Localizing and comparing weight maps generated from linear kernel machine learning models

J. Schrouff*, J. Cremers*, G. Garraux*, L. Baldassarre†, J. Mourão-Miranda‡ and C. Phillips*
*Cyclotron Research Centre, University of Liège, Belgium
†Ecole Polytechnique Federale of Lausanne, Lausanne, Switzerland
‡Computer Science Department, University College London, United Kingdom

***Abstract*—Recently, machine learning models have been applied to neuroimaging data, allowing to make predictions about a variable of interest based on the pattern of activation or anatomy over a set of voxels. These pattern recognition based methods present undeniable assets over classical (univariate) techniques, by providing predictions for unseen data, as well as the weights of each voxel in the model. However, the obtained weight map cannot be thresholded to perform regionally specific inference, leading to a difficult localization of the variable of interest. In this work, we provide local averages of the weights according to regions defined by anatomical or functional atlases (e.g. Brodmann atlas). These averages can then be ranked, thereby providing a sorted list of regions that can be (to a certain extent) compared with univariate results. Furthermore, we defined a "ranking distance", allowing for the quantitative comparison between localized patterns. These concepts are illustrated with two datasets.**

***Index Terms*—machine learning; fMRI; pattern localization; ranking; pattern comparison**

I. Introduction

For the last few years, machine learning models have been applied to neuroimaging data [1], allowing to make predictions about a variable of interest based on the pattern of activation or anatomy over a set of voxels. In addition, they might lead to an increased sensitivity to detect the presence of a particular mental representation compared to univariate methods such as the General Linear Model (GLM, [2]). Application of these methods made it possible to decode the category of an object [3] seen by the subject . They also allowed classification of patients and healthy controls [4], [5] and could therefore be used as a diagnostic tool due to their ability to predict the class of an unseen sample.

The main disadvantage of multivariate machine learning models is that local inference with respect to the brain neuroanatomy is complex: although linear models generate weights for each voxel, the model predictions are based on the whole pattern and therefore one cannot threshold the weights to make regional statistical inferences as in univariate analysis. In the present work, we developed a methodology based on a labelled anatomical template (e.g. AAL [6] or Brodmann) to display a regionally smoothed version of the model weights and compute a ranking of the regions according to their contribution to the predictive model. Furthermore, we defined a "ranking distance", which allows the quantitative comparison of patterns in terms of their localization.

These concepts are illustrated using two datasets: (1) the mental imagery of gait in healthy controls and Parkinsonian patients [7] and (2) structural images from aged healthy controls acquired in different centres (IXI dataset[1]). Using these datasets, we aimed at investigating the following questions: *What is the overlap between the ranking obtained from the regionally averaged weights and univariate results?* and *Do regression models built from data acquired in different centres provide similar patterns in terms of their localization?*

II. Material

A. Dataset 1: Parkinson's disease

The material considered in this work is the same as in [7]. Therefore, only a brief description of the population and experimental design will be provided.

In total, 29 subjects participated in the study: 14 patients (7 males; mean age: 65.1 $\pm$ 9.5 years) diagnosed with IPD with different degrees of severity of gait disturbances and 15 controls matched for age (63.8 6 $\pm$ 8.1 years) and gender (7 males). Written informed consents for this research protocol approved by the local ethics committee were obtained from all participants.

Before fMRI, the subjects were asked to walk comfortably and then briskly on a 25m path. After gait evaluation, they were trained to mentally rehearse themselves walking on the path.

All subjects then underwent an fMRI session comprising three tasks: mental imagery of standing on the path (STAND), walking at a comfortable pace along the path (COMF) and walking briskly along the path (BRISK). The COMF and BRISK conditions were self-paced, subjects indicating when they had completed each trial by a key press, while each trial of the STAND condition was constrained by the duration of the previous COMF trial.

fMRI data preprocessing and univariate analysis were performed using SPM8[2]. Functional images were realigned and co-registered to the structural image before normalisation using DARTEL [8]. Finally, smoothing was applied using a 8mm FWHM Gaussian kernel. A General Linear Model then summarized the time series from each subject by modelling

[1]IXI - Information eXtraction from Images, funded by EPSRC GR/S21533/02, http://www.brain-development.org/

[2]www.fil.ion.ucl.ac.uk/spm

each condition by a boxcar function convoluted with a canonical haemodynamic response function. In the end, three images per subject were considered for further analysis: the parametric maps of STAND, COMF and BRISK representing the BOLD signal activity associated with each condition.

B. Dataset 2: Age regression

The IXI dataset was used to perform age regression. More specifically, as in [9], the older subjects (from 60 to 90 years old) were selected for regression based on the scalar momentum features resulting from the normalization [10], [11].

We used a subset of the data by randomly selecting 54 subjects that were acquired in centre 1, and 54 subjects acquired in centre 2. Please note that there is no significant difference between the two populations in terms of age.

III. Methods

A. Pattern discrimination

For dataset 1, classification was performed using binary SVM [12]. Model performance was then assessed in terms of the balanced and class accuracies, as well as Positive Predictive Values (PPV). The regression of dataset 2 was based on a Relevance Vector Machine (RVM) [13] and assessed in terms of mean squared error (MSE) and correlation between the targets and predictions. In both cases, a leave-one-subject-out cross-validation was performed to compute model performance, its significance being assessed by a non-parametric testing using 1000 random permutations of the training labels. All machine learning based modelling steps have been performed in PRoNTo [9][3].

B. Pattern localization

From the linear models leading to a significant classification (or regression), weight maps were built. The weights were then locally averaged based on labelled atlases of regions (here using the AAL): for each region, we computed its "normalized weight" as the average of the weights (in absolute value) of the voxels comprised in that region.

$$NW_{ROI} = \frac{\sum_{v \in ROI} |W_v|}{m_{ROI}} \qquad (1)$$

with v representing the index of a voxel in the weight image, W_v its weight and m_{ROI}, the number of voxels in region ROI.

Weights that do not correspond to any labelled region are pooled into an additional region, referred to as *others*.

Similar to a principal component analysis, the labelled regions can then be ranked according to the percentage of the total normalized weights that they explain. This results in a sorted list of regions, that can then be compared (to a certain extent) with univariate results.

The labelled regions used to localize the patterns were defined by the AAL atlas from the WFU-PickAtlas [14] toolbox in SPM. The 117 regions from this manually generated atlas are illustrated in Figure 1.

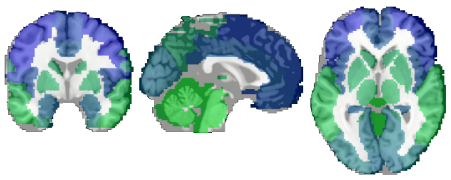

Fig. 1. **AAL atlas.** Views of the 117 labelled regions defined in the AAL atlas (in green and blue). In total, the brain has been parcelled into 117 labelled regions of interest. The additional "others" region comprises all voxels in grey.

C. Pattern comparison

The rankings obtained from different patterns could be quantitatively compared using a measure of distance, inspired from the field of web search [15]. It is computed as:

$$dr(f_1, f_2) = \frac{2}{N * (N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} I_{f_1, f_2}(i, j) \qquad (2)$$

where

$$I_{f_1,f_2}(i,j) = \begin{cases} 1 & \text{if } f_1(i) < f_1(j) \text{ and } f_2(i) > f_2(j) \\ 0 & \text{otherwise} \end{cases}$$

with $dr(f_1, f_2)$, the ranking distance between the models f_1 and f_2 and N, the number of elements in the ranking, which corresponds to the number of regions in the atlas in the present case (i.e. 117 labelled regions). The values of dr range from 0 (exactly the same rankings) to 1 (exactly opposite rankings).

Using permutations, we were able to associate a p-value to the obtained distance, thereby providing significance of dr. More specifically, the labels were permuted before training, enabling the building of "permuted" weights maps, and hence "permuted" rankings. The null hypothesis here is that the labels do not contain any information about the location of the variable of interest, such that the ranking obtained from the model using the permuted labels is random. If the ranking distance between two patterns is significantly small (or large), those two patterns are then significantly (dis-) similar in terms of their localization.

IV. Results

A. Localizing the pattern of Parkinson's Disease

Based on [7], we built the COMF versus STAND comparison for the control and patient groups separately. The different conditions were also combined to compare the two groups (see Table I). The pattern leading to the only significant discrimination between groups was then localized.

Using the ranking of the normalized weights, we can then compare the univariate results based on the same data [7] to the top 10 (arbitrarily fixed for illustration) of the regions list. This is illustrated in Figure 2 for the three considered

[3]Please note that PRoNTo does not provide SVR or RVM classification, which is the reason why the classification and the regression are not based on the same machine.

Univariate

Area	CTRL	IPD	Ctrl>IPD
Lateral premotor	Bi	-	-
Pre-SMA	Bi	R	-
Anterior cingulate	Bi	R	-
Middle frontal (DLPFC)	R	-	-
Inferior frontal	Bi	-	-
Anterior insula	Bi	-	-
Intraparietal sulcus	Bi	-	Bi
Precuneus	Bi	-	R
Parieto-occipital sulcus	Bi	-	Bi
Post hippocampus	Bi	-	L
Parahippocampal gyrus	Bi	-	-
Lingual gyrus	Bi	-	R
Caudate nucleus	R	-	-
Anterior putamen	Bi	-	-
Anterior pallidum	L	-	-
PPN/MLR area	L	-	L
Lateral pons	L	-	-
Cerebellar vermis	Midline	-	Midline
Cerebellar hemisphere	Bi	-	Bi

Multivariate

Rank	COMF vs STAND (CTRL)	COMF vs STAND (IPD)	CTRL vs IPD (COMF+BRISK)
1	Medulla	Olfactory (L)	Caudate (L)
2	Vermis 3	SupraMarginal (L)	Vermis 10
3	Cerebellum 3 (L)	Cerebellum 10	Medulla
4	Vermis 4-5	Vermis 3	Mid Frontal (L)
5	SMA (R)	Rectus (L)	Inf Frontal (R)
6	Sup Temporal (L)	SMA (R)	Rectus (L)
7	Caudate (L)	Rectus (R)	Inf Frontal (L)
8	SMA (L)	Med Frontal (R)	Cerebellum 7 (R)
9	Inf Frontal (L)	Olfactory (R)	Cerebellum 10 (L)
10	Angular (L)	Ant Cingulate (R)	Sup Frontal (R)

Fig. 2. **Univariate and multivariate results** The table on the left presents the univariate results of [7]. The table on the right presents the top 10 ranked regions according to the normalized weights. Any anatomical overlap between the top regions from the univariate and multivariate results for the control (CTRL, blue) or patient (IPD, green) group, or for their comparison (CTRL>IPD, red) is represented by a coloured frame. L (left), R (right) and Bi (left and right) correspond to the lateralization of the considered region.

TABLE I
BALANCED ACCURACY (IN %) FOR THE IPD VS. CTRL CLASSIFICATION FOR EACH COMBINATION OF THE THREE TASKS (ROWS). "ALL" REPRESENTS THE COMBINATION OF THE THREE TASKS. SIGNIFICANT CLASSIFICATION RESULTS ARE DISPLAYED IN BOLD (p-VALUE COMPUTED FROM 1000 PERMUTATIONS).

Condition	Accuracy
STAND	14.3
COMF	58.3
BRISK	59.0
STAND+COMF	36.3
STAND+BRISK	36.7
COMF+BRISK	**62.3**
All	42.9

models. We observe a good overlap between the univariate and multivariate lists of regions, although the nature of the models is different, and the involved comparisons are not exactly the same (COMF> STAND in univariate, COMF+BRISK versus STAND in multivariate). This result suggests that the three multivariate models correctly identified the pattern of neuronal activity underlying the considered discriminations (COMF vs STAND, or CTRL vs IPD) and was not based on noise in the data. This is particularly interesting for the comparison between CTRL and IPD, since the univariate results reported only a few significant areas, and the performance of the classification is quite low (although significantly above chance).

B. Comparing patterns across acquisition centres

Here, we investigated the differences in patterns between the two different centres. If the two centres were leading to similar patterns generated from the regression based on the subjects' age, the ranking distance between the two patterns should be (significantly) small. On the other hand, using a mix of data from the two centres in each regression model (half from centre 1, half from centre 2) should lead to a decrease in the ranking distance between patterns since the variability due to the centre is distributed equivalently in the two patterns.

We therefore built four regression models comprising each 54 subjects: centre1, centre2, mix1 (first half of the 54 subjects from each centre: 27 + 27) and mix2 (second half of the 54 subjects from each centre). The age regression led to significant results (1000 permutations), with correlations (MSE) of 0.45, 0.63, 0.43 and 0.65 (21.16, 20.77, 26.90, 16.19), respectively for each model.

Anatomically labelled (AAL atlas) regions were then ranked for each model, allowing the computation of the ranking distance between the two centres, as well as for the two mixes. As illustrated on Figure 3, the null hypothesis could not be rejected for the comparison between the two centres ($dr = 0.3495$, $p = 0.8518$) nor for the comparison between mixes ($dr = 0.2741$, $p = 0.0686$). However, there is a clear decrease in the ranking distance when mixing the two centres (result close to significance), suggesting that the patterns are more similar in terms of localization than when considering different centres independently[4].

V. DISCUSSION

In this work, we proposed a methodology to ease the interpretation of multivariate patterns obtained from machine learning based modelling. Although the local average of the weights is a simple approach, it allows to rank the regions in terms of their contribution to the model. To validate the proposed methodology, the top ranked regions were qualitatively

[4]This result was replicated by using other mixes of the subjects: subjects with even identifiers from centre 1 and uneven identifiers from centre 2, and conversely for the other mix. The ranking distance was $dr = 0.2840$, which is also smaller than when comparing the two centres.

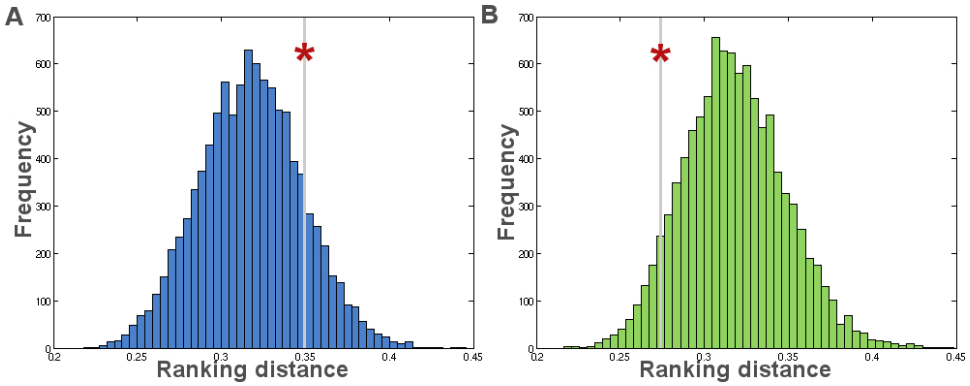

Fig. 3. **Ranking distance A** between centres, **B** when mixing the data from the two centres. The null distributions of the ranking distance are represented by histograms (in blue and green, respectively). The distance between patterns when considering the original labels are represented by red stars. In both cases, the null hypothesis cannot be rejected ($p > 0.05$).

compared to previous univariate results. The ranking of the regions might lead to an easier interpretation of the weights, as well as a way of checking that the model correctly identified the variable of interest to perform the classification/regression. The developed methodology could be applied to any map: the results of sparse models in the voxel space, of multi-kernel machines, of accuracy map from the searchlight approach [16] or using the "source" pattern instead of the weight map [17].

Furthermore, models are usually compared in terms of performance (accuracy, area under the ROC curve, ...) or in terms of model evidence, i.e. the trade-off between model fitting and model complexity. In this work, we presented an approach to compare patterns in terms of their localization. This could be particularly useful to answer neuroscientific and/or clinical questions such as: *Does these two centres lead to the same models in terms of pattern localization*? The results from the IXI dataset showed that training on a mix of images from different centres led to more similar patterns than when training on data acquired in different centres. This result therefore favours the use of multi-centric data when building classifiers aimed for computer-aided diagnosis tools, for example.

VI. Conclusion

The present work investigated pattern localization, as well as quantitative pattern comparison using two datasets[5]. The preliminary results show that the developed methodology seems promising, although more work is needed.

References

[1] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fmri: a tutorial overview." *Neuroimage*, vol. 45, pp. S199–S209, 2009.

[2] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, *Statistical Parametric Mapping: the analysis of functional brain images*, K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds. London: Elsevier Academic Press, 2007.

[5]This work will be distributed in PRoNTo (Pattern Recognition for Neuroimaging Toolbox), a user-friendly toolbox, making machine learning models available to every neuroscientist [9].

[3] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage*, vol. 19, pp. 261–270, 2003.

[4] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease." *Brain*, vol. 131, pp. 681–689, 2008.

[5] C. Phillips, M.-A. Bruno, P. Maquet, M. Boly, Q. Noirhomme, C. Schnakers, A. Vanhaudenhuyse, M. Bonjean, R. Hustinx, G. Moonen, A. Luxen, and S. Laureys, "'Relevance vector machine' consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients," *Neuroimage*, vol. 56, pp. 797–808, 2011.

[6] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, pp. 273–289, 2002.

[7] J. Cremers, K. D'Ostilio, J. Stamatakis, V. Delvaux, and G. Garraux, "Brain activation pattern related to gait disturbances in Parkinson's disease," *Mov. Disord.*, vol. 27, pp. 1498–1505, 2012.

[8] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, pp. 95 – 113, 2007.

[9] J. Schrouff, M. J. Rosa, J. Rondina, A. Marquand, C. Chu, J. Ashburner, C. Phillips, J. Richiardi, and J. Mourão-Miranda, "PRoNTo: Pattern Recognition for Neuroimaging Toolbox," *Neuroinformatics*, pp. 1–19, 2013.

[10] J. Ashburner and K. Friston, "Computing average shaped tissue probability templates," *NeuroImage*, vol. 45, pp. 333–341, 2009.

[11] ——, "Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation," *NeuroImage*, vol. 55, pp. 954–967, 2011.

[12] C. Burges, "A tutorial on support vector machines for pattern recognition." *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[13] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[14] J. A. Maldjian, P. J. Laurienti, J. B. Burdette, and R. A. Kraft, "An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets." *NeuroImage*, vol. 19, pp. 1233–1239, 2003.

[15] R. Lempel and S. Moran, "Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs," *Information Retrieval*, vol. 8, pp. 245–264, 2005.

[16] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *PNAS*, vol. 103, pp. 3863–3868, 2006.

[17] F. Biessmann, S. Dähne, F. C. Meinecke, B. Blankertz, K. Görgen, K.-R. Müller, and S. Haufe, "On the interpretability of linear multivariate neuroimaging analyses: Filters, patterns and their relationship," in *2nd NIPS Workshop on Machine Learning and Interpretation in NeuroImaging*, 2012.