Interpretability of Multivariate Brain Maps in Brain Decoding: Definition and Quantification

Seyed Mostafa Kia[1,2,3,*]

Center for Mind/Brain Sciences (CIMeC), University of Trento, via delle Regole 101, 38123, Mattarello, TN, Italy

Abstract

Brain decoding is a popular multivariate approach for hypothesis testing in neuroimaging. Linear classifiers are widely employed in the brain decoding paradigm to discriminate among experimental conditions. Then, the derived linear weights are visualized in the form of multivariate brain maps to further study the spatio-temporal patterns of underlying neural activities. 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 ratios, and the high dimensionality of 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 paper, first, we present a theoretical definition of interpretability in brain decoding; we show that the interpretability of multivariate brain maps can be decomposed into their reproducibility and representativeness. Second, as an application of the proposed definition, we formalize a heuristic method for approximating the interpretability of multivariate brain maps in a binary magnetoencephalography (MEG) decoding scenario. Third, we pro-

[*]Corresponding author:
Email address: `seyedmostafa.kia@unitn.it` (Seyed Mostafa Kia)
[1]University of Trento, Trento, Italy
[2]Fondazione Bruno Kessler (FBK), Trento, Italy
[3]Centro Interdipartimentale Mente e Cervello (CIMeC), Trento, Italy

pose to combine the approximated interpretability and the performance of the brain decoding into a new multi-objective criterion for model selection. Our results for the 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.

Keywords: MVPA, brain decoding, brain mapping, interpretation, model selection

1. Introduction

Understanding the mechanisms of the brain has been a crucial topic throughout the history of science. Ancient Greek philosophers envisaged different functionalities for the brain ranging from cooling the body to acting as the seat of the rational soul and the center of sensation [1]. Modern cognitive science, emerging in the 20th century, provides better insight into the brain's functionality. 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 to test scientific hypotheses on a large set of independent variables [2, 3]. 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 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. Although some techniques such as the non-parametric cluster-based permutation test [4, 5] offer more sensitivity because of the cluster assumption, they still experience low sensitivity to brain activities that are narrowly distributed in time and space [2, 6]. The multivariate counterparts of mass-univariate analysis, known generally as multivariate pattern analysis (MVPA), have the potential to overcome these deficits. Multivariate approaches are capable of identifying complex spatio-temporal interactions between different brain areas with higher sensitivity and specificity than univariate analysis [7], especially in group analysis of neuroimaging data [8].

Brain decoding [9] is an MVPA technique that delivers a model to predict the mental state of a human subject based on the recorded brain signal. There are two potential applications for brain decoding: 1) brain-computer interfaces (BCIs) [10, 11], and 2) multivariate hypothesis testing [12]. 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.

In brain decoding, generally, linear classifiers are used to assess the relation between independent variables, i.e., features, and dependent variables, i.e., cognitive tasks [13, 14, 15]. This assessment is performed by solving a linear 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) [16, 17, 18, 19] and magnetoencephalogram/electroencephalogram (MEEG) studies [20, 21, 22, 23, 24, 25, 26]. It has been shown that brain decoding can be used in combination with brain encoding [27] to infer the causal relationship between stimuli and responses [28].

Brain mapping [29] is a higher form of neuroimaging that assigns precomputed quantities, e.g., univariate statistics or weights of a linear classifier, to the spatio-temporal representation of neuroimaging data. In MVPA, 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 neuroscientific perspective, a brain map is considered *interpretable* if it enables the scientist to answer *where*, *when*, and *how* questions.

Typically, a trained classifier is a black box that predicts the label of an unseen data point with some accuracy. Valverde-Albacete and Peláez-Moreno [30] 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. It is also shown by Ramdas et al. [31] that in the case of data with small sample size using the classification accuracy as a test statistic for two sample testing should be performed with extra cautious. Beside these limitations of classification accuracy in inference, and considering the fact that the best predictive model might not be the most informative one [32]; a classifier, taken alone, only answers the question of *what* is the most likely label of a given unseen sample [33]. This fact is generally known as knowledge extraction gap [34] in the classification context. Thus far, many efforts have been devoted to filling the knowledge extraction gap of linear and non-linear data modeling methods in different areas such as computer vision [35], signal processing [36], chemometrics [37], bioinformatics [38], and neuroinformatics [39].

Despite the theoretical advantages of MVPA, its practical application to inferences regarding neuroimaging data is limited primarily by a lack of interpretability [40, 41, 42]. Therefore, improving the interpretability of linear brain decoding and associated brain maps is a primary goal in the brain imaging literature [43]. The lack of interpretability of multivariate brain maps is a direct consequence of low signal-to-noise ratios (SNRs), high dimensionality of whole-scalp recordings, high correlations among different dimensions of data, and cross-subject variability [15, 44, 45, 14, 46, 47, 48, 49, 50, 51, 52, 41]. 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 penalty terms for regularization to enhance stability selection.

The first approach to improving the interpretability of brain decoding concentrates on the model selection procedure. Model selection is a procedure in which the best values for the hyper-parameters of a model are determined [14]. The selection process is generally performed by considering the generalization performance, i.e., the accuracy, of a model as the decisive criterion. Rasmussen et al. [53] 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, they incorporated the spatial reproducibility of brain maps in the model selection procedure. An analogous approach, using a different definition of spatial reproducibility, is proposed by Conroy et al. [54]. Beside spatial reproducibility, the stability of the classifiers [55] is another criterion that is used in combination with generalization performance to enhance the interpretability. For example, [56, 57] showed that incorporating the stability of models into cross-validation improves the interpretability of the estimated parameters (by linear models).

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) [58, 50] and 2) combining the structural and functional prior knowledge with the decoding process so as to enhance stability selection. Group Lasso [59] and total-variation penalty [60] are two effective methods using this technique [61, 62]. Sparse penalized discriminant analysis [63], group-wise regularization [7], randomized Lasso [47], smoothed-sparse logistic regression [64], total-variation L1 penalization [65, 66], the graph-constrained elastic-net [67, 68], and randomized structural sparsity [69] are examples of brain decoding methods in which regularization techniques are employed to improve stability selection, and thus, the interpretability of brain decoding.

Recently, taking a new approach to the problem, Haufe et al. questioned the interpretability of weights of linear classifiers because of the contribution of noise in the decoding process [70, 39, 71]. 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 major challenge of estimating the empirical covariance matrix of the small sample size neuroimaging data [72] limits the practical application of this method.

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, [47, 73]). 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, [46, 53, 54, 74]). 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 [53, 74], dice score [46], or parameter variability [39, 54] are used for quantifying the stability of brain maps, each of which considers different aspects of local or global reproducibility.

With the aim of filling this 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. Furthermore, we show that the interpretability can be decomposed into the reproducibility and the representativeness of brain maps. 2) As a proof of the concept, we propose a practical heuristic based on event-related fields for quantifying the interpretability of brain maps in MEG decoding scenarios. 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 show that incorporating the interpretability into the model selection procedure provides more reproducible, more neurophysiologically plausible, and (as a result) more interpretable maps.

2. Methods

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 (iid) 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, and $\mathbf{Y}$ represents the experimental conditions or dependent variables. The goal of brain decoding is to find the function $\Phi_S : \mathbf{X} \rightarrow \mathbf{Y}$ as an estimation of the ideal function $\Phi^* : \mathcal{X} \rightarrow \mathcal{Y}$.

As is a common assumption in the neuroimaging context, we assume 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:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(\Phi(\mathbf{X}), \Phi_S(\mathbf{X}) + \lambda \Omega(\Theta) \tag{1}$$

where $\mathcal{L} : \mathbf{Z} \times \mathbf{Z} \rightarrow \mathbb{R}^+$ is the loss function, $\Omega : \mathbb{R}^p \rightarrow \mathbb{R}^+$ 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, [75, 76, 60, 77]). 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 $\hat{\Theta}$ 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 $\vec{\Theta} = \frac{\Theta}{\|\Theta\|}$ ($\|.\|$ represents the 2-norm of a vector).

As shown in Eq. 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^*) [78]. Therefore, all of the implicit assumptions (such as linearity) regarding Φ^* might not hold on Φ_S, and vice versa (see the supplementary material for a simple illustrative example). The *irreducible error* ε is the direct consequence of sampling; it sets a lower bound on the error, where we have:

$$\Phi_S(\mathbf{X}) = \Phi^*(\mathbf{X}) + \varepsilon \tag{2}$$

The distribution of ε dictates the type of loss function $\mathcal{L}$ in Eq. 1. For example, assuming a Gaussian distribution with mean 0 and variance σ^2 for ε implies the least squares loss function [79].

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. Our definition of interpretability is based on two main assumptions: 1) the brain decoding problem is linearly separable; 2) its *unique* and neurophysiologically *plausible*[1] solution, i.e., Φ^*, is available.

Consider a linearly separable brain decoding problem in an ideal scenario where $\varepsilon = 0$ and $rank(\mathbf{X}) = p$. In this case, Φ^* is linear and its parameters Θ^* are unique and plausible. The unique parameter vector Θ^* can be computed as follows:

$$\Theta^* = \Sigma_{\mathbf{X}}^{-1}\mathbf{X}^T\mathbf{Y} \tag{3}$$

Using Θ^* as the reference, we define the *strong-interpretability* of an MBM as follows:

Definition 1. An MBM $\vec{\Theta}$ associated with a linear function Φ is "strongly-interpretable" if and only if $\vec{\Theta} \propto \Theta^*$.

It can be shown that, in practice, the estimated solution of a linear brain problem (using Eq. 1) is not strongly-interpretable because of the inherent limitations of neuroimaging data, such as uncertainty [80] 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 $\hat{\Theta} \propto [0.1, 1.2]^T$ can be considered more interpretable than a linear classifier where $\hat{\Theta} \propto [2, 1]^T$. This issue raises the following question:

Problem 1. Let $S^1, \ldots, S^m$ be m perturbed training sets drawn from S via a certain perturbation scheme such as jackknife, bootstrapping [81], or cross-validation [82]. Assume $\hat{\vec{\Theta}}^1, \ldots, \hat{\vec{\Theta}}^m$ are m MBMs of a certain Φ (estimated using Eq. 1 for certain $\mathcal{L}$, Ω, and λ) on the corresponding perturbed training sets. How can we quantify the proximity of Φ to the strongly-intrepretable solution of brain decoding problem Φ^*?

[1]Here, neurophysiological plausibility refers to the spatio-temporal chemo-physical constraints of the underlying neural activity that is highly dependent on the acquisition device.

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 α^j $(j = 1, \ldots, m)$ be the angle between $\vec{\hat{\Theta}}^j$ and $\vec{\Theta}^*$. The "interpretability" $(0 \leq \eta_\Phi \leq 1)$ of the MBM derived from a linear function Φ is defined as follows:

$$\forall j \in \{1, \ldots, m\}, \eta_\Phi = \mathbb{E}_S[\cos(\alpha^j)] \tag{4}$$

Empirically, the interpretability is the mean of cosine similarities between Θ^* and MBMs derived from different samplings of the training set. 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. 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., $\mathcal{N}(0, \sqrt{\frac{1}{p}})$. Due to the small variance for a large enough p values, any similarity value that is significantly larger than zero represents a meaningful similarity between two high dimensional vectors (see the supplementary material for more details about the distribution of cosine similarity).

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

2.3. Interpretability Decomposition into Reproducibility and Representativeness

An alternative approach toward quantifying the interpretability is to assess separately 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 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 different estimated parameters on different portions of data. Here, we are interested in assessing this variability as a measure for uniqueness. Let θ_i^j be the ith ($i = 1, \ldots, p$) element of an MBM estimated on the jth ($j = 1, \ldots, m$) perturbed training set. We first define the *main multivariate brain map* as follows:

Definition 3. The "main multivariate brain map" $\vec{\Theta}^\mu \in \mathbb{R}^p$ of a linear function Φ is defined as the sum of estimated MBMs $\hat{\vec{\Theta}}^j$ ($j = 1, \ldots, m$) on the perturbed training sets S^j in the unit hyper-sphere:

$$\vec{\Theta}^\mu = \frac{\left[\sum_{j=1}^m \theta_1^j \quad \sum_{j=1}^m \theta_2^j \quad \ldots \quad \sum_{j=1}^m \theta_p^j\right]^T}{\left\|\left[\sum_{j=1}^m \theta_1^j \quad \sum_{j=1}^m \theta_2^j \quad \ldots \quad \sum_{j=1}^m \theta_p^j\right]^T\right\|} \tag{5}$$

The definition of $\vec{\Theta}^\mu$ is analogous to the main prediction of a learning algorithm [83]; it provides a reference for quantifying the reproducibility of an MBM:

Definition 4. Let $\vec{\Theta}^\mu$ be the main multivariate brain map of Φ. Then, let α^j be the angle between $\hat{\vec{\Theta}}^j$ and $\vec{\Theta}^\mu$. The "reproducibility" ψ_Φ ($0 \leq \psi_\Phi \leq 1$) of an MBM derived from a linear function Φ is defined as follows:

$$\forall j \in \{1, \ldots, m\}, \psi_\Phi = \mathbb{E}_S[\cos(\alpha^j)] \tag{6}$$

In fact, reproducibility provides a measure for quantifying the dispersion of MBMs, computed over different perturbed training sets, from the main multivariate brain map.

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

Definition 5. Let $\vec{\Theta}^\mu$ be the main multivariate brain map of Φ. The "representativeness" ($0 \leq \beta_\Phi \leq 1$) is defined as the cosine similarity between $\vec{\Theta}^\mu$ and $\vec{\Theta}^*$:

$$\beta_\Phi = \frac{|\vec{\Theta}^\mu . \vec{\Theta}^*|}{\left\|\vec{\Theta}^\mu\right\| \left\|\vec{\Theta}^*\right\|} \tag{7}$$

The following proposition shows the relationship between the presented definitions for reproducibility, representativeness, and the interpretability:

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

See Appendix D for a proof. Proposition 1 indicates the interpretability can be decomposed into the representativeness and the reproducibility of a decoding model.

2.4. A Heuristic for Practical Quantification of Interpretability in Time-Domain MEG decoding

In practice, it is impossible to evaluate the interpretability, as Φ^* is unknown. In this study, to provide a practical proof of the mentioned theoretical concepts, we propose the use of contrast event-related fields (cERFs) of MEG data as neurophysiological plausible heuristics for Θ^* in a binary MEG decoding scenario in the time domain.

The EEG/MEG data are a mixture of several simultaneous stimulus-related and stimulus-unrelated brain activities. In general, unrelated-stimulus brain activities are considered as 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. It is expected that the average will converge to the true value of the signal with a variance of $\frac{\sigma^2}{n}$. The result of the averaging process is generally known as ERF in the MEG context; separate interpretation of different ERF components can be performed [84][1].

Assume $\mathbf{X}^+ = \{x_i \in \mathbf{X} \mid y_i = 1\} \in \mathbb{R}^{n^+ \times p}$ and $\mathbf{X}^- = \{x_i \in \mathbf{X} \mid y_i = -1\} \in \mathbb{R}^{n^- \times p}$. Then, the cERF brain map $\vec{\Theta}^{cERF}$ is computed as follows:

$$\vec{\Theta}^{cERF} = \frac{\frac{1}{n^+}\sum_{x_i \in X^+} x_i - \frac{1}{n^-}\sum_{x_i \in X^-} x_i}{\left\|\frac{1}{n^+}\sum_{x_i \in X^+} x_i - \frac{1}{n^-}\sum_{x_i \in X^-} x_i\right\|} \tag{8}$$

Using the core theory presented in [39], it can be shown that cERF is the equivalent generative model for the least squares solution in a binary

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

time-domain MEG decoding scenario (see Appendix A). Using $\vec{\Theta}^{cERF}$ as a heuristic for $\vec{\Theta}^*$, the representativeness can be approximated as follows:

$$\tilde{\beta}_\Phi = \frac{|\vec{\Theta}^\mu.\vec{\Theta}^{cERF}|}{\left\|\vec{\Theta}^\mu\right\| \left\|\vec{\Theta}^{cERF}\right\|} \tag{9}$$

Where $\tilde{\beta}_\Phi$ is an approximation of β_Φ and we have:

$$\beta_\Phi = \Delta_\beta \tilde{\beta}_\Phi \pm \sqrt{(1-\tilde{\beta}_\Phi^2)(1-\Delta_\beta^2)} \tag{10}$$

Δ_β represents the cosine similarity between $\vec{\Theta}^*$ and $\vec{\Theta}^{cERF}$ (see Figures B.8 and Appendix B). If $\Delta_\beta \to 1$ then $\tilde{\beta}_\Phi \to \beta_\Phi$.

In a similar manner, $\vec{\Theta}^{cERF}$ can be used to heuristically approximate the interpretability as follows:

$$\tilde{\eta}_\Phi = \forall j \in \{1, \dots, m\}, \tilde{\eta}_\Phi = \mathbb{E}_S(\cos(\gamma^j)) \tag{11}$$

where $\gamma_1, \dots, \gamma_m$ are the angles between $\hat{\vec{\Theta}}^1, \dots, \hat{\vec{\Theta}}^m$ and $\vec{\Theta}^{cERF}$. The following equality represents the relation between η and $\tilde{\eta}$ (see Figures C.9 and Appendix C).

$$\eta_\Phi = \Delta_\beta \tilde{\eta}_\Phi \pm \frac{\sqrt{1-\Delta_\beta^2}}{m}(\sin\gamma_1 + \cdots + \sin\gamma_m) \tag{12}$$

Again, if $\Delta_\beta \to 1$ then $\tilde{\eta}_\Phi \to \eta_\Phi$. Notice that Δ_β is independent of the decoding approach used; it only depends on the quality of the heuristic. It can be shown that $\tilde{\eta}_\Phi = \tilde{\beta}_\Phi \times \psi_\Phi$.

Eq. 12 shows that the choice of 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 because of the destructive contribution of Δ_β. Therefore, the choice of heuristic should be carefully justified based on accepted and well-defined facts regarding the nature of the collected data (see the supplementary material for the experimental investigation of the limitations of the proposed heuristic).

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* [85]. This procedure generally involves numerical optimization of the model selection criterion. 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 the predictive power as the only decisive criterion in model selection is problematic in terms of interpretability [86, 53, 54]. 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 function Φ, respectively. We propose the use of the *scalarization* technique [87] for combining $\tilde{\eta}_\Phi$ and δ_Φ into one scalar $0 \leq \zeta(\Phi) \leq 1$ as follows:

$$\zeta_\Phi = \begin{cases} \frac{\omega_1 \tilde{\eta}_\Phi + \omega_2 \delta_\Phi}{\omega_1 + \omega_2} & \delta_\Phi \geq \kappa \\ 0 & \delta_\Phi < \kappa \end{cases} \tag{13}$$

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.

It can be shown that the hyper-parameters of a model Φ are optimized based on ζ_Φ are Pareto optimal [88]. In other words, there exist no other Φ' for which we obtain both $\tilde{\eta}_{\Phi'} > \tilde{\eta}_\Phi$ and $\delta_{\Phi'} > \delta_\Phi$. We expect that optimizing the hyper-parameters based on ζ_Φ, rather only δ_Φ, yields more informative MBMs.

2.6. Experimental Materials

2.6.1. Toy Dataset

To illustrate the importance of integrating the interpretability of brain decoding with the model selection procedure, we use simple 2-dimensional toy data presented in [39]. Assume that the true underlying generative function Φ^* is defined by

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

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 is contaminated by Gaussian noise with co-variance $\Sigma = \begin{bmatrix} 1.02 & -0.3 \\ -0.3 & 0.15 \end{bmatrix}$. Figure 1 shows the distribution of the noisy data.

2.6.2. MEG Data

We use the MEG dataset presented in [89][1]. The dataset was also used for the DecMeg2014 competition[2]. In this dataset, visual stimuli consisting of famous faces, unfamiliar faces, and scrambled faces are presented to 16 subjects and fMRI, EEG, and MEG signals are 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 vs. scramble stimuli has been well studied. The event-related potential analysis of EEG/MEG shows that $N170$ occurs $130 - 200ms$ after stimulus presentation and reflects the neural processing of faces [90, 89]. 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) [2, 6]. These facts motivate us to employ multivariate approaches that are more sensitive to these effects.

As in [51], 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 $1Hz$, down-sampled to $250Hz$, and trimmed from $200ms$ before the stimulus onset to $800ms$ after the stimulus. Thus, each trial has 250 time-points for each of the 306 MEG sensors (102 magnetometers and

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

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

204 planar gradiometers)[1]. 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.

2.7. Classification and Evaluation

In all experiments, a least squares classifier with L1-penalization, i.e., Lasso [75], is 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 helps us to better illustrate the importance of including the interpretability in the model selection. Lasso solves the following optimization problem:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \|\Phi(\mathbf{X}) - \Phi_S(\mathbf{X})\|_2^2 + \lambda \|\Theta\|_1 \tag{14}$$

where 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. Here, we try to find the best regularization parameter value among $\lambda = \{0.001, 0.01, 0.1, 1, 10, 50, 100, 250, 500, 1000, 5000, 10000, 15000, 25000, 50000\}$.

We use the out-of-bag (OOB) [91, 92] 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 [81] are used to create perturbed training sets (we set $m = 50$) [2]. 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 Appendix E. Employing OOB provides the possibility of computing the bias and variance of the model as contributing factors in EPE.

To investigate the behavior of the proposed model selection criterion, we benchmark it against the commonly used performance criterion in the 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

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

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

each subject to estimate the parameter of the linear function $\hat{\Phi}_i$, where $\mathbf{Y}_i = \mathbf{X}_i\hat{\Theta}_i$. Let $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ represent the optimized solution based on δ_Φ and ζ_Φ, respectively. We denote the MBM associated with $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ by $\hat{\vec{\Theta}}_i^\delta$ and $\hat{\vec{\Theta}}_i^\zeta$, respectively. Therefore, for each subject, we compare the resulting decoders and MBMs computed based on these two model selection criteria.

3. Results

3.1. Performance-Interpretability Dilemma: A Toy Example

In the definition of Φ^* on the toy dataset discussed in Section 2.6.1, x_1 is the decisive variable and x_2 has no effect on the classification of the data into target classes. Therefore, excluding the effect of noise and based on the theory of the maximal margin classifier [93, 94], $\vec{\Theta}^* \propto [1, 0]^T$ is the true solution to the decoding problem. By accounting for the effect of noise and solving the decoding problem in $(\mathbf{X}, \mathbf{Y})$ space, we have $\vec{\Theta} \propto [\frac{1}{\sqrt{(5)}}, \frac{2}{\sqrt{(5)}}]^T$ as the parameter of the linear classifier. Although the estimated parameters on the noisy data yield 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 [39]. 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 case, because $\vec{\Theta}^*$ is known, the exact value of interpretability can be computed using Eq. 4. Table 1 compares the resultant performance and interpretability from Lasso. Lasso achieves its highest performance ($\delta_\Phi = 0.9884$) at $\lambda = 10$ with $\hat{\vec{\Theta}} \propto [0.4636, 0.8660]^T$ (indicated by the magenta line in Figure 1). 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 is incoherent with its performance. Thus, optimizing the parameter of

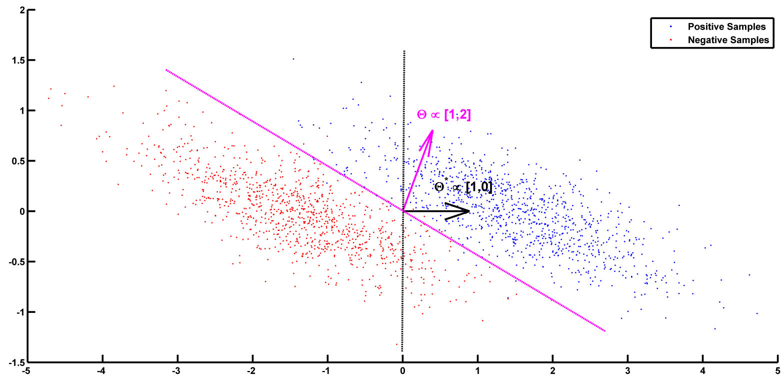


Figure 1: Noisy samples of toy data. The black line shows the true separator based on the generative model (Φ^*). The magenta line shows the most accurate classification solution. Because of the contribution of noise, any interpretation of the parameters of the most accurate classifier yields a misleading conclusion with respect to the true underlying phenomenon [39].

Table 1: Comparison between δ_Φ, η_Φ, and ζ_Φ for different λ values on the toy 2D 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**
$\vec{\hat{\Theta}} \propto$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4521\\0.8919\end{bmatrix}$	$\begin{bmatrix}0.4532\\0.8914\end{bmatrix}$	$\begin{bmatrix}0.4636\\0.8660\end{bmatrix}$	$\begin{bmatrix}0.4883\\0.8727\end{bmatrix}$	$\begin{bmatrix}0.5800\\0.8146\end{bmatrix}$	$\begin{bmatrix}0.99\\0.02\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$

the model based on its performance does not necessarily improve its interpretability. This observation confirms the previous finding by Rasmussen et al. [53] 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 this case) leads to more interpretability for the decoding models.

3.2. Mass-Univariate Hypothesis Testing on MEG Data

Results show that non-parametric mass-univariate analysis is unable to detect narrowly distributed effects in space and time (e.g., an $N170$ component) [2, 6]. To illustrate the advantage of the proposed decoding framework for spotting these effects, we performed a non-parametric cluster-based permutation test [5] on our MEG dataset using Fieldtrip toolbox [95]. 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 hypothesis testing.

3.3. Single-Subject Decoding on MEG Data

In this experiment, we aim to compare the multivariate brain maps of brain decoding models when δ_Φ and ζ_Φ are used as the criteria for model selection. Figure 2(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 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 2(b)].

One possible reason behind the performance-interpretability dilemma is illustrated in Figure 3. The figure shows the mean and standard deviation of bias, variance, and EPE of Lasso across 16 subjects. The plot 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 is not reflected in EPE but Figure 2(b) shows a significant effect on the reproducibility.

Table 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 significantly higher reproducibility, representativeness, and (as a result)

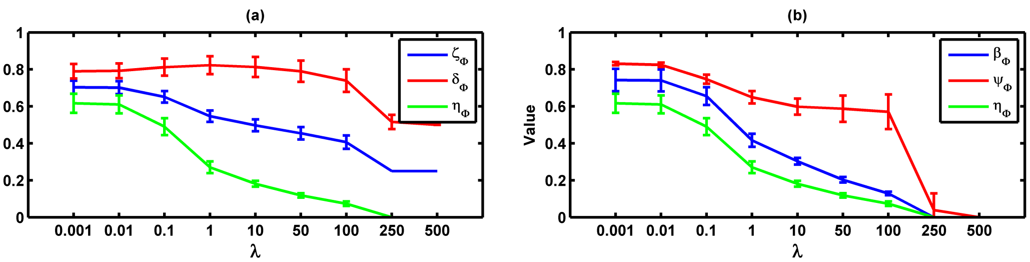


Figure 2: (a) Mean and standard-deviation of the performance, interpretability, and plausibility of Lasso over 16 subjects. The performance and interpretability become incoherent as λ increases. (b) Mean and standard-deviation of the reproducibility, representativeness, and interpretability of Lasso over 16 subjects. The interpretability declines because of the decrease in both reproducibility and representativeness.

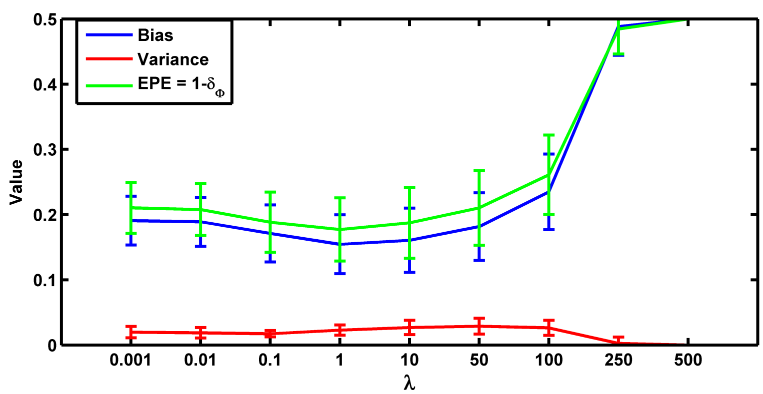


Figure 3: Mean and standard-deviation of the bias, variance, and EPE of Lasso over 16 subjects. The effect of variance on the EPE is overwhelmed by bias.

Table 2: The performance, reproducibility, representativeness, and interpretability of $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$ over 16 subjects.

Subj	Criterion: δ_Φ					Criterion: ζ_Φ				
	δ_Φ	ζ_Φ	$\tilde{\eta}_\Phi$	β_Φ	ψ_Φ	δ_Φ	ζ_Φ	$\tilde{\eta}_\Phi$	β_Φ	ψ_Φ
1	0.81	0.53	0.26	0.42	0.62	0.78	0.70	0.63	0.76	0.83
2	0.80	0.70	0.60	0.72	0.83	0.80	0.70	0.60	0.72	0.83
3	0.81	0.63	0.45	0.64	0.71	0.78	0.71	0.64	0.78	0.83
4	0.84	0.52	0.20	0.31	0.66	0.76	0.70	0.64	0.77	0.83
5	0.80	0.54	0.29	0.44	0.65	0.78	0.69	0.61	0.73	0.83
6	0.79	0.52	0.24	0.39	0.63	0.74	0.67	0.61	0.74	0.82
7	0.84	0.55	0.27	0.40	0.66	0.81	0.70	0.59	0.71	0.84
8	0.87	0.55	0.24	0.35	0.68	0.85	0.68	0.52	0.61	0.84
9	0.80	0.55	0.31	0.46	0.67	0.77	0.67	0.57	0.69	0.82
10	0.79	0.53	0.26	0.41	0.64	0.77	0.68	0.58	0.70	0.83
11	0.74	0.65	0.56	0.68	0.82	0.74	0.65	0.56	0.68	0.82
12	0.80	0.55	0.29	0.46	0.64	0.79	0.70	0.61	0.74	0.83
13	0.83	0.50	0.18	0.29	0.61	0.77	0.70	0.63	0.76	0.82
14	0.90	0.58	0.27	0.39	0.68	0.81	0.78	0.74	0.89	0.84
15	0.92	0.63	0.34	0.48	0.71	0.89	0.78	0.66	0.77	0.86
16	0.87	0.55	0.23	0.37	0.62	0.81	0.74	0.67	0.81	0.83
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**

interpretability compensating for 0.04 of performance.

These results are further analyzed in Figure 4 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 significantly better interpretable models by compensating a negligible degree of performance in 14 out of 16 subjects. Figure 4(a) shows that employing δ_Φ provides on average slightly higher accurate models (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 4(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$= 5.6 \times 10^{-6}$) 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 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 $\tilde{\eta}_\Phi = 0.63$.

The advantage of the exchange between the performance and the interpretability can be seen in the quality of MBMs. Figure 5a and 5b 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 sensors are plotted. Furthermore, the topographic maps represent the spatial pat-

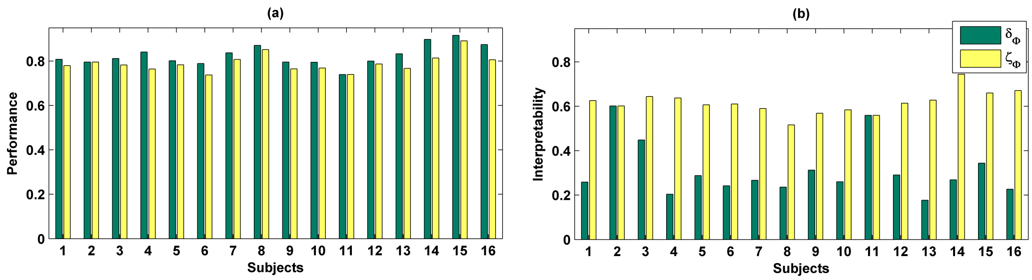
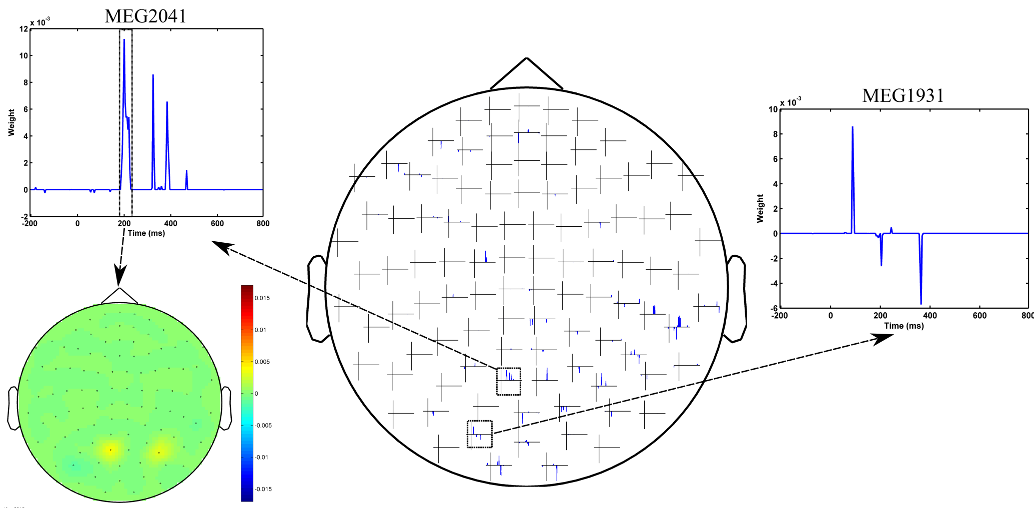


Figure 4: a) Comparison between performance of $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$. Adopting ζ_Φ instead of δ_Φ in model selection yields (on average) 0.04 less accurate classifiers over 16 subjects. b) Comparison between interpretability of $\hat{\Phi}_i^\delta$ and $\hat{\Phi}_i^\zeta$. Adopting ζ_Φ instead of δ_Φ in model selection yields on average 0.31 more interpretable classifiers over 16 subjects.

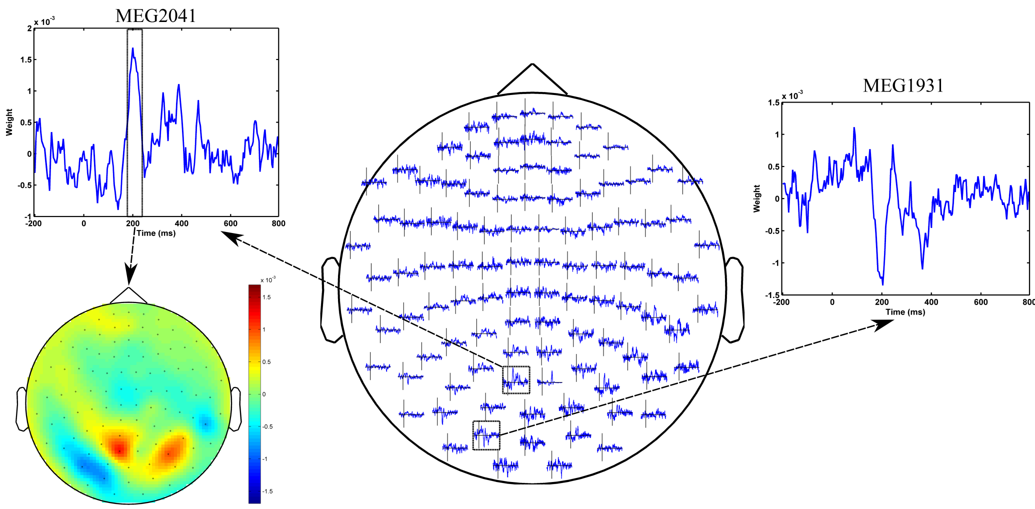
terns of weights averaged between $184ms$ and $236ms$ after stimulus onset[1]. While $\hat{\vec{\Theta}}_1^\delta$ is sparse in time and space, it fails to accurately represent the spatio-temporal dynamic of the N170 component. Furthermore, the multicollinearity problem arising from the correlation between 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, $\hat{\vec{\Theta}}_1^\zeta$ represents the dynamic of the N170 component in time (see Figure 6). In addition, it also shows the spatial pattern of two dipoles in the posterior and temporal areas. In summary, $\hat{\vec{\Theta}}_1^\zeta$ suggests a more representative pattern of the underlying neurophysiological effect than $\hat{\vec{\Theta}}_1^\delta$.

In addition, optimizing the hyper-parameters of brain decoding based on ζ_Φ offers more reproducible brain decoders. According to table 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 7 visualizes $\hat{\vec{\Theta}}_1^\delta$ and $\hat{\vec{\Theta}}_1^\zeta$ over 4 perturbed training sets. The spatial maps [Figure 7(a) and Figure 7(c)] are plotted for the magnetometer sensors averaged in the time interval between $184ms$ and $236ms$ after stimulus onset. The temporal maps [Figure 7(b) and Figure 7(d)] are showing

[1]The bounds of colorbars are symmetrized based on the maximum absolute value of parameters



(a) Spatio-temporal pattern of $\vec{\hat{\Theta}}_1^{\delta}$.



(b) Spatio-temporal pattern of $\vec{\hat{\Theta}}_1^{\zeta}$.

Figure 5: Comparison between spatio-temporal multivariate maps of the most accurate (5a) and the most interpretable (5b) classifiers for Subject 1. $\vec{\hat{\Theta}}_1^{\zeta}$ provides more spatio-temporal representativeness of the N170 effect than $\vec{\hat{\Theta}}_1^{\delta}$.

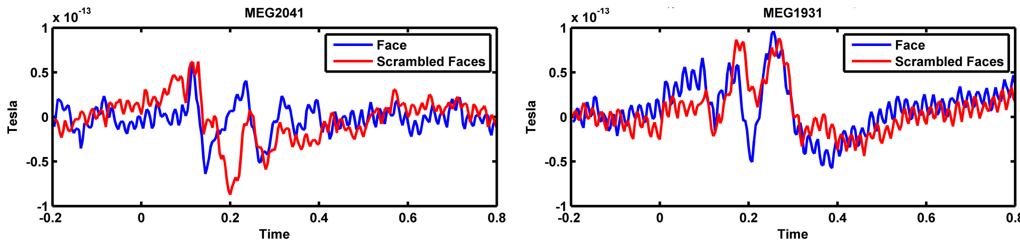


Figure 6: Event related fields (ERFs) of face and scrambled face samples for MEG2041 and MEG1931 sensors.

the multivariate temporal maps of MEG1931 and MEG2041 sensors. While $\vec{\tilde{\Theta}}_1^{\delta}$ is unstable in time and space across the 4 perturbed training sets, $\vec{\tilde{\Theta}}_1^{\zeta}$ provides more reproducible maps.

4. Discussions

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 (see the quick meta-analysis on the literature in the supplementary material); however, a formal formulation of the interpretability is never presented. In this study, our primary interest is to present a 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 time-domain MEG decoding in which we show how the presented definition can be exploited to heuristically approximate the interpretability. This example shows how

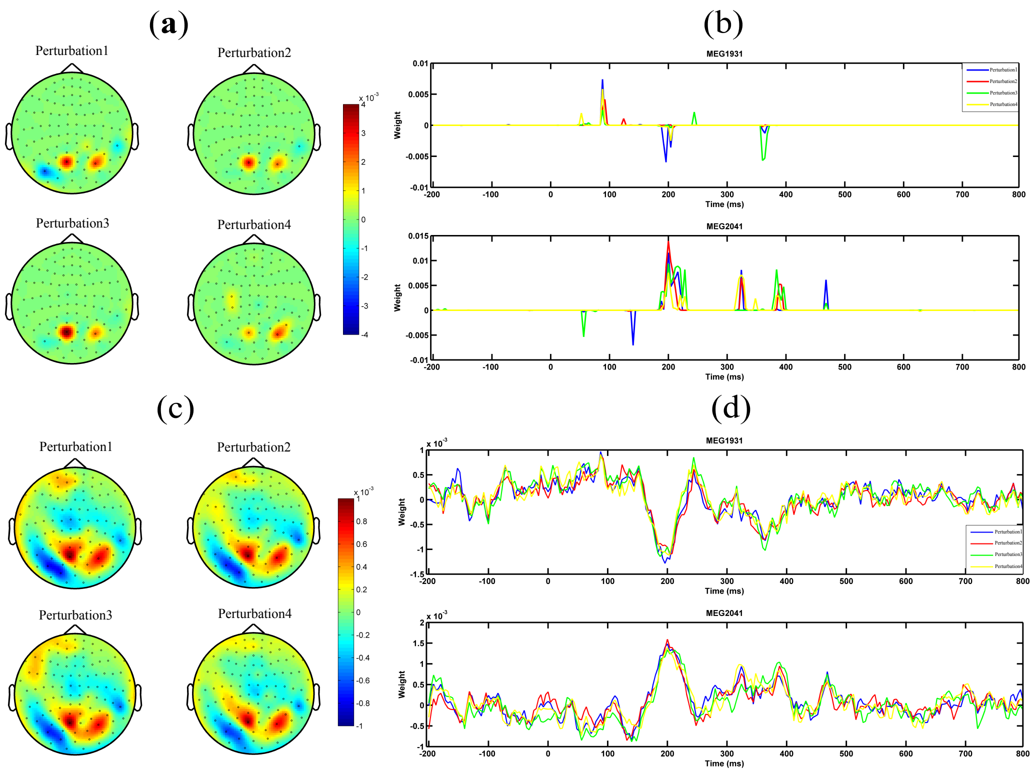


Figure 7: Comparison of the reproducibility of Lasso when δ_Φ and ζ_Φ are used in the model selection procedure. (a) and (b) show the spatio-temporal patterns represented by $\vec{\hat{\Theta}}_1^\delta$ across the 4 perturbed training sets. (c) and (d) show the spatio-temporal patterns represented by $\vec{\hat{\Theta}}_1^\zeta$ across the 4 perturbed training sets. Employing ζ_Φ instead of δ_Φ in the model selection yields more reproducible MBMs.

partial prior knowledge[1] regarding underlying brain 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.

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 [53]. Therefore, considering extra criteria might be required. However, because of the lack of a formal definition for interpretability, different characteristics of brain decoding models are considered as the main objective in improving their interpretability. Reproducibility [53, 54], stability selection [7, 47, 69], sparsity [96], and neurophysiological plausibility [97] 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 [53, 54, 43] or the stability of the classifiers [56, 57]

[1]The partial knowledge can be based on already known facts regarding the timing and location of neural activity.

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 with respect 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 1, there is a trend in research within the brain decoding context in which prior knowledge is injected into the penalization term as a technique to improve the interpretability of decoding models. Thus far, in the literature, there is no ad-hoc method to compare these different methods. 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.

4.3. Regularization and Interpretability

Haufe et al. [39] 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. The problem is primarily caused by the decoding process that minimizes the classification error only considering the uncertainty in the output space [80, 98, 99] and not the uncertainty in the input space (or noise). The authors concluded that the interpretability of brain decoding cannot be improved using regularization. Our experimental results on the toy data (see Section 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 significant role in improving the interpretability of results. For example, in this case, the true generative function behind the sampled data is sparse (see Section 2.6.1), but because of the noise in the data, the sparse model is not the most accurate one. Using a more comprehensive criterion (in this case, ζ_Φ) shows the advantage of selecting correct prior assumptions about the distribution of the data via regularization. This observation encourages the modification of the conclusion in [39] 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.

4.4. Advantage over Mass-Univariate Analysis

Mass-univariate hypothesis testing methods are among the most popular tools in neuroscience research because they provide significance checks and a fair level of interpretability via univariate brain maps. 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 [5] 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 [2]. 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.

4.5. 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 [27]. 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 supplementary material). Finding physiologically relevant heuristics for other acquisition modalities such as fMRI can be also considered in future work.

5. Conclusions

We presented a novel theoretical definition for the interpretability of linear brain decoding and associated multivariate brain maps. We demonstrated how the interpretability relates to the representativeness and reproducibility of brain decoding. Although it is theoretical, the presented definition provides a first step toward practical solution for filling the knowledge extraction gap in linear brain decoding. As an example of this major breakthrough, and to provide a proof of concept, a heuristic approach based on the contrast event-related field is proposed for practical evaluation of the interpretability in time-domain MEG decoding. We experimentally showed that adding the interpretability of brain decoding models as a criterion in the model selection procedure yields significantly higher interpretable models by sacrificing a negligible amount of performance. 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.

Acknowledgments

The author wishes to thank Sandro Vega-Pons and Nathan Weisz for valuable discussions and comments.

Appendix A. cERF and its Generative Nature

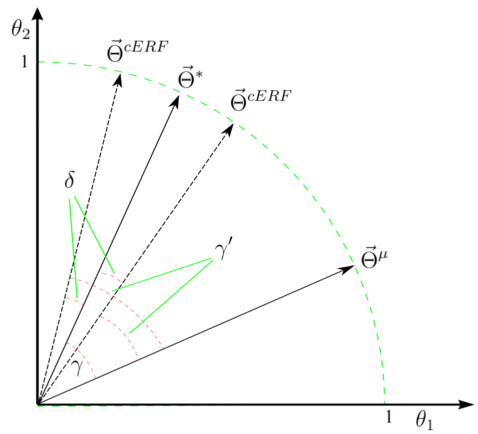
According to [39], for a linear discriminative model with parameters Θ, the unique equivalent generative model can be computed as follows:

$$A \propto \Sigma_{\mathbf{X}} \Theta \tag{A.1}$$

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

$$A \propto \Sigma_{\mathbf{X}} \Sigma_{\mathbf{X}}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y} = \mu^+ - \mu^- \tag{A.2}$$

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,



(a)

Figure B.8: Misrepresentation of $\vec{\Theta}^{cERF}$ with respect to $\vec{\Theta}^{*}$.

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.

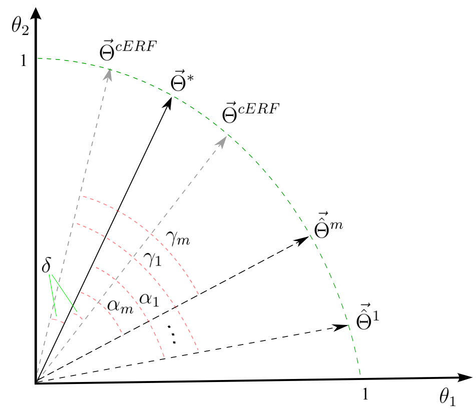
Appendix B. Relation between $\boldsymbol{\beta_\Phi}$ and $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\Phi}}$(Eq. 10)

Let γ be the angle between $\vec{\Theta}^{\mu}$ and $\vec{\Theta}^{*}$. Let γ' be the angle between $\vec{\Theta}^{\mu}$ and $\vec{\Theta}^{cERF}$. Furthermore, assume that δ is the angle between $\vec{\Theta}^{*}$ and $\vec{\Theta}^{cERF}$ and that $\Delta_\beta = \cos(\delta)$. We consider both cases in which β_Φ is underestimated/overestimated by $\tilde{\beta}_\Phi$ (see Figure B.8 as an example in 2-dimensional space). Then, we have:

$$\begin{aligned} \gamma = \gamma' \pm \delta \Rightarrow \cos(\gamma) &= \cos(\gamma' \pm \delta) \\ &= \cos(\gamma)\cos(\delta) \pm \sin(\gamma)\sin(\delta) = \tilde{\beta}_\Phi \Delta_\beta \pm \sqrt{(1-\tilde{\beta}^2)(1-\Delta_\beta^2)} \end{aligned} \tag{B.1}$$

Appendix C. Relation between $\boldsymbol{\eta_\Phi}$ and $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\Phi}}$(Eq. 12)

Let $\alpha_1, \ldots, \alpha_m$ be the angles between $\hat{\vec{\Theta}}^1, \ldots, \hat{\vec{\Theta}}^m$ and $\vec{\Theta}^{*}$, and $\gamma_1, \ldots, \gamma_m$ be the angles between $\hat{\vec{\Theta}}^1, \ldots, \hat{\vec{\Theta}}^m$ and $\vec{\Theta}^{cERF}$. Furthermore, assume that



(a)

Figure C.9: Relation between η_Φ and $\tilde{\eta}_\Phi$.

δ is the angle between $\vec{\Theta}^*$ and $\vec{\Theta}^{cERF}$. We consider both cases in which η_Φ is underestimated/overestimated by $\tilde{\eta}_\Phi$ (see Figure C.9 as an example in 2-dimensional space).

$$\begin{aligned}
\eta_\Phi &= \frac{\cos(\alpha_1) + \cdots + \cos(\alpha_m)}{m} = \frac{\cos(\gamma_1 \pm \delta) + \cdots + \cos(\gamma_m \pm \delta)}{m} \\
&= \frac{\cos(\gamma_1)\cos(\delta) \pm \sin(\gamma_1)\sin(\delta) + \cdots + \cos(\gamma_m)\cos(\delta) \pm \sin(\gamma_m)\sin(\delta)}{m} \\
&\xrightarrow{\Delta_\beta = \cos(\delta)} = \frac{\Delta_\beta[\cos(\gamma_1) + \cdots + \cos(\gamma_m)] \pm \sin(\delta)[\sin(\gamma_1) + \cdots + \sin(\gamma_m)]}{m} \\
&\xrightarrow{\tilde{\eta}_\Phi = \frac{\cos(\gamma_1) + \cdots + \cos(\gamma_m)}{m}} \eta_\Phi = \Delta_\beta \tilde{\eta}_\Phi \pm \frac{\sqrt{1 - \Delta_\beta^2}}{m}(\sin(\gamma_1) + \cdots + \sin(\gamma_m))
\end{aligned} \tag{C.1}$$

Appendix D. Proof of Proposition 1

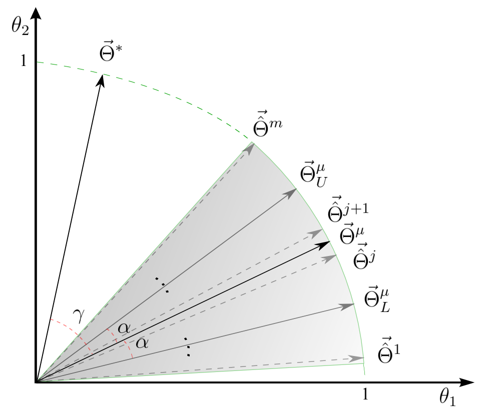
Throughout this proof, we assume that all of the parameter vectors are normalized in the unit hypersphere (see Figure D.10 as an illustrative example in 2 dimensions). Let $T = \{\vec{\hat{\Theta}}^1, \ldots, \vec{\hat{\Theta}}^m\}$ be a set m MBMs, for m perturbed training sets where $\vec{\hat{\Theta}}^i \in \mathbb{R}^p$. 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 $\vec{\Theta}^i \triangledown \vec{\Theta}^k$ 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 $T_U = \{\vec{\Theta}^i \mid \vec{\Theta}^i \triangledown \vec{\Theta}^*\}$ 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. 5, 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. 4 and using a cosine similarity definition, we have:

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

A similar procedure can be used to prove $\tilde{\eta}_\Phi = \tilde{\beta}_\Phi \times \psi_\Phi$ by replacing $\vec{\Theta}^*$ with $\vec{\Theta}^{cERF}$.

Appendix E. Computing the Bias and Variance in Binary Classification

Here, using the out-of-bag (OOB) technique, and based on procedures proposed by [83] and [100], we compute the expected prediction error (EPE) for a linear binary classifier Φ under bootstrap perturbation of the training set. Let m be the number of perturbed training sets resulting from partitioning (X, Y) into (X_{tr}, Y_{tr}) and (X_{ts}, Y_{ts}), i.e., training and test sets. If $\hat{\Phi}^j$ is the linear 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:



(a)

Figure D.10: Relation between representativeness, reproducibility, and interpretability in 2 dimensions.

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

where k_i is the number of times that x_i is present in the test set[1].1

The computation of bias is challenging because the optimal model Φ^* is unknown. According to [101], misclassification error is one of the loss measures that satisfies a Pythagorean-type equality, and:

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\Phi^\mu(\mathbf{x}_i), \Phi^*(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^\mu(\mathbf{x}_i)) - \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^*(\mathbf{x}_i)) \tag{E.2}$$

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:

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\Phi^\mu(\mathbf{x}_i), \Phi^*(\mathbf{x}_i)) \leq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \Phi^\mu(\mathbf{x}_i)) \tag{E.3}$$

[1]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.

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

$$B(\mathbf{x}_i) = \begin{cases} 0 & if \quad \Phi^{\mu}(\mathbf{x}_i) = y_i \\ 1 & \quad otherwise \end{cases} \tag{E.4}$$

Then, the unbiased and biased variances (see [83] for definitions) in each training set can be calculated by:

$$V_u^j(\mathbf{x}_i) = \begin{cases} 1 & if \quad B(\mathbf{x}_i) = 0 \quad and \quad \Phi^{\mu}(\mathbf{x}_i) \neq \hat{\Phi}^j(\mathbf{x}_i) \\ 0 & \qquad otherwise \end{cases} \tag{E.5}$$

$$V_b^j(\mathbf{x}_i) = \begin{cases} 1 & if \quad B(\mathbf{x}_i) = 1 \quad and \quad \Phi^{\mu}(\mathbf{x}_i) \neq \hat{\Phi}^j(\mathbf{x}_i) \\ 0 & \qquad otherwise \end{cases} \tag{E.6}$$

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

$$\begin{aligned} EPE_{\Phi}(X) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} B(\mathbf{x}_i)}_{Bias} + \\ \underbrace{\frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}[V_u^j(\mathbf{x}_i) - V_b^j(\mathbf{x}_i)]}_{Variance} \end{aligned} \tag{E.7}$$

References

[1] E. Crivellato, D. Ribatti, Soul, mind, brain: Greek philosophy and the birth of neuroscience, Brain research bulletin 71 (2007) 327–336.

[2] D. M. Groppe, T. P. Urbach, M. Kutas, Mass univariate analysis of event-related brain potentials/fields i: A critical tutorial review, Psychophysiology 48 (2011) 1711–1725.

[3] E. Maris, Statistical testing in electrophysiological studies, Psychophysiology 49 (2012) 549–565.

[4] E. Bullmore, M. Brammer, S. C. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, P. Sham, Statistical methods of estimation and inference for functional mr image analysis, Magnetic Resonance in Medicine 35 (1996) 261–277.

[5] E. Maris, R. Oostenveld, Nonparametric statistical testing of eeg-and meg-data, Journal of neuroscience methods 164 (2007) 177–190.

[6] D. M. Groppe, T. P. Urbach, M. Kutas, Mass univariate analysis of event-related brain potentials/fields ii: Simulation studies, Psychophysiology 48 (2011) 1726–1737.

[7] M. van Gerven, C. Hesse, O. Jensen, T. Heskes, Interpreting single trial data using groupwise regularisation, NeuroImage 46 (2009) 665–676.

[8] T. Davis, K. F. LaRocque, J. A. Mumford, K. A. Norman, A. D. Wagner, R. A. Poldrack, What do differences between multi-voxel and univariate analysis mean? how subject-, voxel-, and trial-level variance impact fmri analysis, NeuroImage 97 (2014) 271–283.

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

[10] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, T. M. Vaughan, Brain–computer interfaces for communication and control, Clinical neurophysiology 113 (2002) 767–791.

[11] L. F. Nicolas-Alonso, J. Gomez-Gil, Brain computer interfaces, a review, Sensors 12 (2012) 1211–1279.

[12] D. Bzdok, Classical statistics and statistical learning in imaging neuroscience, arXiv preprint arXiv:1603.01857 (2016).

[13] F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: a tutorial overview., NeuroImage 45 (2009) 199–209.

[14] S. Lemm, B. Blankertz, T. Dickhaus, K.-R. Müller, Introduction to machine learning for brain imaging, Neuroimage 56 (2011) 387–399.

[15] M. Besserve, K. Jerbi, F. Laurent, S. Baillet, J. Martinerie, L. Garnero, Classification methods for ongoing eeg and meg signals, Biological research 40 (2007) 415–437.

[16] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex, Science 293 (2001) 2425–2430.

[17] D. D. Cox, R. L. Savoy, Functional magnetic resonance imaging (fmri)brain reading: detecting and classifying distributed patterns of fmri activity in human visual cortex, Neuroimage 19 (2003) 261–270.

[18] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, S. Newman, Learning to decode cognitive states from brain images, Machine Learning 57 (2004) 145–175.

[19] K. A. Norman, S. M. Polyn, G. J. Detre, J. V. Haxby, Beyond mind-reading: multi-voxel pattern analysis of fmri data, Trends in cognitive sciences 10 (2006) 424–430.

[20] L. Parra, C. Alvino, A. Tang, B. Pearlmutter, N. Yeung, A. Osman, P. Sajda, Single-trial detection in EEG and MEG: Keeping it linear, Neurocomputing 52-54 (2003) 177–183.

[21] J. W. Rieger, C. Reichert, K. R. Gegenfurtner, T. Noesselt, C. Braun, H.-J. Heinze, R. Kruse, H. Hinrichs, Predicting the recognition of natural scenes from single trial meg recordings of brain activity, Neuroimage 42 (2008) 1056–1068.

[22] M. K. Carroll, G. A. Cecchi, I. Rish, R. Garg, A. R. Rao, Prediction and interpretation of distributed neural activity with sparse models, NeuroImage 44 (2009) 112–122.

[23] A. M. Chan, E. Halgren, K. Marinkovic, S. S. Cash, Decoding word and category-specific spatiotemporal representations from meg and eeg, Neuroimage 54 (2011) 3028–3039.

[24] H. Huttunen, T. Manninen, J.-P. Kauppi, J. Tohka, Mind reading with regularized multinomial logistic regression, Machine vision and applications 24 (2013) 1311–1325.

[25] D. Vidaurre, C. Bielza, P. Larrañaga, A survey of l1 regression, International Statistical Review 81 (2013) 361–387.

[26] M. Abadi, R. Subramanian, S. Kia, P. Avesani, I. Patras, N. Sebe, Decaf: Meg-based multimodal database for decoding affective physiological responses, IEEE Transactions on Affective Computing 6 (2015) 209–222.

[27] T. Naselaris, K. N. Kay, S. Nishimoto, J. L. Gallant, Encoding and decoding in fmri, Neuroimage 56 (2011) 400–410.

[28] S. Weichwald, T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, M. Grosse-Wentrup, Causal interpretation rules for encoding and decoding models in neuroimaging, NeuroImage 110 (2015) 48–59.

[29] N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping, Proceedings of the National Academy of Sciences of the United States of America 103 (2006) 3863–3868.

[30] F. J. Valverde-Albacete, C. Peláez-Moreno, 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox, PLOS ONE 9 (2014) e84217.

[31] A. Ramdas, A. Singh, L. Wasserman, Classification accuracy as a proxy for two sample testing, arXiv preprint arXiv:1602.02210 (2016).

[32] R. Turner, A model explanation system, 2015.

[33] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. Müller, How to explain individual classification decisions, The Journal of Machine Learning Research 11 (2010) 1803–1831.

[34] A. Vellido, J. Martin-Guerroro, 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, 2012, pp. 163–172.

[35] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015).

[36] G. Montavon, M. Braun, T. Krueger, K.-R. Muller, Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment, Signal Processing Magazine, IEEE 30 (2013) 62–74.

[37] D. Yu, S. J. Lee, W. J. Lee, S. C. Kim, J. Lim, S. W. Kwon, Classification of spectral data using fused lasso logistic regression, Chemometrics and Intelligent Laboratory Systems 142 (2015) 70–77.

[38] K. Hansen, D. Baehrens, T. Schroeter, M. Rupp, K.-R. Müller, Visual interpretation of kernel-based prediction models, Molecular Informatics 30 (2011) 817–826.

[39] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, F. Bießmann, On the interpretation of weight vectors of linear models in multivariate neuroimaging, NeuroImage (2013).

[40] M. R. Sabuncu, A universal and efficient method to compute maps from image-based prediction models, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014 (2014) 353–360.

[41] J.-D. Haynes, A primer on pattern-based approaches to fmri: Principles, pitfalls, and perspectives, Neuron 87 (2015) 257–270.

[42] T. Naselaris, K. N. Kay, Resolving ambiguities of mvpa using explicit models of representation, Trends in cognitive sciences 19 (2015) 551–554.

[43] S. C. Strother, P. M. Rasmussen, N. W. Churchill, K. Hansen, Stability and Reproducibility in fMRI Analysis, New York: Springer-Verlag, 2014.

[44] A. Anderson, J. S. Labus, E. P. Vianna, E. A. Mayer, M. S. Cohen, Common component classification: What can we learn from machine learning?, Neuroimage 56 (2011) 517–524.

[45] K. H. Brodersen, F. Haiss, C. S. Ong, F. Jung, M. Tittgemeyer, J. M. Buhmann, B. Weber, K. E. Stephan, Model-based feature construction for multivariate decoding, NeuroImage 56 (2011) 601–615.

[46] G. Langs, B. H. Menze, D. Lashkari, P. Golland, Detecting stable distributed patterns of brain activation using gini contrast, NeuroImage 56 (2011) 497–507.

[47] G. Varoquaux, A. Gramfort, B. 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 (ICML-12), 2012, pp. 1375–1382.

[48] J.-P. Kauppi, L. Parkkonen, R. Hari, A. Hyvärinen, Decoding magnetoencephalographic rhythmic activity using spectrospatial information, NeuroImage 83 (2013) 921–936.

[49] S. Taulu, J. Simola, J. Nenonen, L. Parkkonen, Novel noise reduction methods, Magnetoencephalography (2014) 35–71.

[50] G. Varoquaux, B. Thirion, How machine learning is shaping cognitive neuroimaging, GigaScience 3 (2014) 28.

[51] E. Olivetti, S. M. Kia, P. Avesani, Meg decoding across subjects, in: Pattern Recognition in Neuroimaging, 2014 International Workshop on, IEEE, 2014.

[52] S. Haufe, S. Dähne, V. V. Nikulin, Dimensionality reduction for the analysis of brain oscillations, NeuroImage (2014).

[53] P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, S. C. Strother, Model sparsity and brain pattern interpretation of classification models in neuroimaging, Pattern Recognition 45 (2012) 2085–2100.

[54] B. R. Conroy, J. M. Walz, P. Sajda, Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fmri decoding models, PloS one 8 (2013) e79271.

[55] O. Bousquet, A. Elisseeff, Stability and generalization, The Journal of Machine Learning Research 2 (2002) 499–526.

[56] B. Yu, Stability, Bernoulli 19 (2013) 1484–1500.

[57] C. Lim, B. Yu, Estimation stability with cross validation (escv), Journal of Computational and Graphical Statistics (2015).

[58] N. Mørch, L. K. Hansen, S. C. Strother, C. Svarer, D. A. Rottenberg, B. Lautrup, R. Savoy, O. B. Paulson, Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover, in: Information processing in medical imaging, Springer Berlin Heidelberg, 1997, pp. 259–270.

[59] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 49–67.

[60] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2005) 91–108.

[61] E. P. Xing, M. Kolar, S. Kim, X. Chen, High-dimensional sparse structured input-output models, with applications to gwas, Practical Applications of Sparse Modeling (2014) 37.

[62] I. Rish, G. A. Cecchi, A. Lozano, A. Niculescu-Mizil, Practical Applications of Sparse Modeling, MIT Press, 2014.

[63] L. Grosenick, S. Greer, B. Knutson, Interpretable classifiers for fmri improve prediction of purchases, Neural Systems and Rehabilitation Engineering, IEEE Transactions on 16 (2008) 539–548.

[64] M. de Brecht, N. Yamagishi, Combining sparseness and smoothness improves classification accuracy and interpretability, NeuroImage 60 (2012) 1550–1561.

[65] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, B. Thirion, Total variation regularization for fmri-based prediction of behavior, Medical Imaging, IEEE Transactions on 30 (2011) 1328–1340.

[66] A. Gramfort, B. Thirion, G. Varoquaux, Identifying predictive regions from fmri with tv-l1 prior, in: Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on, IEEE, 2013, pp. 17–20.

[67] L. Grosenick, B. Klingenberg, S. Greer, J. Taylor, B. Knutson, Whole-brain sparse penalized discriminant analysis for predicting choice, NeuroImage 47 (2009) S58.

[68] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, J. E. Taylor, Interpretable whole-brain prediction analysis with graphnet, NeuroImage 72 (2013) 304–321.

[69] Y. Wang, J. Zheng, S. Zhang, X. Duan, H. Chen, Randomized structural sparsity via constrained block subsampling for improved sensitivity of discriminative voxel identification, NeuroImage (2015).

[70] F. Bießmann, S. Dähne, F. C. Meinecke, B. Blankertz, K. Görgen, K.-R. Müller, S. 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, 2012.

[71] S. Haufe, F. Meinecke, K. Gorgen, S. Dahne, J.-D. Haynes, B. Blankertz, F. Biessmann, Parameter interpretation, regularization and source localization in multivariate linear models, in: Pattern Recognition in Neuroimaging, 2014 International Workshop on, IEEE, 2014, pp. 1–4.

[72] D. A. Engemann, A. Gramfort, Automated model selection in covariance estimation and spatial whitening of meg and eeg signals, NeuroImage 108 (2015) 328–342.

[73] Z. Li, Y. Wang, Y. Wang, X. Wang, J. Zheng, H. Chen, A novel feature selection approach for analyzing high dimensional functional mri data, arXiv preprint arXiv:1506.08301 (2015).

[74] S. M. Kia, S. Vega-Pons, E. Olivetti, P. Avesani, Multi-task learning for interpretation of brain decoding models, in: NIPS Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI), 2014, Springer Lecture Notes on Artificial Intelligence Series, In press.

[75] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[76] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B 67 (2005) 301–320.

[77] R. Jenatton, J.-Y. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms, arXiv preprint arXiv:0904.3523 (2009).

[78] T. Poggio, C. Shelton, On the mathematical foundations of learning, American Mathematical Society 39 (2002) 1–49.

[79] M. C.-K. Wu, S. V. David, J. L. Gallant, Complete functional characterization of sensory neurons by system identification, Annu. Rev. Neurosci. 29 (2006) 477–505.

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

[81] B. Efron, Bootstrap methods: another look at the jackknife, The annals of Statistics (1979) 1–26.

[82] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Ijcai, volume 14, 1995, pp. 1137–1145.

[83] P. Domingos, A unified bias-variance decomposition for zero-one and squared loss, AAAI/IAAI 2000 (2000) 564–569.

[84] M. D. Rugg, M. G. Coles, Electrophysiology of mind: Event-related brain potentials and cognition., Oxford University Press, 1995.

[85] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, volume 2, Springer, 2009.

[86] A. Gramfort, G. Varoquaux, B. Thirion, Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify, in: Machine Learning and Interpretation in Neuroimaging, Springer, 2012, pp. 9–16.

[87] M. Caramia, P. Dell´ Olmo, Multi-objective optimization, Multi-objective Management in Freight Logistics: Increasing Capacity, Service Level and Safety with Optimization Algorithms (2008) 11–36.

[88] R. T. Marler, J. S. Arora, Survey of multi-objective optimization methods for engineering, Structural and multidisciplinary optimization 26 (2004) 369–395.

[89] R. N. Henson, D. G. Wakeman, V. Litvak, K. 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 (2011).

[90] S. Bentin, T. Allison, A. Puce, E. Perez, G. McCarthy, Electrophysiological studies of face perception in humans, Journal of cognitive neuroscience 8 (1996) 551–565.

[91] D. H. Wolpert, W. G. Macready, An efficient method to estimate bagging's generalization error, Machine Learning 35 (1999) 41–55.

[92] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[93] V. N. Vapnik, S. Kotz, Estimation of dependences based on empirical data, volume 40, Springer-verlag New York, 1982.

[94] V. Vapnik, The nature of statistical learning theory, Springer Science & Business Media, 2013.

[95] R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data, Computational intelligence and neuroscience 2011 (2010).

[96] S. Dash, D. M. Malioutov, K. R. Varshney, Learning interpretable classification rules using sequential rowsampling, in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 3337–3341.

[97] B. Afshin-Pour, H. Soltanian-Zadeh, G.-A. Hossein-Zadeh, C. L. Grady, S. C. Strother, A mutual information-based metric for evaluation of fmri data-processing approaches, Human brain mapping 32 (2011) 699–715.

[98] J. B. T. Zhang, Support vector classification with input data uncertainty, Advances in neural information processing systems 17 (2005) 161.

[99] C. Tzelepis, V. Mezaris, I. Patras, Linear maximum margin classifier for learning from uncertain data, arXiv preprint arXiv:1504.03892 (2015).

[100] G. Valentini, T. G. Dietterich, Bias-variance analysis of support vector machines for the development of svm-based ensemble methods, The Journal of Machine Learning Research 5 (2004) 725–775.

[101] R. Tibshirani, Bias, variance and prediction error for classification rules, University of Toronto, Department of Statistics, 1996.