MVPA Permutation Schemes:

Permutation Testing for the Group Level

Joset A. Etzel
Cognitive Control & Psychopathology Lab, Psychology Department
Washington University in St. Louis
St. Louis, MO, USA
jetzel@artsci.wustl.edu

***Abstract*—Permutation tests are widely used for significance testing in fMRI MVPA (multivariate pattern analysis) studies, but the precise way in which the tests are carried out varies, and test design is non-trivial because of complex, autocorrelated, and stratified dataset structures. Previously, we described permutation tests for single-subject datasets, recommending adoption of "dataset-wise" schemes, in which examples are relabeled prior to cross-validation. Here, we extend that work by describing permutation schemes for group analyses: datasets with more than one participant. Group-level MVPA is most often performed with either cross-validation on the subjects or within-subjects cross-validation, each of which requires a different strategy for permutation testing, as illustrated here.**

Keywords- fMRI; classification; significance; permutation; MVPA; cross-validation;

I. Introduction

Permutation testing is a preferred method of establishing statistical significance for task-based fMRI, since these datasets tend to have a complex stratified structure, including groups of participants (e.g. patient or control), multiple scanning sessions, and ordered trials within scanning runs. Additional layers of complexity arise in MVPA (multivariate pattern analysis) studies, due to the use of cross-validated statistics (e.g., classification accuracy from linear support vector machines (SVM)). The multiple layers of dependency in such stratified datasets cannot be disregarded during significance testing without risking false conclusions, since they constrain how the examples can be relabeled [1-5].

The concept of exchangeability is woven into the logic and theoretical foundations of permutation testing, and guides how each test should be designed. Permutation tests estimate significance by creating a null distribution to which the true accuracy (obtained from analyzing the dataset with the true, non-permuted labels) can be compared: if the true accuracy is greater than all accuracies making up this null distribution, its significance is 1/(size of null distribution + 1). The validity of a permutation test is thus tied to the validity of its null distribution: it requires a null distribution that genuinely reflects the accuracies obtained when there is no relationship between the task labels and the voxel values [6-8]. The accuracies when there is no relationship are calculated from datasets in which the task labels have been randomized ("exchanged"), so disrupting any relationship between the labels and voxel values.

However, given the complex structure of the typical MVPA dataset, the labels can generally not be permuted fully at random, but rather within the relevant dataset layers ("exchangeability blocks" [4, 7]); a "stratified" permutation test. Formally, within each exchangeability block the examples can be relabeled without affecting their joint probability distribution [8, 9]; in MVPA terms, we expect the relabeled datasets to classify at chance.

The complex structure of MVPA datasets can lead to ambiguity when designing permutation tests: what are the exchangeability blocks? How, precisely, should the labels be randomized? For example, should the labels be randomized within each participant separately, or across participants? In previous work [3], we described permutation schemes for single-subject datasets, suggesting that the "dataset-wise" permutation scheme is usually most appropriate, since it relabels the examples prior to conducting the cross-validation, thus maintaining more of the true dataset's structure. However, most neuroimaging datasets contain more than one person, and most hypotheses involve inferences at the group, not individual, level. This paper extends our previous work by

sub	run	task	v_1	v_2	...	v_n
1	1	f	0.006	0.484		0.200
1	1	p	0.230	0.305		0.142
1	2	f	0.236	0.159		0.072
1	2	p	0.344	0.157		-0.089
2	1	f	0.155	0.288		0.144
2	1	p	-0.954	0.156		-0.124
2	2	f	0.146	0.011		0.175
2	2	p	-0.004	0.076		-0.104
3	1	f	0.017	0.215		0.047
3	1	p	0.100	0.308		0.053
3	2	f	0.059	0.138		-0.014
3	2	p	0.107	0.147		-0.172

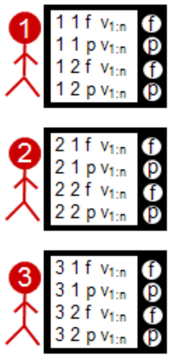

Figure 1. The example dataset. The table (left) is the dataset, with examples in the rows and voxels (v_1, v_2, ... v_n) in the columns. "sub" is the subject identifier, "run" is the scanning run, and "task" is whether the trial was of viewing faces ("f") or places ("p"). The boxes (right) show the data representation used in later figures. Since this is the true-labeled dataset, the circled task labels are white and match the "task" column.

describing permutation testing for group analyses: analyses with more than one person.

For clarity, common MVPA terminology is used in this paper, but its recommendations are not restricted to fMRI, nor to classification-based statistics. So, while the examples and discussion refer to “classification”, “accuracy”, and “voxels”, these terms are intended to apply equally to other cross-validated statistics (e.g., cvMANOVA discriminability [10]), and data from other neuroimaging modalities (e.g., MEG sensors). Similarly, “significance” is intended to refer to statistical significance in the frequentist sense typically used in neuroimaging, not importance more generally.

II. Example Dataset

As in [3], the permutation schemes are illustrated using an example dataset with representative structure (though far smaller than a real study). Concretely, suppose the example dataset is a task-based study performed with three human subjects (Fig. 1). Each person completed two fMRI scanning runs, and each scanning run was made up of two blocks each of two tasks: viewing pictures of places and faces. The task blocks were presented in a random order, separated by enough time to make it reasonable to assume that their labels can be permuted, and with no missings. The hypothesis is that task (viewing a face or place) can be classified (with a linear SVM) in these voxels (corresponding to a small region of interest) and in these participants.

III. Two Forms of Cross-Validation with Groups

In this example experiment we wish to classify task in our group of people. Note that this is not classifying the subjects themselves (e.g., by gender or diagnosis), but rather *task within subject*. Two different forms of cross-validation are common for this type of analysis, which are referred to here as “cross-validation on the subjects” and “within-subjects cross-validation”. Since the dataset is partitioned differently for each form of cross-validation, each has different patterns of dependency and variability, and thus requires a different permutation strategy. These are not the only forms of cross-validation for designs with multiple participants (another is “pooled-subjects”, such as leave-one-run-out cross-validation with data from all subjects at once); the logic presented here is intended to provide guidance on for constructing tests for other designs.

A. *Cross-Validation on the Subjects*

The key characteristic of cross-validation on the subjects is that the subjects’ datasets are the units of cross-validation. Leave-one-subject-out cross-validation (Fig. 2) is common in MVPA. Under this scheme, on each fold data from all but one subject is used for training, with data from the left-out subject used for testing. Each subject is left out in turn, so the number of cross-validation folds is equal to the number of subjects. When the dataset includes larger numbers of subjects significance can be improved by leaving out more than one subject on each fold, such as, for a 120-subject dataset, performing ten-fold cross-validation, leaving twelve subjects out on each fold [11].

Regardless of the number of subjects’ data left out each fold, the critical aspect of cross-validation on the subjects is that an independent accuracy does not result for each subject, but only the accuracy averaged over cross-validation folds: a single value for the dataset as a whole. Thus, unlike within-subjects cross-validation designs (described later), it is not possible to use a simple parametric statistic to establish significance: permutation testing is required.

The structure of the permutation scheme when cross-validation is on the subjects parallels that of performing a leave-one-run-out cross-validation on a single person’s data (see [3]; particularly Figs 2 and 4), with individuals taking the place of runs. The recommendation for designing a permutation test when cross-validation is on the subjects is thus the same as for the analysis of a single subject: the test should follow a dataset-wise permutation scheme.

A single iteration of a permutation test with the example dataset under a dataset-wise scheme is illustrated in Fig. 3. Each iteration begins by relabeling some of the examples, after which the cross-validation and classification is carried

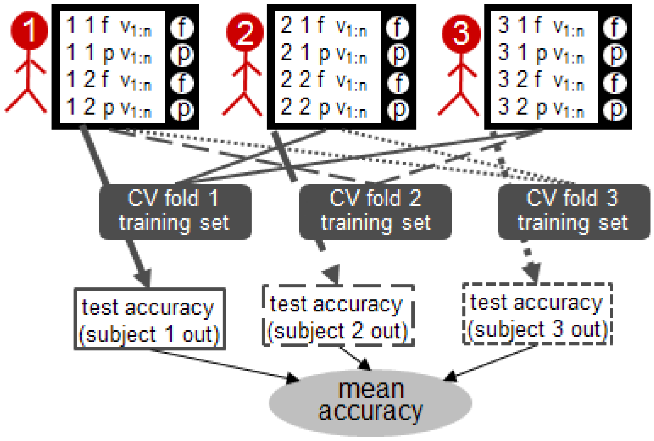

Figure 2. Determing the mean accuracy with cross-validation on the subjects; leave-one-subject-out cross-validation with the example dataset is shown. The subjects do not contribute independent accuracy estimates, but rather make up the cross-validation folds.

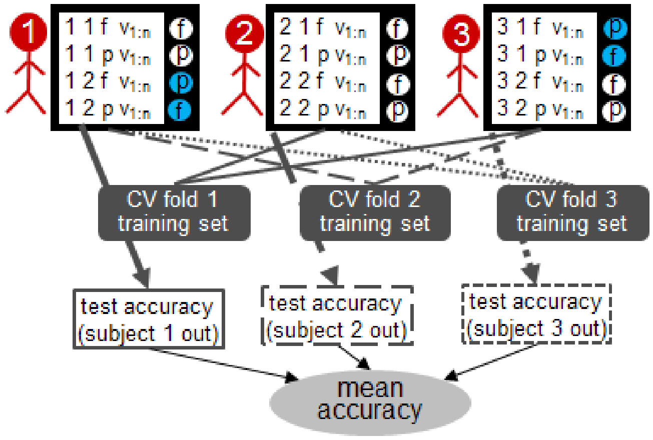

Figure 3. Single iteration of a permutation test when cross-validation is on the subjects. The labels have been permuted within the second run in subject 1 and the first run of subject 3 (highlighted with blue). Note that the analysis structure matches Fig. 2; only labels have changed.

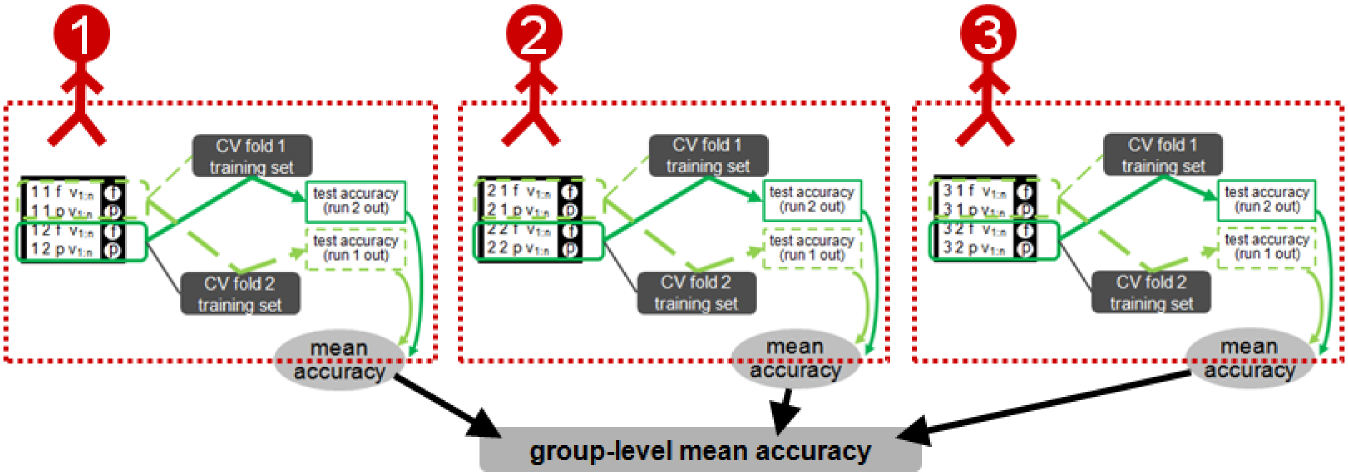

Figure 4. Determining the group-level mean accuracy using within-subjects-cross-validation. Leave-one-run-out cross-validation is performed within each individual participant, shown in the diagram by dotted red lines. These mean accuracies from each person are then averaged, giving the group-level mean accuracy. The group-level mean is thus calculated from statistically independent accuracies.

out on this relabeled dataset in the same way as on the true-labeled dataset (note the similarity between Fig. 2 and Fig. 3: only task labels change). The mean accuracy (averaged over the cross-validation folds) from each iteration is recorded, and used to construct the null distribution.

How should the examples be relabeled? In general, labels should be permuted within the most fundamental unit of the dataset (the exchangeability block), so that the relabeling does not change the structure of the dataset in any other way. For example, in Fig. 3 the relabeling was done within runs, even though partitioning is on the people (not the runs). Often, trials within the same fMRI scanning run are more similar to each other than to trials from other scanning runs, and relabeling examples within each scanning run ensures that this dependency is included in the null distribution.

B. *Within-Subjects Cross-Validation*

The key characteristic of the within-subjects form of cross-validation for group analysis is that cross-validation occurs within the individual participants, with only the resulting statistic combined across participants. Whichever cross-validation technique is appropriate should be performed within the individuals; Figs. 4 and 5 show leave-one-run-out cross-validation. The example dataset has two runs per person, so two cross-validation folds are performed within each person, and so the mean accuracy for each person is calculated by averaging those two accuracies.

Since each person is analyzed separately, these designs produce independent accuracies for each person, which are averaged to obtain the group-level accuracy; we want to establish a significance for this group average (Fig. 4). Since the group-level statistic summarizes independent values, multiple methods can be used to estimate its significance, both parametric and permutation-based. While testing whether the group-level mean is greater than chance with a t-test has been common in the MVPA literature, mixed logit models are likely more appropriate, given that accuracies can only fall between 0 and 1 [12].

Designing the permutation scheme for a within-subjects cross-validation design is different than when cross-validation

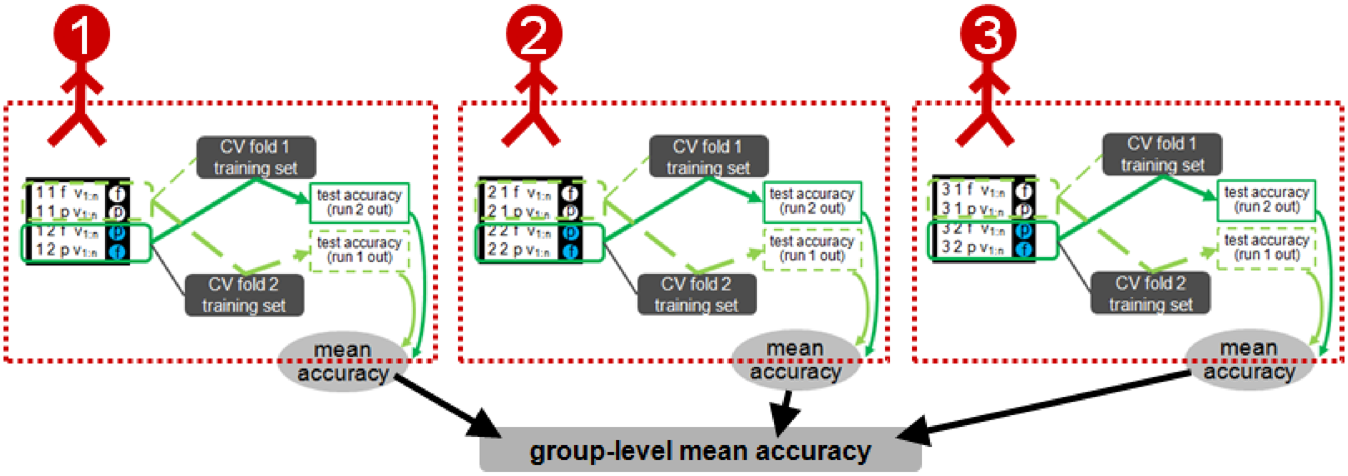

Figure 5. Single iteration of a permutation test with within-subjects cross-validation. In this illustration the labels have been permuted on the second run in all three subjects (blue), so the group-level mean is calculated from means resulting from the same relabeling in all individuals.

is on the subjects, because we are not establishing significance (by creating a null distribution) for the value directly resulting from the cross validation (Fig. 2), but rather for the average of values resulting from independent cross-validated analyses (Fig. 4). Thus, the permutation test requires creating a null distribution for the group-level mean accuracy, which is calculated from individual participants' mean accuracies, and so we need to properly account for both the structure of each individual's dataset and for the relationship between individuals (i.e., that the group-level statistic is calculated from one value from each person).

Generally, the permutation test for each individual should be calculated following a dataset-wise scheme [3], designed with the same considerations as if they were the only person in the dataset. Accordingly, Fig. 5 shows a single iteration of the permutation test in which the examples have been relabeled within the second run only (i.e., runs are the exchangeability blocks).

Since a mean accuracy is calculated for each person independently, the permutation test within each individual could also be carried out independently, with different relabelings for each person. Applying different sets of relabelings to different people may be unavoidable, such as when missing data causes the number of examples in each person to vary, or the number of examples is too small to allow sufficient unique relabelings within each individual. However, when possible, it seems ideal to use the same set of permuted labels within each individual, so that the group-level mean accuracy for each iteration of the permutation test is calculated from the same labels in each person (as shown in Fig. 5: the second run was relabeled in the same way in all subjects). After all, the true-labeled group-level mean accuracy was calculated using the same (true) labels in each person. More research is needed, but there seems little risk of biasing the null distribution if different relabelings are used in each individual when necessary (such as in [5]), so long as each individual contributes one mean accuracy to the group-level mean in each iteration of the permutation test.

IV. Discussion

This paper describes how to perform permutation tests for two common ways of carrying out group-level MVPA: within-subjects cross-validation and cross-validation on the subjects. The diagrams and logic suggested in this paper build upon the recommendations for single-subject analyses made in [3]. It is hoped that, together, these papers clarify the necessary considerations for performing permutation tests, establish a vocabulary for describing the precise manner in which permutation tests are carried out, and provide guidelines for designing permutation tests for arbitrary experimental hypotheses.

While this paper is intended to make the design of permutation tests clear and straightforward, their execution is often complex, with many constraints on how the examples are relabeled. Thus, it is often advisable to precompute the relabelings that will be used for the permutation test, rather than calculate them on the fly. Precomputing the relabelings (e.g., in the first iteration, subject 1's examples will be labeled `f p p f`; in the second iteration, subject 1's examples will be labeled `p f f p`, etc.) has several advantages. First, the validity of each relabeling can be confirmed: is the number of examples of each type equal in each exchangeability block (run, participant, etc.) on each permutation? The relabelings can also be checked for duplicates, and that their distribution is approximately random (if a subset of the possible relabelings are used, rather than a complete permutation test). Finally, precomputing the set of relabelings simplifies running the analyses as individual jobs, so that different machines can run different iterations simultaneously.

Acknowledgment

This work was supported by the National Institutes of Health, grant number R37MH066078 to Todd Braver.

References

[1] K. Schreiber and B. Krekelberg, "The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging," *PLoS One*, vol. 8, p. e69328, 2013.

[2] F. Pereira and M. Botvinick, "Information mapping with pattern classifiers: A comparative study," *NeuroImage*, vol. 56, pp. 476-496, 2011.

[3] 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)*, Philadelphia, PA, USA, 2013, pp. 140-143.

[4] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols, "Permutation inference for the general linear model," *NeuroImage*, vol. 92, pp. 381-397, 2014.

[5] 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.

[6] S. Mukherjee, P. Golland, and D. Panchenko, "Permutation Tests for Classification," in *AI Memo 2003-019*: Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2003.

[7] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Human Brain Mapping*, vol. 15, pp. 1-25, 2001.

[8] P. I. Good, *Resampling methods: a practical guide to data analysis*, 3rd ed. Boston, Basel, Berlin: Birkhäuser Boston, 2006.

[9] E. S. Edgington, *Randomization Tests*, 3rd ed. New York: Marcel Dekker, 1995.

[10] C. Allefeld and J.-D. Haynes, "Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA," *NeuroImage*, vol. 89, pp. 345-357, 2014.

[11] P. Golland and B. Fischl, "Permutation tests for classification: towards statistical significance in image-based studies," in *Information Processing in Medical Imaging*. vol. 2732/2003, C. J. Taylor and J. A. Noble, Eds.: Springer Berlin/Heidelberg, 2003, pp. 330-41.

[12] T. F. Jaeger, "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," *Journal of Memory and Language*, vol. 59, pp. 434-446, 2008.