Cross-modal searchlight classification:

Methodological challenges and recommended solutions

Samuel A. Nastase, Yaroslav O. Halchenko
Department of Psychological and Brain Sciences
Dartmouth College
Hanover, NH, USA
samuel.a.nastase.gr@dartmouth.edu

Ben Davis, Uri Hasson
Center for Mind/Brain Sciences (CIMeC)
University of Trento
Rovereto, Italy
uri.hasson@unitn.it

***Abstract*—Multivariate cross-classification is a powerful tool for decoding abstract or supramodal representations from distributed neural populations. However, this approach introduces several methodological challenges not encountered in typical multivariate pattern analysis and information-based brain mapping. In the current report, we review these challenges, recommend solutions, and evaluate alternative approaches where possible. We address these challenges with reference to an example fMRI data set where participants were presented with brief series of auditory and visual stimuli of varying predictability with the aim of decoding predictability across auditory and visual modalities. In analyzing this data set, we highlight four particular challenges: response normalization, cross-validation, direction of cross-validation, and permutation testing.**

Keywords—cross-classification; cross-modal; fMRI; MVPA

I. Introduction

Multivariate cross-classification has become an increasingly prevalent tool for decoding abstract, cross-modal, or supramodal neural representations [1]. Consider an experimental design with two fully-crossed factors, each with two levels: Factor A (e.g., predictability: high, low) and Factor B (e.g., modality: auditory, visual). In the cross-classification framework, a classifier is trained to discriminate between the levels of, e.g., Factor A (high vs. low predictability) based on data from only one level of Factor B (e.g., the auditory modality), then tested on data from the left-out level of Factor B (the visual modality). Successful cross-classification indicates that the patterns of activation in a given brain region encode information about Factor A that generalizes across Factor B. Note that this approach is a specific case of more generally cross-validating across samples to ensure that information about condition assignments generalizes across, e.g., particular stimuli [2].

This more complex design introduces several analytic challenges not typically encountered in conventional classification analyses, and not explicitly addressed in prior methodological reports [2], [3]. In the following, we outline these challenges in the context of an illustrative example data set where participants were presented with brief series of auditory and visual stimuli of varying levels of temporal predictability. We show in what ways different analytic choices implicitly test different questions, and bring attention to the varying options and their merits to better inform future studies.

II. fMRI Data

Twenty-six participants were presented with brief series of auditory, visual, or audiovisual stimuli of varying predictability while performing a cover task. Predictability was manipulated using four different levels of Markov entropy quantifying the probabilistic transition constraints among elements within a series. This amounts to a 3 (modality) x 4 (entropy) factorial design with 12 conditions total. We focus on classifying the highest and lowest entropy levels of the auditory and visual series. Auditory series consisted of a repeated sampling of four tone tokens, while visual series consisted of four shape tokens presented at four locations surrounding a central fixation cross, with stimuli presented at 3.3 Hz for ~11 s. There were four runs, each containing three instances of each condition. Functional data were preprocessed in FSL [4], and a conventional univariate GLM was then used to estimate responses at each voxel for each experimental condition. Ultimately, this resulted in four samples (one for each run) per each of the eight conditions of interest (four levels of entropy in the auditory and visual modalities). Further analyses were confined to a consensus gray matter mask comprising the union of individual gray matter masks across participants. The following multivariate analyses were performed using the PyMVPA software package [5]. Cross-classification was performed using linear support vector machines (SVMs) and leave-one-participant-out cross-validation (Fig. 1) within spherical searchlights constrained to the union gray matter mask. Each searchlight had a 3-voxel radius (6 mm), and on average included 107 voxels (SD = 21 voxels).

III. Methodological Challenges

A. Response Normalization

Prior to multivariate analysis, response patterns are typically normalized (i.e., z-scored) in one of two ways: *a*) normalized across samples (i.e., conditions) per feature, which alters response patterns but captures relative differences in voxel responses across conditions; or *b*) normalized across features (i.e., voxels) per sample, which alters voxel responses across conditions but preserves the relative shape of the pattern per condition [2]. Normalizing across features ensures that a classifier cannot capitalize on regional-average differences in activation (e.g., within a searchlight) across conditions, while normalizing across samples does not, and emphasizes voxelwise differences in activation across conditions without necessarily

978-1-4673-6530-7/16/$31.00

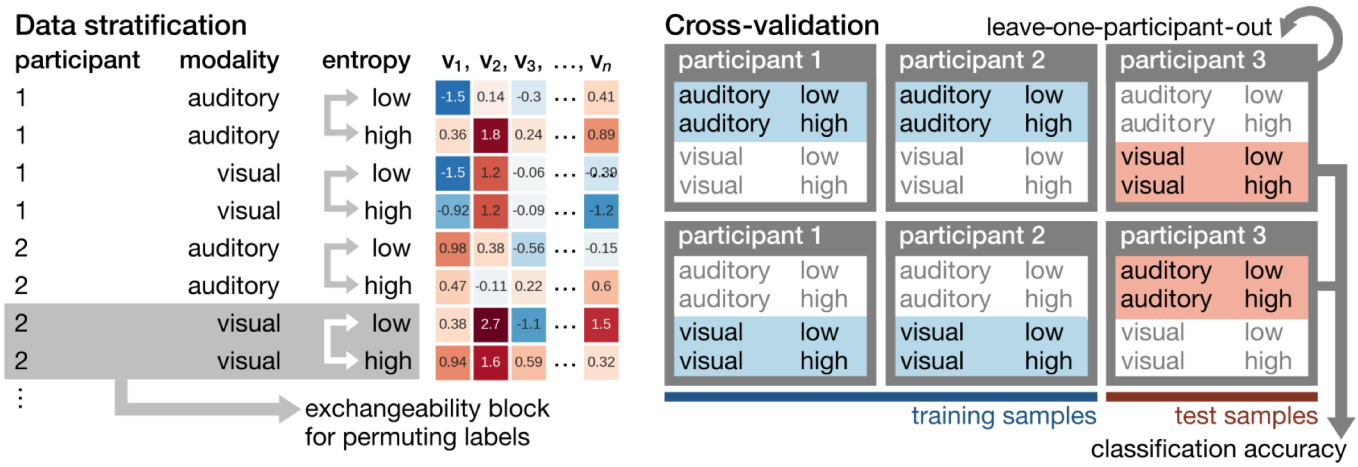

Fig. 1. Cross-validation and label permutation must respect data stratification. Both participant and modality constrain the exchangeability block, and labels of the factor of interest (entropy) must be permuted within levels of the cross-validation factor (modality).

respecting the rank order of response magnitudes in a pattern.

Cross-modal searchlight classification of high vs. low entropy levels with leave-one-participant-out cross-validation (section B) was applied to data preprocessed using three normalization schemes: no normalization, normalization across samples, and normalization across features (Fig. 2). To normalize across samples, we z-scored voxelwise responses for the four original levels of entropy within each level of the cross-validation factor (modality; nested within runs, participants), then averaged these samples across runs. To normalize across features, response patterns were averaged across runs and z-scored per sample within each searchlight.

Compared to no normalization, normalization across features increased searchlight classification accuracies (Wilcoxon signed-rank test; $W = 7.9 \times 10^9$, $p < .001$), while normalization across samples decreased searchlight accuracies ($W = 7.4 \times 10^9$, $p < .001$). Both normalization schemes significantly increased the variance of the distribution of searchlight accuracies (Levene's test; across features: $W = 4.5 \times 10^3$, $p < .001$; across samples: 3.5×10^4, $p < .001$). Both the non-normalized data and data normalized across features produced distributions of searchlight accuracies with a heavy right tail (no normalization: skewness = .049, $p < .001$ against normal distribution; across features: .041, $p < .001$), while normalizing across samples produced a distribution of accuracies not significantly skewed (skewness = .009). In evaluating searchlight classification accuracies, a heavy right tail is desirable as it indicates a greater proportion of searchlights with high accuracies. Finally, we used permutation tests (section D) to assess which searchlights yielded significant classification accuracies at $p < .05$ (uncorrected) for each normalization scheme. Normalization across features resulted in the largest number of significant searchlights (7,899), followed by normalization across samples (6,721) and no normalization (6,034). In section D, we estimate cluster-level significance controlling the familywise error rate.

For the current data set, we opt for normalization across features prior to further analysis, because this yields a distribution of searchlights more similar to the non-normalized data and ensures that the classifier is not capitalizing on regional-average differences in activation. Note however that the commonly-used pattern-correlation classifier implicitly normalizes across features [2]. Furthermore, for certain questions, there may be a theoretical motivation for incorporating regional-average differences in activation into the decoding procedure. For searchlight analysis, normalization across features must be performed per searchlight rather than across the whole brain prior to running the searchlight algorithm. When normalizing across samples, we recommend normalizing responses per feature within each level of the cross-validation factor (nested within runs, participants, etc); this effectively centers voxelwise responses for each level of the cross-validation factor. Normalizing across samples without respecting the boundaries of the cross-validation factor—in this case, modality—will accentuate between-modality response differences that may overpower any information about the factor of interest that is common across modalities. Normalization across samples is only applicable to data sets where the factor of interest comprises several or more samples.

B. Cross-Validation

We implemented leave-one-participant-out cross-validation [6] due to having relatively few independent scanner runs, and because this provides straightforward second-level inference [7]. To perform cross-modal classification, we trained each classifier on samples from one sensory modality in N–1 participants, then tested the classifier on samples from the other sensory modality in the left-out participant (Fig. 1). This procedure was repeated until each participant served as the test participant for both sensory modalities, then the classification accuracies were averaged across left-out participants. Studies with sufficient scanning runs may opt to perform a similar procedure within each participant using leave-one-run-out cross-validation, then proceed to a more conventional group-level analysis [8]; but see [7]. Cross-validation should respect the stratification of the data and ensure that there is a balanced frequency of classes in each fold.

C. Direction of Cross-Validation

In many applications of multivariate pattern analysis,

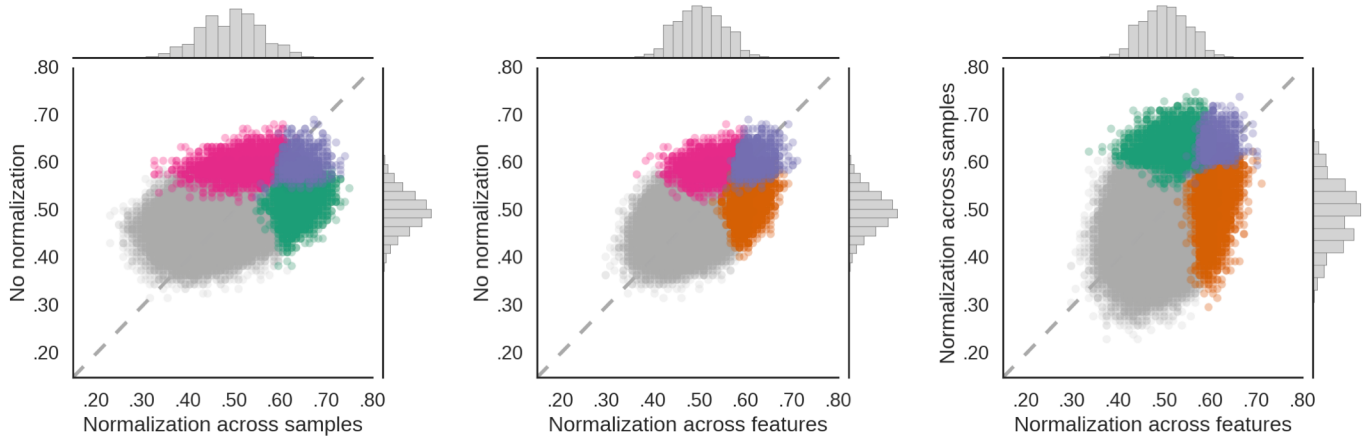

Fig. 2. Effect of response normalization on cross-modal searchlight classification of high vs. low entropy with leave-one-participant-out cross-validation. Each point represents the classification accuracy at a single searchlight for two different normalization schemes. Colors indicate significant searchlight classification accuracies ($p < .05$, uncorrected): pink, significant exclusively with no normalization; green, significant exclusively for normalization across samples; orange, significant exclusively for normalization across features; purple, significant for both compared normalization schemes. Theoretical chance accuracy is .50.

classifiers are cross-validated across scanning runs or participants, and particular training–test splits are not of experimental interest. However, for multivariate cross-classification, asymmetries in training and testing on different levels of the cross-validation factor may be theoretically significant [9]. In practice, significance can be estimated based on each cross-validation direction individually, or the average of both directions. To assess the significance of cross-validation asymmetry (i.e., the difference in accuracy for the two directions), we performed a simple permutation test where the cross-validation direction was randomly permuted in each left out participant for each searchlight, then the difference between directions was computed and averaged across participants (Fig. 3). In our data, only 578 searchlights (0.3% of all searchlights) exhibited significant asymmetry at $p < .05$ (uncorrected); significant asymmetry in this many searchlights may be expected by chance. Although future studies might focus on the classification accuracy averaged across cross-validation folds, we recommend inspecting both directions for asymmetries [1].

D. Permutation Tests

Permutation tests are generally regarded as superior to parametric methods when testing the significance of cross-validated classifier performance on stratified data [8]. Cross-classification introduces an additional level of stratification; namely, the cross-validation factor. For this reason, we recommend permuting the condition labels of interest within each level of the cross-validation factor nested within more typical exchangeability blocks (runs, participants; Fig. 1) [8]. Permuting labels of interest across levels of the cross-validation factor may result in problematic null distributions and inflated p-values. For the present data, we permuted the condition labels (entropy levels) within each sensory modality within each participant, as data were already averaged across runs. After randomly shuffling the labels of interest within these blocks, we recomputed the entire cross-validation procedure per searchlight. This was repeated 1,000 times to construct a null distribution of searchlight accuracy maps. To perform cluster-level inference we first computed residual accuracy maps by subtracting the mean accuracy map from each participant's map. The mean smoothness of these residual accuracy maps was then supplied to a Monte Carlo procedure using random Gaussian noise to estimate cluster-level significance. Reported clusters survived at a cluster-forming threshold of $p < .05$ and a cluster-level threshold of $p < .05$ (controlling the familywise error rate, cluster extent of 123 voxels). Although we estimated significance based on the mean accuracy across both cross-validation directions, searchlights within the surviving clusters may prefer one direction over another (Fig. 4).

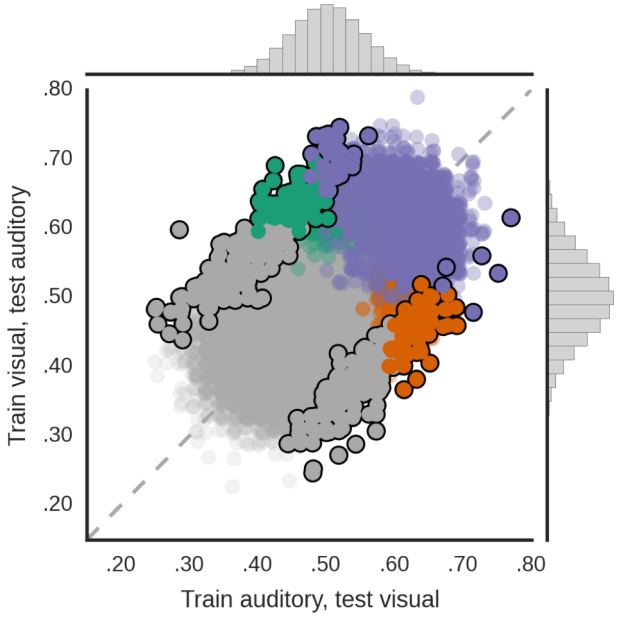

Fig. 3. Cross-validation asymmetry for cross-modal searchlight classification of high vs. low entropy. Color indicates significant classification ($p < .05$, uncorrected): purple, significant when estimated on accuracies averaged across both directions; green, orange, significant when estimated on test auditory and test visual directions, respectively. Searchlights with significant asymmetry at $p < .05$ (uncorrected) are plotted at full opacity with a black border.

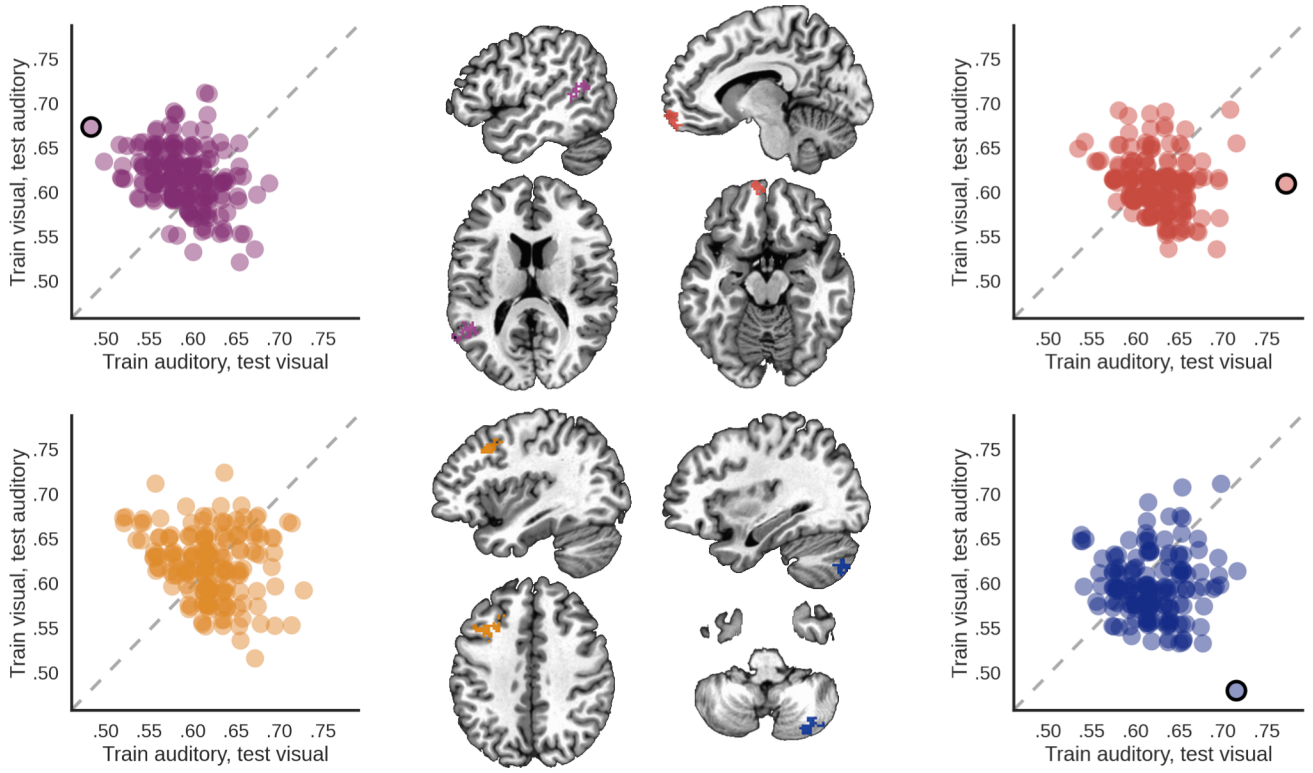

Fig. 4. Significant clusters for cross-modal classification of high vs. low entropy with leave-one-participant-out cross-validation. Scatter plot accuracies are jittered by .01 to visualize overlapping accuracies. Searchlights with significant cross-validation asymmetry at $p < .05$ (uncorrected) are plotted with a black border.

IV. Discussion

We presented four lessons learned in performing multivariate cross-classification. These analytic challenges arise because factorial designs introduce an additional layer of data stratification. Cross-validation must respect data stratification, and in cases where the cross-validation factor has only two levels, we recommend extending cross-validation to include left-out runs or participants. This sort of cross-validation scheme may yield asymmetric classification accuracies, and these asymmetries may be of theoretical interest [9]. A simple permutation-based approach can be used to estimate the significance of asymmetry. To assess the statistical significance of cross-classification accuracies, we recommend using permutation tests that respect the stratification of the data; labels of interest should be permuted within each level of the cross-validation factor to avoid generating overly-permissive null distributions. Finally, we note that these recommendations are provisional—the best choices may vary across data sets and questions. However, each of these analytic decisions must be considered carefully, as different choices imply different assumptions and may impact the significance or interpretability of results.

Acknowledgment

This work was supported by ERC starting grant #263318 (NeuroInt) to U.H. The computations in this work were performed on the Discovery cluster supported by the Research Computing group, ITS at Dartmouth College.

References

[1] J. T. Kaplan, K. Man, and S. G. Greening, "Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations," *Front. Hum. Neurosci.*, vol. 9, p. 151, Mar. 2015.

[2] M. Misaki, Y. Kim, P. A. Bandettini, and N. Kriegeskorte, "Comparison of multivariate classifiers and response normalizations for pattern-information fMRI," *NeuroImage*, vol. 53, no. 1, pp. 103–118, Oct. 2010.

[3] F. Pereira and M. Botvinick, "Information mapping with pattern classifiers: a comparative study," *NeuroImage*, vol. 56, no. 2, pp. 476–496, May 2011.

[4] S. M. Smith *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, suppl. 1, pp. S208–S219, Sep. 2004.

[5] M. Hanke, Y. O. Halchenko, P. B. Sederberg, S. J. Hanson, J. V. Haxby, and S. Pollmann, "PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data," *Neuroinformatics*, vol. 7, no. 1, pp. 37–53, Jan. 2009.

[6] J. A. Clithero, D. V. Smith, R. M. Carter, and S. A. Huettel, "Within- and cross-participant classifiers reveal different neural coding of information," *NeuroImage*, vol. 56, no. 2, pp. 699–708, May 2011.

[7] C. Allefeld, K. Görgen, and J.-D. Haynes, "Valid population inference for information-based imaging: information prevalence inference," *arXiv* preprint, arXiv:1512.00810 [q-bio.NC], Dec. 2015.

[8] J. A. Etzel, "MVPA permutation schemes: permutation testing for the group level," in *2015 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, Jun. 2015.

[9] N. N. Oosterhof, S. P. Tipper, and P. E. Downing, "Visuo-motor imagery of specific manual actions: a multi-variate pattern analysis fMRI study," *NeuroImage*, vol. 63, no. 1, pp. 262–271, Oct. 2012.