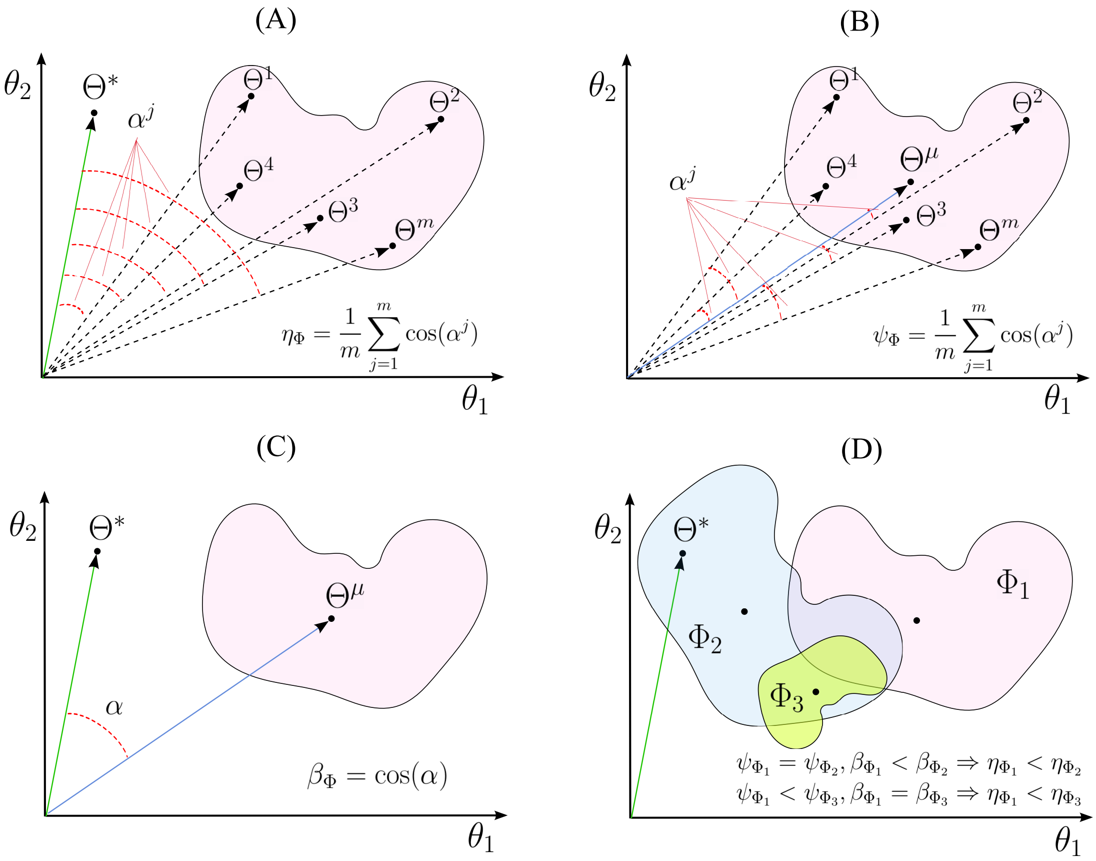

(A)
θ_2
Θ^*
α^j
Θ^1
Θ^2
Θ^4
Θ^3
Θ^m
$\eta_\Phi = \frac{1}{m}\sum_{j=1}^{m}\cos(\alpha^j)$
θ_1
(B)
θ_2
α^j
Θ^1
Θ^2
Θ^4
Θ^μ
Θ^3
Θ^m
$\psi_\Phi = \frac{1}{m}\sum_{j=1}^{m}\cos(\alpha^j)$
θ_1
(C)
θ_2
Θ^*
Θ^μ
α
$\beta_\Phi = \cos(\alpha)$
θ_1
(D)
θ_2
Θ^*
Φ_1
Φ_2
Φ_3
$\psi_{\Phi_1} = \psi_{\Phi_2}, \beta_{\Phi_1} < \beta_{\Phi_2} \Rightarrow \eta_{\Phi_1} < \eta_{\Phi_2}$
$\psi_{\Phi_1} < \psi_{\Phi_3}, \beta_{\Phi_1} = \beta_{\Phi_3} \Rightarrow \eta_{\Phi_1} < \eta_{\Phi_3}$
θ_1

Figure 3.3: A schematic illustrations for **(A)** interpretability (η_Φ), **(B)** reproducibility (ψ_Φ), and **(C)** representativeness (β_Φ) of a linear decoding model in two dimensions. **(D)** The independent effects of the reproducibility and the representativeness of a model on its interpretability.

$$\eta_\Phi = \frac{1}{m}\sum_{j=1}^{m}\cos(\alpha^j). \tag{3.5}$$

Empirically, the interpretability is the mean of cosine similarities between Θ^* and MBMs derived from different samplings of the training set (see Figure 3.3**(A)** for a schematic illustration).

In addition to the fact that employing cosine similarity is a common method for measuring the similarity between vectors, we have another strong motivation for this choice which is elaborated in the next section.

Distribution of Cosine Similarity

It can be shown that, for large values of p, the distribution of the dot product in the unit hyper–sphere, i.e., the cosine similarity, converges to a normal distribution with 0 mean and variance of $\frac{1}{p}$, i.e., $\mathcal{N}(0, \sqrt{1/p})$. Due to the small variance for large enough p values, any similarity value that is significantly larger than zero represents a meaningful similarity between two high–dimensional vectors. In order to analytically demonstrate this fact, we first need to find the distribution of dot product in the uniform unit hyper–sphere. Let a and b be two uniformly–drawn random vectors from a unit hyper–sphere in $\mathbb{R}^p$. Assuming that γ is the angle between a and b, the distribution of cosine similarity is equivalent to the dot product $< a, b >$. Without loss of generality, let b be along the positive x–axis in the coordinate system. Thus, the dot product $< a, b >$ is the projection of a on the x–axis, i.e., x coordinate of a. Therefore, for a certain value of γ, the dot product is a $p-1$–dimensional hyper–sphere that is orthogonal to the x–axis (the red circle in Figure 3.4) and the PDF of the dot product is the surface area of p dimensional hyper–sphere constructed by the dot products for different γ values (the dashed blue sphere in Figure 3.4). To compute the area of this hyper–sphere we take the sum of the surface area of the p dimensional conical frustums over small intervals dx (the gray area in Figure 3.4):

$$Pr(-1 \leq x \leq 1) = 2^{p-2}\pi \int_{-1}^{1} (1-x^2)^{p-2} \frac{dx}{1-x^2} = 2^{p-2}\pi \int_{-1}^{1} (1-x^2)^{p-3} dx \tag{3.6}$$

where $(1-x^2)^{p-2}$ is the surface area of the base of the cone (e.g., the perimeter of the red circle in Figure 3.4) and $\frac{dx}{1-x^2}$ is the slope size. Setting

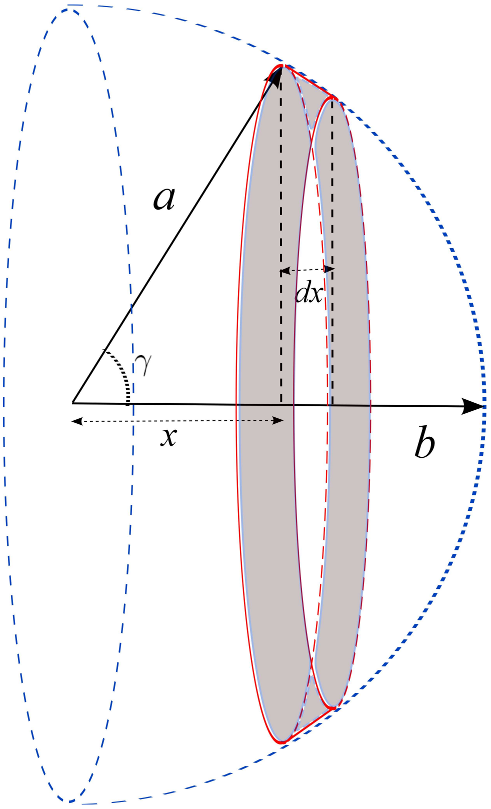


Figure 3.4: Two–dimensional geometrical illustration for computing the PDF of cosine similarity.

$t = \frac{x+1}{2}$ we have:

$$Pr(0 \leq t \leq 1) = 4^{p-2}\pi \int_0^1 t^{\frac{p-3}{2}}(1-t)^{\frac{p-3}{2}}dt \tag{3.7}$$

which is a Beta distribution, where $\alpha = \beta = \frac{p-1}{2}$, i.e., is a symmetric and unimodal distribution with mean 0.5. Because the PDF of $x = 2t - 1$ can be computed using a linear transformation of the above density function, it can be shown that the distribution of the dot product in unit hyper–sphere, i.e., the cosine similarity, has also a symmetric and unimodal distribution with zero mean. Based on the asymptotic assumption of Ref. [176], for large values of p this distribution converges to a normal distribution with $\sigma^2 = \frac{1}{p}$. Therefore, assuming large p, the distribution of cosine similarity for uniformly random vectors drawn from p–dimensional unit hyper–sphere is approximately $\mathcal{N}(0, \sqrt{\frac{1}{p}})$ (see Appendix A.2 for an experimental demonstration).

In what follows, we demonstrate how the definition of interpretability is geometrically related to the uniqueness and plausibility characteristics of the true solution of the brain decoding problem.

3.2.3 Interpretability Decomposition into Reproducibility and Representativeness

The trustworthiness and informativeness of decoding models provide two important motivations for improving the interpretability of models [123]. The trust of a learning algorithm refers to its ability to converge to a unique solution. On the other hand, the informativeness refers to the level of plausible information that can be derived from a model to assist or advise a human expert. Therefore, it is expected that the interpretability can be quantified alternatively by assessing its uniqueness and neurophysiological plausibility. In this section, we firstly define the reproducibility and representativeness as measures for quantifying the uniqueness and neurophysiological plausibility of a brain decoding model, respectively. Then we show how these definitions are related to the definition of interpretability.

The high dimensionality and the high correlations between variables are two inherent characteristics of neuroimaging data that negatively affect the uniqueness of the solution of a brain decoding problem. Therefore, a certain configuration of hyper–parameters may result in different estimated parameters on different portions of data. Here, we are interested in assessing this variability as a measure for uniqueness. We first define the *main multivariate brain map* as follows:

Definition 3. *Let S, $P(S)$, and $\vec{\hat{\Theta}}$ be as defined in Problem 1. The "main* *multivariate brain map" $\vec{\Theta}^{\mu} \in \mathbb{R}^{p}$ of a linear brain decoding model $\hat{\Phi}$ is defined as:*

$$\vec{\Theta}^{\mu} = \frac{\mathbb{E}_{P(S)}[\vec{\hat{\Theta}}]}{\left\|\mathbb{E}_{P(S)}[\vec{\hat{\Theta}}]\right\|_2}. \tag{3.8}$$

Assuming θ_i^j be the ith ($i = 1, \dots, p$) element of an MBM estimated on the jth ($j = 1, \dots, m$) perturbed training set, $\vec{\Theta}^{\mu}$ empirically can be estimated by summing up $\vec{\hat{\Theta}}^j$s (computed on the perturbed training set S^j) in the unit hyper–sphere, and we have:

$$\vec{\Theta}^{\mu} = \frac{\left[\sum_{j=1}^{m} \theta_1^j \;\; \sum_{j=1}^{m} \theta_2^j \;\; \dots \;\; \sum_{j=1}^{m} \theta_p^j\right]^T}{\left\|\left[\sum_{j=1}^{m} \theta_1^j \;\; \sum_{j=1}^{m} \theta_2^j \;\; \dots \;\; \sum_{j=1}^{m} \theta_p^j\right]^T\right\|_2}. \tag{3.9}$$

The main multivariate brain map, $\vec{\Theta}^{\mu}$, provides a reference for quantifying the reproducibility of an MBM:

Definition 4. *Let S, $P(S)$, and $\vec{\hat{\Theta}}$ be as defined in Problem 1, and $\vec{\Theta}^{\mu}$ be the main multivariate brain map of $\hat{\Phi}$. Then, assume α be the angle between $\vec{\hat{\Theta}}^j$ and $\vec{\Theta}^{\mu}$. The "reproducibility" ψ_{Φ} ($0 \leq \psi_{\Phi} \leq 1$) of a linear brain decoding model $\hat{\Phi}$ is defined as*

$$\psi_{\Phi} = \mathbb{E}_{P(S)}[\cos(\alpha)]. \tag{3.10}$$

Let $\vec{\hat{\Theta}}^1, \dots, \vec{\hat{\Theta}}^m$ are m MBMs estimated on the corresponding perturbed training sets, and α^j ($j = 1, \dots, m$) be the angle between $\vec{\hat{\Theta}}^j$ and $\vec{\Theta}^{\mu}$. Then, the empirical version of Eq. 3.10 can be rewritten as

$$\psi_{\Phi} = \frac{1}{m} \sum_{j=1}^{m} \cos(\alpha^j). \tag{3.11}$$

In fact, reproducibility provides a measure for quantifying the dispersion of MBMs, computed over different perturbed training sets, from the main multivariate brain map. Figure 3.3(**B**) shows a schematic illustration for the reproducibility of a linear brain decoding model.

On the other hand, the similarity between the main multivariate brain map of a decoder and the true solution can be employed as a measure for the neurophysiological plausibility of a model. We refer to this similarity as the *representativeness* of a linear brain decoding model:

Definition 5. *Let* $\vec{\Theta}^{\mu}$ *be the main multivariate brain map of* $\hat{\Phi}$*. The "representativeness"* β_{Φ} $(0 \leq \beta_{\Phi} \leq 1)$ *of a linear brain decoding model* $\hat{\Phi}$ *is defined as the cosine similarity between its main multivariate brain map* $(\vec{\Theta}^{\mu})$ *and the parameters of the true solution* $(\vec{\Theta}^{*})$*,*

$$\beta_{\Phi} = \frac{|\vec{\Theta}^{\mu}.\vec{\Theta}^{*}|}{\left\|\vec{\Theta}^{\mu}\right\|_2 \left\|\vec{\Theta}^{*}\right\|_2}. \tag{3.12}$$

Figure 3.3(**C**) schematically illustrates the definition of representativeness.

As discussed before, the notion of interpretabilty is tightly related to the uniqueness and plausibility, and thus to the reproducibility and representativeness, of a decoding model. The following proposition analytically shows this relationship:

Proposition 1. $\eta_{\Phi} = \beta_{\Phi} \times \psi_{\Phi}$.

Proof. Throughout this proof, we assume that all of the parameter vectors are normalized in the unit hypersphere (see Figure 3.5 as an illustrative example in 2 dimensions). Let $T = \{\vec{\hat{\Theta}}^1, \ldots, \vec{\hat{\Theta}}^m\}$ be a set m MBMs, for m perturbed training sets where $\vec{\hat{\Theta}}^i \in \mathbb{R}^p$. Now, consider any arbitrary $p-1$-dimensional hyperplane $\mathcal{A}$ that contains $\vec{\Theta}^{\mu}$. Clearly, $\mathcal{A}$ divides the p-dimensional parameter space into 2 subspaces. Let $\triangledown$ and $\blacktriangledown$ be binary operators where $\vec{\Theta}^i \nabla \vec{\Theta}^k$ indicates that $\vec{\Theta}^i$ and $\vec{\Theta}^k$ are in the same subspace, and $\vec{\Theta}^i \blacktriangledown \vec{\Theta}^k$ indicates that they are in different subspaces. Now, we define $T_U = \{\vec{\Theta}^i \mid \vec{\Theta}^i \nabla \vec{\Theta}^*\}$ and $T_L = \{\vec{\Theta}^i \mid \vec{\Theta}^i \blacktriangledown \vec{\Theta}^*\}$. Let the cardinality of T_L denoted by $n(T_L)$ be j $(n(T_L) = j)$. Thus, $n(T_U) = m - j$. Now, assume that $\measuredangle(\vec{\hat{\Theta}}^i, \mathcal{A}) = \alpha_1, \ldots, \alpha_j$ are the angles between $\vec{\hat{\Theta}}^i \in T_L$ and $\mathcal{A}$, and (similarly) $\alpha_{j+1}, \ldots, \alpha_m$ for $\vec{\hat{\Theta}}^i \in T_U$ and $\mathcal{A}$. Based on Eq. 3.8, let $\vec{\Theta}^\mu_L$ and $\vec{\Theta}^\mu_U$ be the main maps of T_L and T_U, respectively. Therefore, we obtain $\vec{\Theta}^\mu = \frac{\vec{\Theta}^\mu_L + \vec{\Theta}^\mu_U}{\|\vec{\Theta}^\mu_L + \vec{\Theta}^\mu_U\|}$ and $\measuredangle(\vec{\Theta}^\mu_L, \mathcal{A}) = \measuredangle(\vec{\Theta}^\mu_U, \mathcal{A}) = \alpha$. Furthermore, assume $\measuredangle(\vec{\Theta}^*, \mathcal{A}) = \gamma$. As a result, $\psi_\Phi = \cos(\alpha)$ and $\beta_\Phi = \cos(\gamma)$. According to Eq. 3.4 and using a cosine similarity definition, we have:

$$\begin{aligned}
\eta_\Phi &= \frac{1}{m} \sum_{j=1}^{m} \left| \vec{\Theta}^* . \vec{\hat{\Theta}}^j \right| \\
&= \frac{\cos(\gamma + \alpha_1) + \cdots + \cos(\gamma + \alpha_j) + \cos(\gamma - \alpha_{j+1}) + \cdots + \cos(\gamma - \alpha_m)}{m} \\
&= \frac{\cos(\gamma + \alpha) + \cos(\gamma - \alpha)}{2} \\
&= \frac{\cos(\gamma)\cos(\alpha) - \sin(\gamma)\sin(\alpha) + \cos(\gamma)\cos(\alpha) + \sin(\gamma)\sin(\alpha)}{2} \\
&= \cos(\gamma)\cos(\alpha) = \beta_\Phi \times \psi_\Phi .
\end{aligned} \tag{3.13}$$

A similar procedure can be used to prove $\tilde{\eta}_\Phi = \tilde{\beta}_\Phi \times \psi_\Phi$ by replacing $\vec{\Theta}^*$ with $\vec{\Theta}^{cERF}$ (see Section 3.2.4 for the definition of $\vec{\Theta}^{cERF}$). □

Proposition 1 indicates that the interpretability of a linear brain decoding model can be decomposed into its representativeness and reproducibility. Figure 3.3**(D)** illustrates how the reproducibility and the representativeness of a decoding model independently affect its interpretability. Each colored region schematically represents a span of different solutions of the a certain linear model (for example, with a certain configuration for its hyper–parameters) on different perturbed training sets. The area of each

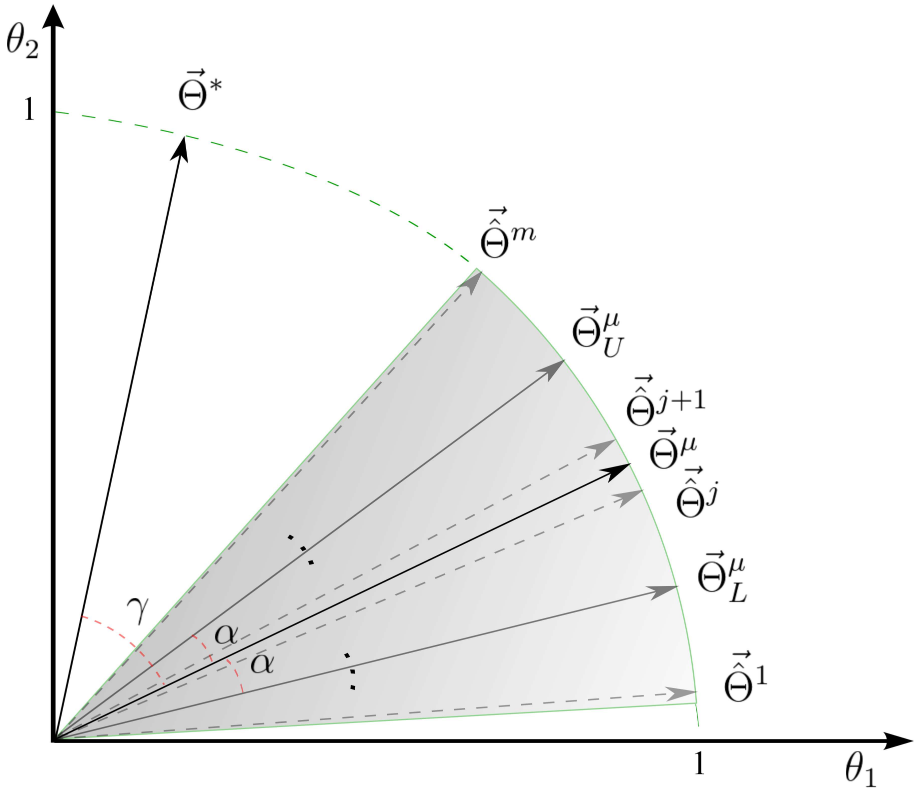


Figure 3.5: Relation between representativeness, reproducibility, and interpretability in 2 dimensions.

region schematically visualizes the reproducibility of each model, i.e., the less is the area, the higher is the reproducibility of a model. Further, the angular distance between the centroid of each region (Θ^μ) and the true solution (Θ^*) visualizes the representativeness of each corresponding model. While Φ_1 and Φ_2 have similar reproducibility, Φ_2 has higher interpretability than Φ_1 because it is more representative of the true solution. On the other hand, Φ_1 and Φ_3 have similar representativeness, however, Φ_3 is more interpretable due to the higher level of reproducibility.

3.2.4 A Heuristic for Practical Quantification of Interpretability in Time–Locked Analysis of MEG Data

In practice, it is impossible to evaluate the interpretability, as the true solution of the brain decoding problem Φ^* is unknown. In this study, to provide a practical proof of the theoretical concepts, we exemplify contrast event–related field (cERF) (see Eq. 3.14 for the definition) as a neurophysiological plausible heuristic for the true parameters of the linear brain decoding problem (Θ^*) in a binary MEG decoding scenario in time domain. Due to the nature of proposed heuristic, its application is limited to the brain responses that are time–locked to the stimulus onset, i.e., the evoked responses.

The MEEG data are a mixture of several simultaneous stimulus–related and stimulus–unrelated brain activitions. Assessing the electro/magneto–physiological changes that are time–locked to events of interest is a common approach in analyzing MEEG data. In general, background brain activity is considered Gaussian noise with zero mean and variance σ^2. One popular approach to canceling the noise component is to compute the average of multiple trials. The assumption is that, when the effect of interest is time–locked to the stimulus onset, the independent noise components can be vanished by means of averaging. It is expected that the average will converge to the true value of the signal with a variance of $\frac{\sigma^2}{n}$ (where n is the number of trials). The result of the averaging process consist of a series of positive and negative peaks occurring at a fixed time relative to the event onset, generally known as ERF in the MEG context. These peaks reflect phasic activity that are indexed with different aspects of cognitive processing [171][2].

[2] The application of the presented heuristic to MEG data can be extended to EEG because of the inherent similarity of the measured neural signals in these two devices. In the EEG context, the ERF can be replaced by the event–related potential (ERP).

Assume $\mathbf{X}^+ = \{x_i \in \mathbf{X} \mid y_i = 1\} \in \mathbb{R}^{n^+ \times p}$ and $\mathbf{X}^- = \{x_i \in \mathbf{X} \mid y_i = -1\} \in \mathbb{R}^{n^- \times p}$ to be sets of positive and negative samples in a binary MEG decoding scenario. Then, the cERF brain map $\vec{\Theta}^{cERF}$ is computed by

$$\vec{\Theta}^{cERF} = \frac{\frac{1}{n^+}\sum_{x_i \in X^+} x_i - \frac{1}{n^-}\sum_{x_i \in X^-} x_i}{\left\|\frac{1}{n^+}\sum_{x_i \in X^+} x_i - \frac{1}{n^-}\sum_{x_i \in X^-} x_i\right\|_2}. \tag{3.14}$$

Generally speaking, $\vec{\Theta}^{cERF}$ is a contrast ERF map between two experimental conditions. Using the core theory presented in [83], the equivalent generative model for the solution of linear brain decoding, i.e., the activation pattern (A), is unique and we have

$$A \propto \Sigma_{\mathbf{X}} \hat{\Theta}. \tag{3.15}$$

Assuming $\hat{\Theta}$ to be the least–squares solution in a binary decoding scenario, the following proposition describes the relation between $\vec{\Theta}^{cERF}$ and the activation pattern A:

Proposition 2. $\vec{\Theta}^{cERF} \propto A$.

Proof. According to [83], for a linear discriminative model with parameters $\hat{\Theta}$, the unique equivalent generative model can be computed as

$$A \propto \Sigma_{\mathbf{X}} \hat{\Theta}. \tag{3.16}$$

In a binary ($\mathbf{Y} = \{1, -1\}$) least–squares classification scenario, we have

$$A \propto \Sigma_{\mathbf{X}} \Sigma_{\mathbf{X}}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y} = \mu^+ - \mu^- \tag{3.17}$$

where $\Sigma_{\mathbf{X}}$ represents the covariance of the input matrix $\mathbf{X}$, and μ^+ and μ^- are the means of positive and negative samples, respectively. Therefore, the equivalent generative model for the above classification problem can be derived by computing the difference between the mean of samples in two classes that is equivalent to the definition of cERF in time–domain MEG data. □

Proposition 2 shows that, in a binary time–domain MEG decoding scenario, cERF is proportional to the equivalent generative model for the solution of a least–squares classifier (see Appendix A.3 for an experimental support on real MEG data). Furthermore, $\vec{\Theta}^{cERF}$ is proportional to the t-statistic that is widely used in the univariate analysis of neuroimaging data. Using $\vec{\Theta}^{cERF}$ as a heuristic for $\vec{\Theta}^{*}$, the representativeness can be approximated as follows:

$$\tilde{\beta}_{\Phi} = \frac{|\vec{\Theta}^{\mu}.\vec{\Theta}^{cERF}|}{\left\|\vec{\Theta}^{\mu}\right\|_2 \left\|\vec{\Theta}^{cERF}\right\|_2} \tag{3.18}$$

where $\tilde{\beta}_{\Phi}$ is an approximation of the actual representativeness β_{Φ}. In a similar manner, $\vec{\Theta}^{cERF}$ can be used to heuristically approximate the interpretability as follows:

$$\tilde{\eta}_{\Phi} = \frac{1}{m}\sum_{j=1}^{m} \cos(\gamma^{j}) \tag{3.19}$$

where $\gamma_1, \ldots, \gamma_m$ are the angles between $\vec{\hat{\Theta}}^{1}, \ldots, \vec{\hat{\Theta}}^{m}$ and $\vec{\Theta}^{cERF}$. It can be shown that $\tilde{\eta}_{\Phi} = \tilde{\beta}_{\Phi} \times \psi_{\Phi}$ (see the proof of Proposition 1).

The proposed heuristic is only applicable to the evoked responses in sensor and source space MEEG data. Despite this limitation, cERF provides an empirical example that shows how the presented theoretical definitions can be applied in a real decoding scenario. The choice of the heuristic has a direct effect on the approximation of interpretability and that an inappropriate selection of the heuristic yields a very poor estimation of interpretability. Therefore, the choice of heuristic should be carefully justified based on accepted and well–defined facts regarding the nature of the collected data.

Since the labels are used in the computation of cERF, a proper validation strategy should be employed to avoid the double–dipping issue [113]. One possible approach is to exclude the entire test set from the model selection procedure using a nested cross–validation strategy. An alternative approach is to employ model–averaging techniques to neatly get advantage of the whole dataset [196]. Since our focus is on the model selection, in the remaining text we implicitly assume that the test data are excluded from the experiments; thus, all the experimental results are reported on the training and validation sets.

3.2.5 Incorporating the Interpretability into Model Selection

The procedure for evaluating the performance of a model so as to choose the best values for hyper–parameters is known as *model selection* [81]. This procedure generally involves numerical optimization of the model selection criterion on the training and validation sets (and not the test set). Let U be a set of hyper–parameters, then the goal of model selection procedure reduces to finding the best model configuration $u^* \in U$ that maximizes the model selection criterion (e.g., generalization performance) on the training set S. The most common model selection criterion is based on an estimator of generalization performance, i.e., the predictive power. In the context of brain decoding, especially when the interpretability of brain maps matters, employing predictive power as the only decisive criterion in model selection is problematic in terms of interpretability of MBMs [40, 63, 166, 196]. Valverde and Moreno [190] experimentally showed that in a classification task optimizing only classification error rate is insufficient to capture the transfer of crucial information from the input to the output of a classifier. This fact highlights the importance of having some control over the estimated model weights in the model selection. Here, we propose a multi–objective criterion for model selection that takes into account both prediction accuracy and MBM interpretability.

Let $\tilde{\eta}_\Phi$ and δ_Φ be the approximated interpretability and the generalization performance of a linear brain decoding model $\hat{\Phi}$, respectively. We propose the use of the *scalarization* technique [33] for combining $\tilde{\eta}_\Phi$ and δ_Φ into one scalar $0 \leq \zeta(\Phi) \leq 1$ as follows:

$$\zeta_\Phi = \begin{cases} \frac{\omega_1 \tilde{\eta}_\Phi + \omega_2 \delta_\Phi}{\omega_1 + \omega_2} & \delta_\Phi \geq \kappa \\ 0 & \delta_\Phi < \kappa \end{cases} \tag{3.20}$$

where ω_1 and ω_2 are weights that specify the level of importance of the interpretability and the performance, respectively. κ is a threshold on the performance that filters out solutions with poor performance. In classification scenarios, κ can be set by adding a small safe interval to the chance level of classification. The hyper–parameters that are optimized based on ζ_Φ are Pareto optimal [128]. We hypothesize that optimizing the hyper–parameters based on ζ_Φ, rather only δ_Φ, yields more informative MBMs.

Algorithm 1 summarizes the proposed model selection scheme. The model selection procedure receives the training set S and a set of possible configurations for hyper–parameters U, and returns the best hyper–parameter configuration u^*.

3.2.6 Experimental Materials

Toy Dataset

We regenerate the simple 2-dimensional toy data presented in [83]. Assume that the true underlying generative function Φ^* is defined by:

Algorithm 1 The model selection procedure.

```
procedure ModelSelection($S$,$U$)
    Compute $\vec{\Theta}^{cERF}$ on $S$.                                  ▷ using Eq. 3.14
    for all $u_i \in U$ do                         ▷ For all hyper–parameter configurations.
        for $j \leftarrow 1, m$ do                         ▷ Data perturbation iterations.
            Partition $S$ into training $S_{tr}$ and validation $S_{vl}$ subsets via a perturbation method.
            Compute $\hat{\Theta}_j$ on $S_{tr}$ using $u_i$ as the hyper–parameter.
        end
        Compute $\delta^i_\Phi$ of $\hat{\Theta}_j$s on $S_{vl}$.
        Compute $\tilde{\eta}^i_\Phi$ of $\hat{\Theta}_j$s using $\vec{\Theta}^{cERF}$.          ▷ using Eq. 3.19
        Compute $\zeta^i_\Phi$.                                             ▷ using Eq. 3.20
    end
    $u^* = \text{argmax}_{u_i \in U}(\zeta_\Phi)$.
    return $u^*$.
```

$$\mathcal{Y} = \Phi^*(\mathcal{X}) = \begin{cases} 1 & if \quad x_1 = 1.5 \\ -1 & if \quad x_1 = -1.5 \end{cases}$$

where $\mathcal{X} \in \{[1.5, 0]^T, [-1.5, 0]^T\}$; and x_1 and x_2 represent the first and the second dimension of the data, respectively. Furthermore, assume the data are contaminated by Gaussian noise with co–variance $\Sigma = \begin{bmatrix} 1.02 & -0.3 \\ -0.3 & 0.15 \end{bmatrix}$. Gaussian noise adds uncertainty to the input space.

Simulated MEG Data

We simulated two classes of MEG data, each of which composed of 250 epochs with length of 330 ms at 300 Hz sampling rate (so that we have 100 time–points). For simplicity, the whole scalp topography was simulated with a single dipole located at -4.7, -3.7, and 5.3 cm in the RAS (right, anterior, superior) coordinate system. The dipole was oriented toward [1,1,0] direction in the RA plane [see Figure 3.6(**A**)]. 102 magnetometer sensors of Elekta Neuromag [3] system were simulated using a standard forward model algorithm implemented in the Fieldtrip toolbox [153]. The epochs of the positive class were constructed by adding three components to the dipole

[3] See `https://www.elekta.com/diagnostic-solutions/elekta-neuromag-triux.html` for more information.

time–course: 1) a time–locked ERF effect with a positive 3 Hz followed by a negative 5 Hz half–cycle sinusoid peaks after 150 ± 10 and 250 ± 10 ms of the epoch onset, respectively; 2) uncorrelated background brain activity that was simulated by summing 50 sinusoids of random frequency from 1 to 125 Hz, and random phase between 0 and 2π. Following the data simulation procedure in [214], the amplitude of any single frequency component of the signal (the ERF effect and the background noise) was set based on the empirical spectral power of human brain activity to mimic the actual MEG signals; and 3) white Gaussian noise scaled with the root mean squared of the signal in each epoch. The epochs of the negative class were constructed without the ERF effect by adding up only the noise components (i.e., the background activity and the white noise). Therefore, the ERF component is considered as the discriminative ground–truth in our experiments [see Figure 3.6(**B**)].

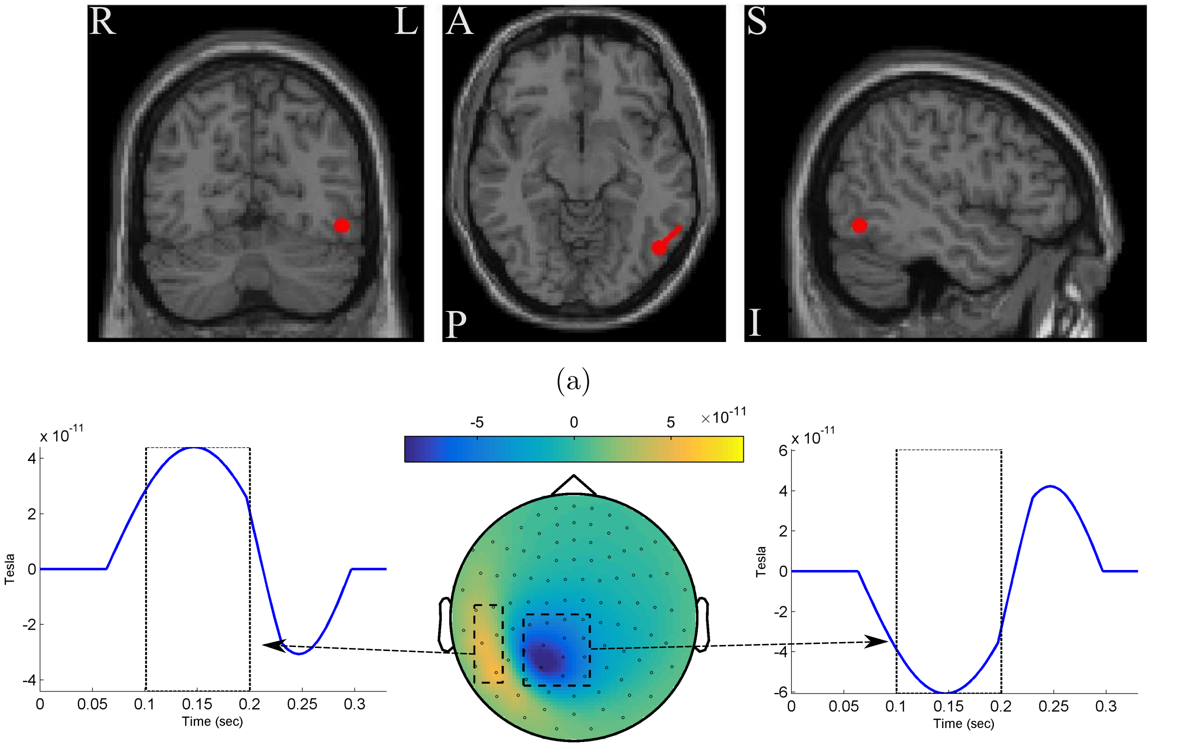
MEG Data

We used the MEG dataset presented in Ref. [90][4]. The dataset was also used for the DecMeg2014 competition[5]. In this dataset, visual stimuli consisting of famous faces, unfamiliar faces, and scrambled faces were presented to 16 subjects and fMRI, EEG, and MEG signals were recorded. Here, we are only interested in MEG recordings. The MEG data were recorded using a VectorView system (Elekta Neuromag, Helsinki, Finland) with a magnetometer and two orthogonal planar gradiometers located at 102 positions in a hemispherical array in a light Elekta–Neuromag magnetically shielded room.

Three major reasons motivated the choice of this dataset: 1) It is publicly available. 2) The spatio–temporal dynamic of the MEG signal for face

[4] The full dataset is publicly available at `ftp://ftp.mrc-cbu.cam.ac.uk/personal/rik.henson/wakemandg_hensonrn/`

[5] The competition data are available at `http://www.kaggle.com/c/decoding-the-human-brain`



(a)

(b)

Figure 3.6: **(A)** The red circles show the dipole position, and the red stick shows the dipole direction. **(B)** The spatio–temporal pattern of the discriminative ground–truth effect.

vs. scramble stimuli has been well studied. The event–related potential analysis of EEG/MEG shows that N170 occurs $130 - 200$ ms after stimulus presentation and reflects the neural processing of faces [19, 90]. Therefore, the N170 component can be considered the ground truth for our analysis. 3) In the literature, non–parametric mass–univariate analysis such as cluster–based permutation tests is unable to identify narrowly distributed effects in space and time (e.g., an N170 component) [64, 65]. These facts motivate us to employ multivariate approaches that are more sensitive to these effects.

Similar to Ref. [151], we created a balanced face vs. scrambled MEG dataset by randomly drawing from the trials of unscrambled (famous or unfamiliar) faces and scrambled faces in equal number. The samples in the face and scrambled face categories are labeled as 1 and -1, respectively. The raw data is high–pass filtered at 1 Hz, down–sampled to 250 Hz, and trimmed from 200 ms before the stimulus onset to 800 ms after the stimulus. Thus, each trial has 250 time–points for each of the 306 MEG sensors (102 magnetometers and 204 planar gradiometers)[6]. To create the feature vector of each sample, we pooled all of the temporal data of 306 MEG sensors into one vector (i.e., we have $p = 250 \times 306 = 76500$ features for each sample). Before training the classifier, all of the features are standardized to have a mean of 0 and standard–deviation of 1.

3.2.7 Classification and Evaluation

In all experiments, Lasso [185] classifier with ℓ_1 penalization was used for decoding. Lasso is a very popular classification method in the context of brain decoding, mainly because of its sparsity assumption. The choice of Lasso, as a simple model with only one hyper–parameter, helps us to better illustrate the importance of including the interpretability in the model selection. The solution of decoding is computed by solving the following optimization problem:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \, \mathcal{L}(\mathbf{X}\Theta, \mathbf{Y}) + \lambda \left\| \Theta \right\|_1 \tag{3.21}$$

where $\|.\|_1$ represents the ℓ_1-norm, and λ is the hyper–parameter that specifies the level of regularization. Therefore, the aim of the model selection is to find the best value for λ on the training set S. Here, we try to find the best regularization parameter value among $\lambda = \{0.001, 0.01, 0.1, 1, 10, 50, 100, 250, 500, 1000\}$.

[6] The preprocessing scripts in python and MATLAB are available at: `https://github.com/FBK-NILab/DecMeg2014/`

As a complementary experiment, we repeated the single–subject decoding on the real MEG data also using an elastic–net classifier [223]. Elastic–net combines ℓ_1 and ℓ_2 penalization methods. Thus it has two hyper–parameters, λ and α, to control the amount of regularization, and the weights on the types of penalization, respectively. We have:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \mathcal{L}(\mathbf{X}\Theta, \mathbf{Y}) + \lambda[\alpha \left\|\Theta\right\|_1 + (1-\alpha) \left\|\Theta\right\|_2^2] \tag{3.22}$$

where $\|.\|_1$ and $\|.\|_2$ represent ℓ_1-norm and ℓ_2-norm, respectively. Therefore, the aim of the model selection is to find the best value for both λ and α. Here, we try to find the best hyper–parameter values among $\lambda = \{0.001, 0.01, 0.1, 1, 10, 50, 100, 250, 500, 1000\}$ and $\alpha = \{0, 0.0001, 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 1\}$.

We used the out–of–bag (OOB) [29,210] method for computing δ_Φ, ψ_Φ, $\tilde{\beta}_\Phi$, $\tilde{\eta}_\Phi$, and ζ_Φ for different values of λ. In OOB, given a training set $(\mathbf{X}, \mathbf{Y})$, m replications of bootstrap [51] are used to create perturbed training and validation sets (we set $m = 50$) [7]. In all of our experiments, we set $\omega_1 = \omega_2 = 1$ and $\kappa = 0.6$ in the computation of ζ_Φ. Furthermore, we set $\delta_\Phi = 1 - EPE$ where EPE indicates the expected prediction error; it is computed using the procedure explained in Section 2.4.4. Employing OOB provides the possibility of computing the bias and variance of the model as contributing factors in EPE.

[7] The MATLAB code used for experiments is available at `https://github.com/smkia/interpretability/`

Table 3.1: Comparison between δ_Φ, η_Φ, and ζ_Φ for different λ values on the toy example shows the performance–interpretability dilemma, in which the most accurate classifier is not the most interpretable one.

λ	0	0.001	0.01	0.1	1	10	50	100	250	500	1000
$\delta(\Phi)$	0.9883	0.9883	0.9883	0.9883	0.9883	**0.9884**	0.9880	0.9840	0.9310	0.9292	0.9292
$\eta(\Phi)$	0.4391	0.4391	0.4391	0.4392	0.4400	0.4484	0.4921	0.5845	0.9968	**1**	**1**
$\zeta(\Phi)$	0.7137	0.7137	0.7137	0.7137	0.7142	0.7184	0.7400	0.7842	0.9639	**0.9646**	**0.9646**
$\hat{\vec{\Theta}} \propto$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4521\\0.8919\end{bmatrix}$	$\begin{bmatrix}0.4532\\0.8914\end{bmatrix}$	$\begin{bmatrix}0.4636\\0.8660\end{bmatrix}$	$\begin{bmatrix}0.4883\\0.8727\end{bmatrix}$	$\begin{bmatrix}0.5800\\0.8146\end{bmatrix}$	$\begin{bmatrix}0.99\\0.02\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$

3.3 Results

3.3.1 Performance–Interpretability Dilemma: A Toy Example

In the definition of Φ^* on the toy dataset discussed in Section 3.2.6, x_1 is the decisive variable and x_2 has no effect on the classification of samples into target classes. Therefore, excluding the effect of noise and based on the theory of the maximal margin classifier [194], $\vec{\Theta}^* \propto [1, 0]^T$ is the true solution to the decoding problem. By accounting for the effect of noise, solving the decoding problem in $(\mathbf{X}, \mathbf{Y})$ space yields $\hat{\vec{\Theta}} \propto [1/\sqrt{5}, 2/\sqrt{5}]^T$ as the parameters of the linear classifier. Although the estimated parameters on the noisy data provide the best generalization performance for the noisy samples, any attempt to interpret this solution fails, as it yields the wrong conclusion with respect to the ground truth (it says x_2 has twice the influence of x_1 on the results, whereas it has no effect). This simple experiment shows that the most accurate model is not always the most interpretable one, primarily because the contribution of the noise in the decoding process [83]. On the other hand, the true solution of the problem $\vec{\Theta}^*$ does not provide the best generalization performance for the noisy data.

To illustrate the effect of incorporating the interpretability in the model selection, a Lasso model with different λ values is used for classifying the toy data. In this example, because $\vec{\Theta}^*$ is known, the exact value of in-

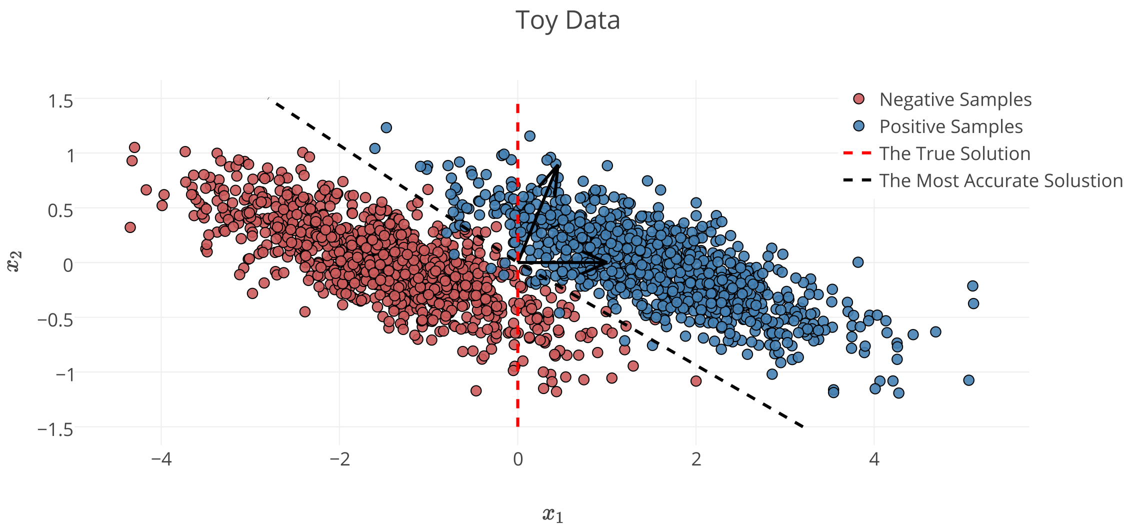


Figure 3.7: Noisy samples of toy data. The dotted line shows the true separator based on the generative model (Φ^*). The dashed line shows the most accurate classification solution. Because of the contribution of noise, any interpretation of the parameters of the most accurate classifier yields a misleading conclusion with respect to the true underlying phenomenon [83].

terpretability can be computed using Eq. 3.5. Table 3.1 compares the resultant performance and interpretability from Lasso. Lasso achieves its highest performance ($\delta_\Phi = 0.9884$) at $\lambda = 10$ with $\vec{\tilde{\Theta}} \propto [0.4636, 0.8660]^T$ (indicated by the black dashed line in Figure 3.7). Despite having the highest performance, this solution suffers from a lack of interpretability ($\eta_\Phi = 0.4484$). By increasing λ, the interpretability improves so that for $\lambda = 500, 1000$ the classifier reaches its highest interpretability by compensating for 0.06 of its performance. Our observation highlights two main points:

1. In the case of noisy data, the interpretability of a decoding model can be possibly incoherent with its performance. Thus, optimizing the parameter of the model based on its performance does not necessarily improve its interpretability. This observation confirms the previous finding by Rasmussen et al. [166] regarding the trade–off between the

spatial reproducibility (as a measure for the interpretability) and the prediction accuracy in brain decoding.

2. If the right criterion is used in the model selection, employing proper regularization technique (sparsity prior, in the case of toy data) leads to more interpretable decoding models.

3.3.2 Decoding on Simulated MEG Data

With the main aim of comparing the quality of the heuristically approximated interpretability with respect to its actual value, we solve the decoding problem on the simulated MEG data where the ground–truth discriminative effect is known. The ground truth effect $\vec{\Theta}^*$ is used to compute the actual interpretability of the decoding model. On the other hand, interpretability is approximated by means of $\vec{\Theta}^{cERF}$. The whole data simulation and decoding processes are repeated 25 times and the results are summarized in Figure 3.8. Figure 3.8**(A)** and 3.8**(B)** show the actual (η_Φ) and the approximated ($\tilde{\eta}_\Phi$) interpretability for different λ values. Even though $\tilde{\eta}_\Phi$ consistently overestimates η_Φ, there is a significant co–variation (Pearson's correlation p-value $= 9 \times 10^{-4}$) between two measures as λ increases. Thus, despite overestimation problem of the heuristically approximated interpretability values, they are still reliable measures for quantitative comparison between interpretability level of brain decoding models with different hyper–parameters. For example, both η_Φ and $\tilde{\eta}_\Phi$ suggest the decoding model with $\lambda = 50$ as the most interpretable model.

Figure 3.8**(C)** shows brain decoding models at $\lambda = 10$ and $\lambda = 50$ yield equivalent generalization performances (Wilcoxon rank sum test p-value $= 0.08$), while the MBM resulted from $\lambda = 50$ has significantly higher interpretability (Wilcoxon rank sum test p-value $= 4 \times 10^{-9}$). The advantage of this difference in interpretability levels is visualized in Figure 3.9 where

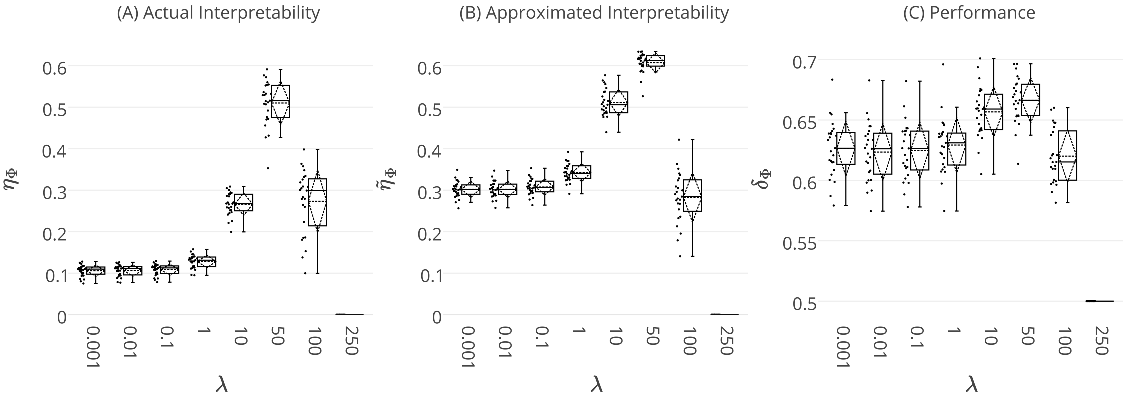


Figure 3.8: **(A)** The actual η_Φ, and **(B)** the heuristically approximated interpretability $\tilde{\eta}_\Phi$ of decoding models across different λ values. There is a significant co–variation (Pearson's correlation p-value $= 9 \times 10^{-4}$) between η_Φ and $\tilde{\eta}_\Phi$. **(C)** The generalization performance of decoding models. The box gives the quartiles, while the whiskers give the 5 and 95 percentiles.

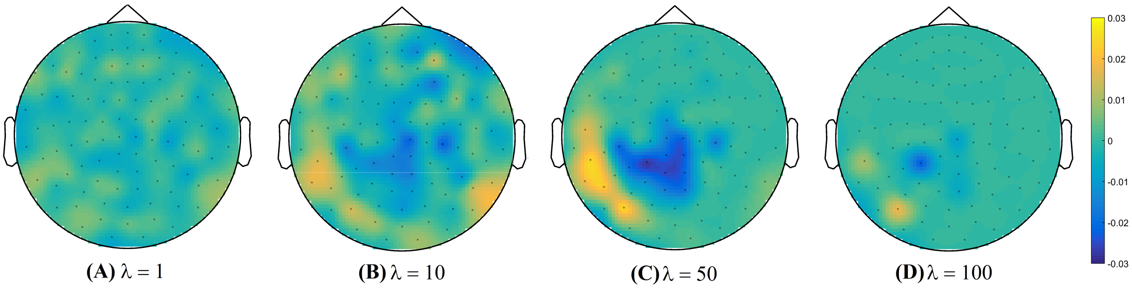


Figure 3.9: Topographic maps of weights of brain decoding models with different λ values.

topographic maps are plotted for the weights of brain decoding models with different λ values by averaging the classifier weights in the time interval of 100 to 200 ms. The visual comparison shows MBM at $\lambda = 50$ is more similar to the ground–truth map [see Figure 3.6**(B)**] than the MBMs computed at other λ values. This superiority is well–reflected in the corresponding approximated interpretability values, that confirms the effectiveness of the interpretability criterion in measuring the level of information in the MBMs.

The results of this experiment confirm again the fact that the generalization performance is not a reliable criterion to measure the level of information learned by a linear classifier. For example consider the decoding model with $\lambda = 1$ in which the performance of the model is significantly above the chance level [see Figure 3.8(**C**)] while the corresponding MBM [Figure 3.9(**A**)] is completely misrepresents the ground–truth effect [Figure 3.6(**B**)].

3.3.3 Single–Subject Decoding on MEG Data

Lasso

To investigate the behavior of the proposed model selection criterion ζ_Φ, we benchmark it against the commonly used performance criterion δ_Φ in a single–subject decoding scenario. Assuming $(\mathbf{X}_i, \mathbf{Y}_i)$ for $i = 1, \ldots, 16$ are MEG trial/label pairs for subject i, we separately train a Lasso model for each subject to estimate the parameter of the linear function $\hat{\Phi}_i$, where $\mathbf{Y}_i = \mathbf{X}_i\hat{\Theta}_i$. We represent the optimized solution based on δ_Φ and ζ_Φ by $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$, respectively. We also denote the MBM associated with $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ by $\vec{\hat{\Theta}}_i^\delta$ and $\vec{\hat{\Theta}}_i^\zeta$, respectively. Therefore, for each subject, we compare the resulting decoders and MBMs computed based on these two model selection criteria.

Figure 3.10(**A**) represents the mean and standard–deviation of the performance and interpretability of Lasso across 16 subjects for different λ values. The performance and interpretability curves further illustrate the performance–interpretability dilemma of Lasso classifier in the single–subject decoding scenario, in which increasing the performance delivers less interpretability. The average performance across subjects is improved when λ approaches 1, but on the other side, the reproducibility and the representativeness of models declines significantly [see Figure 3.10(**B**)] (Wilcoxon rank sum test p-value= 9×10^{-4} and 8×10^{-7}, respectively). In fact, in

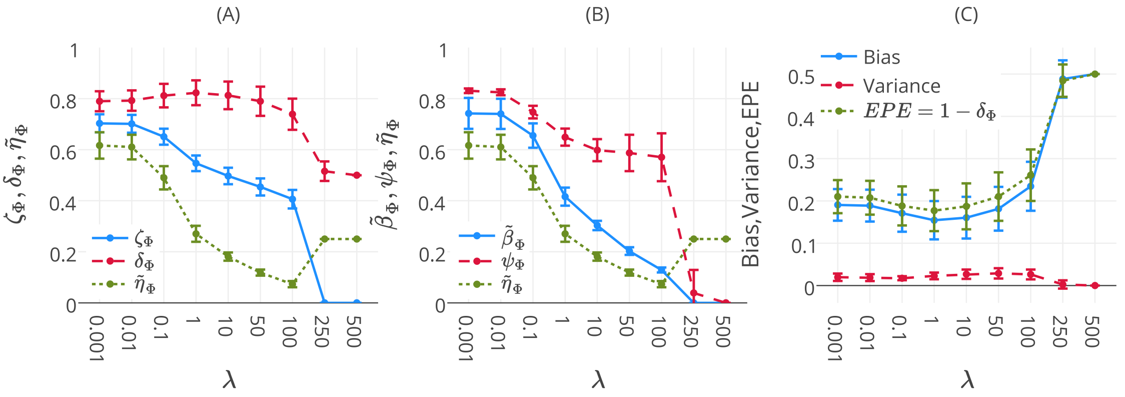


Figure 3.10: **(A)** Mean and standard–deviation of the performance (δ_Φ), interpretability (η_Φ), and ζ_Φ of Lasso over 16 subjects. **(B)** Mean and standard–deviation of the reproducibility (ψ_Φ), representativeness (β_Φ), and interpretability (η_Φ) of Lasso over 16 subjects. The interpretability declines because of the decrease in both reproducibility and representativeness (see Proposition 1). **(C)** Mean and standard–deviation of the bias, variance, and EPE of Lasso over 16 subjects. While the change in bias is correlated with that of EPE (Pearson's correlation coefficient= 0.9993), there is anti–correlation between the trend of variance and EPE (Pearson's correlation coefficient= −0.8884).

this dataset a higher amount of sparsity increases the gap between the generalization performance and interpretability.

One possible reason behind the performance–interpretability dilemma in this experiment is illustrated in Figure 3.10**(C)**. The figure shows the mean and standard deviation of bias, variance, and EPE of Lasso across 16 subjects. The plot shows while the change in bias is correlated with that of EPE (Pearson's correlation coefficient= 0.9993), there is anti–correlation between the trends of variance and EPE (Pearson's correlation coefficient= −0.8884). Furthermore, it proposes that the effect of variance is overwhelmed by bias in the computation of EPE, where the best performance (minimum EPE) at $\lambda = 1$ has the lowest bias, its variance is higher than for $\lambda = 0.001, 0.01, 0.1$. While this tiny increase in the variance has negligible effect on the EPE of the model, Figure 3.10**(B)** shows its significant (Wilcoxon rank sum test p-value= 8×10^{-7}) negative effect on the

Table 3.2: The performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$ over 16 subjects.

Subs	Criterion: $\delta(\Phi)$					Criterion: $\zeta(\Phi)$					δ_{cERF}
	$\delta(\Phi)$	$\zeta(\Phi)$	$\tilde{\eta}(\Phi)$	$\tilde{\beta}(\Phi)$	$\psi(\Phi)$	$\delta(\Phi)$	$\zeta(\Phi)$	$\tilde{\eta}(\Phi)$	$\tilde{\beta}(\Phi)$	$\psi(\Phi)$	
1	0.81	0.53	0.26	0.42	0.62	0.78	0.70	0.63	0.76	0.83	0.56
2	0.80	0.70	0.60	0.72	0.83	0.80	0.70	0.60	0.72	0.83	0.54
3	0.81	0.63	0.45	0.64	0.71	0.78	0.71	0.64	0.78	0.83	0.57
4	0.84	0.52	0.20	0.31	0.66	0.76	0.70	0.64	0.77	0.83	0.55
5	0.80	0.54	0.29	0.44	0.65	0.78	0.69	0.61	0.73	0.83	0.54
6	0.79	0.52	0.24	0.39	0.63	0.74	0.67	0.61	0.74	0.82	0.57
7	0.84	0.55	0.27	0.40	0.66	0.81	0.70	0.59	0.71	0.84	0.56
8	0.87	0.55	0.24	0.35	0.68	0.85	0.68	0.52	0.61	0.84	0.56
9	0.80	0.55	0.31	0.46	0.67	0.77	0.67	0.57	0.69	0.82	0.57
10	0.79	0.53	0.26	0.41	0.64	0.77	0.68	0.58	0.70	0.83	0.59
11	0.74	0.65	0.56	0.68	0.82	0.74	0.65	0.56	0.68	0.82	0.53
12	0.80	0.55	0.29	0.46	0.64	0.79	0.70	0.61	0.74	0.83	0.58
13	0.83	0.50	0.18	0.29	0.61	0.77	0.70	0.63	0.76	0.82	0.59
14	0.90	0.58	0.27	0.39	0.68	0.81	0.78	0.74	0.89	0.84	0.62
15	0.92	0.63	0.34	0.48	0.71	0.89	0.78	0.66	0.77	0.86	0.63
16	0.87	0.55	0.23	0.37	0.62	0.81	0.74	0.67	0.81	0.83	0.65
Mean	**0.83±0.05**	0.57 ± 0.05	0.31 ± 0.12	0.45 ± 0.13	0.68 ± 0.07	0.79 ± 0.04	**0.70± 0.04**	**0.62±0.05**	**0.74±0.06**	**0.83±0.01**	0.58 ± 0.03

reproducibility of maps from $\lambda = 0.1$ to $\lambda = 1$.

Table 3.2 summarizes the performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$ for 16 subjects. The average result over 16 subjects shows that employing ζ_{Φ} instead of δ_{Φ} in model selection provides higher reproducibility, representativeness, and (as a result) interpretability compensating for 0.04 of performance. The last column of table (δ_{cERF}) summarizes the performance of decoding models over 16 subjects when $\vec{\Theta}^{cERF}$ is used as classifier weights. The comparison illustrates a significant difference (Wilcoxon rank sum test p-value= 1.5×10^{-6}) between δ_{cERF} and $\delta(\Phi)$s. These facts demonstrate that $\vec{\hat{\Theta}}^{\zeta}$ is a good compromise between $\vec{\hat{\Theta}}_{\delta}$ and $\vec{\Theta}^{cERF}$ in terms of classification performance and model interpretability.

These results are further analyzed in Figure 3.11 where $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$ are compared subject–wise in terms of their performance and interpretability. The comparison shows that adopting ζ_{Φ} instead of δ_{Φ} as the criterion for model selection yields higher interpretable models by compensating a negligible degree of performance in 14 out of 16 subjects. Figure 3.11**(A)** shows that employing δ_{Φ} provides on average slightly higher accurate mod-

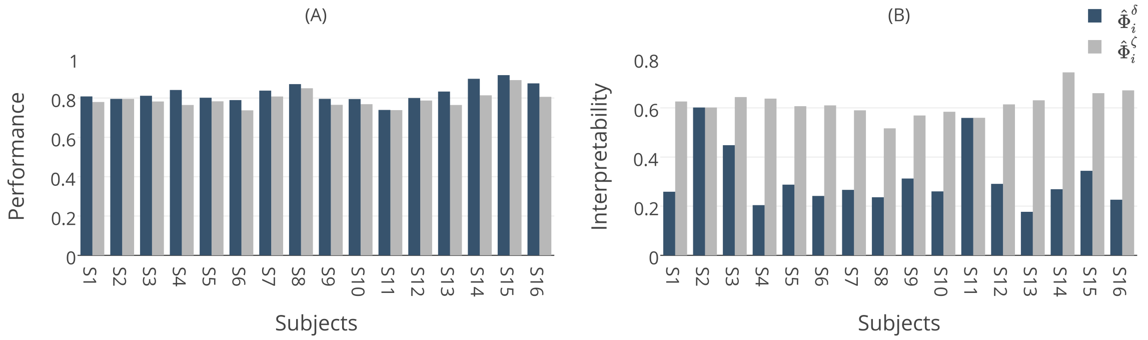
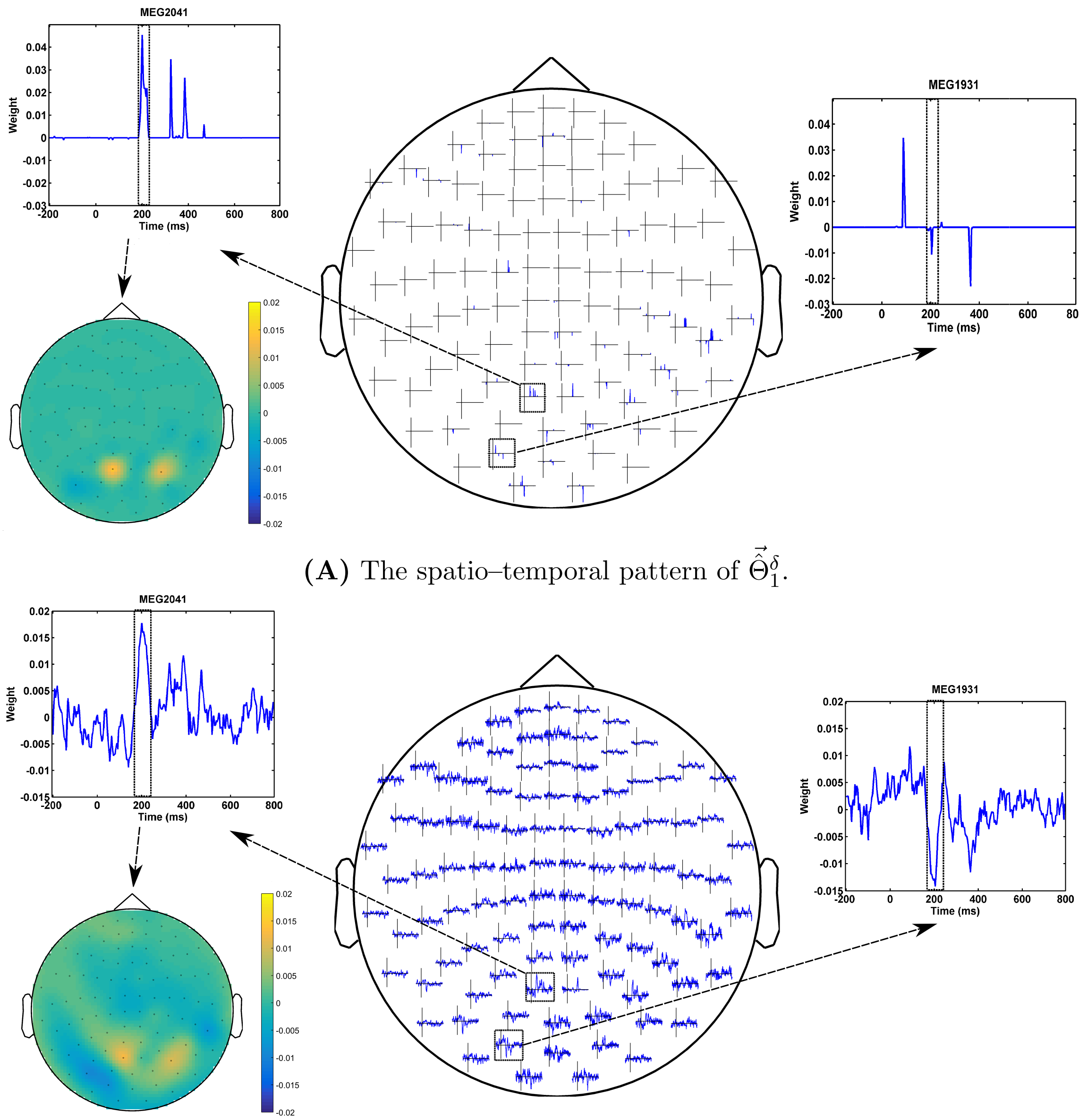


Figure 3.11: **(A)** Comparison between generalization performances of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$. Adopting ζ_Φ instead of δ_Φ in model selection yields (on average) 0.04 less accurate classifiers over 16 subjects. **(B)** Comparison between interpretabilities of $\hat{\Phi}_i^{\delta}$ and $\hat{\Phi}_i^{\zeta}$. Adopting ζ_Φ instead of δ_Φ in model selection yields on average 0.31 more interpretable classifiers over 16 subjects.

els (Wilcoxon rank sum test p-value= 0.012) across subjects (0.83 ± 0.05) than using ζ_Φ (0.79 ± 0.04). On the other side, Figure 3.11**(B)** shows that employing ζ_Φ and compensating by 0.04 in the performance provides (on average) substantially higher (Wilcoxon rank sum test p-value= 5.6×10^{-6}) interpretability across subjects (0.62 ± 0.05) compared to δ_Φ (0.31 ± 0.12). For example, in the case of subject 1 (see Table 3.2), using δ_Φ in model selection to select the best λ value for the Lasso yields a model with $\delta_\Phi = 0.81$ and $\tilde{\eta}_\Phi = 0.26$. In contrast, using ζ_Φ delivers a model with $\delta_\Phi = 0.78$ and $\tilde{\eta}_\Phi = 0.63$. This inverse relationship between performance and interpretability is direct consequence of over–fitting of model to the noise structure in a small–sample–size brain decoding problem (see Section 3.3.1).

The advantage of the exchange between the performance and the interpretability can be seen in the quality of MBMs. Figure 3.12**(A)** and 3.12**(B)** show $\vec{\hat{\Theta}}_1^{\delta}$ and $\vec{\hat{\Theta}}_1^{\zeta}$ of subject 1, i.e., the spatio–temporal multivariate maps of the Lasso models with maximum values of δ_Φ and ζ_Φ, respectively. The maps are plotted for 102 magnetometer sensors. In each case, the time course of weights of classifiers associated with the MEG2041 and MEG1931



(A) The spatio–temporal pattern of $\vec{\hat{\Theta}}^{\delta}_{1}$.

(B) The spatio–temporal pattern of $\vec{\hat{\Theta}}^{\zeta}_{1}$.

Figure 3.12: Comparison between spatio–temporal multivariate maps of **(A)** the most accurate, and **(B)** the most interpretable classifiers for Subject 1. $\vec{\hat{\Theta}}^{\zeta}_{1}$ provides a better spatio–temporal representation of the N170 effect than $\vec{\hat{\Theta}}^{\delta}_{1}$.

sensors are plotted. Furthermore, the topographic maps represent the spatial patterns of weights averaged between 184 and 236 ms after the stimulus onset. While $\vec{\hat{\Theta}}^{\delta}_{1}$ is sparse in time and space, it fails to accurately represent the spatio–temporal dynamic of the N170 component. Furthermore, the multicollinearity problem arising from the correlation between the time course of the MEG2041 and MEG1931 sensors causes extra attenuation of the N170 effect in the MEG1931 sensor. Therefore, the model is unable to capture the spatial pattern of the dipole in the posterior area. In contrast, $\vec{\hat{\Theta}}_1^{\zeta}$ represents the dynamic of the N170 component in time. In addition, it also shows the spatial pattern of two dipoles in the posterior and temporal areas. In summary, $\vec{\hat{\Theta}}_1^{\zeta}$ suggests a more representative pattern of the underlying neurophysiological effect than $\vec{\hat{\Theta}}_1^{\delta}$.

In addition, optimizing the hyper–parameters of brain decoding based on ζ_{Φ} offers more reproducible brain decoders. According to Table 3.2, using ζ_{Φ} instead of δ_{Φ} provides (on average) 0.15 more reproducibility over 16 subjects. To illustrate the advantage of higher reproducibility on the interpretability of maps, Figure 3.13 visualizes $\vec{\hat{\Theta}}_1^{\delta}$ and $\vec{\hat{\Theta}}_1^{\zeta}$ over 4 perturbed training sets. The spatial maps [Figure 3.13(**A**) and Figure 3.13(**C**)] are plotted for the magnetometer sensors averaged in the time interval between 184 and 236 ms after stimulus onset. The temporal maps [Figure 3.13(**B**) and Figure 3.13(**D**)] are showing the multivariate temporal maps of MEG1931 and MEG2041 sensors. While $\vec{\hat{\Theta}}_1^{\delta}$ is unstable in time and space across the 4 perturbed training sets, $\vec{\hat{\Theta}}_1^{\zeta}$ provides more reproducible maps.

Elastic–Net

Figure 3.14 summarizes the mean and standard–deviation of the performance and interpretability of elastic–net across 16 subjects for different levels of regularization and sparsity. The results illustrate that increasing the amount of sparsity, by increasing α, increases the chance of performance–interpretability dilemma. While for a ridge model, with $\alpha = 0$, the performance and interpretability are consistent, by increasing the sparsity they

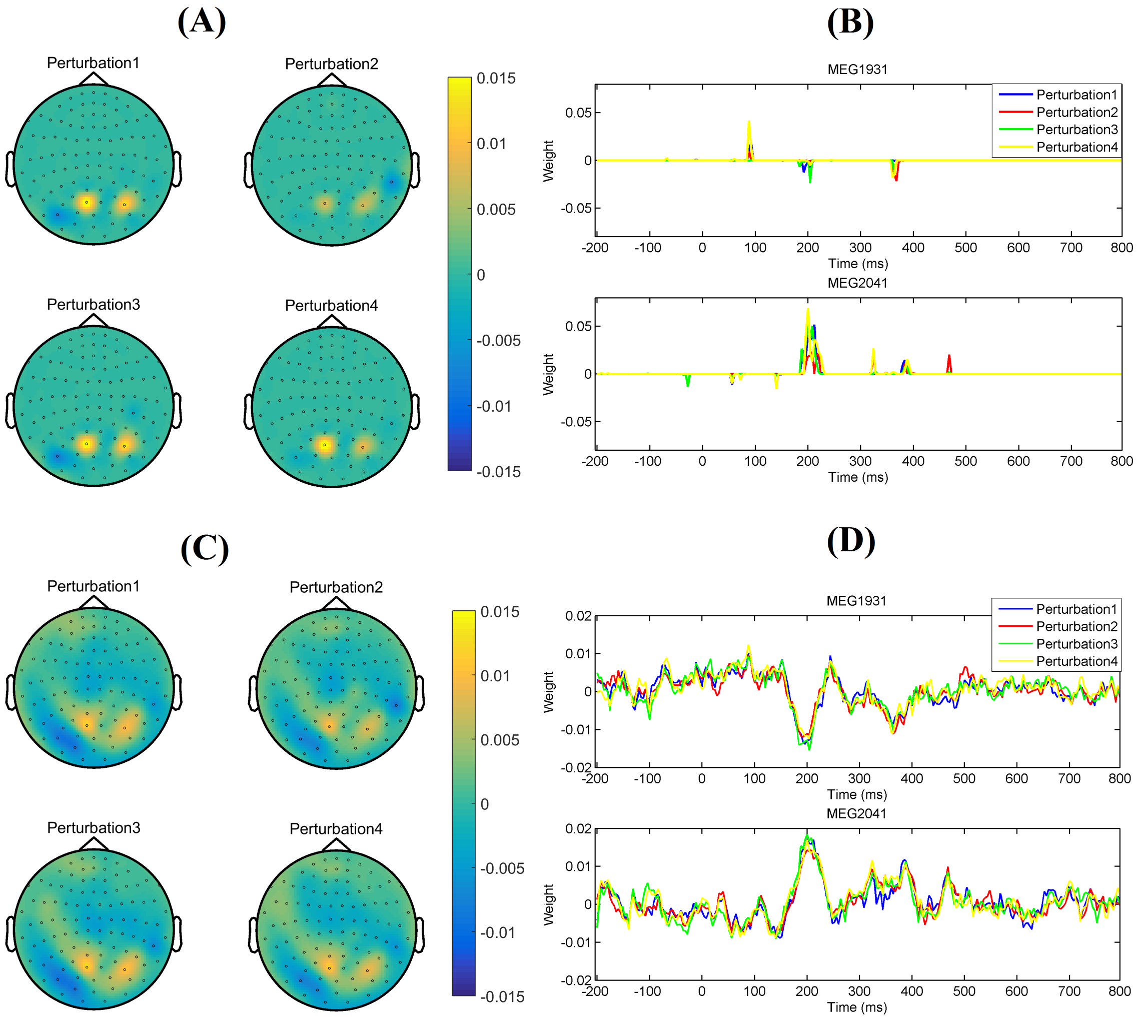


Figure 3.13: Comparison of the reproducibility of Lasso when δ_Φ and ζ_Φ are used in the model selection procedure. **(A)** and **(B)** show the spatio–temporal patterns represented by $\hat{\vec{\Theta}}_1^\delta$ across the 4 perturbed training sets. **(C)** and **(D)** show the spatio–temporal patterns represented by $\hat{\vec{\Theta}}_1^\zeta$ across the 4 perturbed training sets. Employing ζ_Φ instead of δ_Φ in the model selection yields on average 0.15 more reproduciblilty of MBMs.

show a divergent behavior. This observation illustrates the smooth, rather sparse, nature of the underlying effect in space and time.

These results are further analyzed in Figure 3.15 where $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ are compared subject–wise in terms of their performance and interpretability. Similar to the Lasso model in the main text, the comparison shows that

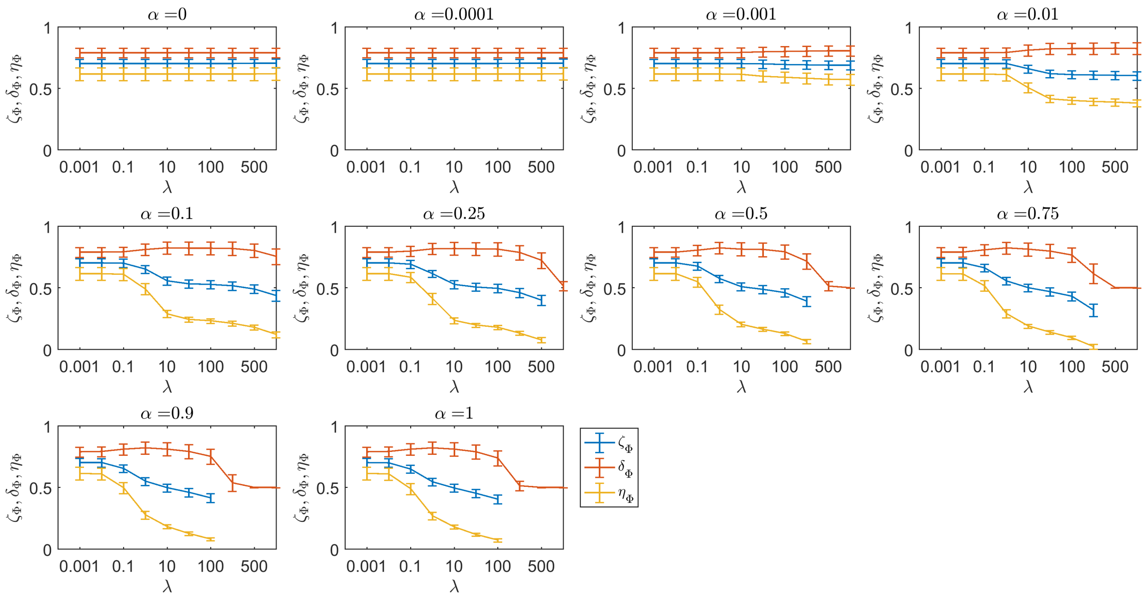


Figure 3.14: The mean and standard–deviation of the performance (δ_Φ), interpretability (η_Φ), and ζ_Φ of the elastic–net model over 16 subjects. In this dataset, increasing the amount of sparsity increases the chance of performance–interpretability dilemma.

adopting ζ_Φ instead of δ_Φ as the criterion for model selection yields higher interpretable models by compensating a negligible degree of performance across all subjects. Figure 3.15**(A)** shows that employing δ_Φ provides on average slightly higher accurate models across subjects (0.83 ± 0.05) than using ζ_Φ (0.79 ± 0.04). On the other side, Figure 3.15**(B)** shows that employing ζ_Φ and compensating by 0.04 in the performance provides (on average) substantially higher level of interpretability across subjects (0.62 ± 0.05) compared to δ_Φ (0.34 ± 0.11). The results obtained using elastic–net classifier are very similar to the ones of Lasso in the main text.

3.3.4 Mass–Univariate Hypothesis Testing on MEG Data

It is shown by [64, 65] that non–parametric mass–univariate analysis is unable to detect narrowly distributed effects in space and time (e.g., an N170 component). To illustrate the advantage of the proposed decod-

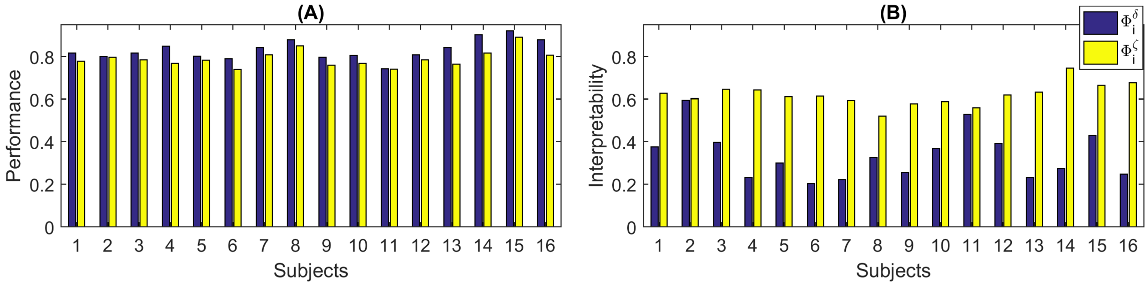


Figure 3.15: **(A)** Comparison between generalization performances of $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ using elastic–net as the classifier. **(B)** Comparison between the interpretability of $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ using elastic–net as the classifier. The results obtained by the elastic–net classifier are very similar to the Lasso model.

ing framework for spotting these effects, we performed a non–parametric cluster–based permutation test [127] on our MEG dataset using Fieldtrip toolbox [153]. In a single subject analysis scenario, we considered the trials of MEG recordings as the unit of observation in a between–trials experiment. Independent–samples t–statistics are used as the statistics for evaluating the effect at the sample level and to construct spatio–temporal clusters. The maximum of the cluster–level summed t–value is used for the cluster level statistics; the significance probability is computed using a Monte Carlo method. The minimum number of neighboring channels for computing the clusters is set to 2. Considering 0.025 as the two–sided threshold for testing the significance level and repeating the procedure separately for magnetometers and combined–gradiometers, no significant result is found for any of the 16 subjects. This result motivates the search for more sensitive (and, at the same time, more interpretable) alternatives for univariate hypothesis testing.

3.3.5 Across–Subject Decoding of MEG Data

As demonstrated in our results in Section 3.3.3, in the single–subject decoding of MEG data the performance and the interpretability of a Lasso

Table 3.3: The performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}^{\delta}$ and $\hat{\Phi}^{\zeta}$ in the across–subject decoding scenario.

	λ	δ_{Φ}	ψ_{Φ}	$\tilde{\beta}_{\Phi}$	$\tilde{\eta}_{\Phi}$	ζ_{Φ}
$\hat{\Phi}^{\delta}$	0.1	0.7277	0.7841	0.4597	0.3605	0.5441
$\hat{\Phi}^{\zeta}$	0.01	0.7275	0.7853	0.4596	0.3609	0.5442

classifier are not consistent. In this experiment the aim is to assess the relation between interpretability and generalization performance in the across–subject decoding scenario. To perform across–subject analysis we performed the decoding and evaluation phases on the pooled samples of all subjects. Table 3.3 summarizes the performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}^{\delta}$ and $\hat{\Phi}^{\zeta}$ in the across–subject decoding scenario.

The comparison of results illustrates a negligible difference between $\hat{\Phi}^{\delta}$ and $\hat{\Phi}^{\zeta}$ in terms of ζ_{Φ} and δ_{Φ} in the across–subject decoding. In other words, in this case the interpretability and performance of the model are consistent and the most accurate model is very close to the most interpretable one. Therefore in across–subject decoding, using merely the generalization performance as the dicisive criterion in the model selection procedure would be enough for drawing interpretable brain maps. One possible explanation behind this observation can be the increase in the sample size in the across–subject decoding scenario.

Figure 3.16 shows the spatio–temporal multivariate brain map of $\hat{\Phi}^{\zeta}$ in the across–subject decoding scenario. The resulting multivariate brain map represents the feedforward and feedback information flow in visual cortical areas [132]. The 3 dipoles in $184 - 236$ ms time interval [Figure 3.16(**B**)] show the feedforward information flow from the posterior area to the parietal and ventral areas. The topographic maps in the two following time intervals [Figure 3.16(**C**) and 3.16(**D**)] show the spatial dynamic of face processing from posterior to temporal lobs. Finally Figure 3.16(**E**) shows

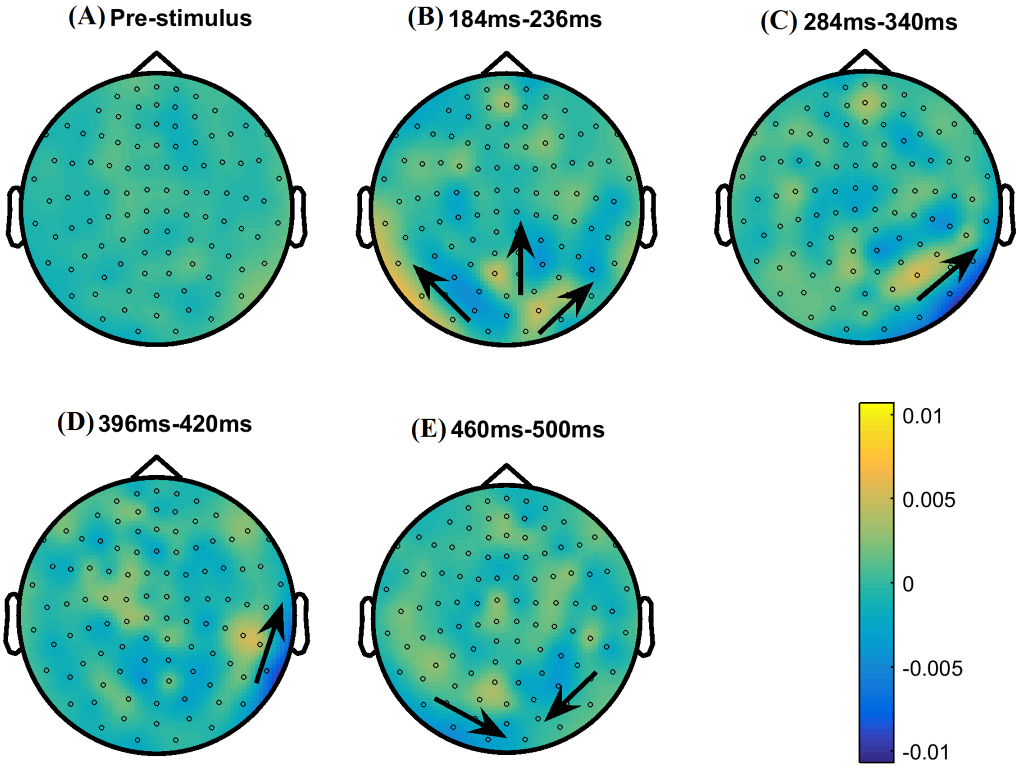


Figure 3.16: The spatio–temporal MBM of face processing in the across–subject decoding scenario: **(A)** before the stimulus onset, **(B)** 3 occipo–parietal dipoles 200 ms after the stimulus onset, **(C)** and **(D)** the forward ventral information flow from 300 to 400 ms after the stimulus onset, **(E)** the backward information flow from temporal areas to occipital area 500 ms after the stimulus onset.

a weak but still visible backward information flow from temporal lobes to the posterior area 500 ms after the stimulus onset.

3.4 Discussions

3.4.1 Defining Interpretability: Theoretical Advantages

An overview of the brain decoding literature shows frequent co–occurrence of the terms interpretation, interpretable, and interpretability with the terms model, classification, parameter, decoding, method, feature, and pattern; however, a formal formulation of the interpretability is never presented. In this study, our primary interest is to present a simple and theoretical definition of the interpretability of linear brain decoding models and their corresponding MBMs. Furthermore, we show the way in which interpretability is related to the reproducibility and neurophysiological representativeness of MBMs. Our definition and quantification of interpretability remains theoretical, as we assume that the true solution of the brain decoding problem is available. Despite this limitation, we argue that the presented definition provides a concrete framework of a previously abstract concept and that it establishes a theoretical background to explain an ambiguous phenomenon in the brain decoding context. We support our argument using an example in the time–domain MEG decoding in which we show how the presented definition can be exploited to heuristically approximate the interpretability. Our experimental results on MEG data shows accounting for the approximated measure of interpretability has a positive effect on the human interpretation of brain decoding models. This example shows how partial prior knowledge regarding the timing and location of neural activity can be used to find more plausible multivariate patterns in data. Furthermore, the proposed decomposition of the interpretability of MBMs into their reproducibility and representativeness explains the relationship between the influential cooperative factors in the interpretability of brain decoding models and highlights the possibility of indirect and partial evaluation of interpretability by measuring these effective factors.

3.4.2 Application in Model Evaluation

Discriminative models in the framework of brain decoding provide higher sensitivity and specificity than univariate analysis in hypothesis testing of neuroimaging data. Although multivariate hypothesis testing is performed based solely on the generalization performance of classifiers, the emergent need for extracting reliable complementary information regarding the underlying neuronal activity motivated a considerable amount of research on improving and assessing the interpretability of classifiers and their associated MBMs. Despite ubiquitous use, the generalization performance of classifiers is not a reliable criterion for assessing the interpretability of brain decoding models [166, 170]. Therefore, considering extra criteria might be required. However, because of the lack of a formal definition for interpretability, different characteristics of linear classifiers are considered as the decisive criterion in assessing their interpretability. Reproducibility [14, 40, 166], stability selection [195, 201], sparsity [44, 175], and neurophysiological plausibility [4] are examples of related criteria.

Our definition of interpretability helped us to fill this gap by introducing a new multi–objective model selection criterion as a weighted compromise between interpretability and generalization performance of linear models. Our experimental results on single–subject decoding showed that adopting the new criterion for optimizing the hyper–parameters of brain decoding models is an important step toward reliable visualization of learned models from neuroimaging data. It is not the first time in the neuroimaging context that a new metric is proposed in combination with generalization performance for the model selection. Several recent studies proposed the combination of the reproducibility of the maps [40, 166, 178] or the stability of the classifiers [122, 196, 215] with the performance of discriminative models to enhance the interpretability of decoding models. Our definition of interpretability supports the claim that the reproducibility is not the only effective factor in interpretability. Therefore, our contribution can be considered a complementary effort to the state of the art of improving the interpretability of brain decoding at the model selection level. Furthermore, this work presents an effective approach for evaluating the quality of different regularization strategies for improving the interpretability of MBMs. As briefly reviewed in Section 3.1, there is a trend of research within the brain decoding context in which the prior knowledge is injected into the decoding process via the penalization term in order to improve the interpretability of decoding models. Thus far, in the literature, there is no ad–hoc method to directly compare the interpretability of MBMs resulting from different penalization techniques. Our findings provide a further step toward direct evaluation of interpretability of the currently proposed penalization strategies. Such an evaluation can highlight the advantages and disadvantages of applying different strategies on different data types and facilitates the choice of appropriate methods for a certain application.

3.4.3 Regularization and Interpretability

Haufe et al. [83] demonstrated that the weight in linear discriminative models are unable to accurately assess the relationship between independent variables, primarily because of the contribution of noise in the decoding process. They concluded that the interpretability of brain decoding cannot be improved using regularization. The problem is primarily caused by the decoding process *per se*, where it minimizes the classification error only considering the uncertainty in the output space [5, 187, 218] and not the uncertainty in the input space (or noise).

Our experimental results on the toy data (see Section 3.3.1) shows that if the right criterion is used for selecting the best values for hyper–parameters, appropriate choice of the regularization strategy can still play a significant role in improving the interpretability of results. For example, in the case of toy data, the true generative function behind the sampled data is sparse (see Section 3.2.6), but because of the noise in the data, the sparse model is not the most accurate one.

On the other hand, a more comprehensive criterion (in this case, ζ_Φ) that considers also the interpretability of model parameters facilitates the selection of correct prior assumptions about the distribution of the data via regularization. This observation encourages a modification of the conclusion of Haufe et al. [83] as follows: if the performance of the model is the only criterion in the model selection, then the interpretability cannot necessarily be improved by means of regularization. This modification offers a practical shift in methodology, where we propose to replace the post–processing of weights with refinement of hyper–parameter selection based on the newly developed model selection criterion.

3.4.4 The Performance–Interpretability Dilemma

The performance–interpretability dilemma refers to the trade–off between the generalization performance and the interpretability of a decoding model. In some applications of brain decoding, such as BCI, a more accurate model (even with no interpretability) is desired. On the other hand, when the brain decoding is employed for hypothesis testing purpose, an astute balance between two factors is more favorable. The presented metric for model selection (ζ_Φ) provides the possibility to maintain this balance. An important question at this point is on the nature of the performance–interpretability dilemma, whether it is model–driven or data–driven? In other words, whether some decoding models (e.g., sparse models) suffer from this deficit, or it is independent from the decoding model and depends on the distribution of data rather assumptions of the decoding model.

Our experimental observations shed light on the fact that the performance–interpretability dilemma is driven by the *uncertainty* [5] in data. The uncertainty in data refers to the difference between the true solution of decoding Φ^* and the solution of decoding in sampled data space Φ_S, and is generally consequence of noise in the input or/and output spaces (see Appendix A.1 for a simple illustration about the effect of uncertainty in the input space on the learning process). This gap between Φ^* and Φ_S is also known as irreducible error (see Eq. 3.2) in the learning theory, and it cannot fundamentally be reduced by minimizing the error. Therefore, any attempt toward improving the classification performance in the sampled data space might increase the irreducible error. As an example, our experiment on the toy data (see Section 3.3.1) shows the effect of noise in input space on the performance–interpretability dilemma. Improving the performance of the model (i.e., fitting to Φ_S) diverges the estimated solution of decoding $\hat{\Phi}$ from its true solution Φ^*, thus reduces the interpretability of the decoding model. Furthermore, our experiments demonstrate that incorporating the interpretability of decoding models in model selection facilitates finding the best match between the decoding model and the distribution of data. For example in classification of toy data, the new model selection metric ζ_Φ selects the more sparse model with a better match to the true distribution of data, despite worse generalization performance.

3.4.5 Advantage over Mass–Univariate Analysis

Mass–univariate hypothesis testing methods are among the most popular tools for forward inference on neuroimaging data in cognitive neuroscience field. Mass–univariate analyses consist of univariate statistical tests on single independent variables followed by multiple comparison correction. Generally, multiple comparison correction reduces the sensitivity of mass–univariate approaches because of the large number of univariate tests involved. Cluster–based permutation testing [127] provides a more sensitive univariate analysis framework by making the cluster assumption in the multiple comparison correction. Unfortunately, this method is not able to detect narrow spatio–temporal effects in the data [64]. As a remedy, brain decoding provides a very sensitive tool for hypothesis testing; it has the ability to detect multivariate patterns, but suffers from a low level of interpretability. Our study proposes a possible solution for the interpretability problem of classifiers, and therefore, it facilitates the application of brain decoding in the analysis of neuroimaging data. Our experimental results for the MEG data demonstrate that, although the non–parametric cluster–based permutation test is unable to detect the N170 effect in MEG data, employing ζ_Φ instead of δ_Φ in model selection not only detects the stimuli-relevant information in the data, but also assures both reproducible and representative spatio–temporal mapping of the timing and the location of underlying neurophysiological effect.

3.4.6 Limitations and Future Directions

Despite theoretical and practical advantages, the proposed definition and quantification of interpretability suffer from some limitations. All of the presented concepts are defined for linear models, with the main assumption that $\Phi^* \in \mathcal{H}$ (where $\mathcal{H}$ is a class of linear functions). This fact highlights the importance of linearizing the experimental protocol in the data collection phase [143]. Extending the definition of interpretability to non–linear models demands future research into the visualization of non–linear models in the form of brain maps. Currently, our findings cannot be directly applied to non–linear models. Furthermore, the proposed heuristic for the time–domain MEG data applies only to binary classification. One possible solution in multiclass classification is to separate the decoding problem into several binary sub–problems. In addition the quality of the proposed heuristic is limited for the small sample size datasets (see Appendix A.4 for an experimental illustration). Of course the proposed heuristic is just an example of possible options for assessing the neurophysiological plausibility of MBMs in time–locked analysis of MEG data, thus, improving the quality of heuristic would be of interest in future researches. Finding physiologically relevant heuristics for other acquisition modalities such as fMRI, or frequency domain MEEG data, can be also considered as possible directions in future work.

Chapter 4

Multi–Task Joint Feature Learning for Group MEG Decoding

4.1 Introduction

A common approach in cognitive neuroscience is to record brain activity, and to correlate that activity with behavioral parameters in order to discover *where*, *when*, and *how* a brain region participates in a particular cognitive process. In functional neuroimaging research, scientists often employ mass–univariate hypothesis testing methods, i.e., methods which have been designed to test scientific hypotheses on a large set of independent variables [64, 126]. Mass–univariate hypothesis testing is based on performing multiple (generally thousands) univariate tests, which most commonly involves performing a t–test, for each independent variable, e.g., each voxel. The statistical results for each voxel can then be projected onto a structural image to form a brain map, that provides information about which region in the brain is related to the experimental conditions. For instance, in a common paradigm used to investigate the neural correlates of face perception, participants see either intact faces or scrambled faces. A univariate contrast is then run in each voxel and clusters of voxels that are significantly more active for intact faces inform us of where in the brain holistic face processing occurs.

While mass–univariate analyses can at times be useful, there are a number of problematic aspects. Here we outline three major problematic aspects: 1) due to its univariate nature, the interaction between different independent variables cannot be exploited [41]; 2) the high dimensionality of neuroimaging data requires a large number of tests, but running this many tests requires multiple comparison correction, and current multiple comparison correction at the voxel level is overly conservative, increasing type II errors and decreasing sensitivity [52]. Although some techniques, such as the non–parametric cluster–based permutation test [31, 127] offer more sensitivity by weakly controlling the family–wise error rate, they still experience low sensitivity to brain activities that are narrowly distributed in time and space due to the cluster assumption [64, 65]; 3) because of inter–subject differences (in time and space), it is likely that univariate statistical tests fail to find significant effects [126] as these tests implicitly assume a one–to–one correspondence between independent variables across different subjects.

A potentially more promising approach to overcome the shortcomings of mass–univariate hypothesis testing is *Brain decoding* [89, 154]. Brain Decoding is a multivariate pattern analysis (MVPA) technique that attempts to predict the mental state of a human subject based on the recorded brain signal. More specifically, brain decoding involves training an algorithm to classify a number of samples of labeled brain data, and testing it on unseen data. The generalization performance of a brain decoding model is used as a measure for performing inference on neuroimaging data, or in other words for concluding that a certain area or set of areas are important for a specific cognitive process, or a certain class of stimuli. Brain decoding is capable of capturing complex spatio–temporal interactions between different brain areas with higher sensitivity and specificity than univariate analysis [85]. Moreover, it avoids the multiple comparison problem, as it deals with the whole set of independent variables at once.

Due to the high dimensionality and limited number of samples typically associated with neuroimaging data [41, 114], generally in brain decoding, the linear classifiers are used to assess the relation between spatio–temporal brain measurements and cognitive tasks [22, 118, 157]. This assessment is performed by solving an optimization problem that minimizes a loss function by learning linear weights associated with each independent variable. These learned linear weights can then be visualized in the form of a *brain map*, in which the engagement of different brain areas in a cognitive task is illustrated. In fact, brain mapping via brain decoding can be viewed as a *pattern recovery* problem, where the goal is to recover spatio–temporal patterns of the discriminative brain activity involved in the cognitive processing of external stimuli. If successful, brain maps created by brain decoding can provide a comprehensive and interpretable explanation regarding the nature of neural representations and brain states, and may be more informative for cognitive science than a numerical decoding accuracy measurement, as is currently commonly used [154]. Currently, brain decoding is a gold standard in multivariate analysis of functional magnetic resonance images (fMRI) [41, 86, 135, 149] and magnetoencephalogram/electroencephalogram(MEG/EEG) data [3, 34, 36, 93, 156, 167, 199]. However a number of challenges still remain, particularly regarding the interpretability of recovered brain maps at the individual or group level.

4.1.1 Group–level Brain Decoding: Approaches and Challenges

Group–level analyses are extremely important, as they allow for results to be generalized to new individuals. In brain decoding, an ideal group–level approach should be able to recover both structural and functional similarities and dissimilarities across different individuals. These similarities and dissimilarities generally occur at both a coarse and a fine level in space and time, and can provide valuable spatio–temporal information about both the macro and micro–structures underlying the cognitive function in question. For example, visual stimuli in general evoke a coarsely similar effect in early visual brain areas across different subjects, but the response to different types, or categories of visual stimuli can differ from subject to subject at the finer level (see Ref. [87] for more examples). This across–subject functional variability makes group–level inference on neuroimaging data challenging, particularly since there is also substantial across–subject variability in brain structure composition (e.g., the different size and shape of brains) [129, 164, 165, 180, 181]. This problem is even more pronounced when one takes into account the difference in the spatio–temporal structure of noise, that commonly occurs due to different sources of the external and internal noise, or to preprocessing errors. These variations not only negatively affect the generalization power of brain decoding, but they also make post–hoc interpretation of the derived brain maps more challenging, due to concerns about lack of reproducibility and plausibility. For these reasons, it is crucial to explore more effective decoding methods that are capable of recovering structural and functional similarities and dissimilarities in a group–level analysis of neuroimaging data.

There are two main approaches to group–level inference in brain decoding: 1) A decoding model is trained and tested for each subject independently, and then generalization performance is averaged across subjects; 2) A single decoding model is trained and tested on the pooled samples of all subjects. While the first approach does not take advantage of similarities between different subjects, the second method incorrectly assumes that the brain recordings of all subjects are drawn from the same distribution. These subtle assumptions may lead to impaired predictive performance [129, 151] and to complications in interpreting results. Therefore, it is highly important to develop a principled approach that enables identification of common features across subjects [192] while accounting for inter–subject differences that result from variations in structural and/or functional anatomy.

4.1.2 Contribution

In this chapter, we present an application of multi–task joint feature learning [9] which allows for accurate spatio–temporal pattern recovery at the group–level decoding of MEG data. Multi–task learning [35] (MTL) is a machine learning technique in which a number of related problems with salient shared properties is simultaneously solved (see Section 2.4.5). Previous work has shown that MTL has some benefits over the trivial single–task setting, especially in terms of specificity and stability of feature maps [107, 131]. In our proposed framework, we consider the data of each subject as a task in MTL framework, and, we simultaneously train only one decoding model over all subjects. Further, we employ $\ell_{2,1}$ regularization [124] to learn sparse patterns consistently across different subjects, i.e., to jointly learn the features across different subjects. This learning process facilitates consistent sparse pattern recovery across individual subjects while at the same time preserving idiosyncratic structural and functional properties within each individual.

To evaluate the effectiveness of the multi–task joint feature decoding algorithm, we compared its performance against number of currently popular single–subject and pooled decoding approaches. We used three criteria in our comparisons: 1) generalization performance, 2) reproducibility of brain maps, and 3) the quality of pattern recovery. All analyses were run on both synthetic and real MEG datasets. We chose MEG data because its complex wealth of spatiotemporal information poses a particular challenge in recovering multi–way patterns in space and time. Our results demonstrate the potential of multi–task joint feature learning in recovering the similarity and dissimilarity of brain activities across different subjects in group–level MEG decoding, while still maintaining competitive performance and high reproducibility with respect to single–subject and pooling approaches. Such an approach can lead to more interpretable decoding models in group–level multivariate analysis of MEG data. To our knowledge, the present work is the first to use multi–task joint feature learning in the context of group–level MEG decoding. Considering the fact that, only EEG and MEG can non–invasively record brain activity at a high temporal resolution [75, 78], the proposed approach provides the possibility for recovering temporal brain dynamics within the millisecond time scale, a crucial task if we hope to understand the human brain function in real–time [77, 79].

In the remaining text, we first review the basic concepts of discriminative linear brain decoding, and then formally elaborate the pros and cons of single–subject and pooling approaches for group–level brain decoding. We then present the multi–task joint–feature learning approach as a possible alternative to single–subject and pooling approaches, and we formally show the that multi–task joint feature learning provides significant benefits over the currently popular approaches. Finally, we discuss the significance of our work in improving the interpretability of brain maps derived from group brain decoding analyses, its position with respect to the related works, the limitations of our approach, and possible future directions.

4.2 Materials and Methods

4.2.1 Notation

In the following text, we denote scalar numbers by lower case italic letters, vectors by lower case boldface letters, and matrices by capital boldface letters. We use $\mathbf{a}_{i,:}$ and $\mathbf{a}_{:,i}$ to refer to the ith row and column of matrix $\mathbf{A}$, respectively. We use $\|.\|_2$ to denote the ℓ_2-norm of a vector, $\|.\|_1$ the ℓ_1-norm of a vector, and $[.,.]$ for the row–wise vector concatenation operator.

4.2.2 Brain Decoding for Brain Mapping: The Pattern Recovery Problem

The aim of brain decoding is to learn the function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^p$ represents the space of neural activity, and $\mathcal{Y} \in \{1, 2, \dots, c\}$ represents the categorical output space, i.e., the target classes of the stimuli. In this paper, for sake of simplicity, we focus on the binary brain decoding problem where $\mathcal{Y} \in \{-1, 1\}$. Let $(\mathbf{x}_j, y_j) \in \mathbb{R}^p \times \{-1, 1\}$ be the jth sample, $\forall j \in \{1, 2, \dots, n\}$, that is, independently and identically distributed (*iid*), drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, based on an unknown Borel probability measure P, and we have $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{n \times p}$ and $\mathbf{y} = [y_1, y_2, \dots, y_n] \in \{-1, 1\}^n$. In the neuroimaging context, it is commonly assumed that the solution of a brain decoding problem is among a family of linear functions $\mathcal{H}$. Therefore, the aim of brain decoding reduces to finding an empirical linear approximation of $\mathcal{F}$ in $\mathcal{H}$. This approximation can be obtained via a maximum a–posteriori estimation, or alternatively, by solving a regularized empirical risk minimization (rERM) problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{y}, \mathbf{X}\mathbf{w}) + \lambda\Omega(\mathbf{w}) \tag{4.1}$$

where $\hat{\mathbf{w}} \in \mathbb{R}^p$ represents the weight vector of the linear classifier and $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$, $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$ is the loss function, $\Omega : \mathbb{R}^p \to \mathbb{R}^+$ is the regularization term, and $\lambda \geq 0$ is a hyper–parameter that controls the amount of regularization. There are various choices for Ω, each of which reduces the hypothesis space $\mathcal{H}$ to $\mathcal{H}' \subseteq \mathcal{H}$ by enforcing different functional or structural constraints on the parameters of the linear decoding model.

The ℓ_1 and the squared ℓ_2 penalizations where $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$ and $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$, respectively, are two common choices for the regularization terms. The ℓ_1 regularization, also known as Lasso [185], promotes sparsity in the parameter space, while ℓ_2 enforces a Gaussian prior on the distribution of parameters.

The generalization performance of the decoding model can be estimated via data perturbation techniques, such as cross–validation [111] or bootstrapping [51], both of which evaluate the quality of predictions in $\hat{\mathbf{y}}$ with respect to the actual target classes in $\mathbf{y}$. The learned parameters of the decoding model $\hat{\mathbf{w}}$ can be possibly used in the form of a brain map in order to visualize the discriminating brain activity between different stimulus categories. This inverse inference approach for multivariate analysis of neuroimaging data is generally known as neural *pattern recovery* [195] and has many applications in medical diagnosis and hypothesis testing.

4.2.3 Group–Level Brain Decoding

Let $(\mathbf{x}_j^{(i)}, y_j^{(i)})$ be the jth, $\forall j \in \{1, 2, \ldots, n^{(i)}\}$, *iid* sample that is drawn from the joint distribution of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, based on an unknown Borel probability measure $P^{(i)}$, where $i \in \{1, 2, \ldots, s\}$ denotes the neural recordings for the ith subject. In this study, we are interested in MEG data decoding, thus here $\mathbf{x}_j^{(i)}$ refers to the jth trial of MEG recording on subject i. The sampling probability measures, i.e., $P^{(i)}$, are subject–specific and they depend on the device used to measure the neural activity. These probability measures are partially different from subject to subject due to structural and functional variability across individuals, as well as different levels and types of internal and external noise contamination. While the difference in noise levels is uninformative and should be ignored, the structural and functional differences might reflect valuable and meaningful information regarding the different cognitive processes across individuals. In the remaining text, we use $D = \{(\mathbf{X}^{(i)}, \mathbf{y}^{(i)}) \mid i \in \{1, 2, \ldots, s\}\}$ to denote the training set composed of s subjects neural recordings, where $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{n^{(i)}}^{(i)}] \in \mathbb{R}^{n^{(i)} \times p}$ and $\mathbf{y}^{(i)} = [y_1^{(i)}, y_2^{(i)}, \ldots, y_{n^{(i)}}^{(i)}] \in \{-1, 1\}^{n^{(i)}}$.

A successful group–level pattern recovery via brain decoding should reflect similarities and dissimilarities in neural correlates across different subjects, while ignoring the uninformative noise patterns. There are two common approaches used to solve the brain decoding problem at the group–level [129]:

1. **Single–Subject Decoding:** In single–subject decoding the rERM problem is solved independently for each subject in order to find linear functions $F^{(i)} : \mathbf{X}^{(i)} \to \mathbf{y}^{(i)}$ as linear estimations of the solution to the brain decoding problem $\mathcal{F}$:

$$\hat{\mathbf{w}}^{(i)} = \underset{\mathbf{w}}{\operatorname{argmin}} \, \mathcal{L}(\mathbf{y}^{(i)}, \mathbf{X}^{(i)}\mathbf{w}) + \lambda^{(i)}\Omega(\mathbf{w}) \tag{4.2}$$

   where $\hat{\mathbf{w}}^{(i)} \in \mathbb{R}^p$ is the recovered brain map for subject i, and we have $\hat{\mathbf{y}}^{(i)} = \mathbf{X}^{(i)}\hat{\mathbf{w}}^{(i)}$. Even though single–subject decoding is based on the correct assumption of heterogeneity of $P^{(i)}$ across different subjects, and therefore accounts for variability in structure, functional profile, and noise of $\mathbf{X}^{(i)}$ for different individuals, its solutions tend to overfit to the noise patterns [83], due to the high–dimensionality of data where $n^{(i)} \ll p$. Consequently, there is high variability between recovered brain maps from different perturbed training sets (for example folds of k-fold cross–validation). This variability makes the post–hoc interpretation of results cumbersome. Furthermore, single–subject decoding relies only on the idiosyncratic brain activity patterns, and thus does not take advantage of coarse–level similarities across different brains [129].

2. **Pooling:** In the pooling scenario, it is assumed that the data for all subjects are generated by the same probability distribution, i.e., $P^{(1)} = P^{(2)} = \cdots = P^{(s)}$, therefore the rERM problem is solved on the pooled samples of all subjects $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(s)}] \in \mathbb{R}^{n \times p}$, $\mathbf{y} = [\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(s)}] \in \{-1, 1\}^n$, where $n = \sum_{i=1}^{s} n^{(i)}$, and we have
$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \, \mathcal{L}(\mathbf{y}, \mathbf{X}\mathbf{w}) + \lambda\Omega(\mathbf{w}). \tag{4.3}$$
Even though the pooling scenario alleviates, to some degree, the overfitting problem ($n > n^{(i)}$), it suffers from the subtle assumption of the homogenity of $P^{(i)}$ across different subjects and, consequently it ignores the various sources of inter–subject variability. This sub–optimal assumption has a negative effect on generalization performance [151]. In addition, the pooling approach recovers only a single brain map for all subjects, thus it is unable to recover the possible structural and functional differences across different individuals.

4.2.4 Multi–Task Joint Feature Learning for Group–Level Decoding

As a compromise between the two aforementioned extremes in multi–subject brain decoding, we propose the *multi–task joint feature learning* paradigm [9] for solving the brain decoding problem at the group–level. In this approach, the brain recording of each subject is considered as a task, and the rERM problem is optimized simultaneously across subjects as follows:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{i=1}^{s} \mathcal{L}(\mathbf{y}^{(i)}, \mathbf{X}^{(i)}\mathbf{w}_{:,i}) + \lambda \left\| \mathbf{W} \right\|_{2,1} \tag{4.4}$$

where $\|\mathbf{W}\|_{2,1} = \sum_{j=1}^{p} \|\mathbf{w}_{j,:}\|_2$ is $\ell_{2,1}$-norm of $\mathbf{W} \in \mathbb{R}^{p \times s}$. The $\ell_{2,1}$-norm is a non–smooth regularizer which encourages learning sparse common features across multiple tasks, i.e., subjects. However, solving the above rERM optimization problem is challenging due to non–smoothness of $\ell_{2,1}$ term. Several algorithms are proposed in the literature for solving this problem or equivalent constrained versions. In this paper, we adopt the accelerated group sparsity learning algorithm [124] for solving Eq. (4.4). This algorithm reformulates the non–smooth $\ell_{2,1}$ as a constrained convex optimization problem with a smooth objective function. This problem is then solved using Nesterov's accelerated projected gradient descent method [144] which provides a superior worst–case convergence rate than standard projected gradient descent, and is much faster than sub–gradient descent and gradient descent algorithms.

In practice, the $\ell_{2,1}$-norm encourages group sparsity over the features across different tasks. The sparse feature selection over the groups of spatio–temporal features is induced by the summation over ℓ_2-norms. As schematically shown in Figure 4.1, the resulting weight matrix is expected to have a similar sparse pattern across different tasks. This is while, inside each selected group of features, the features can have different weights from task to task. This property is especially beneficial for representing the differences in behavior of similar features across different tasks.

The proposed approach has three advantageous characteristics for group–level pattern recovery: 1) it simultaneously optimizes the loss function across subjects. This characteristic, similar to single–subject decoding, and unlike the pooling approach, provides the possibility of subject–specific pattern recovery, while, similar to pooling and unlike the single–subject approach, it learns the underlying patterns of neurophysiological activity on a larger sample size (on all subjects). In addition, the simultaneous optimization provides the infrastructure to learn the shared spatio–temporal patterns across different individuals; 2) it accounts for different noise distributions in the recorded data across subjects, i.e., $\mathbf{X}^{(i)}$, thus enhances

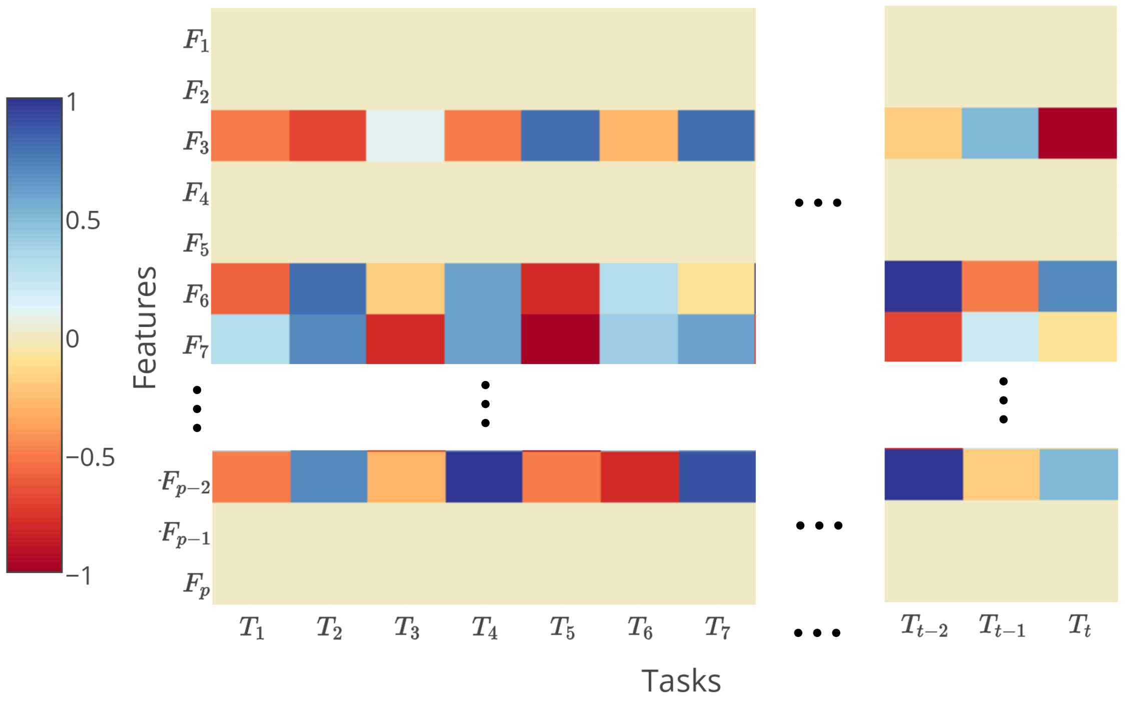


Figure 4.1: A schematic illustration for multi–task joint feature learning via $\ell_{2,1}$-norm. The resulting weight matrix has a similar sparse pattern across different tasks while each feature can have different weights on different tasks.

the subject–specific pattern recovery; 3) it encourages similar sparse pattern recovery across subjects. This characteristic provides the possibility of joint feature learning as it accounts for the similarity of neural responses to a similar stimulus across individuals. We hypothesize that the combination of the proposed multi–task learning and $\ell_{2,1}$ penalization provides a fair compromise in recovering the similarities and dissimilarities of the underlying neurophysiological activations across different subjects.

4.2.5 Experimental Materials

Simulated MEG Data

To evaluate the performance of the multi–task joint feature learning for spatio–temporal pattern recovery in a group–level MEG decoding scenario, we benchmarked it against common off–the–shelf approaches on a simulated MEG dataset. As the ground–truth effect is known in the simulations, we can reliably compare the quality of pattern recovery in different group–level decoding scenarios. To achieve this goal, we simulated sensor–space MEG data for 7 subjects. For each subject, we simulated two classes of MEG trials, each of which was composed by 250 epochs with a length of 330 ms at a 300 Hz sampling rate (so that we have 100 time–points for each MEG sensor). For all subjects, the whole scalp topography was simulated with a single dipole located at -4.7, -3.7, and 5.3 cm in the RAS (Right, Anterior, Superior) coordinate system [Figure 4.2(A)]. The position of the dipole location was arbitrary , but was close enough to the surface of the brain to provide stronger sensor–level patterns. To construct the temporal pattern of the target activity, the epochs of the positive class are, similarly across subject, constructed by adding up 3 components to the dipole time–course:

1. A time–locked effect composed of a positive 3 Hz half–cycle sinusoid peak, followed by a negative 5 Hz half–cycle sinusoid peak. The peaks are set 150 ± 3 and 250 ± 3 ms after the epoch onset, respectively [Figure 4.2(B)].

2. Uncorrelated background brain activity was simulated by summing 50 sinusoids with a random frequency from 1 to 125 Hz, and a random phase varied between 0 and 2π [Figure 4.2(C)]. In order to better mimic the actual magnetic features of the scalp surface, following the data simulation procedure described in Ref. [214], the amplitude of any single frequency component of the signal (the time–locked effect and the background noise) was set based on the empirically estimated spectral power of human brain activity.

3. White Gaussian noise was scaled with respect to the root mean square

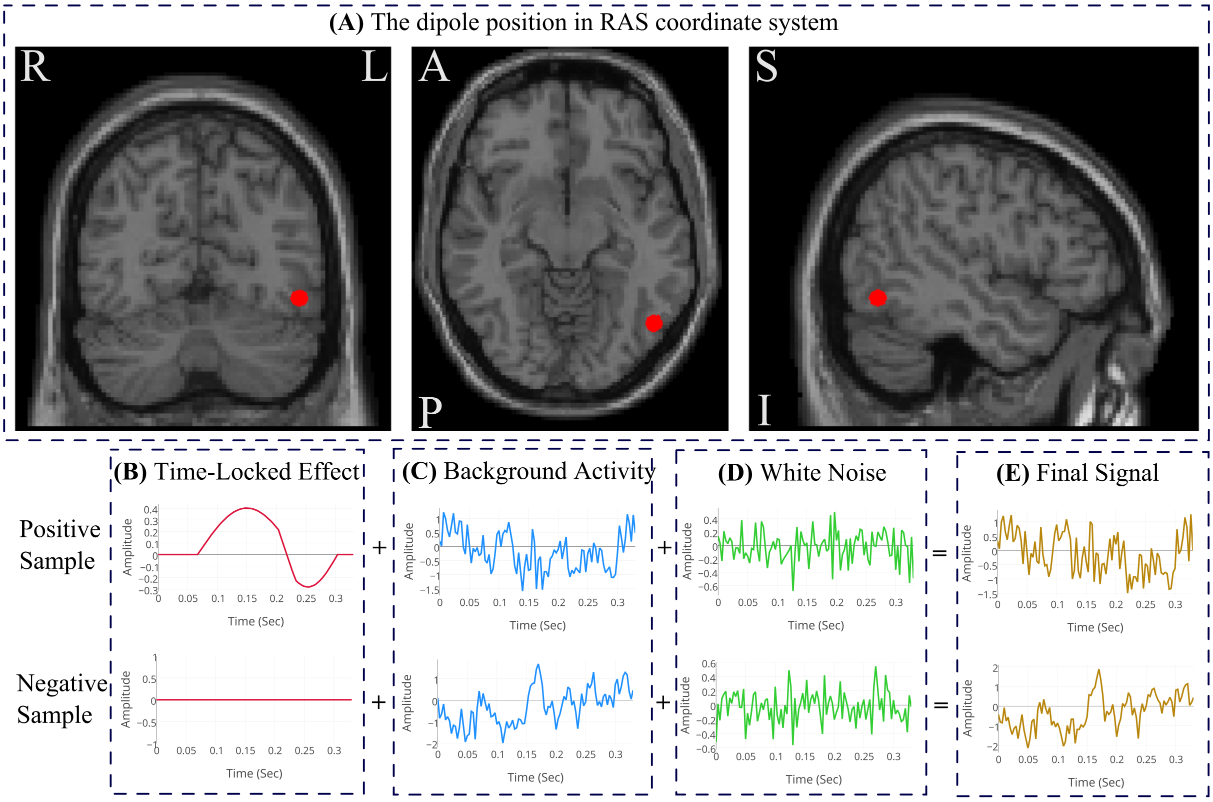


Figure 4.2: **(A)** The dipole position in the RAS coordinate system (the red circle). **(B)** The time–locked target effect is only present in the trials of the positive class. **(C)** The background brain activity is present in all simulated trials. **(D)** All trials are contaminated with white Gaussian noise. **(E)** An example of simulated trials in the positive and negative classes.

of the amplitude of signal in each epoch [Figure 4.2(D)].

The epochs of the negative class were constructed without the time–locked effect and by merely adding up the noise components (i.e., the background activity and the white noise). Therefore, the time–locked component is considered as the discriminating ground–truth pattern in our experiments.

To simulate the sensor–level variability across individuals, for each subject we used different orientation for the dipole in the source space. This variability in orientation of dipoles simulates directly dissimilar formations of gray matter, and indirectly simulates different head shapes and the position of the head inside the MEG helmet for a group of subjects. We set the orientation of dipoles as $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$, $[1, 1, 0]$, $[0, 1, 1]$, $[1, 0, 1]$, and $[1, 1, 1]$ for simulated subject 1 to 7, respectively. These differences in orientation are expected to provide different sensor–level spatial patterns across subjects. In the final step, the signal of 102 magnetometer sensors of the Elekta Neuromag system are simulated using a standard forward model algorithm implemented in the Fieldtrip toolbox [153]. Using brain decoding on the sensor–level simulated MEG data, a successful group–level pattern recovery approach should be able to recover the similar temporal pattern of the time–locked effect in 7 subjects despite the different topological distribution across sensors.

Real MEG Data

In order to evaluate the proposed method on real data, we employed the MEG dataset that is collected by Henson et al. [90][1]. This dataset includes MEG recordings for 16 subjects. In the experimental protocol, visual stimuli consisting of famous, unfamiliar, and scrambled faces are presented to subjects in a random order. MEG data were recorded using a Elekta Neuromag VectorView system. As in Ref. [151], we used the balanced face vs. scrambled dataset where the samples in the face category were randomly drawn from the trials of famous or unfamiliar faces [2]. The samples in the face and scrambled face categories are labeled as 1 and -1, respectively. The raw data was high–pass filtered at 1 Hz, down–sampled to 250 Hz, and trimmed from 200 ms before the stimulus onset to 800 ms after the stimulus onset. Thus, each trial has 250 time–points for each 306 MEG sensor (102 magnetometers and 204 planar gradiometers)[3]. To create the

[1] The full dataset is publicly available at `ftp://ftp.mrc-cbu.cam.ac.uk/personal/rik.henson/wakemandg_hensonrn/`

[2] The extracted dataset is used in DecMeg2014 competition and is publicly available at `https://www.kaggle.com/c/decoding-the-human-brain/data`

[3] The preprocessing scripts in python and MATLAB are available at: `https://github.com/FBK-NILab/DecMeg2014/`

feature vector of each sample, we pooled all of the temporal data of the 306 MEG sensors into one vector (i.e., we have $p = 250 \times 306 = 76,500$ features for each sample). Before training the classifier, the features were standardized to have a mean of 0 and standard–deviation of 1.

4.2.6 Classification and Evaluation

We compared our multi–task joint feature learning algorithm with single–subject decoding and pooling approaches in terms of decoding performance, reproducibility of brain maps, and quality of spatio–temporal pattern recovery. ℓ_1 and ℓ_2 penalization terms are used in both single–subject decoding and pooling scenarios [4]. Considering these 3 group–decoding approaches, and different penalization schemes, in total, 5 decoding methods are benchmarked on the simulated and real MEG datasets, namely: SS-L1, SS-L2, Pooling-L1, Pooling-L2, and MT-L21, respectively, single–subject decoding with ℓ_1 regularization, single–subject decoding with ℓ_2 regularization, pooling with ℓ_1 regularization, pooling with ℓ_2 regularization, and multi–task learning with $\ell_{2,1}$ regularization. We employ the implementation presented in MALSAR toolbox [220] for multi–task joint feature learning [5]. Algorithm 2 summarizes the pseudo–code for optimizing Eq. 4.4. In this algorithm, the outer *while* loop performs Nesterov's optimization [144] which is an optimal first–order black box method for smooth convex optimization. The inner *while* loop performs the efficient Euclidean projection onto a set of convex solutions [124].

The out–of–bag (OOB) [189] method with 50 bootstrap replications was used for computing the expected prediction error (EPE) at different regularization levels $\lambda = \{0.001, 0.1, 1, 5, 10, 25, 50, 100, 200, 300\}$. Then

[4] The MATLAB codes that are used for our experiments are made publicly available at `https://github.com/smkia/MTJFL_MEG`.

[5] See `https://github.com/jiayuzhou/MALSAR` for open-source toolbox implementation and documentation.

$1 - EPE$ is used as a measure for the generalization performance. To evaluate the reproducibility of brain maps, we adopt the reproducibility measure introduced in Ref. [108] (see Section 3.2.3).

Algorithm 2 The pseudo–code for optimizing Eq. 4.4. Let $\mathbf{X}_1, \dots, \mathbf{X}_s$, $\mathbf{y}_1, \dots, \mathbf{y}_s$, and λ be as defined in Section 4.2.3. The algorithm receives also *tol* and *maxIter*, i.e., the tolerance and the maximum iteration, as two stopping criteria. The algorithm return the weight matrix $\mathbf{W}_{MT} \in \mathbb{R}^{p \times s}$ as output. In addition to the notation in Section 4.2.1, $\mathbf{A}'$ represents the transpose of matrix $\mathbf{A}$, $\odot$ represents the element–wise matrix multiplication, and $a^{(i,j)}$ denotes the element of matrix A at the ith row and jth column. In all of our experiments, we set $tol = 10^{-4}$ and $maxIter = 1000$.

```
Input: $\mathbf{X}_1, \dots, \mathbf{X}_s; \mathbf{y}_1, \dots, \mathbf{y}_s; \lambda; tol; maxIter$
Output: $\mathbf{W}_{MT} \in \mathbb{R}^{p \times s}$
Initialize: $\mathbf{W}_0, \mathbf{W}_1, = 0^{p \times s}; v_0 = 0, v_1 = +\infty; \alpha_0 = 0; \alpha_1 = 1; \gamma = 1; iter = 1; \Delta = +\infty^{p \times s};$
while $\left| v_1 + \lambda \sum_{i=1}^{p} \left\| \mathbf{w}_1^{(i,:)} \right\|_2 - v_0 - \lambda \sum_{i=1}^{p} \left\| \mathbf{w}_0^{(i,:)} \right\|_2 \right| > tol \times (v_0 + \lambda \sum_{i=1}^{p} \left\| \mathbf{w}_0^{(i,:)} \right\|_2)$ & $iter \leq maxIter$ do
    $\mathbf{S} = \mathbf{W}_1 + \frac{\alpha_0 - 1}{\alpha_1} \times (\mathbf{W}_1 - \mathbf{W}_0)$
    for $t \leftarrow 1, s$ do
        $\mathbf{g}^{(:,t)} = \mathbf{X}_t'(\mathbf{X}_t \mathbf{s}^{(:,t)} - \mathbf{y}_t)$
    end of for
    $f = 0.5 \times \sum_{t=1}^{s} \left\| \mathbf{X}_t \mathbf{s}^{(:,t)} - \mathbf{y}_t \right\|_2^2$
    while $\|\Delta\|_F^2 > 10^{-20}$
        $\mathbf{U} = \mathbf{S} - \frac{\mathbf{G}}{\gamma}$
        $\eta = \frac{\lambda}{\gamma}$
        for $i \leftarrow 1, p$ do
            $\mathbf{l}^{(i,:)} = \max(0^{1 \times s}, 1 - \frac{\eta}{\|\mathbf{u}^{(i,:)}\|_2})$
        end of for
        $\mathbf{L} = \mathbf{L} \odot \mathbf{U}$
        $\Delta = \mathbf{L} - \mathbf{S}$
        $v_0 = v_1$
        $v_1 = 0.5 \times \sum_{t=1}^{s} \left\| \mathbf{X}_t \mathbf{l}^{(:,t)} - \mathbf{y}_t \right\|_2^2$
        if $v_1 > f + \sum_{t=1}^{s} \sum_{i=1}^{p} (\Delta^{(i,t)} \times g^{(i,t)}) + \frac{\gamma}{2} \times \|\Delta\|_F^2$ then do
            break the inner while loop.
        end of if
        $\gamma = \gamma \times 2$
    end of while
    $\mathbf{W}_0 = \mathbf{W}_1$
    $\mathbf{W}_1 = \mathbf{L}$
    $\alpha_0 = \alpha_1$
    $\alpha_1 = 0.5 \times (1 + \sqrt{1 + 4 \times \alpha_1^2})$
    $iter = iter + 1$
end of while
$\mathbf{W}_{MT} = \mathbf{W}_1$
```

4.3 Results

In this section, we compare the proposed multi–task joint feature learning with traditional single–subject and pooling approaches in a group multivariate analysis of MEG data. The comparisons are made based on the decoding performance, reproducibility of brain maps, and quality of the recovered spatio–temporal brain maps. Figure 4.3 shows generalization performance and the reproducibility of the 5 different methods on the simulated and real MEG data. In the case of simulated data the bar diagrams depict the average performance and reproducibility over 10 simulation runs and 7 simulated subjects. The results of the real MEG data are averaged over 16 subjects.

4.3.1 Simulated Data

Single–Subject Decoding

In the single–subject decoding scenario, ℓ_1 penalization provides higher generalization performance than ℓ_2, but this slight advantage in decoding performance leads to a substantial drop in the level of reproducibility of brain maps. The multicollinearity in the MEG data is the main reason behind this observation. The ℓ_1 penalization enforces strong sparsity on the parameters of the decoding model that makes the decoding process highly unstable, especially on the multicollinear input space. Due to the nature of the MEG signal, the independent variables are highly correlated in space and time. Therefore, slight changes in the training set (for example using perturbation techniques such as cross–validation) results in high variation on the weights of the classifier. Furthermore, it increases the chance of *miss–fitting* the classifier to spurious noise components that are partially correlated with informative components of the signal.

The sensor maps in Figure 4.4 depict the spatial distribution of the re-

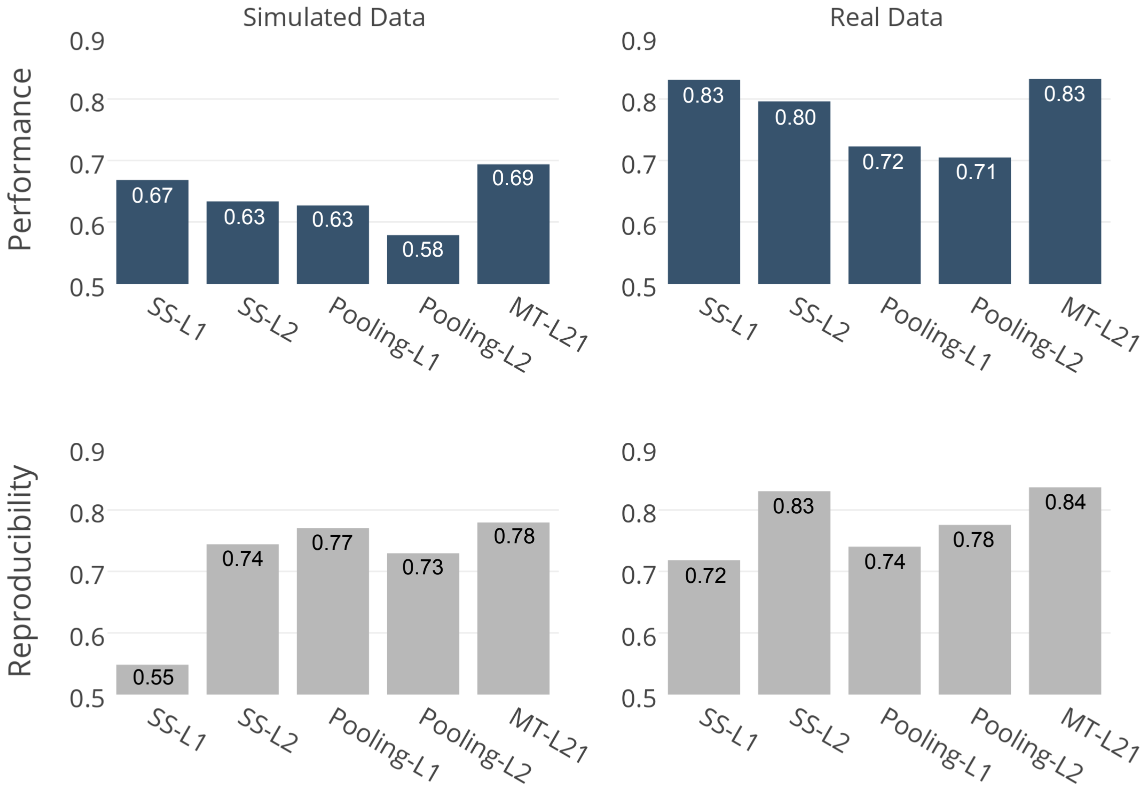


Figure 4.3: Comparison between the generalization performance and the reproducibility of the 5 different methods on the simulated and real MEG data. The results on the simulated data are averaged over 10 simulation runs and 7 simulated subjects. The results on the real MEG data are averaged over 16 subjects. MT-L21 provides the best decoding performance, while preserving the highest reproducibility level among other competing methods.

covered patterns for the 5 different methods tested on data from 7 simulated subjects. A comparison between the first (the ground–truth maps) and the second (the SS-L1 maps) columns of topographic sensor maps illustrates how the pattern recovery by means of ℓ_1 regularized classifier is affected by these deficits. The recovered maps via SS-L1 are over–attenuated in space compared to the ground–truth effect, because the correlated sensors are ignored by ℓ_1 penalized classifier, as they do not provide extra information for decoding. Further, SS-L1 recovers some extra spurious spatial patterns that are not present in the ground–truth maps.

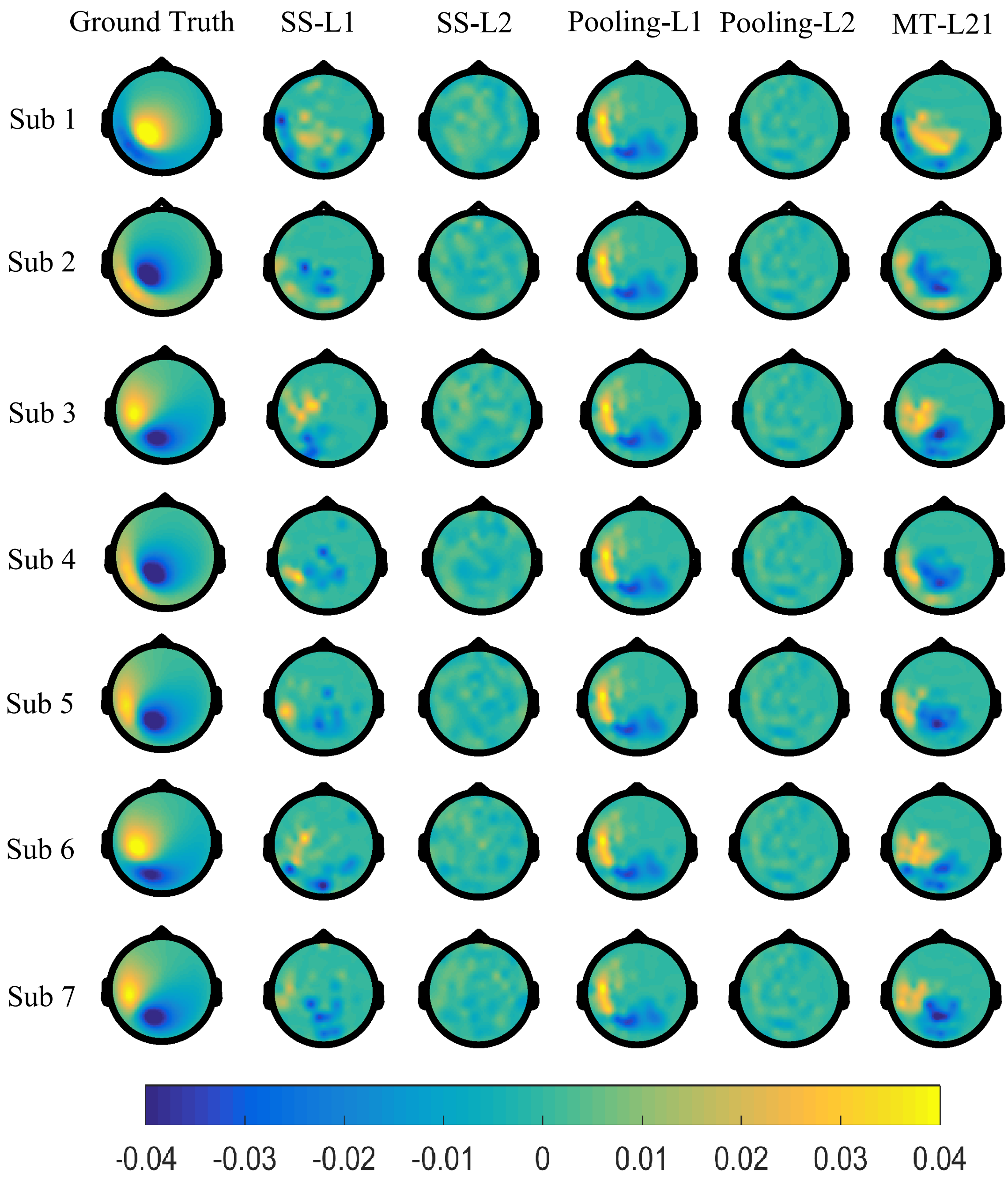


Figure 4.4: Topographic sensor maps of the ground–truth effect and the weight vectors computed using 5 different decoding approaches (columns) on 7 simulated subjects (rows). The weight vectors are normalized in the unit hyper–sphere. The maps show the averaged weights in 100 ms interval from 100 to 200 ms after the stimulus onset.

The same conclusions can be made for the temporal pattern recovery based on the temporal maps. Figure 4.5 shows the temporal maps of 5 different methods for the first three simulated subjects (see Appendix A.5 for similar maps on simulated subjects 4 to 7). The temporal patterns show the averaged classifier weights over the highlighted channels. The channels are selected based on the spatial distribution of the dipole in the ground–truth effect. Again the temporal pattern recovered by SS-L1 (the blue dashed line) has much less expansion in time compared to the ground–truth effect (the red line).

On the other hand, SS-L2 with a Gaussian prior assumption on the distribution of weights, provides a higher level of reproducibility than SS-L1, however it completely fails to recover the spatio–temporal pattern of the ground–truth effect (see the third column of Figure 4.4 and the dotted purple line in Figure 4.5). This fact is also well–reflected in Table 4.1, where the recovered maps using SS-L2 show substantially less cosine similarity with the ground–truth effect compared to SS-L1.

Pooling

The pooling method generally shows a lower performance than the single–subject and multi–task approaches (see Figure 4.3). This loss in performance is expected due to the wrong assumption on the similarity of P_i across different subjects [151]. On the other hand, both Pooling-L1 and Pooling-L2 approaches provide higher reproducibility than SS-L1. Putting the performance aside, the main problem of the pooling approach arises when the quality of pattern recovery matters. In pooling, since we come up with only one model for all subjects, the subject specific pattern recovery is impossible. In other words, the pooling approach ignores across–subject structural and functional differences, and provides only one brain map for all subjects. The fourth and the fifth columns of Figure 4.4 show the sim-

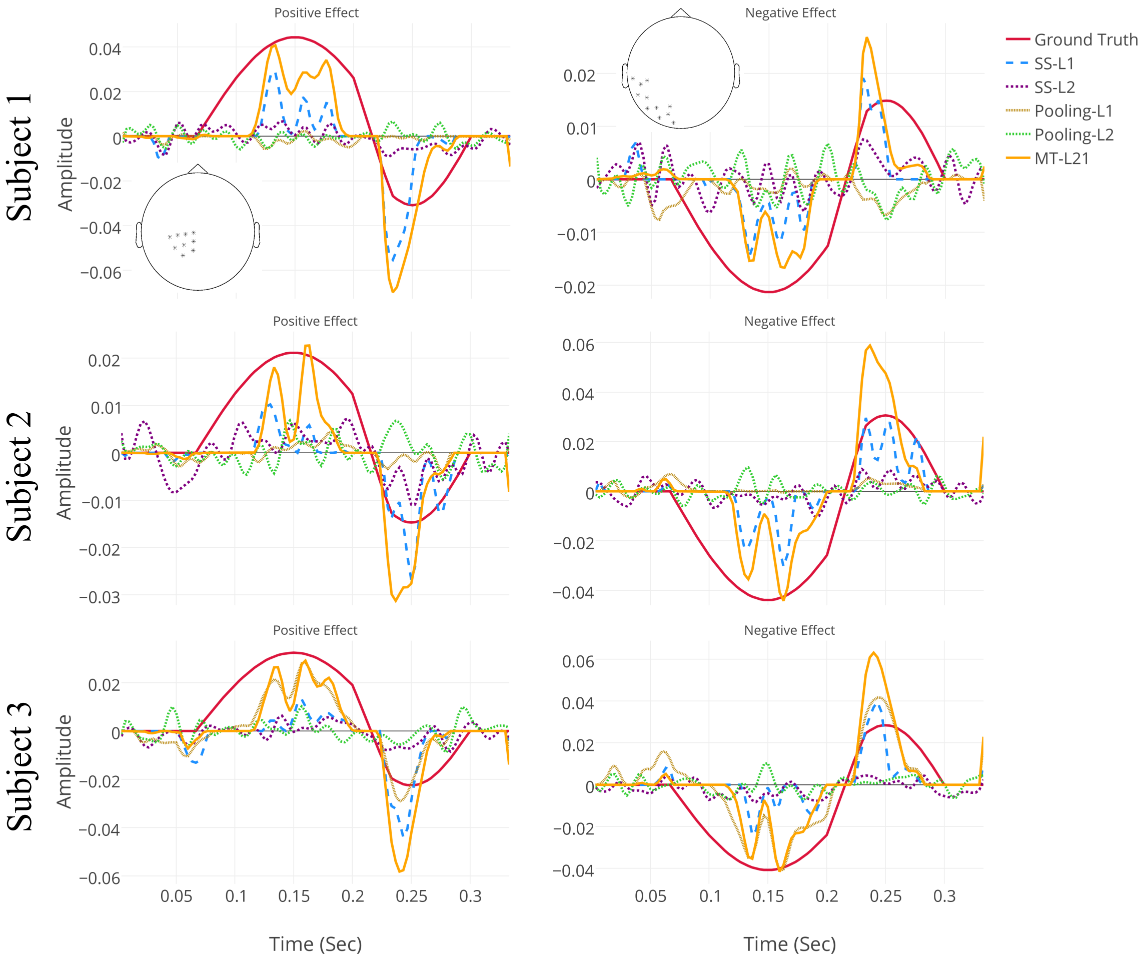


Figure 4.5: Comparison between the temporal maps of the 5 different decoding methods with the ground–truth effect, on data from the first three simulated subjects. The time courses are showing the temporal patterns of the recovered effect computed by averaging the weights of the classifier over the highlighted channels. The channels are selected based on the spatial distribution of the dipole in the ground–truth effect (see Figure 4.4).

ilar recovered spatial patterns of simulated MEG data for seven subjects. While in some subjects the recovered pattern by Pooling-L1 provides a fair representation of the ground–truth effect, in some subjects (for example subject 1) it completely misrepresents the ground–truth (see also Figure 4.5 for temporal patterns). Similar to single–subject decoding, the ℓ_2 penalization in the pooling scenario fails completely in spatio–temporal pattern recovery (see Table 4.1 for quantitative comparison).

Table 4.1: Cosine similarity between the recovered patterns for the 5 decoding methods and the ground truth effect. The numbers show the average and the standard deviation of cosine similarities between the ground–truth and brain maps in 10 simulation runs. The bold faced numbers show the best method for each subject. The last row of the table shows the mean similarity across subjects. MT-L21 maps are significantly more representative of the ground–truth effect than other benchmarked approaches.

	SS-L1	**SS-L2**	**Pooling-L1**	**Pooling-L2**	**MT-L21**
Sub1	0.36 ± 0.07	0.10 ± 0.01	-0.08 ± 0.02	0 ± 0.01	$\mathbf{0.62 \pm 0.05}$
Sub2	0.37 ± 0.07	0.10 ± 0.01	0.13 ± 0.02	0 ± 0.01	$\mathbf{0.63 \pm 0.05}$
Sub3	0.33 ± 0.03	0.11 ± 0.01	$\mathbf{0.60 \pm 0.03}$	0.06 ± 0.00	0.59 ± 0.03
Sub4	0.38 ± 0.04	0.11 ± 0.01	0.30 ± 0.02	0.02 ± 0.01	$\mathbf{0.64 \pm 0.05}$
Sub5	0.35 ± 0.05	0.10 ± 0.01	0.55 ± 0.03	0.05 ± 0.00	$\mathbf{0.62 \pm 0.05}$
Sub6	0.32 ± 0.03	0.11 ± 0.01	0.47 ± 0.02	0.05 ± 0.00	$\mathbf{0.57 \pm 0.03}$
Sub7	0.38 ± 0.05	0.11 ± 0.01	$\mathbf{0.61 \pm 0.03}$	0.06 ± 0.00	$\mathbf{0.61 \pm 0.03}$
Mean	0.36	0.11	0.37	0.04	**0.61**

Multi–Task Joint Feature Learning

The proposed multi–task joint feature learning method, MT-L21, achieves as high of performance as the single–subject decoding, while preserving high reproducibility like in the pooling approach. More importantly, it enables reliable subject–specific pattern recovery in time and space. This fact is well reflected in the topological plots in the sixth column of Figure 4.4. The recovered maps show a fair overlap with the ground–truth effect. This overlap is also reflected in the cosine similarity between the recovered maps and the ground–truth map in Table 4.1, where the MT-L21 map has a 0.61 average similarity across the 7 simulated subjects. The superiority of MT-L21 in decoding performance, reproducibility, and pattern recovery can be explained by its three main characteristics: 1) unlike the pooling method, and similar to single–subject decoding, it correctly assumes a different sampling distribution (and thus different noise distribution) across MEG recordings of different subjects, and therefore provides a higher generalization capability; 2) unlike the single–subject method, and

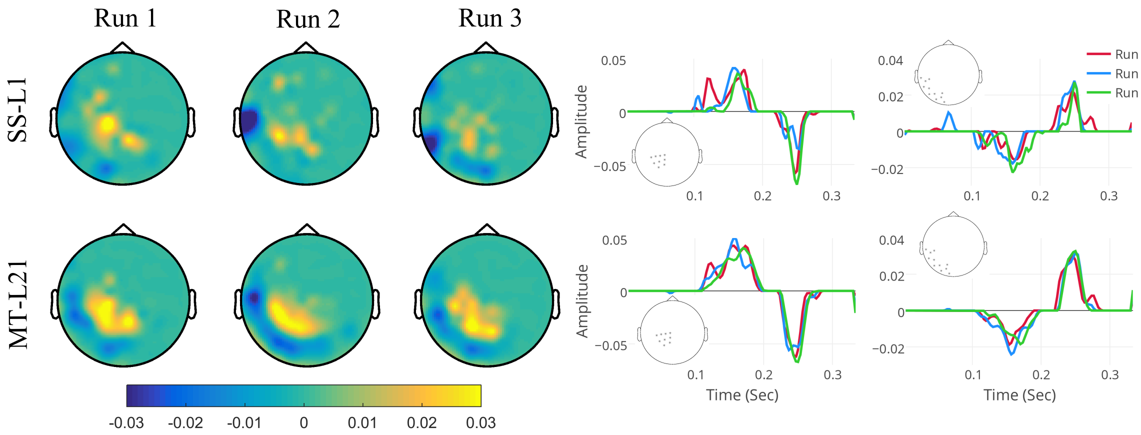


Figure 4.6: A comparison between the reproducibility of spatio–temporal maps in the SS-L1 and MT-L21 decoding approaches. The topographic maps are plotted by averaging the weights of the classifier between 100 and 200 ms in 3 simulation runs of simulated subject 1. The recovered time courses are plotted by averaging the weights over the highlighted channels. MT-L21 is more stable in recovering the spatio–temporal maps.

similar to the pooling approach, the classifier is trained simultaneously on all subjects. This specification alleviates the high dimensionality problem (as we train on more samples), and therefore provides more highly reproducible brain maps; 3) $\ell_{2,1}$ regularization enforces group sparsity in weight distributions. Thus the recovered maps are sparser than ℓ_2 regularization, and at the same time are more consistent in space and time from subject to subject than ℓ_1 regularization.

To compare the effect of the reproducibility of pattern recovery on the final interpretation of brain maps, we conducted a decoding experiment on the simulated MEG data. In this experiment, we compared pattern recovery in SS-L1 and MT-L21 (as they have the best generalization performance among the other methods) in three simulation runs. As described in Section 4.2.5, the distribution of noise in the simulated data is different from run to run. In fact, this difference simulates the across–session variation of the MEG data on a single subject.

The recovered patterns for SS-L1 and MT-L21 of simulated subject 1 across 3 simulation runs are shown in the first and the second row of Figure 4.6, respectively. The SS-L1 maps show higher run–to–run variation than the MT-L21 maps in both space and time. These kind of variations make the post–hoc interpretation of maps cumbersome, and they might lead to misinterpretation of results with respect to the actual underlying effect (see Figure 4.4 and Figure 4.5 for the ground–truth). On the other hand, MT-L21 shows a more stable pattern recovery in both space and time, where it consistently recovers the same correct dipole in the sensor space.

In common practice, generalization performance is the only criterion in model selection. This means that the hyper–parameters of decoding models are decided based only on model accuracy, rather than its ability for reliable pattern recovery. This approach may be shortsighted, especially when interpreting the spatio–temporal source of discriminative brain activity is desired [108]. Therefore, adopting decoding methods that show higher reproducibility of brain maps in addition to higher generalization performance facilitates the further interpretation of recovered maps.

4.3.2 Real MEG Data

Figure 4.7(A) depicts the scatter plot of the quality of 16 decoding models (for 16 subjects of real MEG data) in the performance–reproducibility plane. The distribution of the generalization performances across 16 subjects [Figure 4.7(B)] shows no statistically significant differences in performances for SS-L1 and MT-L21 (Wilcoxon's rank sum test p-value = 0.6538), while SS-L2 has significantly lower performance than the other two approaches (Wilcoxon's rank sum test p-value = 0.0125 and 0.0035, respectively). On the other side, SS-L1 has substantially lower reproducibility than SS-L2 and MT-L2 (Wilcoxon's rank sum test p-value = 8×10^{-7}, see [Figure 4.7(C)]). These results confirm the capability of MT-L21 in deliv-

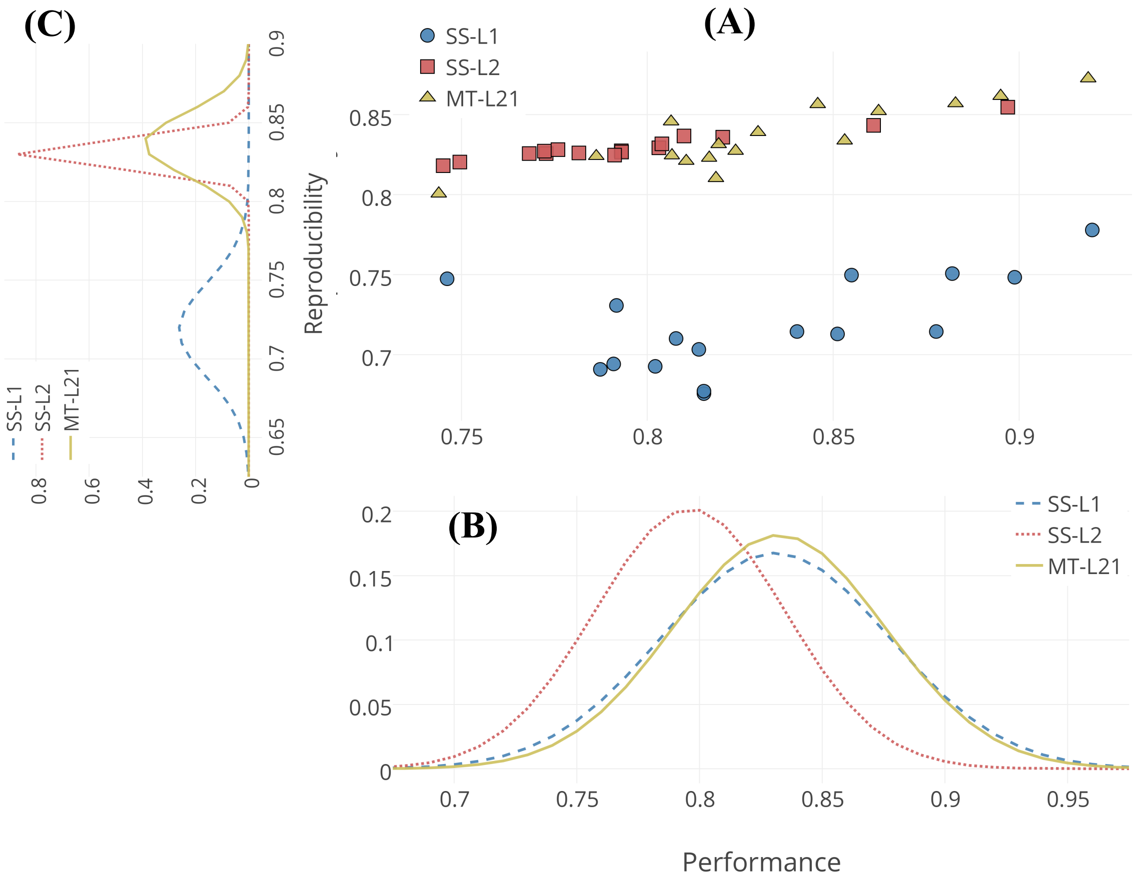


Figure 4.7: Comparison between the performance and reproducibility of SS-L1, SS-L2, and MT-L21 across 16 subjects of real MEG data. **(A)** The scatter plot of 16 decoding models in the performance–reproducibility plane. The circles represent subjects and the colors denote different methods. **(B)** The fitted normal distributions on the performance of 16 decoding models for 3 different approaches. **(C)** The fitted normal distributions on the reproducibility of 16 decoding models for 3 different approaches.

ering highly accurate individual models (same as SS-L1) while preserving the reproducibility of decoding models, across subjects.

Figure 4.8 illustrates the recovered spatio–temporal patterns by MT-L21 across 16 subjects of real MEG data (see Appendix A.6 for other methods). In almost all subjects, MT-L21 is able to spot an occipo–temporal dipole in the sensor space. The different position of the dipole from subject to subject is expected due to differences in head shapes, anatomical properties, and position of the head in the MEG helmet. Despite meaningful

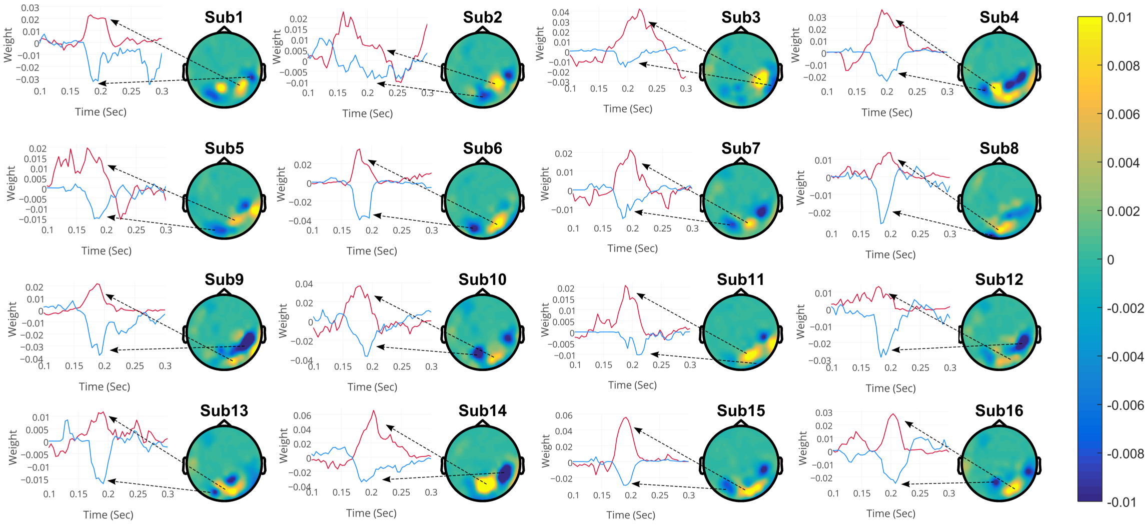


Figure 4.8: The recovered spatio–temporal representation of the N170 effect in 16 subjects from the real MEG dataset. The topoplots show the classifier weights for magnetometer sensors averaged in the 150 to 250 ms time period after stimulus onset. The corresponding plots represent the temporal dynamic of the dipole (red for the positive effect and blue for the negative effect) in the time dimension.

but different spatial patterns, the N170 effect is robustly recovered across almost all subjects around 200 ms after the stimulus onset. These results confirm the previous event–related potential/field analysis of EEG/MEG that shows that N170 occurs $130-200$ ms after the stimulus presentation, and reflects the neural processing of faces [19, 90].

4.4 Discussion

4.4.1 Higher Interpretability of Brain Maps in Multi–Subject Brain Decoding

The learned parameters of linear decoding models can be visualized in the form of brain maps. These brain maps can be used to explore the spatio–temporal origin of the underlying neurophysiological discriminating activity among two or several cognitve tasks, or types of stimuli. Despite theoretical advantages of brain mapping via brain decoding, such as higher sensitivity and specificity than the alternative univariate approaches, its application to inference on neuroimaging data is limited, primarily due to the lack of interpretability [88, 142, 172, 197]. From a cognitive neuroscience perspective, reproduciblility and neurophysiological plausiblilty of a brain map [108] are two necessary conditions for *interpretability* of its corresponding brain decoding model. There are two main reasons behind the interpretability problem: 1) the ill–posed nature of the brain decoding problem, where we have huge number of spatio–temporal features (order of 10^5) while the number of samples is limited (order of 10^2), this causes the generalization problem of over–fitting the model on the training set [30,118,133]; 2) multicollinearity [195] among predictors, where the strong correlation between spatio–temporal measurements of brain activity yields coefficient instability in linear brain decoding models [67]. Therefore, there is an emergent need to incorporate structural and functional prior knowledge on brain segregation and integration, in order to achieve stable, reliable, and interpretable brain maps. There are two main directions in the literature toward this goal: 1) employing structured penalization techniques; 2) reducing the variance of feature selection via enhanced stability selection.

Structured regularization approaches combine intelligently the basic regularizers (such as ℓ_1 and ℓ_2) in order to take advantage of any prior-knowledge the experimenter may have about the correlational structure in the neuroimaging data. Group Lasso [217] and total–variation penalty [186] are two effective methods for enforcing spatio–temporal structure of covariance between predictors into the regularization [168,212]. Group–wise regularization [192], smoothed–sparse logistic regression [46], total–variation ℓ_1 penalization [62, 133], and the graph–constrained elastic–net [67], are examples of structured regularization methods that are used effectively in the brain decoding context. On the other hand, stability selection is an ensemble learning method to reduce the variance of feature selection for high dimensional data analysis [173] that has recently received high attention in the context of multivariate analysis of brain recordings (see for example Refs. [195] and [201]).

Despite the aforementioned efforts, so far less attention is devoted to improving the interpretability of brain decoding models in group–level brain decoding. In fact, employing structured regularization or stability selection approaches in a single–task brain decoding framework still suffers from the inadequacy of single–subject or pooling methods in multi–subject multivariate analysis of neuroimaging data. On the other hand, multi–task joint feature learning provides the infrastructure for combining structured regularization with stability selection in group–level multivariate analysis. While $\ell_{2,1}$ penalty combines ℓ_2 and ℓ_1 norms to enforce group sparsity, its integration with simultaneous optimization in multi–task learning also offers a variant of stability selection across a group of subjects. By taking into account the inter–subject spatio–temporal similarities and dissimilarities of brain activity, multi–task joint feature learning provides higher interpretability for multivariate brain maps at the group–level, as supported by our experimental results.

4.4.2 Related Work

The problem of recovering subject–specific brain responses is discussed in several recent neuroimaging studies. In this section, we review some related studies that tried to handle this problem at the preprocessing or decoding stages.

The problem of characterizing the fine–level distinctive patterns in population response topographies was first elucidated by Haxby et al. [87], where the authors presented a novel functional alignment method, called *hyper–alignment*, in order to derive a set of basis functions that are common across different individuals. They then modeled individual cortical response patterns as weighted sums of these basis functions. In a single–subject fMRI decoding scenario, they showed the superiority of the decoding performance in the hyper–aligned common space over the anatomically aligned data. Further, they showed the back–projection of predictive basis functions in the common space to the subject native space provides subject–specific distinctive spatial maps. One practical limitation of this method is in estimating the parameters of hyper–alignment, i.e., basis functions, that should be performed on a separate dataset (preferably in response to natural complex stimuli such as a movie). In addition, to the best of our knowledge, at this time the application of hyper–alignment remains limited to fMRI data.

To overcome inter–subject variability in group analysis of fMRI data, Takerkart et al. [180] introduced a graph–based support vector classification approach for across–subject multivariate pattern analysis. In this method, an unsupervised learning approach is employed to construct attribute graphs for fMRI data, where each node has two attributes, namely location and activation. A support vector machine classifier with graph kernel is used for classification in graph space. The authors hypothesized that the inter–subject variability can be characterized based on different node attributes across subjects. Despite high generalization performance in a pooling decoding scenario, their proposed method lacks transparency in their model, due to the non–linear nature of the employed classifier. In addition, due to the single–task nature, it suffers from all of the limitations of the single–subject and pooling decoding scenarios. Another study attempted to take into account both the similarity in macro–structures, and the dissimilarity in micro–structures of brain activity across different individuals' brains. Rao et al. [165] proposed a sparse overlapping group lasso (SOGLasso) method to learn both the commonalities and the differences across brains in an fMRI study. To do this, the authors introduced a new penalty by combining ℓ_2 and ℓ_1 to promote both inter and inner group sparsity. In spite of higher flexibility of SOGLasso over the pure group–wise regularization in feature selection, no practical solution is suggested for group–level decoding of neuroimaging data.

The idea of recasting the multi–subject brain decoding problem to a MTL framework was first presented by Marquand et al. [129] where the authors defined the input data of each subject as a task. The Gaussian process MTL was employed in order to model the relationship, and to induce coupling between tasks. To visualize the discriminative brain maps, a transformation from function space to weight space is used to compute the predictive weight vectors in the input space. Then, a procedure called predictive mapping, in combination with permutation one–sample t–tests is used to identify discriminating regions. On an fMRI dataset, they showed the employed MTL approach provides more accurate and reproducible models than single–subject and pooling strategies. This work shares similarities with the approach presented in this paper, given that both adopt an MTL framework, however our approach offers a major advantage because our method allows for sparse pattern recovery by applying $\ell_{2,1}$ penalization. Importantly, the sparse recovered patterns provide a more convenient post–hoc interpretation of brain maps. Another minor advantage of our method is its computational simplicity, where unlike in Ref. [129], there is no need for estimating covariance matrices on a small number of samples of high dimensional input data in the decoding or visualization phase.

4.4.3 Limitation and Future work

The proposed multi–task joint feature learning framework uses $\ell_{2,1}$ regularization to impose structured group sparsity in brain pattern recovery.

Despite the experimental success presented in this study, one challenge is that, $\ell_{2,1}$ blindly encourages similar sparse patterns across different subjects. In other words, it does not consider the possibility of different inter and inner group sparsity profiles across different subjects. Enforcing extra prior information regarding the structure of data in time and space can provide these possibilities and lead to a new generation of enriched pattern recovery methods. One possible way to move toward this goal is to encode higher level prior information in the form of a graph, and then add an extra graph–fused penalty term to the current optimization scheme. This change can lead to a convex minimization problem involving the sum of a smooth function (the loss function), a non–smooth proximable function ($\ell_{2,1}$ penalty term), and the composition of a proximable function with a linear operator (the graph–fused term) that can be solved using the first–order splitting algorithm proposed in Ref. [39]. Implementation of the first–order splitting algorithm for the multi–task learning, and comparing its quality in pattern recovery with $\ell_{2,1}$ is an important future direction for our work.

Chapter 5

Conclusions

The primary goal of this thesis was to reduce the knowledge extraction gap in multivariate analysis of neuroimaging data by improving the interpretability of linear brain decoding models. Considering the importance of group–level inference in cognitive neuroscience studies and numerous challenges in this direction, the secondary goal was focused on exploring more effective decoding methods that are capable of recovering structural and functional similarities and dissimilarities in a group–level analysis of neuroimaging data.

To this end, first we presented a novel theoretical definition for the interpretability of linear brain decoding and associated multivariate brain maps. We demonstrated how the interpretability can be decomposed to the representativeness and reproducibility of a linear brain decoding model. This decomposition explains the relationship between the influential cooperative factors in the interpretability of brain decoding models and highlights the possibility of indirect and partial evaluation of interpretability by measuring these effective factors. The presented definition provides a first step toward practical solution for filling the knowledge extraction gap in linear brain decoding and it provides a theoretical background to explain a previously ambiguous concept in the brain decoding context. To provide a proof of concept, a heuristic approach based on the contrast event–related field is exemplified for practical evaluation of the interpretability in multivariate recovery of evoked MEG responses. We further proposed to combine the interpretability and the performance of the brain decoding as a new Pareto optimal multi–objective criterion for model selection. We experimentally showed that considering the interpretability of brain decoding models in the model selection procedure has a positive effect on the human interpretation of multivariate brain maps compensating a negligible amount of performance. Collectively, our methodological and experimental achievements can be considered a complementary theoretical and practical effort that contributes to researches on enhancing the interpretability of multivariate pattern analysis. Despite theoretical and practical advantages, the proposed definition and quantification of interpretability only applies to linear models, therefore extending the definition of interpretability to non–linear models demands future research into the visualization of non–linear models in the form of brain maps. Furthermore, the application of the proposed heuristic for approximating the interpretability is limited to the time–locked MEG responses, thus finding physiologically relevant heuristics for other acquisition modalities such as fMRI, and other brain responses such as induced responses can be also considered as possible directions in future work.

Second, we presented an application of multi–task joint feature learning in multi–subject decoding of MEG data where the MEG recording of each subject is defined as a task in the multi–task classification paradigm and $\ell_{2,1}$ regularization is used to recover sparse heterogeneous patterns of brain activity across different individuals. The proposed framework provides the possibility of consistent sparse pattern recovery across different individuals while at the same time preserving idiosyncratic structural and functional properties, thus yields higher interpretability for multivariate brain maps at the group–level. In addition, multi–task joint feature learning provides the infrastructure for combining structured regularization with stability selection in group–level multivariate analysis. While $\ell_{2,1}$ penalty combines ℓ_2 and ℓ_1 norms to enforce group sparsity, its integration with simultaneous optimization in multi–task learning also offers a variant of stability selection across a group of subjects. Considering the importance of group–level inference in neuroimaging context, and inadequacy of classical univariate and multivariate approaches in group–level analysis, the proposed approach can provide a methodological shift toward higher sensitive and at the same time higher interpretable brain decoding models. Our experiments on synthetic and real MEG data demonstrated the superiority of proposed approach in reproducibility and quality of recovered patterns over the traditional single–subject and pooling approaches. To the best of our knowledge, our effort for the first time addresses the problem of across subject pattern recovery in MEG decoding. Considering the high temporal and spatial resolution of MEG brain recordings, the proposed approach provides the possibility for recovering temporal brain dynamics within the millisecond time scale with a fair spatial granularity. Our future plan is to improve group–level pattern recovery by enforcing extra structural spatio–temporal prior knowledge via adding a graph–fused penalty term to the current optimization scheme. Hopefully this addition provides the possibility of accounting for different inter and inner group sparsity profiles across different individuals.

Our contributions aimed to extend the state of the art in multi–disciplinary researches for reliable, reproducible, and plausible inference on neuroimaging data, by facilitating the application of *brain decoding for brain mapping*. We hope this thesis contributes a tiny step toward answering historical questions in understanding the brain and its functions.

Bibliography

[1] Mojtaba Khomami Abadi, Seyed Mostafa Kia, Ramanathan Subramanian, Paolo Avesani, and Nicu Sebe. Decoding affect in videos employing the MEG brain signal. In *International Conference and Workshops on Automatic Face and Gesture Recognition*. IEEE, 2013.

[2] Mojtaba Khomami Abadi, Seyed Mostafa Kia, Ramanathan Subramanian, Paolo Avesani, and Nicu Sebe. User–centric affective video tagging from MEG and peripheral physiological responses. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013.

[3] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. DECAF: MEG–based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, July 2015.

[4] Babak Afshin-Pour, Hamid Soltanian-Zadeh, Gholam-Ali Hossein-Zadeh, Cheryl L Grady, and Stephen C Strother. A mutual information–based metric for evaluation of fMRI data–processing approaches. *Human brain mapping*, 32(5):699–715, 2011.

[5] Charu C Aggarwal and Philip S Yu. A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623, 2009.

[6] Antii I. Ahonen, Matti S. Hämäläinen, Risto J. Ilmoniemi, Matti J. Kajola, Jukka E. T. Knuutila, Juha T. Simola, and Vias A. Vilkman. Sampling theory for neuromagnetic detector arrays. *IEEE Transactions on Biomedical Engineering*, 40(9):859–869, Sept 1993.

[7] Ariana Anderson, Jennifer S Labus, Eduardo P Vianna, Emeran A Mayer, and Mark S Cohen. Common component classification: What can we learn from machine learning? *Neuroimage*, 56(2):517–524, 2011.

[8] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi–task feature learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 41–48, Cambridge, MA, USA, 2006. MIT Press.

[9] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi–task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[10] Yaniv Assaf and Ofer Pasternak. Diffusion tensor imaging DTI–based white matter mapping in brain research: a review. *Journal of Molecular Neuroscience*, 34(1):51–61, 2008.

[11] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel–wise explanations for non–linear classifier decisions by layer–wise relevance propagation. *PloS one*, 10(7), 2015.

[12] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[13] Dale L Bailey, David W Townsend, Peter E Valk, and Michael N Maisey. *Positron emission tomography.* Springer, 2005.

[14] Luca Baldassarre, Massimiliano Pontil, and Janaina Mourao-Miranda. Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding. *Frontiers in Neuroscience*, 11(62):1–15, 2017.

[15] Yoav Benjamini. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[17] Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step–up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[18] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[19] Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological studies of face perception in humans. *Journal of cognitive neuroscience*, 8(6):551–565, 1996.

[20] Hans Berger. Über das elektrenkephalogramm des menschen. *European Archives of Psychiatry and Clinical Neuroscience*, 87(1):527–570, 1929.

[21] James O. Berger. Could fisher, jeffreys and neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 02 2003.

[22] Michel Besserve, Karim Jerbi, Francois Laurent, Sylvain Baillet, Jacques Martinerie, and Line Garnero. Classification methods for ongoing EEG and MEG signals. *Biological research*, 40(4):415–437, 2007.

[23] Felix Bießmann, Sven Dähne, Frank C Meinecke, Benjamin Blankertz, Kai Görgen, Klaus-Robert Müller, and Stefan Haufe. On the interpretability of linear multivariate neuroimaging analyses: filters, patterns and their relationship. In *Proceedings of the 2nd NIPS Workshop on Machine Learning and Interpretation in Neuroimaging*, Harrahs and Harveys, Lake Tahoe, 2012.

[24] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single–trial analysis and classification of erp componentsa tutorial. *NeuroImage*, 56(2):814–825, 2011.

[25] L Bokobza. Near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 6:3–18, 1998.

[26] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[27] Nashaat Boutros, Michael W Torello, Elizabeth M Burns, Shu-Shieh Wu, and Henry A Nasrallah. Evoked potentials in subjects at risk for alzheimer's disease. *Psychiatry research*, 57(1):57–63, 1995.

[28] Michael Brammer. The role of neuroimaging in diagnosis and personalized medicine–current position and likely future directions. *Dialogues in Clinical Neuroscience*, 11(4):389–396, 2009.

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] Kay H. Brodersen, Florent Haiss, Cheng S. Ong, Fabienne Jung, Marc Tittgemeyer, Joachim M. Buhmann, Bruno Weber, and Klaas E. Stephan. Model–based feature construction for multivariate decoding. *NeuroImage*, 56(2):601–615, May 2011.

[31] Edward Bullmore, Michael Brammer, Steve CR Williams, Sophia Rabe-Hesketh, Nicolas Janot, Anthony David, John Mellers, Robert Howard, and Pak Sham. Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277, 1996.

[32] Danilo Bzdok, Gaël Varoquaux, and Bertrand Thirion. Neuroimaging research from null–hypothesis falsification to out–of–sample generalization. *Educational and Psychological Measurement*, pages 1–13, 2016.

[33] Massimiliano Caramia and Paolo Dell'Olmo. *Multi–objective management in freight logistics: Increasing capacity, service level and safety with optimization algorithms*. Springer London, 2008.

[34] Melissa K Carroll, Guillermo A Cecchi, Irina Rish, Rahul Garg, and A Ravishankar Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112–122, 2009.

[35] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[36] Alexander M Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S Cash. Decoding word and category–specific spatiotemporal representations from MEG and EEG. *Neuroimage*, 54(4):3028–3039, 2011.

[37] David Cohen. Magnetoencephalography: evidence of magnetic fields produced by alpha–rhythm currents. *Science*, 161(3843):784–786, 1968.

[38] David Cohen. Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. *Science*, 175(4022):664–666, 1972.

[39] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[40] Bryan R. Conroy, Jennifer M. Walz, and Paul Sajda. Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fmri decoding models. *PLOS ONE*, 8(11):1–11, 11 2013.

[41] David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fMRI)brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.

[42] Enrico Crivellato and Domenico Ribatti. Soul, mind, brain: Greek philosophy and the birth of neuroscience. *Brain research bulletin*, 71(4):327–336, 2007.

[43] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

[44] Sanjeeb Dash, Dmitry M Malioutov, and Kush R Varshney. Learning interpretable classification rules using sequential rowsampling. In

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3337–3341, South Brisbane, 2015.

[45] Tyler Davis, Karen F LaRocque, Jeanette A Mumford, Kenneth A Norman, Anthony D Wagner, and Russell A Poldrack. What do differences between multi–voxel and univariate analysis mean? how subject–, voxel–, and trial–level variance impact fMRI analysis. *NeuroImage*, 97:271–283, 2014.

[46] Matthew de Brecht and Noriko Yamagishi. Combining sparseness and smoothness improves classification accuracy and interpretability. *NeuroImage*, 60(2):1550–1561, 2012.

[47] 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.

[48] Joseph Dien and Alecia M Santuzzi. Application of repeated measures ANOVA to high–density ERP. In *Event–related potentials: A methods handbook*, chapter 4, pages 57–81. MIT press, Cambridge, Massachusetts, 2004.

[49] Pedro Domingos. A unified bias–variance decomposition for zero–one and squared loss. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 564–569. AAAI Press, 2000.

[50] Emanuel Donchin, Kevin M Spencer, and Ranjith Wijesinghe. The mental prosthesis: assessing the speed of a P300–based brain–computer interface. *IEEE transactions on rehabilitation engineering*, 8(2):174–179, 2000.

[51] Bradley Efron. *Bootstrap Methods: Another Look at the Jackknife*, pages 569–593. Springer New York, New York, NY, 1992.

[52] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

[53] Kara D Federmeier and Marta Kutas. Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, 40(7):730–747, 2002.

[54] Aaron G Filler. The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTI. *Nature Precedings*, 7(1):1–69, 2009.

[55] Ronald Aylmer Fisher. *Statistical methods for research workers.* Genesis Publishing Pvt Ltd, 1925.

[56] Cynthia HY Fu, Robin Murray, Tamara Russell, Carl Senior, and Daniel Roy Weinberger. *Neuroimaging in psychiatry.* Martin Dunitz London, 2003.

[57] Michael S Gazzaniga. *The cognitive neurosciences.* MIT press, 2004.

[58] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[59] Pouya Ghaemmaghami, Mojtaba Khomami Abadi, Seyed Mostafa Kia, Paolo Avesani, and Nicu Sebe. Movie genre classification by exploiting MEG brain signals. In *International Conference on Image Analysis and Processing*, pages 683–693. Springer International Publishing, 2015.

[60] Steven N. Goodman. Multiple comparisons, explained. *American Journal of Epidemiology*, 147(9):807–812, 1998.

[61] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013.

[62] Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Identifying predictive regions from fMRI with TV–L1 prior. In *International Workshop on Pattern Recognition in Neuroimaging*, pages 17–20, Philadelphia, PA, 2013.

[63] Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify. In Georg Langs, Irina Rish, Moritz Grosse-Wentrup, and Brian Murphy, editors, *Machine Learning and Interpretation in Neuroimaging*, pages 9–16. Springer, Berlin, 2012.

[64] David M Groppe, Thomas P Urbach, and Marta Kutas. Mass univariate analysis of event–related brain potentials/fields i: A critical tutorial review. *Psychophysiology*, 48(12):1711–1725, 2011.

[65] David M Groppe, Thomas P Urbach, and Marta Kutas. Mass univariate analysis of event–related brain potentials/fields ii: Simulation studies. *Psychophysiology*, 48(12):1726–1737, 2011.

[66] Logan Grosenick, Stephanie Greer, and Brian Knutson. Interpretable classifiers for fMRI improve prediction of purchases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(6):539–548, 2008.

[67] Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole–brain prediction analysis with graphnet. *NeuroImage*, 72:304–321, 2013.

[68] Matti Hämäläinen and Riitta Hari. Magnetoencephalographic characterization of dynamic brain activation: Basic principles and methods of data collection and source analysis. *Brain mapping: The methods*, pages 227–254, 2002.

[69] Matti Hämäläinen, Riitta Hari, Risto J. Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. Magnetoencephalography¯theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:413–497, Apr 1993.

[70] Matti S Hämäläinen and Risto J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32(1):35–42, 1994.

[71] Todd C Handy. *Event–related potentials: A methods handbook*. MIT press, 2005.

[72] Katja Hansen, David Baehrens, Timon Schroeter, Matthias Rupp, and Klaus-Robert Müller. Visual interpretation of kernel–based prediction models. *Molecular Informatics*, 30(9):817–826, 2011.

[73] Riitta Hari. *Activation of the Human Auditory Cortex by Various Sound Sequences : Neuromagnetic Studies*, pages 87–92. Springer US, Boston, MA, 1989.

[74] Riitta Hari. Magnetoencephalography studies of action observation. *New Frontiers in Mirror Neurons Research*, page 58, 2015.

[75] Riitta Hari, Sari Levänen, and Tommi Raij. Timing of human cortical functions during cognition: role of MEG. *Trends in Cognitive Sciences*, 4(12):455–462, 2000.

[76] Riitta Hari and Lauri Parkkonen. MEG in the study of higher cortical functions. *Progress in epileptic disorders*, 5:103–112, 2008.

[77] Riitta Hari and Lauri Parkkonen. The brain timewise: how timing shapes and supports brain function. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1668), 2015.

[78] Riitta Hari, Lauri Parkkonen, and Cathy Nangini. The brain in time: insights from neuromagnetic recordings. *Annals of the New York Academy of Sciences*, 1191(1):89–109, 2010.

[79] Riitta Hari and Riitta Salmelin. Magnetoencephalography: from SQUIDs to neuroscience: Neuroimage 20th anniversary special edition. *Neuroimage*, 61(2):386–396, 2012.

[80] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning, 2001.

[81] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer, New York, 2009.

[82] Stefan Haufe, Sven Dähne, and Vadim V. Nikulin. Dimensionality reduction for the analysis of brain oscillations. *NeuroImage*, 101:583–597, July 2014.

[83] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2013.

[84] Stefan Haufe, Frank Meinecke, Kai Gorgen, Sven Dahne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biessmann. Parameter interpretation, regularization and source localization in multivariate linear models. In *International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4, Tubingen, 2014.

[85] James V Haxby. Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage*, 62(2):852–855, 2012.

[86] James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539):2425–2430, September 2001.

[87] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high–dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[88] John-Dylan Haynes. A primer on pattern–based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, 87(2):257–270, 2015.

[89] John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, July 2006.

[90] Richard N. Henson, Daniel G. Wakeman, Vladimir Litvak, and Karl J. Friston. A Parametric Empirical Bayesian framework for the EEG/MEG inverse problem: generative models for multisubject and multimodal integration. *Frontiers in Human Neuroscience*, 5(76), 2011.

[91] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[92] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[93] Heikki Huttunen, Tapio Manninen, Jukka-Pekka Kauppi, and Jussi Tohka. Mind reading with regularized multinomial logistic regression. *Machine vision and applications*, 24(6):1311–1325, 2013.

[94] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[95] Harold Jeffreys. *The theory of probability.* OUP Oxford, 1998.

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

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

[98] David D Jensen and Paul R Cohen. Multiple comparisons in induction algorithms. *Machine Learning*, 38(3):309–338, 2000.

[99] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

[100] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven A Siegelbaum, and AJ Hudspeth. *Principles of neural science*, volume 4. McGraw–hill New York, 2000.

[101] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[102] Jukka-Pekka Kauppi, Lauri Parkkonen, Riitta Hari, and Aapo Hyvärinen. Decoding magnetoencephalographic rhythmic activity using spectrospatial information. *NeuroImage*, 83:921–936, 2013.

[103] DL Keene, S Whiting, and ECG Ventureyra. Electrocorticography. *Epileptic Disorders*, 2(1):57–64, 2000.

[104] Seyed Mostafa Kia, Emanuele Olivetti, and Paolo Avesani. Discrete cosine transform for MEG signal decoding. In *International Workshop on Pattern Recognition in Neuroimaging*, pages 132–135. IEEE, 2013.

[105] Seyed Mostafa Kia and Andrea Passerini. Interpretability in linear brain decoding. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.

[106] Seyed Mostafa Kia, Fabian Pedregosa, Anna Blumenthal, and Andrea Passerini. Group–level spatio–temporal pattern recovery in MEG decoding using multi–task joint feature learning. Submitted, 2017.

[107] Seyed Mostafa Kia, Sandro Vega-Pons, Emanuele Olivetti, and Paolo Avesani. Multi–task learning for interpretation of brain decoding models. In Irina Rish, Georg Langs, Leila Wehbe, Guillermo Cecchi, Kai-min Kevin Chang, and Brian Murphy, editors, *Machine Learning and Interpretation in Neuroimaging: 4th International Workshop, MLINI 2014, Held at NIPS 2014, Montreal, QC, Canada, December 13, 2014, Revised Selected Papers*, pages 3–11. Springer International Publishing, Cham, 2016.

[108] Seyed Mostafa Kia, Sandro Vega Pons, Nathan Weisz, and Andrea Passerini. Interpretability of multivariate brain maps in linear brain decoding: Definition, and heuristic quantification in multivariate analysis of MEG time–locked effects. *Frontiers in Neuroscience*, 10(619):1–22, 2017.

[109] S Knake, E Halgren, H Shiraishi, K Hara, HM Hamer, PE Grant, VA Carr, D Foxe, S Camposano, E Busa, et al. The value of multichannel MEG and EEG in the presurgical evaluation of 70 epilepsy patients. *Epilepsy research*, 69(1):80–86, 2006.

[110] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[111] Ron Kohavi. A study of cross–validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[112] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information–based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868, 2006.

[113] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.

[114] Stephen LaConte, Stephen Strother, Vladimir Cherkassky, Jon Anderson, and Xiaoping Hu. Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, 26(2):317–329, 2005.

[115] Georg Langs, Bjoern H Menze, Danial Lashkari, and Polina Golland. Detecting stable distributed patterns of brain activation using gini contrast. *NeuroImage*, 56(2):497–507, 2011.

[116] Celeste D Lefebvre, Yannick Marchand, Gail A Eskes, and John F Connolly. Assessment of working memory abilities using an event–related brain potential (ERP)–compatible digit span backward task. *Clinical Neurophysiology*, 116(7):1665–1680, 2005.

[117] Erich Leo Lehmann. The fisher, Neyman–Peerson theories of testing hypotheses: One theory or two? In *Selected Works of EL Lehmann*, pages 201–208. Springer, 2012.

[118] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[119] Johannes Lenhard. Models and statistical inference: The controversy between fisher and neyman–pearson. *The British journal for the philosophy of science*, 57(1):69–91, 2006.

[120] Eric C Leuthardt, Gerwin Schalk, Jonathan R Wolpaw, Jeffrey G Ojemann, and Daniel W Moran. A brain–computer interface using electrocorticographic signals in humans. *Journal of neural engineering*, 1(2):63–71, 2004.

[121] Caiyan Li and Hongzhe Li. Network–constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[122] Chinghway Lim and Bin Yu. Estimation stability with cross validation (ESCV). *Journal of Computational and Graphical Statistics*, 25:464–492, 2015.

[123] Zachary C Lipton, David C Kale, Charles Elkan, Randall Wetzell, Sharad Vikram, Julian McAuley, Randall C Wetzell, Zhanglong Ji, Balakrishnan Narayaswamy, and Cheng-I Wang. The mythos of model interpretability. *IEEE Spectrum*, 2016.

[124] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi–task feature learning via efficient $\ell_{2,1}$-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press, 2009.

[125] Steven J Luck. *An introduction to the event–related potential technique.* MIT press, 2014.

[126] Eric Maris. Statistical testing in electrophysiological studies. *Psychophysiology*, 49(4):549–565, 2012.

[127] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of EEG–and MEG–data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[128] R Timothy Marler and Jasbir S Arora. Survey of multi–objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395, 2004.

[129] Andre F Marquand, Michael Brammer, Steven CR Williams, and Orla M Doyle. Bayesian multi–task learning for decoding multi-subject neuroimaging data. *NeuroImage*, 92:298–311, 2014.

[130] Donald W Marquardt. An algorithm for least–squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[131] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[132] Georgios Michalareas, Julien Vezoli, Stan van Pelt, Jan-Mathijs Schoffelen, Henry Kennedy, and Pascal Fries. Alpha–beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. *Neuron*, 2016.

[133] Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fMRI–based prediction of behavior. *IEEE Transactions on Medical Imaging*, 30(7):1328–1340, 2011.

[134] Rupert G. Miller. *Simultaneous statistical inference.* Springer Science & Business Media, 2012.

[135] Tom M Mitchell, Rebecca Hutchinson, Radu S Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004.

[136] Partha P Mitra and Bijan Pesaran. Analysis of dynamic brain imaging data. *Biophysical journal*, 76(2):691–708, 1999.

[137] Gregoire Montavon, Martin Braun, Thomas Krueger, and Klaus-Robert Muller. Analyzing local structure in kernel–based learning: Explanation, complexity, and reliability assessment. *Signal Processing Magazine, IEEE*, 30(4):62–74, 2013.

[138] Niels Mørch, Lars K Hansen, Stephen C Strother, Claus Svarer, David A Rottenberg, Benny Lautrup, Robert Savoy, and Olaf B Paulson. Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In James Duncan and Gene Gindi, editors, *Information processing in medical imaging*, pages 259–270. Springer Berlin Heidelberg, 1997.

[139] John C Mosher, Paul S Lewis, and Richard M Leahy. Multiple dipole modeling and localization from spatio–temporal MEG data. *IEEE Transactions on Biomedical Engineering*, 39(6):541–557, 1992.

[140] Nadia Müller, Sabine Leske, Thomas Hartmann, Szabolcs Szebényi, and Nathan Weisz. Listen to yourself: The medial prefrontal cortex modulates auditory alpha power during speech preparation. *Cerebral Cortex*, pages 4029–4037, 2015.

[141] Risto Näätänen and Terence Picton. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425, 1987.

[142] Thomas Naselaris and Kendrick N Kay. Resolving ambiguities of MVPA using explicit models of representation. *Trends in cognitive sciences*, 19(10):551–554, 2015.

[143] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

[144] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[145] Jerzy Neyman and Egon S Pearson. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in Statistics*, pages 73–108. Springer, 1992.

[146] Thomas Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446, 2003.

[147] Raymond S Nickerson. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2):241–301, 2000.

[148] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[149] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind–reading: multi–voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[150] Silva M.V. Nunes, Fernando Maestú, and Castro A. Caldas. The role of MEG in unveiling cognition, 2011.

[151] Emanuele Olivetti, Seyed Mostafa Kia, and Paolo Avesani. MEG decoding across subjects. In *International Workshop on Pattern Recognition in Neuroimaging*, Tubingen, Germany, 2014.

[152] Emanuele Olivetti, Sandro Vega-Pons, and Paolo Avesani. The kernel two–sample test for brain networks. *arXiv preprint arXiv:1511.06120*, 2015.

[153] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2010.

[154] Alice J O'Toole, Fang Jiang, Hervé Abdi, Nils Pénard, Joseph P Dunlop, and Marc A Parent. Theoretical, statistical, and practical perspectives on pattern–based classification approaches to the analysis of functional neuroimaging data. *Journal of cognitive neuroscience*, 19(11):1735–1752, 2007.

[155] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[156] Lucas Parra, Chris Alvino, Akaysha Tang, Barak Pearlmutter, Nick Yeung, Allen Osman, and Paul Sajda. Single–trial detection in EEG and MEG: Keeping it linear. *Neurocomputing*, 52-54:177–183, June 2003.

[157] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):199–209, March 2009.

[158] Francesco Piccione, Flavio Giorgi, P Tonin, K Priftis, S Giove, S Silvoni, G Palmas, and F Beverina. P300–based brain computer interface: reliability and performance in healthy and paralysed participants. *Clinical neurophysiology*, 117(3):531–537, 2006.

[159] Maria Rita Piras, Immacolata Magnano, Edoardo Domenico Giorgio Canu, Kai Stephan Paulus, Wanda Maria Satta, A Soddu, Maurizio Conti, Antonio Achene, G Solinas, and I Aiello. Longitudinal

study of cognitive dysfunction in multiple sclerosis: neuropsychological, neuroradiological, and neurophysiological findings. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(7):878–885, 2003.

[160] Russell A. Poldrack. Region of interest analysis for fMRI. *Social cognitive and affective neuroscience*, 2(1):67–70, 2007.

[161] John Polich. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

[162] Karl Popper. *The logic of scientific discovery.* Routledge, 2005.

[163] S Prabhakar, P Syal, and T Srivastava. P300 in newly diagnosed non-dementing Parkinson's disease: effect of dopaminergic drugs. *Neurology India*, 48(3):239–242, 2000.

[164] Nikhil Rao, Christopher Cox, Rob Nowak, and Timothy T Rogers. Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In *Advances in neural information processing systems*, pages 2202–2210, 2013.

[165] Nikhil Rao, Robert Nowak, Christopher Cox, and Timothy Rogers. Classification with the sparse group lasso. *IEEE Transactions on Signal Processing*, 64(2):448–463, 2016.

[166] Peter M Rasmussen, Lars K Hansen, Kristoffer H Madsen, Nathan W Churchill, and Stephen C Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, 45(6):2085–2100, 2012.

[167] Jochem W Rieger, Christoph Reichert, Karl R Gegenfurtner, Toemme Noesselt, Christoph Braun, Hans-Jochen Heinze, Rudolf Kruse, and Hermann Hinrichs. Predicting the recognition of natural

scenes from single trial MEG recordings of brain activity. *Neuroimage*, 42(3):1056–1068, 2008.

[168] Irina Rish, Guillermo A Cecchi, Aurelie Lozano, and Alexandru Niculescu-Mizil. *Practical Applications of Sparse Modeling.* MIT Press, Cambridge, Massachusetts, 2014.

[169] Cristina Rosazza and Ludovico Minati. Resting–state brain networks: literature review and clinical applications. *Neurological Sciences*, 32(5):773–785, 2011.

[170] Jonathan Rosenblatt, Roee Gilron, and Roy Mukamel. Better–than–chance classification for signal detection. *arXiv preprint arXiv:1608.08873*, 2016.

[171] Michael D Rugg and Michael GH Coles. *Electrophysiology of mind: Event–related brain potentials and cognition.* Oxford University Press, 1995.

[172] Mert R Sabuncu. A universal and efficient method to compute maps from image–based prediction models. *Medical Image Computing and Computer–Assisted Intervention–MICCAI 2014*, 8675:353–360, 2014.

[173] Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

[174] Gordon M Shepherd. *The synaptic organization of the brain.* Oxford University Press, 2003.

[175] Nino Shervashidze and Francis Bach. Learning the structure for structured sparsity. *IEEE Transactions on Signal Processing*, 63(18):4894–4902, 2015.

[176] Marcus C Spruill. Asymptotic distribution of coordinates on high dimensional spheres. *Electronic communications in probability*, 12:234–247, 2007.

[177] John D Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of statistics*, pages 2013–2035, 2003.

[178] Stephen C Strother, Peter M Rasmussen, Nathan W Churchill, and KL Hansen. *Stability and Reproducibility in fMRI Analysis.* New York: Springer-Verlag, 2014.

[179] P. C. Sundgren, Q. Dong, D. Gómez-Hassan, S. K. Mukherji, P. Maly, and R. Welsh. Diffusion tensor imaging of the brain: review of clinical applications. *Neuroradiology*, 46(5):339–350, 2004.

[180] Sylvain Takerkart, Guillaume Auzias, Bertrand Thirion, and Liva Ralaivola. Graph–based inter–subject pattern analysis of fMRI data. *PLOS ONE*, 9(8):1–14, 08 2014.

[181] Sylvain Takerkart and Liva Ralaivola. Multiple subject learning for inter–subject prediction. In *International Workshop on Pattern Recognition in Neuroimaging 2014*, pages 1–4. IEEE, 2014.

[182] Catherine Tallon-Baudry and Olivier Bertrand. Oscillatory gamma activity in humans and its role in object representation. *Trends in cognitive sciences*, 3(4):151–162, 1999.

[183] Samu Taulu, Juha Simola, Jukka Nenonen, and Lauri Parkkonen. Novel noise reduction methods. In Selma Supek and Cheryl J. Aine, editors, *Magnetoencephalography: From Signals to Dynamic Cortical Networks*, pages 35–71. Springer, Berlin, Heidelberg, 2014.

[184] Robert Tibshirani. *Bias, variance and prediction error for classification rules.* University of Toronto, Department of Statistics, Toronto, 1996.

[185] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

[186] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[187] Christos Tzelepis, Vasileios Mezaris, and Ioannis Patras. Linear maximum margin classifier for learning from uncertain data. *arXiv preprint arXiv:1504.03892*, 2015.

[188] Peter J Uhlhaas and Wolf Singer. Neural synchrony in brain disorders: relevance for cognitive dysfunctions and pathophysiology. *Neuron*, 52(1):155–168, 2006.

[189] Giorgio Valentini and Thomas G Dietterich. Bias–variance analysis of support vector machines for the development of SVM–based ensemble methods. *The Journal of Machine Learning Research*, 5:725–775, 2004.

[190] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. 100information transfer factor explains the accuracy paradox. *PLOS ONE*, 9(1):1–10, 01 2014.

[191] Freek van Ede and Eric Maris. Physiological plausibility can increase reproducibility in cognitive neuroscience. *Trends in cognitive sciences*, 20:567–569, 2016.

[192] Marcel van Gerven, Christian Hesse, Ole Jensen, and Tom Heskes. Interpreting single trial data using groupwise regularisation. *NeuroImage*, 46(3):665–676, 2009.

[193] Steven Van Voorhis and Steven A Hillyard. Visual evoked potentials and selective attention to points in space. *Perception & Psychophysics*, 22(1):54–62, 1977.

[194] Vladimir Naumovich Vapnik and Samuel Kotz. *Estimation of dependences based on empirical data*, volume 40. Springer-verlag New York, 1982.

[195] Gael Varoquaux, Alexandre Gramfort, and Bertrand Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1375–1382, Edinburgh, Scotland, 2012.

[196] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross–validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2016.

[197] Gael Varoquaux and Bertrand Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3(1):1–7, 2014.

[198] Alfredo Vellido, JD Martin-Guerroro, and P Lisboa. Making machine learning models interpretable. In *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). Bruges, Belgium*, pages 163–172, 2012.

[199] Diego Vidaurre, Concha Bielza, and Pedro Larrañaga. A survey of L1 regression. *International Statistical Review*, 81(3):361–387, 2013.

[200] Wei Wang, Alan D Degenhart, Jennifer L Collinger, Ramana Vinjamuri, Gustavo P Sudre, P David Adelson, Deborah L Holder, Eric C Leuthardt, Daniel W Moran, Michael L Boninger, et al. Human motor cortical activity recorded with Micro–ECoG electrodes, during individual finger movements. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 586–589. IEEE, 2009.

[201] Yilun Wang, Junjie Zheng, Sheng Zhang, Xunjuan Duan, and Huafu Chen. Randomized structural sparsity via constrained block subsampling for improved sensitivity of discriminative voxel identification. *NeuroImage*, 117:170–183, 2015.

[202] Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015.

[203] Peter H. Westfall and S. Stanley Young. *Resampling–based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

[204] Wikipedia. Cerebral cortex — Wikipedia, the free encyclopedia, 2017. [Online; accessed 16-March-2017].

[205] Wikipedia. Magnetoencephalography — Wikipedia, the free encyclopedia, 2017. [Online; accessed 16-March-2017].

[206] Wikipedia. Neuron — Wikipedia, the free encyclopedia, 2017. [Online; accessed 16-March-2017].

[207] Jonathan R Wolpaw, Niels Birbaumer, William J Heetderks, Dennis J McFarland, P Hunter Peckham, Gerwin Schalk, Emanuel Donchin, Louis A Quatrano, Charles J Robinson, Theresa M Vaughan, et al. Brain–computer interface technology: a review of the first international meeting. *IEEE transactions on rehabilitation engineering*, 8(2):164–173, 2000.

[208] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[209] Jonathan R Wolpaw, Dennis J McFarland, Gregory W Neat, and Catherine A Forneris. An EEG–based brain–computer interface for cursor control. *Electroencephalography and clinical neurophysiology*, 78(3):252–259, 1991.

[210] David H Wolpert and William G Macready. An efficient method to estimate bagging's generalization error. *Machine Learning*, 35(1):41–55, 1999.

[211] Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience.*

[212] Eric P Xing, Mladen Kolar, Seyoung Kim, and Xi Chen. *High–Dimensional Sparse Structured Input–Output Models, with Applications to GWAS*, chapter 4, pages 37–64. MIT Press, Cambridge, Massachusetts, 2014.

[213] Brian S. Yandell. *Practical data analysis for designed experiments*, volume 39. Crc Press, 1997.

[214] Nick Yeung, Rafal Bogacz, Clay B Holroyd, and Jonathan D Cohen. Detection of synchronized oscillations in the electroencephalogram: an evaluation of methods. *Psychophysiology*, 41(6):822–832, 2004.

[215] Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

[216] Donghyeon Yu, Seul Ji Lee, Won Jun Lee, Sang Cheol Kim, Johan Lim, and Sung Won Kwon. Classification of spectral data using fused lasso logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 142:70–77, 2015.

[217] 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.

[218] Jinbo Bi Tong Zhang. Support vector classification with input data uncertainty. In Lawrence K. Saul, Yair Weiss, and Lon Bottou, editors, *Advances in neural information processing systems*, volume 17, pages 161–168. The MIT Press, Cambridge, Massachusetts, US, 2005.

[219] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[220] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Malsar: Multi–task learning via structural regularization. *Arizona State University*, 2011.

[221] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi–task learning formulation for predicting disease progression. In *Proceedings of international conference on Knowledge discovery and data mining*, pages 814–822. ACM, 2011.

[222] JE Zimmerman, Paul Thiene, and JT Harding. Design and operation of stable rf–biased superconducting point–contact quantum devices,

and a note on the properties of perfectly clean metal contacts. *Journal of Applied Physics*, 41(4):1572–1580, 1970.

[223] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

Appendix A

Appendices

A.1 Uncertainty in Input Space and Learning

Here we present a simple example to illustrate the effect of uncertainty in the input space on the learning process and interpretation of results. Let $\mathcal{X} \in [0,1] \times [0,1]$, $\mathcal{Y} \in \{-1,1\}$, and the distribution of input space $\rho_{\mathcal{X}}$ be a 2D-uniform distribution. If (a,b) be a random sample, we have:

$$\mathcal{Y} = sgn(\Phi^*(\mathcal{X})) = \begin{cases} 1 & if \quad a < b \\ -1 & if \quad a \geq b \end{cases}.$$

In this example two classes are linearly separable and we have $\vec{\Theta}^* \propto [-0.71, 0.71]^T$ [see Figure A.1(A)]. We add Gaussian noise with co–variance $\Sigma = \begin{bmatrix} 0.02 & -0.01 \\ -0.01 & 1 \end{bmatrix}$ to the sampled data. Figure A.1(B) shows the new distribution of samples after adding noise to the data. After noise contamination, the positive and negative classes are no longer linearly separable (Φ_S is not linear). Using the true solution for classifying the noisy data yields 0.711 ± 0.003 classification accuracy. This is while solving the least squares problem on the noisy samples provides $\hat{\vec{\Theta}} \propto [-0.97, 0.25]^T$ as the linear approximation of Φ_S with accuracy rate of 0.747 ± 0.003. Any model selection approach based on the generalization performance promotes the

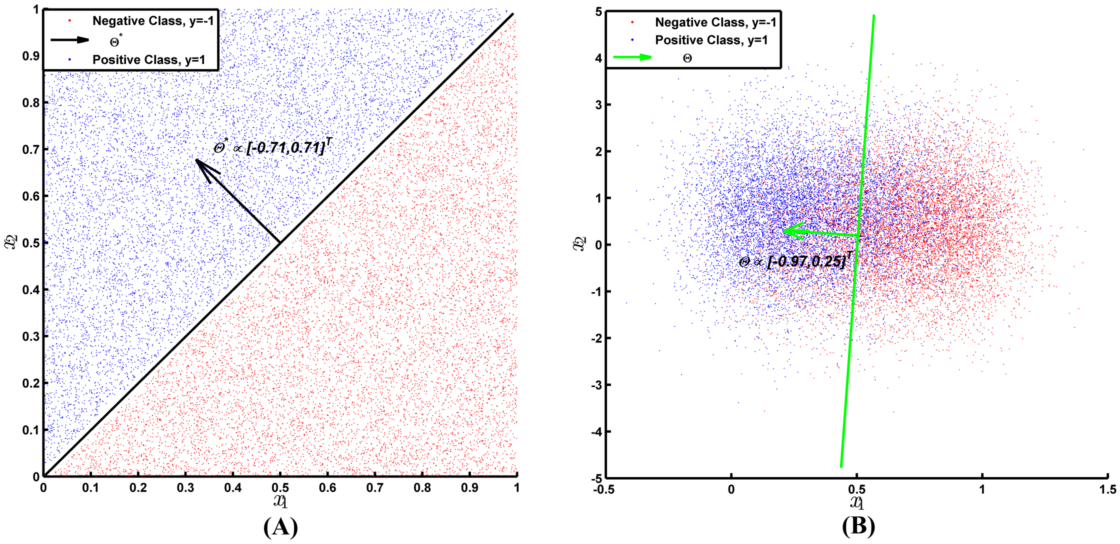


Figure A.1: **(A)** The distribution of the sampled data without noise and the true solution. **(B)** The distribution of sampled data after noise contamination and the estimated solution of least squares.

solution of least squares over the true solution. The extra 0.04 improvement of the performance of least square over the true solution can be considered as over–fitting to noise and it is the source of misinterpretation of results. Any attempt to interpret the $\vec{\hat{\Theta}}$ leads to a misleading conclusion with respect to the actual underlying function.

Table A.1: Distribution of cosine similarity between two random p-dimensional vectors.

$p =$	**5**	**10**	**50**	**100**	**500**	**1000**	**5000**	**10000**
Fitted μ	-0.00016	0.0012	0.00071	-0.00079	0.00075	-0.00017	-0.00021	-0.000006
Fitted σ	0.4492	0.3189	0.1411	0.0999	0.0450	0.0316	0.0143	0.0099
$\sqrt{\frac{1}{p}}$	0.4472	0.3162	0.1414	0.1	0.0447	0.0316	0.0141	0.010
Anderson–Darling test	1	1	0	0	0	0	0	0
Critical Value	0.8123	0.6070	0.2761	0.1946	0.0884	0.0621	0.0284	0.0193

A.2 The Distribution of Cosine Similarity: an Experimental Support

To experimentally illustrate the characteristics of the distribution of cosine similarity, 10000 random vectors for $p = 5, 10, 50, 100, 500, 1000, 5000, 10000$ are drawn from uniform distribution in $[-1, 1]$. Then histogram of similarity between each random vector with a random reference vector is computed separately for each value of p. Figure A.2 shows the histograms where the red curve in each histogram represents the normal distribution fitted to the histogram. The mean and standard deviation of the fitted normal distributions are summarized in Table A.1. We tested the normality of the distributions using Anderson–Darling test. Table A.1 shows the results of tests where 1 means the null–hypothesis is rejected (the distribution is different from normal). The comparison between the fitted standard deviation with $\sigma = \sqrt{\frac{1}{p}}$ experimentally confirms our initial expectation on the standard deviation of distribution of cosine similarity. The critical values for different p are shown in Table A.1 by calculating 95% percentile of the distribution. For a large enough p the critical value is very close to zero and therefore any value significantly larger than zero represents a meaningful similarity between two high dimensional vectors.

A.3 Experimental Comparison Between the Activation Patterns and cERF

As shown in Section 3.2.4, cERF is the equivalent generative model for the least squares solution in a binary time–domain MEG decoding scenario. The aim of this appendix is to provide an experimental support for Proposition 2. To achieve this goal, we experimentally compare cERFs and activation patterns (APs) in the experiment on the real MEG data

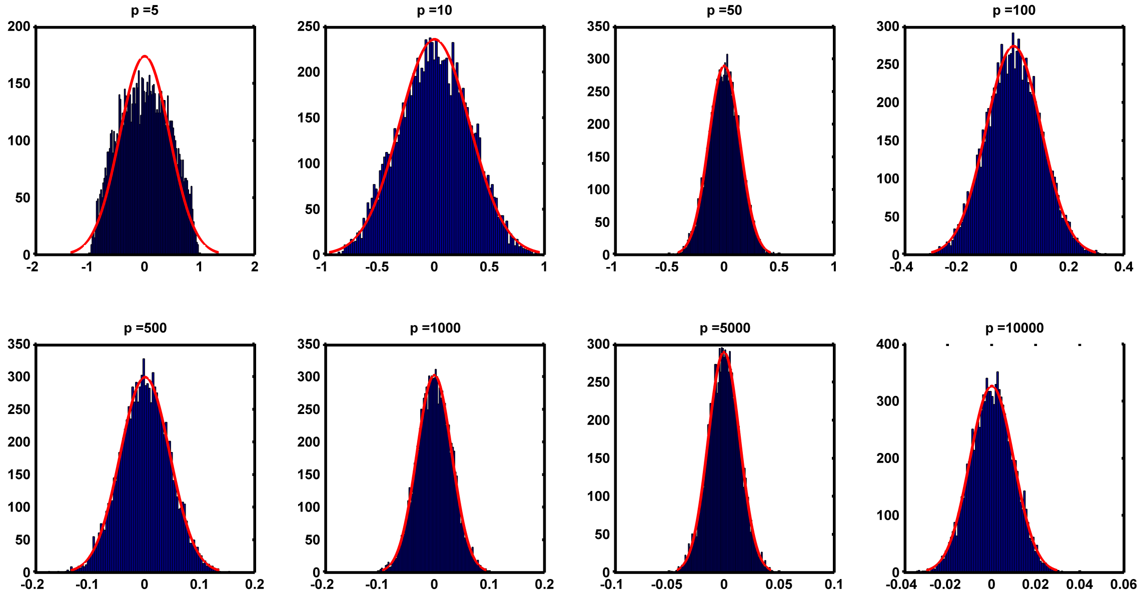


Figure A.2: Histograms of cosine similarity between 10000 random vectors with a random reference vector in p dimensional space.

(see Section 3.2.6 for the explanation of data and Section 4.2.6 for the decoding process). The APs are computed based on the weight vector of the most accurate decoding models (i.e., $\hat{\vec{\Theta}}^{\delta}$) using the proposed approach in Ref. [83]. Table A.2 summarizes the cosine similarity between cERFs and APs across 16 subjects. In addition, it compares the generalization performance of cERFs (denoted by δ_{cERF}) and APs (denoted by δ_{AP}) with that of the weights of the decoding model selected based on the proposed criterion ζ_{Φ} (denoted by δ_{ζ}).

The results experimentally confirm the validity of Proposition 2 as the cosine similarity between cERFs and APs are very close to 1 for all subjects. Furthermore, while cERFs and APs show equal prediction power (Wilcoxon rank sum test p-value$= 0.84$), they are significantly less predictive than the weights of the selected model by ζ_{Φ} criterion (Wilcoxon rank sum test p-value$= 1.5 \times 10^{-6}$).

A.4 Limitations of the Proposed Heuristic

In this appendix the goal is to experimentally investigate the limitations of the proposed heuristic based on contrast event–related fields in approximating the representativeness and interpretability of brain decoding models. Here we examine the effect of the sample size and the uncertainty in input and output spaces on the quality of this approximation.

In our experiments, following the data simulation procedure in Ref. [46], we simulated samples of the positive class as a 2-dimensional 100×100 pattern of a $5Hz$ sine wave [see Figure A.3(A)]. All samples in the positive class are corrupted with Gaussian noise with 0 mean and σ standard deviation [see Figure A.3(B)]. The value of σ is used to control the level of uncertainty in the input space. The samples in the negative class are constructed by drawing 100×100 random patterns from the Gaussian distribution with 0 mean and σ standard deviation [see Figure A.3(C)]. Similar to Θ^{cERF},

Table A.2: Cosine similarity between cERFs and APs across 16 subjects and comparison between the generalization performance of cERFs (δ_{cERF}), APs (δ_{AP}), and the weights of the decoding model selected based on the proposed criterion (δ_{ζ}).

Subjects	cERF-AP similarity	δ_{cERF}	δ_{AP}	δ_{ζ}
1	1	0.56	0.56	0.78
2	0.9998	0.54	0.54	0.80
3	0.9998	0.57	0.57	0.78
4	0.9970	0.55	0.55	0.76
5	0.9999	0.54	0.54	0.78
6	1	0.57	0.57	0.74
7	1	0.56	0.56	0.81
8	1	0.56	0.56	0.85
9	0.9999	0.57	0.57	0.77
10	0.9999	0.59	0.59	0.77
11	0.9997	0.53	0.53	0.74
12	0.9999	0.58	0.58	0.79
13	0.9973	0.59	0.58	0.77
14	1	0.62	0.61	0.81
15	1	0.63	0.62	0.89
16	1	0.65	0.65	0.81
Mean	0.9996	0.58 ± 0.03	0.57 ± 0.03	0.79 ± 0.04

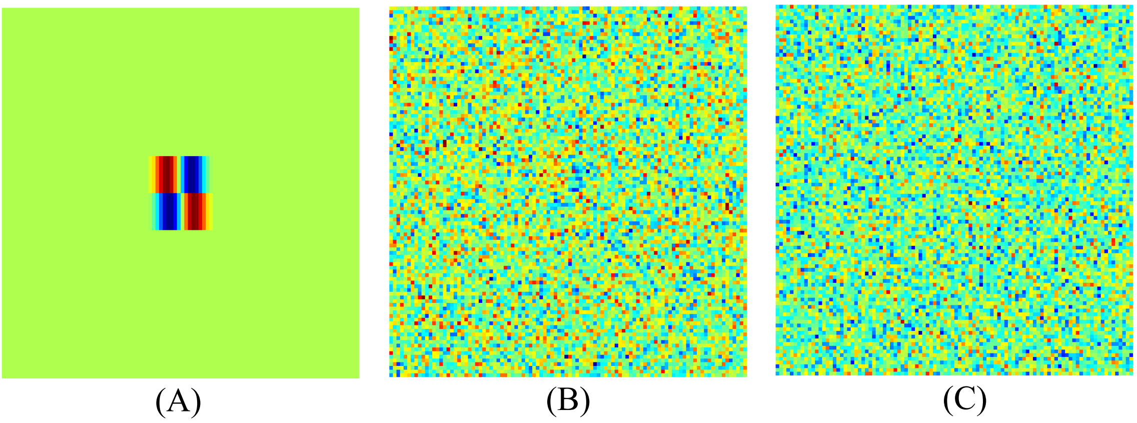


Figure A.3: **(A)** The clean positive sample. **(B)** A noisy positive sample **(C)** A negative sample.

here we use $\mu^+ - \mu^-$ as an heuristic approximation for Θ^* (where $\mu+$ and μ^- are averages of positive and negative samples, respectively).

To create the feature vectors, we rearranged the 2D-patterns into a 1D-vector (i.e., we have $p = 100 \times 100 = 10000$ features for each sample). Then the ordinary least–squares (OLS) classifier is used to classify the data into positive and negative classes. To evaluate the effect of sample size we repeated the experiment for $n = 20, 200, 500, 1000, 2000, 5000, 10000, 15000$ balanced samples of positive and negative classes. The parameter ϵ that shows the ratio of miss–labeled data is used to control the level of uncertainty in the output space.

In the first experiment the level of uncertainty in the input space is kept fixed $\sigma = 1$, and we use $\epsilon = 0, 0.01, 0.05, 0.1, 0.2, 0.3$ to control the uncertainty in the output space for different sample sizes. All the procedures (data simulation and classification) are repeated 15 times to estimate the errorbars. Figure A.4 summarizes the result. Figure A.4(A) shows the positive effect of sample size on the quality of heuristic in presence of uncertainty in the data. As the sample size increases the Δ_β, which measures the cosine similarity between Θ^* and $\mu^+ - \mu^-$, approaches to 1. The clean pattern of positive class is used as Θ^* (ground–truth) in computation of

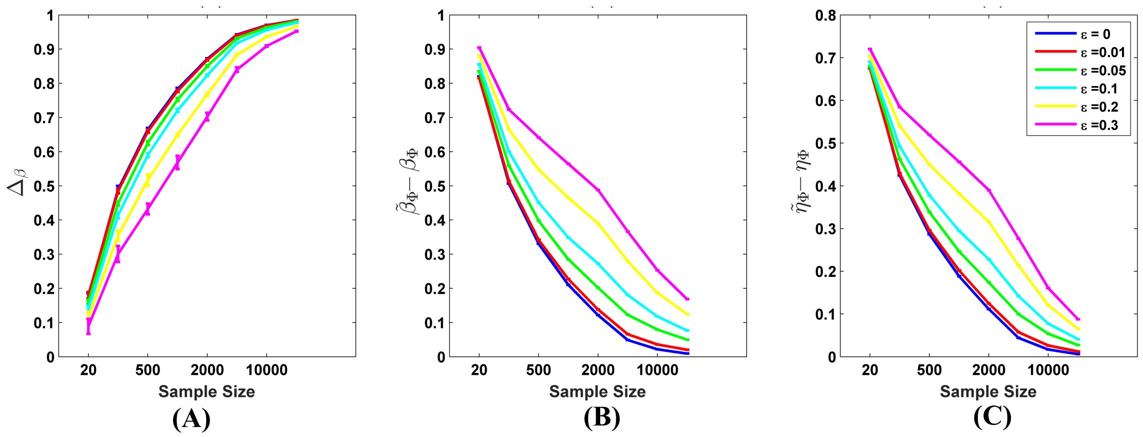


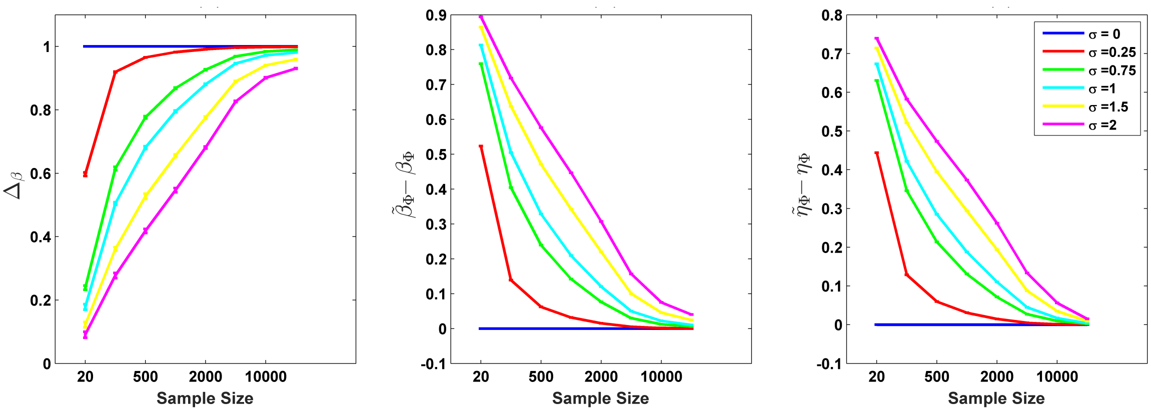
Figure A.4: **(A)** The effect of sample size and ϵ on Δ_β. Increase in sample size and decrease in ϵ improves our approximation of **(B)** representativeness and **(C)** interpretability.

Δ_β and β_Φ. Furthermore, it shows the higher uncertainty in the output space yields lower Δ_β. This fact is well reflected in Figure A.4(B) and Figure A.4(C) where the difference between actual and approximated representativeness and interpretability are plotted for different sample size and ϵ values.

To further analyze the quality of heuristic, in the second experiment we change the level of uncertainty in the input space by changing $\sigma = 0, 0.25, 0.75, 1, 1.5, 2$ and keeping fixed $\epsilon = 0$. Figure A.5 summarizes the result. Again the increase in sample size improves the quality of approximation. Our experiments highlights the effect of sample size on the quality of the proposed heuristic in the presence of uncertainty in input and output spaces. This fact limits the application of the proposed heuristic on the small sample size datasets.

A.5 Recovered Time Courses on Simulated Data

This appendix presents the complementary figures for Section 4.3.1. Figure A.6- A.9 compare the temporal maps of 5 different decoding methods with the ground–truth effect on simulated subject 4-7, respectively. The



(A) **(B)** **(C)**

Figure A.5: **(A)** The effect of sample size and σ on Δ_β. Increase in sample size and decrease in σ improves our approximation of **(B)** representativeness and **(C)** interpretability.

time courses show the temporal patterns of the recovered effect computed by averaging the weights of the classifier over the effective channels. The effective channels are selected based on the spatial distribution of the dipole in the ground–truth effect.

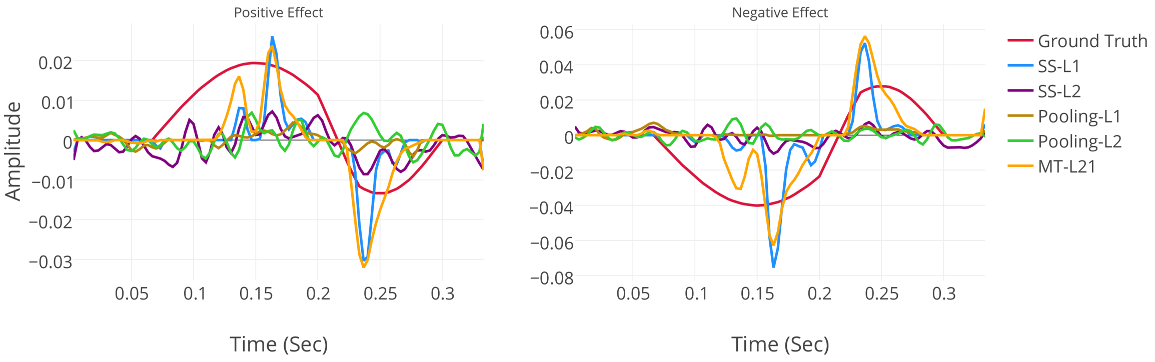


Figure A.6: Recovered time course for simulated subject 4 using 5 different methods.

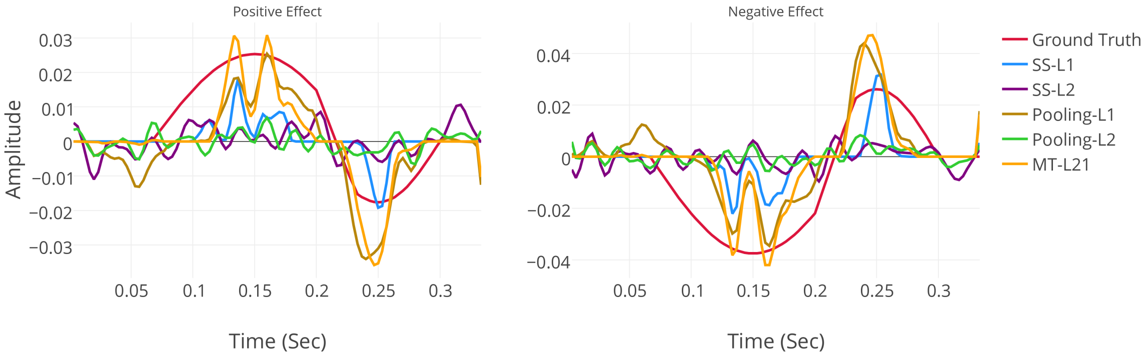


Figure A.7: Recovered time course for simulated subject 5 using 5 different methods.

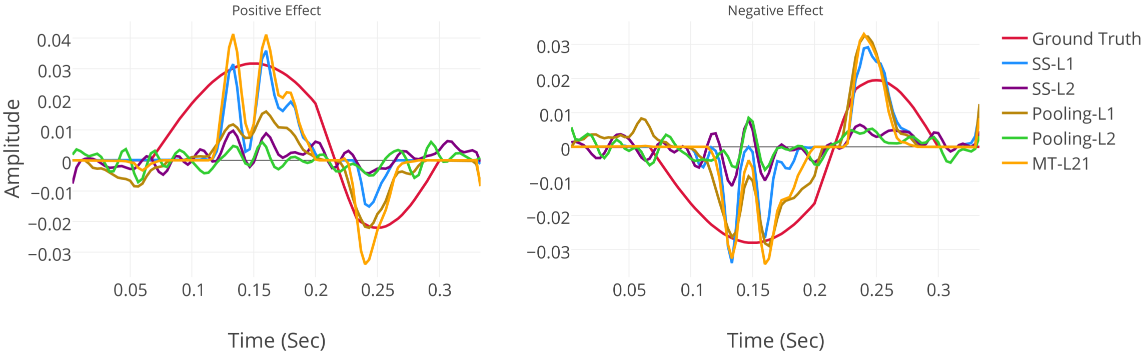


Figure A.8: Recovered time course for simulated subject 6 using 5 different methods.

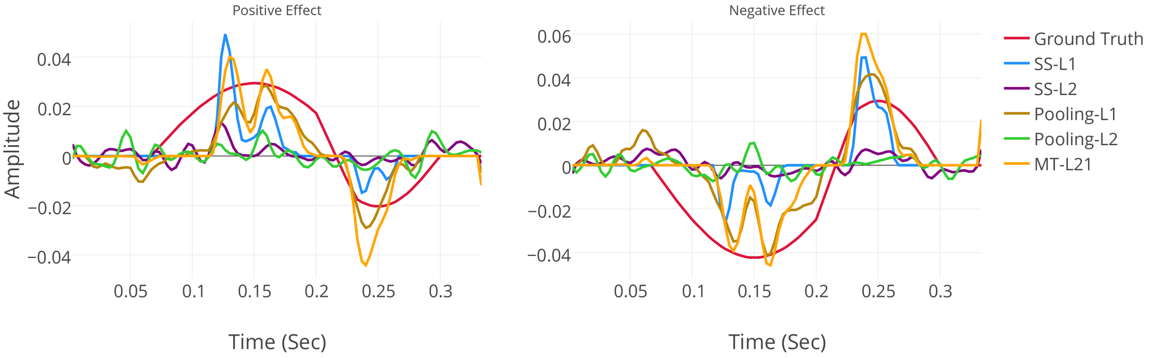


Figure A.9: Recovered time course for simulated subject 7 using 5 different methods.

A.6 Recovered Topoplots on Real Data

Here we present complementary figures for Section 4.3.2. Figure A.10-A.12 show the recovered topological maps from the real MEG dataset for 16 subjects using SS-L1, SS-L2, and pooling approaches. The topoplots show the classifier weights for magnetometer sensors averaged in 150 to 250 ms after the stimulus onset.

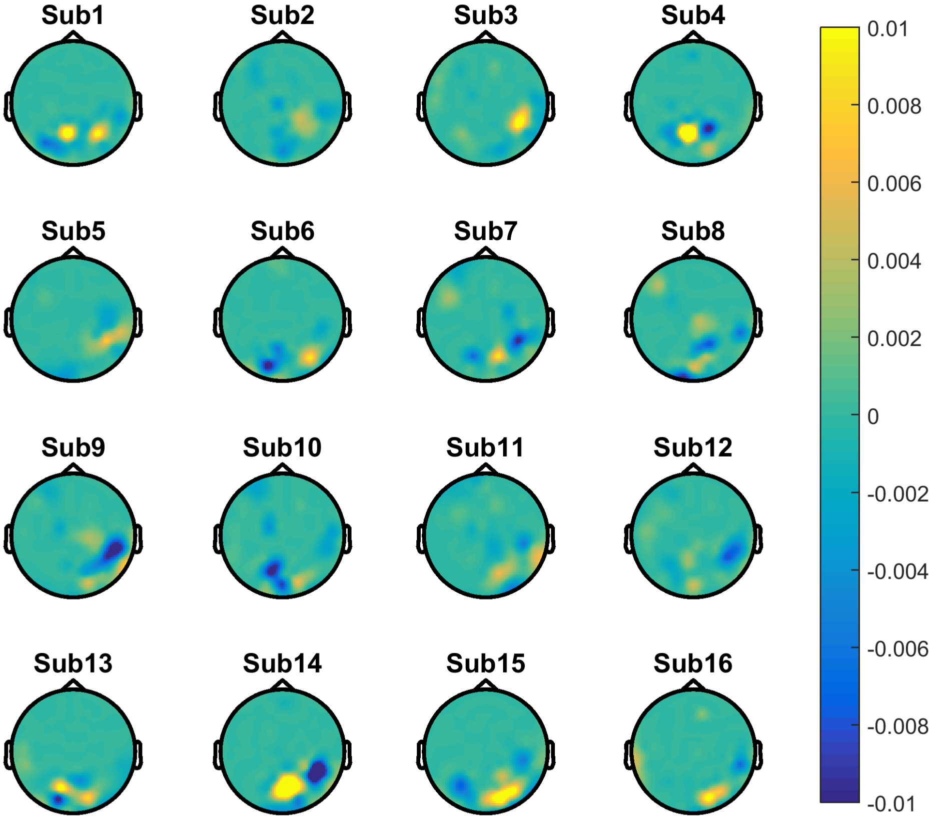


Figure A.10: Recovered topological maps using SS-L1 method from the real MEG dataset across all 16 subjects.

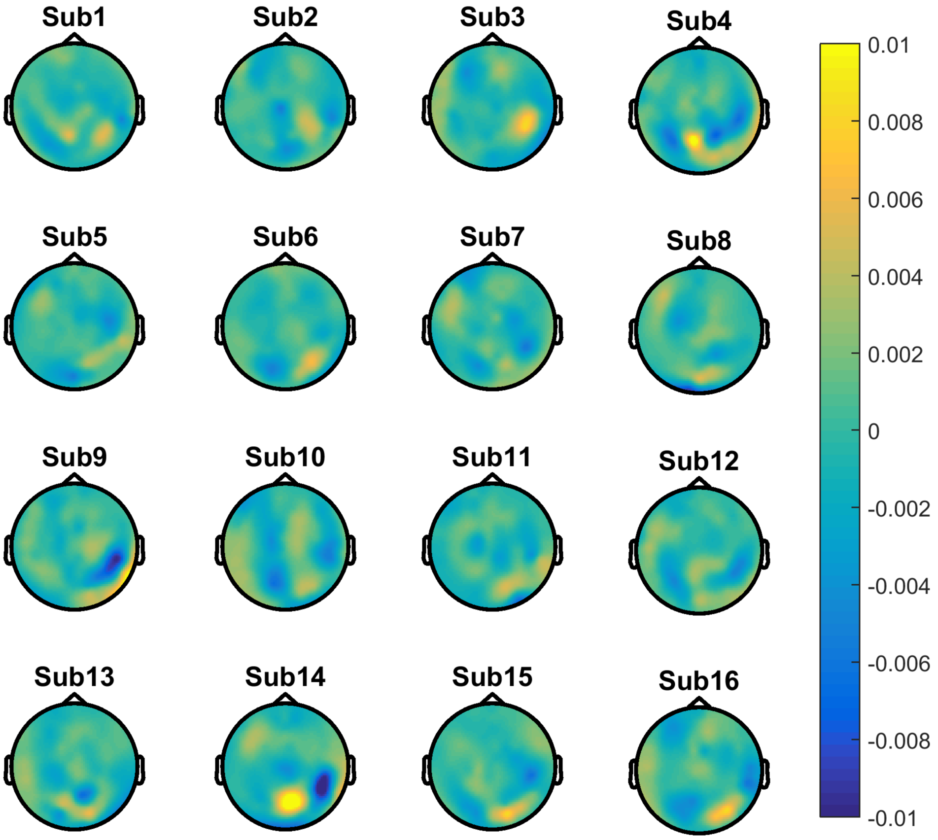


Figure A.11: Recovered topological maps using SS-L2 method from the real MEG dataset across all 16 subjects.

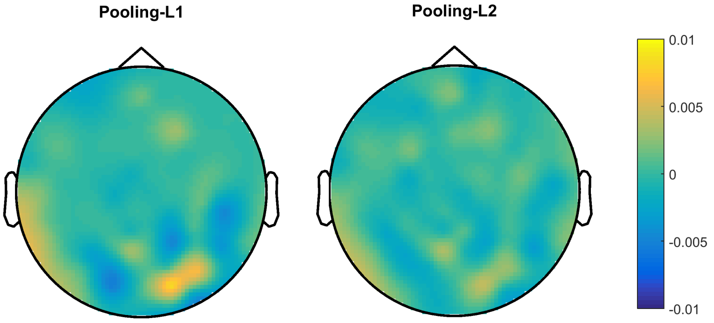


Figure A.12: Recovered topological maps using Pooling-L1 (left) and Pooling-L2 (Right) methods from the real MEG dataset.