MVPA Significance Testing When Just Above Chance, and Related Properties of Permutation Tests

Joset A. Etzel
Cognitive Control & Psychopathology Lab, Psychological & Brain Sciences
Washington University in St. Louis
Saint Louis, MO, USA
jetzel@wustl.edu

***Abstract*— Parametric statistical tests (e.g., t-tests) can sometimes return highly significant results in cases that would be considered uninformative, such as when the individuals' accuracies are just above chance. This paper demonstrates that permutation tests can produce the expected non-significant results in these datasets. The properties of null distributions underlying this difference in significance are illustrated: their relative insensitivity to dataset information content, but sensitivity to dataset characteristics such as number of participants, examples, and runs.**

Keywords— fMRI, classification, significance, permutation, MVPA, cross-validation

I. Introduction

Permutation (randomization) tests have become the standard for significance testing in multivariate pattern analyses (MVPA), and increasingly, neuroimaging in general [1]–[3]. In neuroimaging MVPA, the overarching purpose of significance testing is to allow an informed guess as to whether there is sufficient information contained in the brain activity data to indicate which cognitive state or task condition the participant is in. Rephrased, we are generally asking whether the decoding classification accuracies actually obtained in the experiment are significantly greater than what could be expected by chance.

To make the discussion concrete, imagine a dataset with 20 participants, in which we want to test whether two conditions (looking at pictures of houses or faces) were classified more accurately than chance (0.5). Assume that the datasets were collected, processed, and analyzed properly, yielding an estimate of performance (e.g., classification accuracy) for each person. In many neuroimaging datasets (and in this example), the participants are independent (e.g., tested separately, not genetically related), so it is not unreasonable (if perhaps less than ideal) to consider parametric testing, such as a t-test, for the group level [4].

However, the distributional assumptions built into t-tests can lead to non-intuitive (and undesired) results, such as when individuals' accuracies are clustered at and just above chance levels. For example, the R [5] code in Fig. 1 assigns each of the 20 participants an accuracy between 0.49 (just below chance) and 0.54. Most researchers would not consider such a distribution meaningfully better than chance, but the t-test is highly significant, with t=3.8 and $p < 0.001$. This is not an error, per se, but a consequence of the mechanics of the t-test: the individual accuracies were generated to be tightly clustered, and mostly above chance, so the t-test properly indicates that a group mean accuracy below chance is highly unlikely. However, this does not match our interpretation of the test results: we take the amount an accuracy is above chance into account, not merely that it is above chance; we interpret an accuracy of 0.8 differently than 0.505.

This paper demonstrates that permutation tests can produce meaningful significance values, even when individual accuracies are just above chance. While this exact scenario does not occur often, it provides a case study for the different assumptions and properties of permutation and parametric statistical testing. Further, the insensitivity of permutation test null distributions to information content, but sensitivity to dataset structure, is illustrated and explained. Code to reproduce the simulations, results, and figures is available in an Open Science Foundation project, https://osf.io/c3s75/.

II. Methods

To simplify examples and discussion, this paper refers to "classification accuracy" and uses fMRI terminology (voxels, runs), but is intended to apply to other measures of quantifying performance (e.g., error rate), and to other imaging modalities (or fMRI projected to the surface). Linear support vector machines (SVMs; c=1; e1071 R interface to libsvm) were used for classification in all simulations, and simulations always had two stimulus classes. The results described here should apply if using other linear algorithms, but may not if using nonlinear classifiers or nested procedures (e.g., for feature selection or parameter optimization).

```
> set.seed(350);  # for reproducibility
> ds <- runif(20, min=0.49, max=0.54);
> round(ds, 3); # show dataset, rounded
  0.495 0.498 0.531 0.505 0.526 0.526 0.501 0.510
  0.539 0.517 0.526 0.513 0.507 0.497 0.504 0.502
  0.509 0.496 0.512 0.504
> t.test(ds, mu=0.5, alternative='greater');
  t = 3.8488, df = 19, p-value = 0.0005411
  95 percent confidence interval: 0.5060498  Inf
> mean(ds)
  [1] 0.5109849
```

Fig. 1. R code generating a set of 20 individual "accuracies" clustered near chance, with high t-test significance. R output edited for brevity.

This research was supported by NIH R37MH066078 to Todd Braver.

978-1-5386-3159-1/17/$31.00 ©2017 IEEE

A. Dataset Simulation

The motivating situation for this paper is when the group accuracy is just above chance (and individual accuracies clustered near chance); particularly cases in which t-tests perform non-intuitively, such as reporting that a mean accuracy of 0.51 is highly significant (Fig. 1). Exploring these properties requires multiple identically-structured datasets with different levels of information content (i.e., producing low to high accuracies), which necessitates simulated datasets.

Linear SVM classifiers (as used here) need bias (an activation difference between the classes) in individual voxels in order to distinguish the classes. Thus, the simulation created voxel "activity" values for one class by sampling from a uniform distribution, then added a random amount (indicated as "Signal" in the results) to each to make the examples for the other class. Varying the amount of added signal changes the difficulty of distinguishing the two classes, with larger amounts making easier classification.

The simulated datasets have 20 people and a single 50-voxel ROI. Unless otherwise noted, each person has data for 4 scanning runs, each of which has 10 examples of each of the two classes (80 examples total). We expect actual datasets to have dependencies within each run (examples collected closer in time to be more similar). While such dependency was not put into the simulated datasets, cross-validation was always performed on the runs (i.e., 4-fold cross-validation), as if it existed. Replication datasets were created when noted by changing the random seeds.

B. Permutation Testing

This paper concerns cases in which analyses are carried out within each individual separately and then combined at the group level. Note that this is different from intrinsically group-focused analyses, such as when images from different people are combined in the training sets and leave-one-subject-out cross-validation is used; different significance testing strategies are required for those analyses. Here, the permutation tests were always carried out within each individual person (of the simulated datasets) separately, permuting class labels within each run, dataset-wise [6]; all runs were relabeled in each permutation iteration (not just training or testing). A single set of 1000 random label rearrangements were generated, and these rearrangements were used for each participant (since there were no missings). The group-level null distribution is the average across participants for each of the 1000 relabelings (plus the true-labeled dataset). Thus, each of the 1001 group-level accuracies making up the group-level null distribution was the average accuracy of the 20 people with a particular dataset labeling; no person contributed an accuracy from a single relabeling (permutation iteration) to more than one group average (unlike [1]).

TABLE I. GROUP-LEVEL RESULTS FROM REPLICATION 1.

Signal	Accuracy	Permutation p	t-test p	t-test t
0.25	0.508	0.294	< 0.001	4.95
0.5	0.512	0.219	< 0.001	4.87
0.75	0.52	0.102	< 0.001	5.29
1	0.525	0.050	< 0.001	6.16
4	0.587	0.001	< 0.001	9.96
6	0.636	0.001	< 0.001	15.34
10	0.722	0.001	< 0.001	20.33
15	0.834	0.001	< 0.001	32.14

III. RESULTS

A. Null distributions normal and insensitive to signal level

The group-level accuracy, permutation test p-value, t-test t and p-values at increasing signal levels for the first simulated dataset are listed in Table 1. Note that the permutation test is not significant for low signal levels, though the t-test is highly significant throughout. The null distributions for each signal strength are shown in Fig. 2 (corresponding p-values are listed in Table 1). The null distributions are nearly identical across replications (different random seeds) and signal levels, and approximately normal. This may be surprising, since the signal strength in the datasets varies considerably, producing group-level mean accuracies (for Replication 1) of 0.51 (for signal of 0.25, plotted in red) to 0.83 (for signal of 15, plotted in black). However, null distributions should not be affected by signal strength: once the labels are permuted all class information is gone, so how much information was present in the true-labeled dataset should not affect the accuracy of the permuted-label datasets. Null distributions that are not approximately normal, or are strongly affected by the signal strength of the dataset, are a sign of an error in the permutation scheme, such as variance or structure in the actual dataset that is not present in the relabeled datasets (e.g., randomizing labels within the training

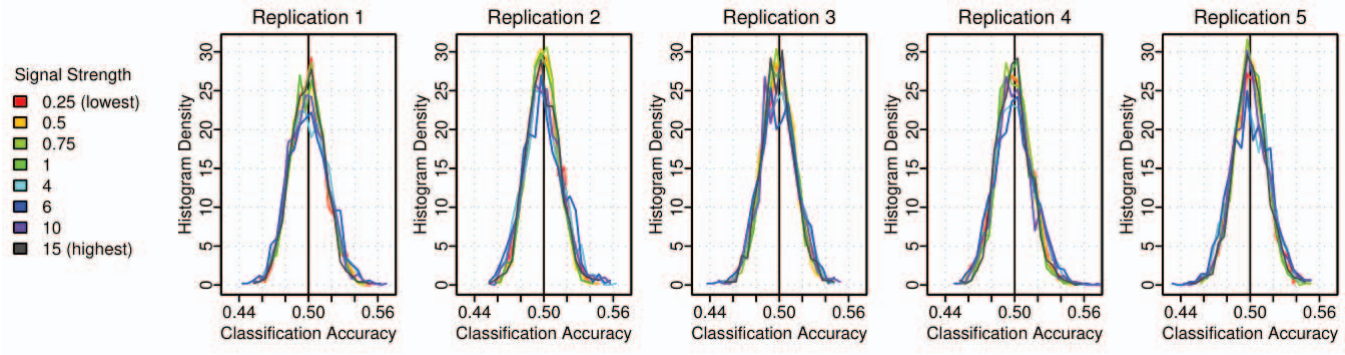

Fig. 2. Group-level null distributions for five replications of the simulation at each signal strength.

set as a whole, rather than within individual training set runs).

Since the permutation test relabelings were carried out within each individual, we can examine the null distribution (and so significance) of each person separately, providing additional information about dataset properties, as shown in Fig. 3. Note that the individuals' accuracies (red x) increase as the simulated datasets' signal level increases, but the null distributions (boxplots) are nearly identical, paralleling what was observed at the group level (Fig. 2). Thus, the individual subjects' classification accuracies are non-significant with signal strength 0.25 (the accuracies are near the median of the null distribution), far into the tails of the null distributions with signal strength 6, and well above the null distributions ($p < 0.001$) at signal strength 15.

B. Null distributions change with dataset structure

While the previous example showed that null distributions are fairly insensitive to the amount of signal in a dataset, they are sensitive to changes in the dataset structure, such as the number of subjects, runs, and examples.

The effect of changing the number of subjects is shown in Fig. 4. The group-level null distribution for the full dataset (20 subjects), signal strength 1, repetition 1, is shown in the far right pane. The group-level null distributions for the same dataset, but taking only the first 5, 10, or 15 subjects, are shown in the first three panes of Fig. 4. All four null distributions are approximately normal and centered on chance (0.5), but their variance decreases as the number of subjects included in the group increases. This pattern is expected, since the group-level null distribution was generated by averaging each subject's accuracy on each relabeled dataset; all things being equal (as they are in these simulations), we expect the variance of the group mean to decrease as members are added.

The effect of changing the number of runs and examples is illustrated in the first three panes of Fig. 5. These simulations used 20 subjects and signal strength 1, but differing numbers of runs and examples: 5 examples of each class in each of 3 runs in the first pane, 10 examples of each class but only 2 runs in the second pane, and 10 examples of each class with 8 runs in the third pane. Since the signal strength in these simulated datasets was kept constant, the effect on the classification accuracy and null distribution can be understood: more runs, and more examples in each run, make the datasets larger, and since every example is informative (the simulation was constructed with most voxels containing information to distinguish the classes), larger datasets are more informative. Conversely, smaller datasets (fewer runs and fewer examples) are less informative, and so more variable. Accordingly, the null distributions for 2 runs of 10 examples and 3 runs of 5 examples are wider (higher variance) than that for 4 runs of 10 examples, while the null distribution for 8 runs of 10 examples is narrower (lower variance).

C. Null distributions change with permutation scheme

Null distributions are also affected by the permutation scheme, generally narrowing (less variance, improperly higher significance for a particular accuracy) if the permutation scheme does not include all of the stratification and dependencies in the dataset. For example, the far-right pane of Fig. 5 shows the null distribution created from a fold-wise permutation scheme ([2]; labels randomized within each run, but separately on each cross-validation fold) of the same dataset (signal strength 1, replication 1) whose null distribution from a dataset-wise permutation scheme is shown in the far-right pane of Fig. 4. As can be seen by comparing the far-right panes of Figs. 4 and 5, the fold-wise scheme produces a lower variance null distribution. Accordingly, the same group-level accuracy of 0.525 has a significance level of 0.05 (50/1001)

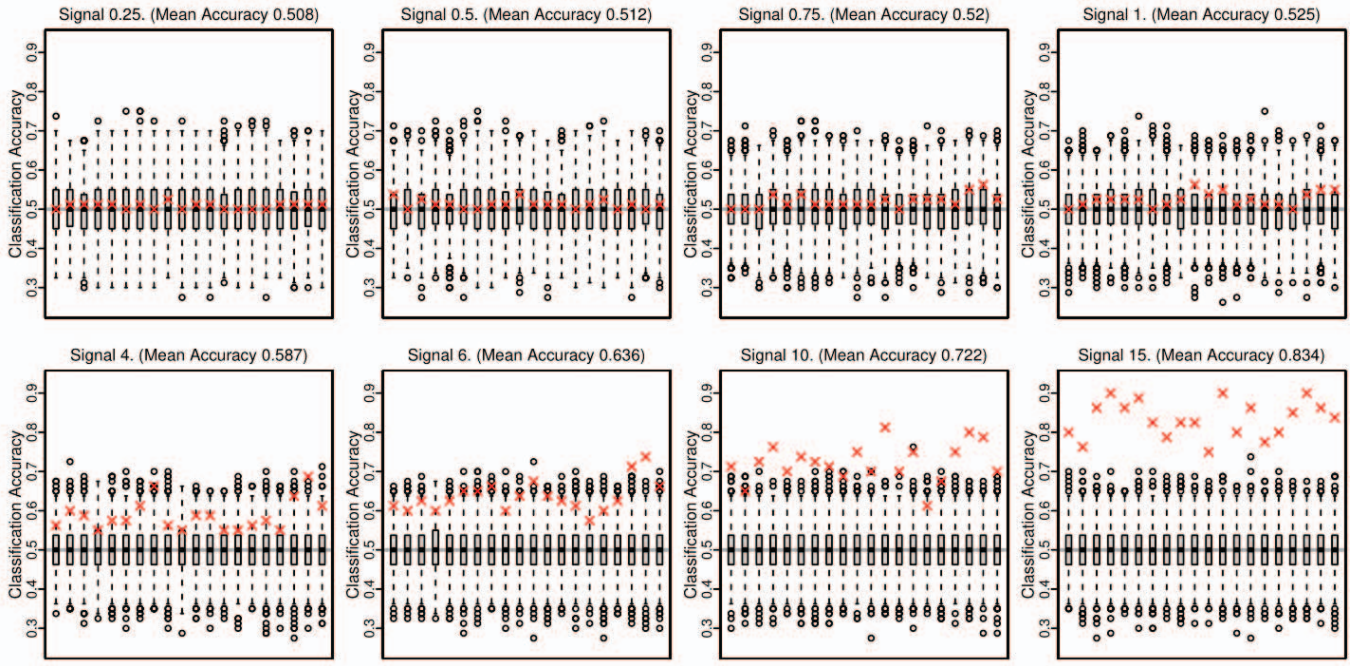

Fig. 3. Null distributions (boxplots) and classification accuracies (red x) for each subject in the first replication of the simulations at different signal levels. Group-level mean accuracies are listed in the plot titles; statistics in Table 1, and group-level null distributions in the first pane of Fig. 2.

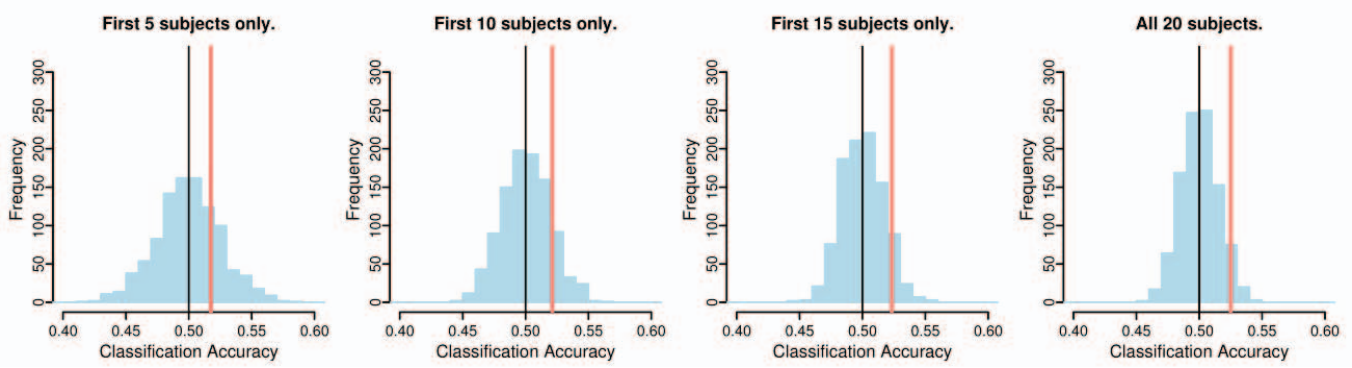

Fig. 4. Null distributions change when structural dataset characteristics change: number of subjects. These null distributions (blue) and accuracies (salmon lines) were calculated from the simulation with signal level 1, first replication. The far-right pane shows the null distribution for the full simulated dataset of 20 subjects, while the other panes show the null distribution when including only the first 5, 10, or 15 subjects (each subject is shown separately in Fig. 3). All histograms are with 0.01 accuracy bins for direct comparison.

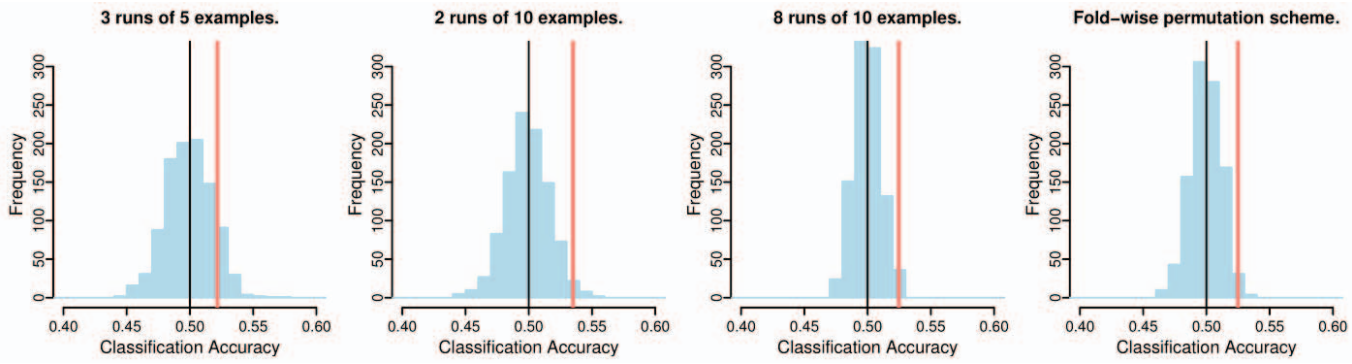

Fig. 5. Null distributions change when structural dataset characteristics change: varying numbers of runs and examples. All datasets in this figure have 20 subjects, signal level 1. The first three panes have the number of runs and examples per run listed in the plot titles. The rightmost pane has use the same dataset as the rightmost pane of Fig. 4 (4 runs of 10 examples each, replication 1), but with a fold-wise permutation scheme, rather than the dataset-wise scheme used in all other simulations.

with the dataset-wise scheme, but p=0.016 with the fold-wise scheme (16/1001); more significant, but also more likely to reflect an inflated false positive rate.

IV. Conclusions

This paper demonstrates several properties of permutation tests, including a case in which permutation and t-tests can differ widely in significance values: when subjects' accuracies are clustered at or just above chance. Properties of null distributions underlying that difference in significance values were illustrated, particularly their relative insensitivity to dataset information content, but sensitivity to other dataset characteristics, such as number of participants, examples, runs, and permutation scheme.

References

[1] J. Stelzer, Y. Chen, and R. Turner, "Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control," *Neuroimage*, vol. 65, pp. 69–82, 2013.

[2] J. A. Etzel, "MVPA Permutation Schemes: Permutation Testing for the Group Level," in *2015 International Workshop on Pattern Recognition in NeuroImaging*, 2015, pp. 65–68.

[3] A. M. Winkler, M. A. Webster, D. Vidaurre, T. E. Nichols, and S. M. Smith, "Multi-level block permutation," *Neuroimage*, vol. 123, pp. 253–268, 2015.

[4] J.-D. Haynes, "A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives," *Neuron*, vol. 87, pp. 257–270, 2015.

[5] R Development Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2015.

[6] J. A. Etzel and T. S. Braver, "MVPA Permutation Schemes: Permutation Testing in the Land of Cross-Validation," in *3rd International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, 2013, pp. 140–143.