Multi-Task Learning for Interpretation of Brain Decoding Models

Seyed Mostafa Kia[1,2,3], Sandro Vega-Pons[2,3], Emanuele Olivetti[2,3], and Paolo Avesani[2,3]

[1] University of Trento, Trento, Italy
[2] NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy
[3] Centro Interdipartimentale Mente e Cervello (CIMeC), University of Trento, Italy
seyedmostafa.kia@unitn.it

Abstract. Improving the interpretability of multivariate models is of primary interest for many neuroimaging studies. In this study, we present an application of multi-task learning (MTL) to enhance the interpretability of linear classifiers once applied to neuroimaging data. To attain our goal, we propose to divide the data into spatial fractions and define the temporal data of each spatial unit as a task in MTL paradigm. Our result on magnetoencephalography (MEG) data reveals preliminary evidence that, 1)dividing the brain recordings into spatial fractions based on spatial units of data and 2)considering each spatial fraction as a task, are two factors that provide more stability and consequently more interpretability for brain decoding models.

1 Introduction

Cognitive neuroscientists are generally concerned with discovering answer of *where*, *when* and *how* a certain brain activity contributes to a particular cognitive process. Hence, a multivariate analysis of recorded brain activity, e.g., Electroencephalography (EEG), Magnetoencephalography (MEG), or functional Magnetic Resonance Imaging (fMRI), is considered *interpretable* if it can find accurate and stable answer to *where*, *when* and *how* questions. Therefore, improving the interpretability of multivariate analysis is of high interest in the brain imaging literature [24].

Nowadays, mass-univariate hypothesis testing methods are widely employed to make inference on neuroimaing data [11, 17, 18]. Despite popularity of these univariate methods, they are generally unable to spot complex interactions between different brain areas [7]. Recent studies tried to find multivariate alternatives to univariate hypothesis testing [16, 20], however, classification-based approaches are still the most popular tools for multivariate analysis of neuroimaging data [9]. These approaches go under the name of *brain decoding* and generally use linear classifiers to find evidence for stimulus related information in neuroimaging data. The weights of linear classifiers provide quantitative measurements to assess the relation between each dimension of data, i.e., features, and the underlying cognitive task. However, these approaches suffer from lack

of interpretability due to the high dimensionality of data and high correlation between features [3, 12, 13].

Currently, there are two main directions in neuroimaging literature to improve the interpretability of multivariate linear models. The first concentrates on model selection in order to increase the stability of brain decoding model. This approach suggests taking into account the stability of models in model selection procedure. For example, [22] computed the correlation between weights of models across different cross-validation runs, and utilized it besides accuracy for model selection in joint accuracy-reproducibility space. Analogous approaches have been proposed in [1, 4, 6, 26].

The second approach focuses on the underlying mechanism of regularization to enhance the interpretability of weights of classifier. The main idea is two-fold: 1)customizing the regularization terms to address the ill-posed nature of brain decoding problems (where the number of samples are much less than the number of features); and 2)to incorporate structural or functional prior knowledge into the regularization procedure. Group Lasso [29] and total-variation penalty [25] are tow effective methods in this direction [23, 28]. As an example in the neuroimaging context, [9] by modifying the regularization term of logistic regression, proposed a group-wise regularization term for finding sparse and easy to interpret models. Elsewhere, [10] used total-variation penalty to inject a spatial segmentation prior into the sparse model with Lasso penalty. Similar efforts have been made in [3, 12, 27].

Despite the mentioned efforts, recently [13, 14] questioned the interpretability of linear discriminative models, i.e., weights of linear classifiers, due to the contribution of noise to the amplitude of weights. To address this problem, they proposed a procedure to transform discriminative models into equivalent generative models by multiplying linear classifier weights by the covariance matrix of the input features (see 2.2). Their experiments on simulated, EEG, and fMRI data illustrated that, whereas direct interpretation of linear classifier weights may cause misinterpretation of results, their proposed solution effectively solves the problem.

In this study, we approach the problem of interpretability by employing a multi-task learning (MTL) framework in order to improve the stability and as a result the interpretability of brain decoding models. We are willing to stress two key advantages of MTL over single-task learning in brain decoding interpretation: 1)reformulating the brain decoding problem into a multi-task problem, by defining each spatial unit of data as a task, provides more stability for brain decoding models; 2)learning the pattern of activities simultaneously over spatial units increases the performance of decoding compared to the single-task learning where a number of classifiers are trained separately on each spatial unit.

The rest of this paper is organized as follows: in section 2 we introduce multi-task elastic-net and we show how a brain decoding problem can be recast into the MTL paradigm. Then, in section 3, we present our experimental results on an MEG dataset by comparing the performance and the stability of MTL with single-task learning. Finally, section 4 concludes this paper.

2 Methods

2.1 Notation

Let $(X, Y) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \in \mathbb{R}^{n \times p} \times \mathbb{N}^n$ be the n samples of neuroimaging data, e.g., MEG data, where each $\mathbf{x}_i$ is a p dimensional vector of spatio-temporal features throughout presentation of stimulus of class y_i. The goal of brain decoding is to find a function Φ such that $Y = \Phi(X)$. In the linear case $Y = XW$ where $W \in \mathbb{R}^p$ represents the weights associated by a linear classifier to every corresponding element of $\mathbf{x}_i$.

2.2 From Classifier Weights to Activation Patterns

Recently, [13] showed the weights of a linear classifier, i.e, W, are not neurophysiologically interpretable. They illustrated that any interpretation based on W can cause wrong conclusions with respect to the spatio-temporal source of signal of interest. As a solution, showing that for every discriminative model there exists an equivalent generative model, they proposed a procedure to transform the weights of linear classifiers to *activation patterns* A:

$$A = \Sigma_X W \Sigma_{\hat{S}}^{-1} \tag{1}$$

where Σ_X and $\Sigma_{\hat{S}}^{-1}$ represent covariance matrix of X and $\hat{S}$, respectively, and $\hat{S}$ is latent factor representing estimated neural sources.

In fact, an activation pattern is the solution of the equivalent generative model that encodes the strength and polarity of the activity of interest in each dimension of data. Therefore, there is a clear physiological interpretation for activation patterns. In the binary classification setting where there is just one latent factor $\hat{Y}$ estimated by the model, the Eq.1 can be rewritten as:

$$A = \frac{\Sigma_X W}{\sigma_{\hat{Y}}^2} \propto \Sigma_X W \tag{2}$$

2.3 Multi-task Elastic-net

Multi-task learning (MTL) has recently received particular attention in machine learning and computer vision literature [30]. MTL tries to learn the underlying relation between tasks simultaneously by extracting common information across them. It has been shown that, in some applications, the simultaneous learning procedure of MTL is advantageous over learning each task independently [8]. Furthermore, splitting a single-task problem into a multi-task problem can effectively change the relative size of samples to features for each task. Thus MTL can provide higher stability by reducing the degree-of-freedom of the solution space.

In this study, we first define a *spatial fraction* as a time-series of each spatial unit of neuroimaging data. For example in the case of MEG data, the time-series measured by each MEG sensor is defined as one spatial fraction of data.

Then, we define each spatial fraction as a *task* in the MTL framework. We consider the MTL scenario of having the same outputs and different inputs for each task [2, 8]. Thus, a brain decoding problem can be reformulated as $(X, Y) = \{(X^1, Y), \ldots, (X^\tau, Y)\}$; where each pair of (X^i, Y) defines a traditional brain decoding problem (see 2.1) on just one spatial fraction of data, $X^i \in \mathbb{R}^{n \times p^*}$ represents n samples of data at ith spatial fraction, τ represents number of tasks; and $p^* = p/\tau$ is the number of temporal features at each spatial fraction.

Using this new representation of brain decoding, the multi-task elastic-net (MTEN) optimization problem, as an instance of MTL, can be formulated as follows [5, 31]:

$$\hat{W}^{MTEN} = \underset{W \in \mathbb{R}^{p^* \times \tau}}{\operatorname{argmin}} \sum_{i=1}^{\tau} \left\| X^i W^i - Y \right\|_F^2 + \rho_1 \left\| W \right\|_1 + \rho_2 \left\| W \right\|_F^2 \tag{3}$$

where $\|.\|_1$ and $\|.\|_F^2$ are representing the $l1$ and $l2$ penalties respectively, and $W \in \mathbb{R}^{p^* \times \tau}$ is the MTEN weight matrix. The regularization parameters ρ_1 and ρ_2 control sparsity and smoothness over temporal patterns of spatial fractions, respectively.

The MTEN optimization problem can be considered as an extension of single-task regression with elastic-net regularization [32]. A general specification of MTEN is its shared $l1$ and $l2$ penalties among all tasks. Furthermore, in this setting, the number of temporal features of each task (p^*) is reduced by factor of the number of tasks (τ) with respect to that of the original feature space (p). In practice and using common down-sampling techniques even $p^* < n$ is achievable. Therefore, the input data of each task can be a full rank matrix.

To compute the final prediction of the MTL model, we use a simple averaging mechanism among the tasks. We first define a *decoding-related task* (DRT) set D, as a set of tasks which provide decoding performance over a certain threshold ϕ in the training-set. The threshold ϕ can be decided using nested cross-validation or can be fixed based on some heuristics. After finding DRT members, to compute the final prediction for every sample in the test-set, we compute the mean over predictions of classifiers in D.

Furthermore, considering the fact that decoding models with below chance performance are not interpretable under any circumstances, those spatial fractions that are not effective in decoding should be filtered out from the joint activation patterns. Therefore, we merely use the weights of classifiers in D to compute activation patterns of MTEN. The activation patterns associated to unrelated tasks are set to zero when constructing the full spatio-temporal activation pattern A. To compute the activation pattern of each member of DRT set A^{i^*} ($i^* \in D$), we adopt Eq.2 as follows:

$$A^{i^*} \propto \Sigma_{X^{i^*}} W^{i^*} \tag{4}$$

3 Experiments

3.1 Material and Experimental Setup

We tested the proposed method on the first 5 subjects of an MEG dataset where visual stimuli consisting of famous faces, unfamiliar faces and scrambled faces are presented to subjects. The original dataset consists of 16 subjects and it is described in [15][1]. This dataset is also used for DecMeg2014 competition[2]. Same as [19], we created a balanced face vs. scramble dataset by drawing at random from the trials of famous and unfamiliar faces in equal number to that scrambled faces. The raw data is high-pass filtered at 1Hz, down-sampled to 250Hz, and epoched from 200ms before the stimulus onset to 800 ms after the stimulus. Thus each trial has 250 time-points for each of the 306 MEG sensors (102 magnetometers and 204 planar gradiometers)[3].

To illustrate the advantage of MTEN in improving the interpretability of brain decoding model, we conduct three different experiments. These three settings help us to examine the impact of division of data into spatial fractions and employing the MTL paradigm separately:

1. We first pool all temporal data of 306 MEG sensors into one vector (i.e., we have 250*306=76500 features for each sample) and then we use the linear regression with elastic-net regularization to solve the brain decoding problem (we refer to this experiment as EN).
2. We divide the data into spatial fractions, then we employ the linear regression with elastic-net regularization to train a model on each spatial fraction separately (we refer to this experiment as STEN).
3. After dividing data into spatial fractions, we use MTEN to train the decoding model (we refer to this experiment as MTEN).

For selecting DRT members in the second and third experiments , the threshold ϕ (see 2.3) is set to $\mu_{perf} + \sigma_{perf}$, where μ_{perf} and σ_{perf} are respectively mean and standard-deviation of performances computed over all spatial fractions (tasks) on the training set. In all settings, the best values for ρ_1 and ρ_2 were decided using nested cross-validation (CV) to ensure unbiased error estimation [21]. In the inner loop of CV, a grid search on $[0, 0.001, 0.01, 0.1, 1, 10, 50, 100]$ is used to find optimal values for both ρ_1 and ρ_2. MALSAR [31] toolbox is used for training the models. The MATLAB code for all experiments is available at `https://github.com/smkia/MTL_Interpretation`.

[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`

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

3.2 Results and Discussions

Fig.1 compares the performance and the stability of EN, STEN, and MTEN experiments. The performance of classifiers is measured based on the area under the ROC curve (AUC). The stability is quantified by computing the pair-wise correlation between weight matrices across 10 folds of CV (see [22]). The bars and the error-bars are showing the mean and the standard deviation of AUC and correlations over 10 folds of CV.

The annotations below each group of bars are showing the result of two-sample t-test between each pair of benchmarked methods, where $-$, $*$, and $**$ are representing *not significant*, *significant* with $p-value < 0.05$, and *significant* with $p-value < 0.001$, respectively. All the results are corrected for multiple-comparison using Bonferroni correction. Excluding the second subject which shows completely different behaviour, Fig.1 highlights the following points:

1. While MTEN and EN have more or less the same performance, MTEN provides significantly better stability than EN.
2. STEN and MTEN provide more stability than EN, supporting the idea that dividing the data into spatial fractions improves the stability of models by reducing the degree of freedom of solution space.
3. Despite their similar stability, MTEN provides better performance than STEN illustrating the advantage of learning all tasks simultaneously in MTL framework.

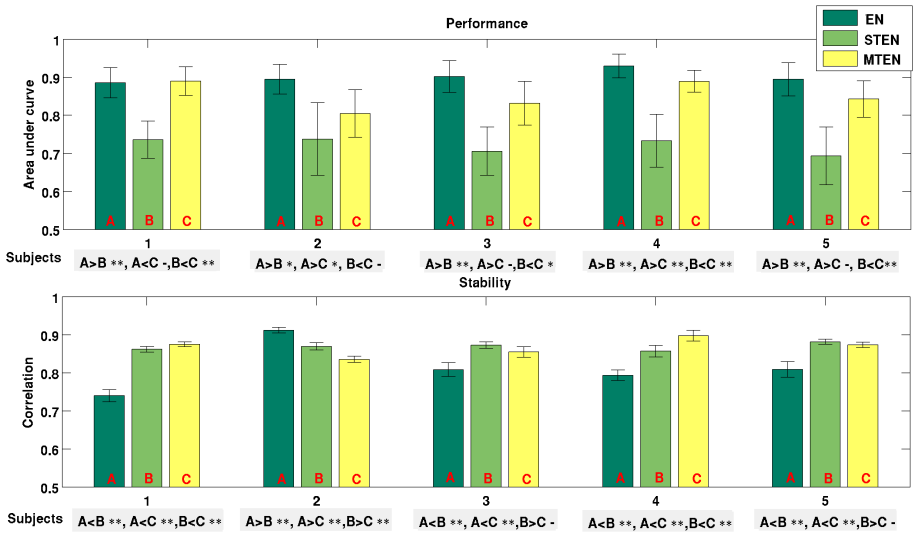

Fig. 1: Comparison between performance (upper diagram) and stability (lower diagram) of EN, STEN, and MTEN for 5 subjects.

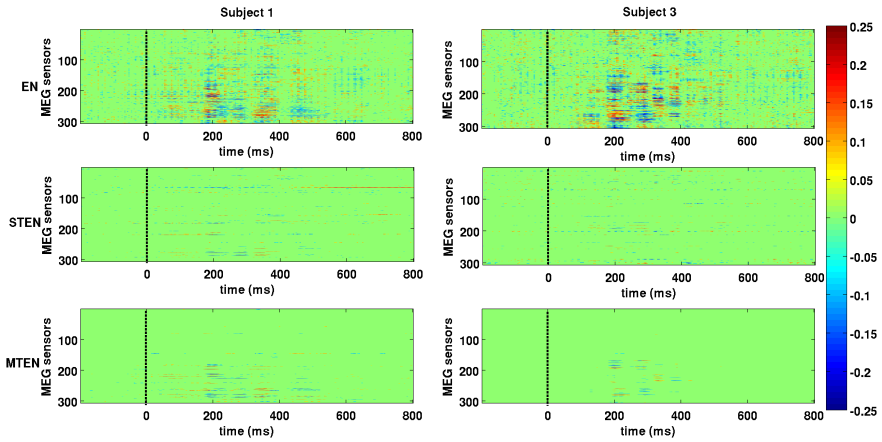

Fig. 2: Spatio-temporal activation patterns of 2 subjects computed by EN, STEN, and MTEN.

Fig. 2 elaborates more the advantage of MTL paradigm in improving the interpretability of results. This figure shows mean activation patterns of MTEN, STEN, and EN over 10 folds of CV for two subjects (other subjects show similar behaviour). These activation patterns are computed using Eq.2 in EN case, and using Eq.4 in STEN and MTEN cases. The horizontal and vertical axes represent time and sensors dimensions respectively, and the dashed line shows the stimulus onset. Comparison between these activation patterns suggests:

1. MTEN and STEN provide more sparse solution than EN.
2. Activation patterns of MTEN show no stimulus related activity before stimulus onset, in contrast to EN. Considering the experiment design used for data acquisition (see 3.1), any discriminating activity before stimulus onset has no scientific interpretation. These activations before stimulus in EN case can be consequence of overfitting of the model to noise.
3. Pre-stimulus activation in EN case rises the question that the transformation proposed by [13] might not guarantee the interpretability of brain decoding models, and the regularization strategy beside learning algorithm are still playing important roles.

4 Conclusion

In this paper, we introduced a new application of MTL to enhance the interpretability of brain decoding models. Our results on an MEG dataset show that recasting the brain decoding problem into the MTL framework is an effective technique to achieve more stable and consequently more interpretable models. These characteristics of the proposed method makes it more appropriate for making inference in cognitive neuroscience studies. Replacing elastic-net with a new penalization method in the MTL paradigm can be considered a possible future extention for our work.

References

1. Afshin-Pour, B., Soltanian-Zadeh, H., Hossein-Zadeh, G.A., Grady, C.L., Strother, S.C.: A mutual information-based metric for evaluation of fmri data-processing approaches. Human brain mapping 32(5), 699–715 (2011)
2. Ben-David, S., Gehrke, J., Schuller, R.: A theoretical framework for learning from a pool of disparate data sources. In: international conference on Knowledge discovery and data mining. pp. 443–449. ACM (2002)
3. de Brecht, M., Yamagishi, N.: Combining sparseness and smoothness improves classification accuracy and interpretability. NeuroImage 60(2), 1550–1561 (2012)
4. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44(1), 112–122 (2009)
5. Chen, X., Kim, S., Lin, Q., Carbonell, J.G., Xing, E.P.: Graph-structured multi-task regression and an efficient optimization method for general fused lasso. arXiv preprint arXiv:1005.3579 (2010)
6. Conroy, B.R., Walz, J.M., Sajda, P.: Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fmri decoding models. PloS one 8(11), e79271 (2013)
7. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fmri)brain reading: detecting and classifying distributed patterns of fmri activity in human visual cortex. Neuroimage 19(2), 261–270 (2003)
8. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117. ACM (2004)
9. van Gerven, M., Hesse, C., Jensen, O., Heskes, T.: Interpreting single trial data using groupwise regularisation. NeuroImage 46(3), 665–676 (2009)
10. Gramfort, A., Thirion, B., Varoquaux, G.: Identifying predictive regions from fmri with tv-l1 prior. In: Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on. pp. 17–20. IEEE (2013)
11. Groppe, D.M., Urbach, T.P., Kutas, M.: Mass univariate analysis of event-related brain potentials/fields i: A critical tutorial review. Psychophysiology 48(12), 1711–1725 (2011)
12. Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E.: Interpretable whole-brain prediction analysis with graphnet. NeuroImage 72, 304–321 (2013)
13. Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage (2013)
14. Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.D., Blankertz, B., Biessmann, F.: Parameter interpretation, regularization and source localization in multivariate linear models. In: Pattern Recognition in Neuroimaging, 2014 International Workshop on. pp. 1–4. IEEE (2014)
15. Henson, R.N., Wakeman, D.G., Litvak, V., Friston, K.J.: A Parametric Empirical Bayesian framework for the EEG/MEG inverse problem: generative models for multisubject and multimodal integration. Frontiers in Human Neuroscience (76)
16. Kia, S.M.: Mass-Univariate Hypothesis Testing on MEEG Data using Cross-Validation. Master's thesis, University of Trento (2013)
17. Maris, E.: Statistical testing in electrophysiological studies. Psychophysiology 49(4), 549–565 (2012)

18. Maris, E., Oostenveld, R.: Nonparametric statistical testing of eeg-and meg-data. Journal of neuroscience methods 164(1), 177–190 (2007)
19. Olivetti, E., Kia, S.M., Avesani, P.: Meg decoding across subjects. In: Pattern Recognition in Neuroimaging, 2014 International Workshop on (2014)
20. Olivetti, E., Kia, S.M., Avesani, P.: Sensor-level maps with the kernel two-sample test. In: Pattern Recognition in Neuroimaging, 2014 International Workshop on. pp. 1–4. IEEE (2014)
21. Olivetti, E., Mognon, A., Greiner, S., Avesani, P.: Brain decoding: biases in error estimation. In: Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010 First Workshop on. pp. 40–43. IEEE (2010)
22. Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C.: Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45(6), 2085–2100 (2012)
23. Rish, I., Cecchi, G.A., Lozano, A., Niculescu-Mizil, A.: Practical Applications of Sparse Modeling. MIT Press (2014)
24. Strother, S.C., Rasmussen, P.M., Churchill, N.W., Hansen, K.: Stability and Reproducibility in fMRI Analysis. New York: Springer-Verlag (2014)
25. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(1), 91–108 (2005)
26. Valverde-Albacete, F.J., Peláez-Moreno, C.: 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. PLOS ONE 9(1), e84217 (2014)
27. Varoquaux, G., Gramfort, A., Thirion, B.: Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. arXiv preprint arXiv:1206.6447 (2012)
28. Xing, E.P., Kolar, M., Kim, S., Chen, X.: High-dimensional sparse structured input-output models, with applications to gwas. Practical Applications of Sparse Modeling p. 37 (2014)
29. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)
30. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via structured multi-task sparse learning. International journal of computer vision 101(2), 367–383 (2013)
31. Zhou, J., Chen, J., Ye, J.: MALSAR: Multi-task Learning via Structural Regularization. Arizona State University (2011), `http://www.public.asu.edu/~jye02/Software/MALSAR`
32. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67(2), 301–320 (2005)