Use of Pattern-Information Analysis in Vision Science: A Pragmatic Examination

Mathieu J. Ruiz[1,2], Jean-Michel Hupé[2], and Michel Dojat[1]

[1] GIN - INSERM U836 & Université J. Fourier, 38700 La Tronche, France
{mathieu.ruiz, michel.dojat}@ujf-grenoble.fr
[2] CerCo - CNRS UMR 5549 & Université de Toulouse, 31300 Toulouse, France
{jean-michel.hupe}@cerco.ups-tlse.fr

Abstract. MultiVoxel Pattern Analysis (MVPA) is presented as a successful alternative to the General Linear Model (GLM) for fMRI data analysis. We report different experiments using MVPA to master several key parameters. We found that 1) different feature selections provide similar classification accuracies with different interpretation depending on the underlying hypotheses, 2) paradigms should be created to maximize both Signal to Noise Ratio (SNR) and number of examples and 3) smoothing leads to opposite effects on classification depending on the spatial scale at which information is encoded and should be used with extreme caution.

Keywords: Machine Learning, SVM, Neurosciences, Brain.

1 Introduction

MVPA is presented as a successful alternative to GLM. It outperforms GLM for detecting low SNR brain activation [1], [2], [3] while yielding similar results on data correctly analyzed by GLM [3], [4]. Thanks to methodologists´ efforts, various toolboxes are today freely available, offering a simple, friendly user interface to novices for processing with MVPA. In parallel, some brain researches give the naïve impression that it is simple to obtain outstanding results using MVPA. In this paper, we report different experiments we performed using MVPA to better understand the effects of several key parameters on our data analysis and consequently our final interpretation. We focus on three methodological points that, if better mastered, will help neuroscientists make the right choices for a proper use of MVPA:

- **Feature selection** is of paramount importance to avoid the curse of dimensionality: the use of too many features compared to the set of available examples leads to a bad hyperplane computation and then hampers class separation. In many studies, features are selected in Regions Of Interest (ROIs) defined in independent experiments. Most of the time, a second feature selection inside those ROIs is required: for instance voxels are selected based on T values obtained by a GLM [1], [2], [5]. This leads to an additional (implicit) hypothesis about the data and prevents generalization to the whole ROI. We propose an alternative method.

F. Wang et al. (Eds.): MLMI 2012, LNCS 7588, pp. 103–110, 2012.

- **The example set size** is a problem in fMRI experiments because few examples (typically $\leq$100) can be acquired compared to the number of voxels per acquisition (typically $\geq 10^5$). Moreover, Rapid Event-Related Paradigms (RERP), i.e. an Inter Stimulus Interval (ISI) $\leq$ Hemodynamic Response Function (HRF) peak (about 6s), provide less examples for classification compared to Block Paradigms (BP). From the psychophysics point of view RERP should nevertheless be preferred because they are less cognitively biased than BP. RERP prevent subjects from habituating to the stimuli, mind wandering or being less attentive. We explored methods to increase the number of examples provided by RERP, either by discretizing parameter estimates or by performing supplementary sessions.
- **Spatial smoothing** of the data is customary in GLM analyses in order to increase SNR and reinforce the normality of the voxels distribution. Oppositely since MVPA relies on information stored in activation patterns, smoothing could blur subtle differences between patterns and may better be avoided. We explored the effects of the kernel size of spatial smoothing on the classification accuracy.

2 Materials and Methods

Subjects

Data were collected on two subjects (MR and JMH, authors) during two different sessions each. MR performed twice the same orientation session (S1-O1 and S1-O2). JMH performed an orientation session (S2-O) and a color session (S2-C).

Protocol

Orientation and color sessions contained 4 stimulation conditions, with respectively oriented bars or color concentric sinusoidal gratings visual stimuli, with an additional fixation-only condition for the HRF return to the baseline. Similarly to [6], each condition lasted 1.5s, appeared 6 times and was randomly distributed. ISI was randomly set to 3, 4.5 or 6s.

o *Oriented bars*: Stimuli were black and white sinusoidal gratings taking one out of 4 possible orientations (0, pi/2, pi/4 and 3pi/4 rad.) and displayed within a circular 4.5deg. radius aperture. They were projected on a neutral gray background (CIE xyY=0.32, 0.41, 406). At the center of the screen, a dark gray dot (0.45deg. radius) had to be fixated for the whole duration of the experiment.

o *Color concentric sinusoidal gratings:* Stimuli were similar to [6]. They subtended a maximum 5deg. radius with a square wave luminance profile alternating (1.2 cycles/deg.) between one out of four colors (red, green, blue and yellow) and isoluminant gray. Colors were approximately equalized in luminance using a classical flicker experiment.

o *Attentional task*: Each subject performed a one-back task to maintain his attentional workload constant throughout the experiment. Letters (A, T, O, M and X) appeared one at a time over the central dot in a pseudorandom order. Letters were 1deg. wide, lasted 1s with an ISI of 300ms. The subject had to press a button when the same letter appeared twice in a row (ten times in a run).

MRI Acquisition

We acquired structural and EPI functional images on a Philips 3T Achieva whole body scanner using an 8-channel head coil. For structural image we used a T1-3D FFE sequence (TR/TE: 25/15ms, flip angle: 15deg., acquisition matrix: 180x240x256 (x, y, z), BW=191Hz/pixel, and total measurement time=9min40s). Functional images were acquired using a gradient-echo EPI sequence (TR/TE: 2500/30ms, flip angle: 80deg., acquisition matrix: 80x80, 38 adjacent slices, spatial resolution 2.75x2.75x3mm^3). For each session, 12 functional runs were acquired containing 78 TRs. We used an eye-tracker (ASL 6000) for eye movements recording.

MR Data Processing

Standard preprocessing steps were applied to fMRI data using SPM8[1]: slice timing correction and realignment. We extracted Rigid-Head Motion Parameters (RHMPs). Anatomical T1 images were coregistered to the mean functional image. To address our methodological questions, a smoothing step could be added. We either used a 4mm, 6mm or 8mm 3D FWHM Gaussian filter. Parameter Estimate Images (PEIs, also called images of beta weights) were computed by fitting a boxcar function convolved with HRF. A GLM was created based on 12 predictors (5 conditions, subjects blinks and 6 RHMPs). PEIs discretization was modeled with 9 predictors. The first predictor was a single event; the second predictor contained the 29 other events; 1 predictor contained subjects blinks and 6 predictors were RHMPs. There were as many GLMs as there were stimulation conditions during the whole sessions (i.e. 24 x 12 runs). In order to decrease the noise in the data, PEIs were divided by the square root of the GLM residuals.

Classification

o *Classifiers*: We performed our classification analyses using a Support Vector Classification (SVC, linear kernel, C parameter default value=1) provided by the python module scikit-learn (v.0.10)[2] [7]. We also used a Recursive Feature Elimination (RFE) method that performs a feature selection through iterative classification [8]. The first classification uses all the features of the ROI; at each iteration the worst features (here 1% of the features with the lowest weights) are eliminated, until reaching the chosen final feature set size (set to 288 voxels, RFE performed around 90 iterations).

o *Examples*: PEIs were used as examples for the classifiers. The non-discretized models produced 12 PEIs per condition (or 48 examples in total). The discretized models produced 72 PEIs per condition (or 288 examples in total).

o *Cross-Validation*: We used a leave-one-run-out cross-validation procedure: a classifier was trained on 11 runs (4x11 examples) and tested on the remaining one (4 examples). The accuracy score was the average accuracy of the 12 possible classifiers. For the PEI discretization method we also computed the mean

[1] `http://www.fil.ion.ucl.ac.uk/spm/`
[2] `http://scikit-learn.org/stable/`

accuracy over 12 classifiers, but this time each classifier was trained over 264 examples (24x11) and tested on the 24 examples of the remaining run (examples within a run may not be fully independent so they cannot be separated to be used for training or testing).

o *Feature Selection*: We constrained our analyses inside a ROI that contained most of the visual areas with 3424 and 3313 voxels for subjects MR and JMH respectively (Fig. 1, Left). We performed a second feature selection inside those ROIs by selecting 288 voxels. We chose this number in order to reach optimal conditions when performing the classification with discretized PEIs, by having as many features as examples. Similar amounts of voxels were also used in the literature as a second feature selection [1], [2]. The second feature selection could be based on 1) the best SNR as assessed by T values of an activation contrast (all active conditions against the rest), 2) the best GLM model fitting the data as assessed by F values of all active condition pair contrasts, 3) their importance in the classification as computed by RFE, using the voxels with the best weights and 4) the bootstrap random selection of 288 voxels in the ROI (10 000 iterations).

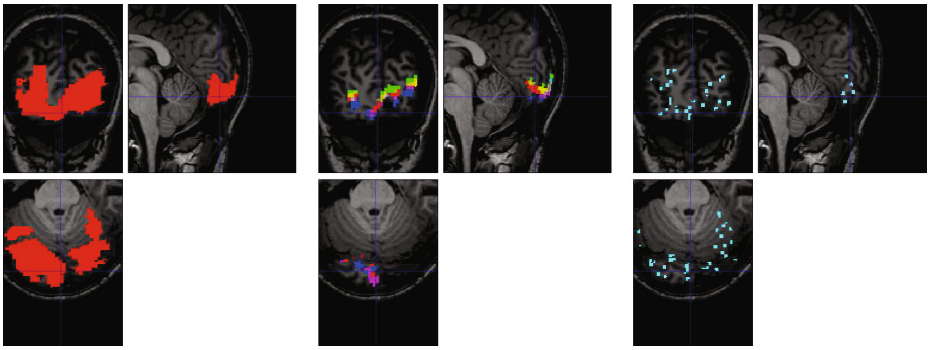

Fig. 1. Left: ROI for subject MR. Middle: voxels selected (n=288) by the best T, RFE and best F methods (red, green and blue respectively). Right: Randomly selected voxels (n=288) that gave the best performance.

3 Results

Feature Selection

In our experiments, chance level and the binomial law significance level (p=0.05) were respectively 25% (four conditions) and 36%. For orientation (S1-O1, S1-O2 and S2-O), information was robust enough (distributed and redundant) in the considered ROI, to obtain significant classification accuracy (Confidence Interval (CI) minimum=43% in average) with 288 voxels randomly selected with bootstrap as well as by using all the voxels of the ROI (Fig. 2A). All methods (but best F for S2-O) yielded classification accuracy higher than the upper

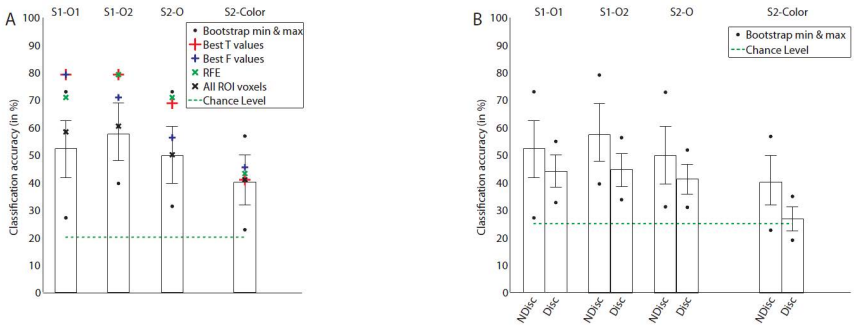

Fig. 2. A. Classification accuracy for all methods used with a 4mm kernel smoothing. B. Classification accuracy for non-discretized and discretized PEIs with a 4mm kernel smoothing. Bars denote the classification mean accuracy for 10 000 iterations with random selection of 288 voxels in the ROI. Error bars denote a 95% CI.

limit of the 95% CI of the bootstrap. Their accuracy, but S1-01, was lower or equal to the maximum accuracy provided by bootstrap method. This indicates that random selection can outperform *a priori* selection criteria. This raises the question of the population of voxels selected by each method. Clearly, (Fig. 1 Middle, Right) indicates differences both in density and localization. With bootstrap method, the best classification was obtained with highly sparse voxels. The overlap was respectively of 48%, 44% and 8% between best T/RFE, best T/best F and bootstrap/other methods (on average).

To assert the significance of these results, we shuffled the set of labels to generate incoherent patterns of information (data not shown). We obtained classification accuracy at the chance level for all methods. Note that for bootstrap method, the binomial law significance level, Bonferroni-corrected for multiple comparisons, was 54% (10000 iterations). With randomized labels, the maximum classification accuracy obtained was below (e.g. 48% for S1-O2). Compared to other methods, bootstrap 1) gives the global performance for the ROI, 2) provided in our case similar results and 3) allows for robust statistics thanks to larger samples. Only this method is considered in the following sections.

Example Set Size

We found a significant decrease (paired T-test, $p \leq 0.0001$, mean 11%) in accuracy for discretized PEIs (Fig. 2B).

We found a significant increase (paired T-test, ≤ 0.0001, 6% to the best single session) in accuracy for the combined session (Fig. 3A).

Smoothing

We found (Fig. 3B) a significant main effect of smoothing (paired t-test: $p \leq 0.0001$). Post-hoc analysis showed significant differences between all conditions (Tukey-HSD test $p \leq 0.0001$ all kernel pairs, but S2-O, 4mm vs. 8mm = $p \leq 0.01$,

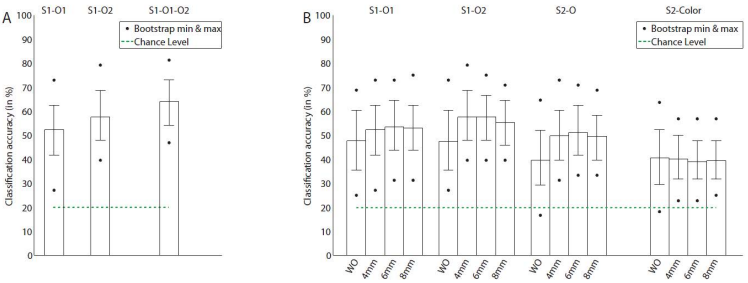

Fig. 3. A. Classification accuracy for two sessions and their combination with a 4mm kernel smoothing. B. Classification accuracy for different smoothing kernel sizes. WO: without smoothing. See Figure 2 for legend.

and S1-O2 4mm vs. 6mm not significant). For orientation sessions, 4mm kernel smoothing systematically provided a significant increase in classification accuracy compared to no blurring, while a small decrease in accuracy (0.5%) was reported when using 4mm compared to no smoothing for color session.

4 Discussion

We investigated the impact on classification accuracy of 1) feature selection 2) increasing example set size either by PEI discretization or additional sessions and 3) kernel smoothing size.

Feature Selection

MVPA requires selecting features (in our case voxels) on which classification will be performed. No feature selection (i.e. using all voxels in the brain) led to poor results (25, 23, 27 and 30% for S1-O1, S1-O2, S2-O and S2-C respectively), not significantly different from chance with the binomial law. Features can be selected through independent experiments such as retinotopic mapping but a second feature selection step can be required. It often relies on a univariate selection, like voxels with the best T values of an activation contrast [2], [9]. Such selection has been criticized because voxels with low T values can be relevant for classification [10]. It requires hypothesis about which voxels are relevant to the classification. Here we compared a bootstrap classification method with two univariate- (GLM) and a multivariate- (RFE) based feature selection in a ROI. We found that the best classification accuracy obtained with bootstrap was similar, or even higher, to accuracy obtained with the best T, F and RFE methods. However, the voxel populations selected by each method were not the same and their repartition was different. Each selected population was relevant (classification obtained by chance led to bad results) and correctly localized in grey matter (see for instance Fig. 1, Right). In consequence, feature selection method may not influence accuracy but importantly can favor a distributed or a localized interpretation of the results.

Example Set Size

Classifiers need enough examples to perform accurately. When using RERP, two solutions were proposed to increase the amount of examples: compute many estimates with low SNR [11], or perform multiple sessions [6]. Mumford et al. [11] suggested to estimate each event in a separate GLM: the first regressor is the onset of a single event; the second regressor contains onsets of all other events. They found that discretized PEIs yielded better classification accuracies than other methods when short ISI were used. At higher ISI, similar classification accuracies were found. Our results showed significant decreases in accuracy (up to 13.5%) with discretized PEIs. This might be due to a lower SNR of the PEIs because fewer events were used compared to [11]. Consequently, a large amount of examples is not enough to increase classification accuracy. In [6], 5 subjects performed 3 to 5 sessions composed of 8 to 10 event-related runs. Each type of stimulus was presented 8 times in a run, which provided them many examples while keeping well estimated PEIs. Classification accuracy was nevertheless highly different between subjects (40-70% in retinotopic V1; chance=12.5%). When using two sessions, we found only a small significant increase in accuracy (6%). First, accuracy was already high in a single session and we might have reached an upper limit. This limit could depend on information encoding (e.g. color session had poorer classification accuracy than orientation session). Second, sessions were realigned and then images interpolated which might have decreased SNR. Note that performing multiple sessions is not always possible with untrained volunteers. Based on these results, we suggest to perform many short runs in the same session, as proposed in [12] using a BP. It remains to be shown whether RERP also benefit from this method.

Smoothing

Several articles investigated smoothing effect using BP or slow ERP with various conclusions [5], [13], [14]. [13] found that to some extent high smoothing tended to provide better SVM (i.e. with fewer support vectors). Using PEIs [14] found that high smoothing led to a significant decrease in accuracy (5%). [5] reported a significant 1-1.5% increase in accuracy with high smoothing for emotional prosody classification. We tested 3 smoothing kernel sizes and found a significant increase in classification accuracy (up to 10%) for the orientation sessions with smoothing. [15] showed that classification of oriented bars came more likely from a radial bias in the periphery of V1 (bias sampling by averaging within voxel boundaries [1], [16]). The radial bias is expressed at a large scale so smoothing probably increased SNR and exacerbated the pattern. Classification accuracy of the color session was hampered by smoothing maybe because information is encoded at a fine-grained level. This would be in line with the idea that improvement due to smoothing reflects the spatial scale at which the information is encoded [5]. We therefore suggest to systematically compare classification accuracies with and without smoothing. It might increase classification accuracy and indirectly reveal how information is encoded.

Acknowledgments. This work is supported by Agence Nationale de Recherche ANR-11-BSH2-010. M. Ruiz is recipient of a PhD grant from MENRT (France).

References

1. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8(5), 679–685 (2005)
2. Gerardin, P., Kourtzi, Z., Mamassian, P.: Prior knowledge of illumination for 3D perception in the human brain. Proc. Natl. Acad. Sci. U.S.A. 107(37), 16309–16314 (2010)
3. Reddy, L., Tsuchiya, N., Serre, T.: Reading the mind's eye: decoding category information during mental imagery. Neuroimage 50(2), 818–825 (2010)
4. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293(5539), 2425–2430 (2001)
5. Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P.: Decoding of emotional information in voice-sensitive cortices. Curr. Biol. 19(12), 1028–1033 (2009)
6. Brouwer, G.J., Heeger, D.J.: Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. 29(44), 13992–14003 (2009)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. 46(1-3), 389–422 (2002)
9. De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E.: Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage 43(1), 44–58 (2008)
10. Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10(9), 424–430 (2006)
11. Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A.: Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59(3), 2636–2643 (2012)
12. Coutanche, M.N., Thompson-Schill, S.L.: The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. Neuroimage 61(4), 1113–1119 (2012)
13. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X.: Support vector machines for temporal classification of block design fMRI data. Neuroimage 26(2), 317–329 (2005)
14. Etzel, J.A., Valchev, N., Keysers, C.: The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. Neuroimage 54(2), 1159–1167 (2011)
15. Freeman, J., Brouwer, G.J., Heeger, D.J., Merriam, E.P.: Orientation decoding depends on maps, not columns. J. Neurosci. 31(13), 4792–4804 (2011)
16. Kriegeskorte, N., Cusack, R., Bandettini, P.: How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage 49(3), 1965–1976 (2010)