Interpretability in Linear Brain Decoding

Seyed Mostafa Kia SEYEDMOSTAFA.KIA@UNITN.IT
Andrea Passerini ANDREA.PASSERINI@UNITN.IT

Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 9 I-38123 Povo, Trento, Italy

Abstract

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 brain decoding models. As a consequence, there is no quantitative measure for evaluating the interpretability of different brain decoding methods. In this paper, we present a simple definition for interpretability of linear brain decoding models. Then, we propose to combine the interpretability and the performance of the brain decoding into a new multi-objective criterion for model selection. Our preliminary results on the toy data show that optimizing the hyper-parameters of the regularized linear classifier based on the proposed criterion results in more informative linear models. The presented definition provides the theoretical background for quantitative evaluation of interpretability in linear brain decoding.

1. Introduction

In cognitive science, researchers usually analyze recorded brain activity 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 (Groppe et al., 2011). 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. The multivariate counterparts of mass-univariate analysis, known generally as multivariate pattern analysis (MVPA), have the potential to overcome this deficit.

Brain decoding (Haynes & Rees, 2006) is an MVPA technique that delivers a model to predict the mental state of a human subject based on the recorded brain signal. From the neuroscientific perspective, a brain map resulting from weight of linear brain decoding model is considered *interpretable* if it enables the scientist to answer *where*, *when*, and *how* questions. But typically a classifier, taken alone, only answers the question of *what* is the most likely label of a given unseen sample. This fact is generally known as knowledge extraction gap (Vellido et al., 2012) 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 (Bach et al., 2015), signal processing (Montavon et al., 2013), chemometrics (Yu et al., 2015), bioinformatics (Hansen et al., 2011), and neuroinformatics (Haufe et al., 2013).

Despite the theoretical advantages of MVPA, its practical application to inferences regarding neuroimaging data is limited primarily due to the knowledge extraction gap (Sabuncu, 2014). Therefore, improving the interpretability of linear brain decoding and associated brain maps is a primary goal in the brain imaging literature (Strother et al., 2014). 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. At present, two main approaches are proposed to enhance the interpretability of multivariate brain maps: 1) introducing new metrics, such as reproducibility of maps or stability of models, into the model selection procedure (Rasmussen et al., 2012; Conroy et al., 2013; Yu, 2013), and 2) introducing new hybrid penalty terms for regularization to incorporate spatio-temporal prior knowledge in the learning (van Gerven et al., 2009; Michel et al., 2011; de Brecht & Yamagishi, 2012; Grosenick et al., 2013).

In spite of the aforementioned efforts to improve the interpretability, 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. With the aim of filling this gap, our contribution is two-fold: 1) assuming that the true solution of brain decoding is available, we present a simple definition of the interpretability in linear brain decoding; 2) 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 a toy dataset, show that incorporating the interpretability into the model selection procedure provides more interpretable models [1].

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 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. In the neuroimaging context, $\mathbf{X}$ indicates the trials of brain recording, and $\mathbf{Y}$ represents the experimental conditions. The goal of brain decoding is to find the function $\Phi_S : \mathbf{X} \to \mathbf{Y}$ as an estimation of the ideal function $\Phi^* : \mathcal{X} \to \mathcal{Y}$.

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 $\mathcal{H}$. 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 solving a risk minimization problem:

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

where Θ denotes the parameters of the linear model, $\mathcal{L} : \mathbf{Z} \times \mathbf{Z} \to \mathbb{R}^+$ 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. λ is generally decided using cross-validation or other data perturbation methods in the model selection procedure.

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. 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 $\vec{\Theta} = \frac{\Theta}{\|\Theta\|}$ ($\|.\|$ is the 2-norm).

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^*) (Poggio & Shelton, 2002). The *irreducible error* ε is the direct consequence of sampling; it sets a lower bound on the error, where we have:

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

2.2. Theoretical Definition

In this section, we present a 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* 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 by:

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

$\Sigma_{\mathbf{X}}$ represents the covariance of $\mathbf{X}$. Using Θ^* as the reference, we can define the *strong-interpretability*:

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

In practice, the estimated solution of a linear brain problem is not strongly-interpretable because of the inherent limitations of neuroimaging data, such as uncertainty (Aggarwal & Yu, 2009) 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 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. *Let $S^1, \ldots, S^m$ be m perturbed training sets drawn from S via a certain perturbation scheme such as bootstrapping, or cross-validation. Assume $\vec{\hat{\Theta}}^1, \ldots, \vec{\hat{\Theta}}^m$ are m MBMs of a certain Φ on the corresponding perturbed training sets. How can we quantify the proximity of Φ to the strongly-intrepretable solution of brain decoding problem Φ^*?*

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:

$$\eta_\Phi = \frac{1}{m} \sum_{j=1}^{m} \cos(\alpha^j) \qquad (4)$$

In fact, the interpretability is the average cosine similarities between Θ^* and MBMs derived from different samplings of the training set. Even though, in practice, the exact computation of η_Φ is unrealistic (as Θ^* is not available), the interpretability of the decoding model can be approximated based on ad-hoc heuristics (see (Kia, 2016) for an example in the magnetoenecephalography decoding). The approximated interpretability can be incorporated in the model selection procedure in order to find more reproducible and plausible decoding models.

2.3. Interpretability in 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* (Hastie et al., 2009). 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. 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 (Rasmussen et al., 2012; Conroy et al., 2013). Here, we propose a multi-objective criterion for model selection that takes into account both prediction accuracy and MBM interpretability.

Let η_Φ and δ_Φ be the interpretability and the generalization performance of a linear function Φ, respectively. We propose the use of the *scalarization* technique (Caramia & Dell´ Olmo, 2008) for combining η_Φ and δ_Φ into one scalar $0 \leq \zeta(\Phi) \leq 1$ as follows:

$$\zeta_\Phi = \begin{cases} \frac{\omega_1 \eta_\Phi + \omega_2 \delta_\Phi}{\omega_1 + \omega_2} & \delta_\Phi \geq \kappa \\ 0 & \delta_\Phi < \kappa \end{cases} \qquad (5)$$

where ω_1 and ω_2 are weights that specify the importance of the interpretability and the performance, respectively. κ is a threshold that filters out solutions with poor performances. In classification scenarios, κ can be set by adding a small safe interval to the chance level. It can be shown that the hyper-parameters of a model Φ are optimized based on ζ_Φ are Pareto optimal (Marler & Arora, 2004).

2.4. Classification and Evaluation

In our experiment, a least squares classifier with L1-penalization, i.e., Lasso (Tibshirani, 1996), is used for decoding. Lasso is a popular classification method in brain decoding, mainly because of its sparsity assumption. The
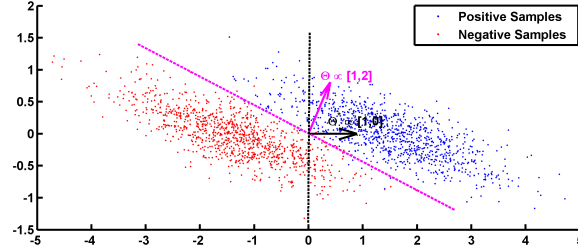

Figure 1. Noisy samples of toy data. The black line shows the true separator based on the generative model (Φ^*). The magenta line shows the most accurate classification solution.

choice of Lasso helps us to better illustrate the importance of including the interpretability in the model selection. Lasso solves the following optimization problem:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \|\Phi(\mathbf{X}) - \Phi_S(\mathbf{X})\|_2^2 + \lambda \|\Theta\|_1 \qquad (6)$$

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 $\lambda = \{0.001, 0.01, 0.1, 1, 10, 50, 100, 250, 500, 1000\}$.

We use the out-of-bag (OOB) (Breiman, 2001) method to compute δ_Φ, η_Φ, and ζ_Φ for different values of λ. In OOB, given a training set $(\mathbf{X}, \mathbf{Y})$, m replications of bootstrap are used to create perturbed training sets (we set $m = 50$) [1]. 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.

3. Experiment

3.1. Experimental Material

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 (Haufe et al., 2013). 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.

---

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

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**
$\vec{\Theta} \propto$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4520\\0.8920\end{bmatrix}$	$\begin{bmatrix}0.4521\\0.8919\end{bmatrix}$	$\begin{bmatrix}0.4532\\0.8914\end{bmatrix}$	$\begin{bmatrix}0.4636\\0.8660\end{bmatrix}$	$\begin{bmatrix}0.4883\\0.8727\end{bmatrix}$	$\begin{bmatrix}0.5800\\0.8146\end{bmatrix}$	$\begin{bmatrix}0.99\\0.02\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$

3.2. Results

In the definition of Φ^* on the toy dataset, 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, $\vec{\Theta}^* \propto [1,0]^T$ 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 (Haufe et al., 2013). 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 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 $\vec{\Theta} \propto [0.4636, 0.8660]^T$ (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 the model based on its performance does not necessarily improve its interpretability. This observation confirms the previous finding by Rasmussen et al. (2012) 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.

4. Discussions

In this study, our primary interest was to present a definition of the interpretability of linear brain decoding models. 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 simple 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.

Despite ubiquitous use, the generalization performance of classifiers is not a reliable criterion for assessing the interpretability of brain decoding models (Rasmussen et al., 2012). 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. Our definition of interpretability helped us to fill this gap by introducing a new multi-objective criterion as a weighted compromise between interpretability and generalization performance. Furthermore, this work presents an effective approach for evaluating the quality of different regularization strategies for improving the interpretability of MBMs. Our findings provide a further step toward direct evaluation of interpretability of the currently proposed penalization strategies.

Despite theoretical advantages, the proposed definition of interpretability suffer from some limitations. 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). Extending the definition of interpretability to non-linear models demands future research in visualization of non-linear models in the form of brain maps.

References

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

Bach, Sebastian, Binder, Alexander, Montavon, Grégoire,

Klauschen, Frederick, Müller, Klaus-Robert, and Samek, Wojciech. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

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

Caramia, Massimiliano and Dell´ Olmo, Paolo. Multi-objective optimization. *Multi-objective Management in Freight Logistics: Increasing Capacity, Service Level and Safety with Optimization Algorithms*, pp. 11–36, 2008.

Conroy, Bryan R, Walz, Jennifer M, and Sajda, Paul. Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fmri decoding models. *PloS one*, 8(11):e79271, 2013.

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

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

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

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

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The elements of statistical learning*, volume 2. Springer, 2009.

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

Haynes, John-Dylan and Rees, Geraint. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, July 2006. ISSN 1471-003X. doi: 10.1038/nrn1931.

Kia, Seyed Mostafa. Interpretability of multivariate brain maps in brain decoding: Definition and quantification. *bioRxiv*, 2016. doi: 10.1101/047522.

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

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

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

Poggio, T and Shelton, CR. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

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

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

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

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

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

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

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

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