

Better-than-chance classification for signal detection

Jonathan Rosenblatt Roei Gilron Roy Mukamel

July 28, 2016

Abstract

[TODO]

1 Introduction

A common workflow in genetics or neuroimaging consists of fitting a classifier, and estimating its predictive accuracy using cross validation. Given that the cross validated accuracy is a random quantity, it is then common to test if the cross validated accuracy is significantly better than chance using a permutation test. Examples in the genetics literature include Jiang et al. [2008], Radmacher et al. [2002] [TODO: elaborate]. Example sin the neuroscientific literature include [Golland and Fischl, 2003, Kriegeskorte et al., 2006, Pereira et al., 2009, Varoquaux et al., 2016]. The number of citations¹ of these papers attest to the popularity of the above workflow: 956 for Kriegeskorte et al. [2006], and 274 for Radmacher et al. [2002], as examples.

To fix ideas, we will adhere to a neuroscientific example: In Gilron et al. [2016], the authors seek to detect brain regions which encode differences between vocal and non-vocal stimuli. According to the MVPA analysis workflow, the localization problem is cast as a supervised learning problem: if the type of the stimulus can be predicted from the spatial activation pattern, significantly better than chance, then a region is declared to encode vocal/non-vocal information. We call this an *accuracy test*, a.k.a. *class prediction* in Simon et al. [2003].

This same signal detection task can be also approached as a two-group multivariate test: Inferring that a region encodes vocal/non-vocal information, is essentially inferring that the spatial distribution of brain activations

¹Based on GoogleScholar. Accesses on 26.7.2016.

25 is different given a vocal/non-vocal stimulus. A practitioner may then call
 26 upon a two-group location test such as Hotelling’s T^2 [Fujikoshi et al., 2011].
 27 Alternatively, if the size of the brain region is too large compared to the num-
 28 ber of observations, so that the spatial covariance cannot be fully estimated,
 29 then a high dimensional version of Hotelling’s test can be called upon, such
 30 as in Srivastava [2013] or Schäfer et al. [2005]. In contrast to *accuracy tests*,
 31 we call these *location tests*, a.k.a. *class comparison* in Simon et al. [2003].

32 At this point, it becomes unclear which is the preferred test. The compar-
 33 ison between location and accuracy tests was precisely the topic of Ramdas
 34 et al. [2016], who compared the Hotelling location test to the accuracy of
 35 *Fisher’s linear discriminant analysis* classifier (LDA) [Hastie et al., 2003].
 36 Using an asymptotic analysis, Ramdas et al. [2016] concluded that accuracy
 37 and location tests are equivalent with respect to their order of convergence
 38 to a consistent test, while they may differ in constants. Put differently, the
 39 (asymptotic) relative efficiency of the tests is not trivially 0 nor ∞ .

40 The relative efficiency, governing the power of the tests, may prove crucial
 41 when dealing with the finite sample sizes in neuroscience and genetics, and
 42 thus the focus of this study. We thus seek to study which test is to be
 43 preferred in finite samples? Our conclusion will be quite simple: *location*
 44 *tests almost always have more power than accuracy tests*.

45 The main argument for our statement rests upon the observation that
 46 with typical sample sizes, the accuracy test statistic is highly discrete. Dis-
 47 crete test statistics are known to be conservative [Hemerik and Goeman,
 48 2014], since they cannot exhaust the permissible false positive rate. For accu-
 49 racy tests, the degree of discretization is governed by the number of samples.
 50 In our running neuroscience example [Gilron et al., 2016], the classification
 51 is performed based on 40 trials, so that the test statistic may assume only 40
 52 possible values. This number of examples is not unusual if considering this
 53 is the number of subject in a genetic study, or the number of trial-repeats in
 54 an fMRI brain scan.

55 The discretization effect is aggravated if the test statistic is highly concen-
 56 trated. For an intuition consider the usage of the *train* accuracy test statistic
 57 (i.e., not cross validated). In Section 4 we then address our main question-
 58 which test has more power? Based on the finding that the location test is
 59 typically more powerful, we try to offer an intuition for this phenomenon in
 60 the Discussion section.

61 2 Problem setup

62 Adhering to our neuroscientific example, we now formalize terminology and
63 notation. Let $y \in \mathcal{Y}$ be a class encoding. In our vocal/non-vocal example
64 we have $\mathcal{Y} = \{-1, 1\}$. Let $x \in \mathcal{X}$ be a p dimensional feature vector. In our
65 vocal/non-vocal example p is the number of voxels in a brain region. We
66 thus have $\mathcal{X} = \mathbb{R}^{27}$.

67 Given n pairs of (x_i, y_i) , typically assumed i.i.d., a location test amounts
68 to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$ (or
69 at least the same location). I.e., the multivariate voxel activation pattern
70 has the same distribution when given a vocal stimulus, as when given a non-
71 vocal stimulus. An accuracy test amounts to learning a predictive model $\hat{f}(x)$
72 from some assumed model class $\hat{f} \in \mathcal{F}$. The prediction accuracy, denoted
73 $T_{\hat{f}}^{acc}$, is defined as the probability of a given classifier \hat{f} of making a correct
74 prediction $T_{\hat{f}}^{acc} := Prob(\hat{f}(x) = y)$ when given a new, randomly drawn data
75 point, (x, y) . A statistically significant “better than chance” estimate of $T_{\hat{f}}^{acc}$
76 is evidence that the classes are distinct.

77 2.1 Candidate Tests

78 The design of a permutation test using the prediction accuracy, requires the
79 following design choices:

- 80 1. How to estimate accuracy?
- 81 2. Is the statistic cross validated or not?
- 82 3. For a K-fold cross validated test statistic: should the data be refolded
83 in each permutation?
- 84 4. Permute labels of features?
- 85 5. For a K-fold cross validated test statistic: should the data folding bal-
86 anced? (a.k.a. stratified).
- 87 6. How many folds?

88 We will now address these questions while bearing in mind that unlike the
89 typical supervised learning setup, we are not interested in an unbiased esti-
90 mate of the prediction error, but rather in the mere detection of a difference
91 between two groups, leading to a better-than-chance accuracy.

92 **How to estimate accuracy?** Given a predictor \hat{f} , a natural test statis-
 93 tic is some estimate of its accuracy $T_{\hat{f}}^{acc}$. Complicating matters: very low
 94 accuracies, even 0, is evidence that the classes are separated, and we only
 95 need to invert the predictions. We can thus consider $|T_{\hat{f}}^{acc} - 0.5|$ as the test
 96 statistic. This, however, implies that if the classes are identical, random
 97 guessing has a 0.5 accuracy. This is not true if the classes are not balanced.
 98 The chance level in which case is the prevalence of the dominant class, we
 99 denote by \hat{p}_{max} . This suggests the following test statistic $|T_{\hat{f}}^{acc} - \hat{p}_{max}|$. Since
 100 we will be aggregating these statistic over random data sets where the dom-
 101 inant class may have varying frequencies, it seems appropriate to standard-
 102 ize the scale of this statistic. We thus also consider the z-scored accuracy:
 103 $|T_{\hat{f}}^{acc} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$.

104 **Cross validated or not?** Were we interested in an unbiased estimator of
 105 the prediction error, there is no question that some independent validation
 106 is in order. Since we are merely interested in detecting a difference between
 107 classes, a biased error estimate is not an issue provided that bias is consistent
 108 over all permutations. The underlying intuition is that if the exact same
 109 computation is performed over all permutations, then a permutation test
 110 will be “fair”, i.e., will not inflate the false positive rate. We will thus be
 111 considering both cross validated accuracies, and *train* accuracies as our test
 112 statistics.

113 **Refolding?** The standard practice in neuroimaging is to refold the data
 114 after each permutation. This is imperative if permuting labels while aiming
 115 at balanced data folds. This is not, however, imperative in general. In this
 116 work, we will adhere to the standard practice of refolding the data within
 117 each permutation.

118 **Permute labels of features?** While seemingly identical, the compound-
 119 ing of permutations with data foldings renders these two approaches distinct.
 120 As an example, consider balanced (stratified) K-fold cross validation where
 121 the initial data folding is balanced. After a label permutation, the folds will
 122 probably not be balanced, and will thus have to be refolded. If the features
 123 are permuted, then the labels conserve their original fold assignments, and
 124 the data need not be refolded. Since we only report results while refolding the
 125 data in each permutation, then the only difference between permuting labels
 126 and permuting features seems to be a computational one. We thus adhere
 127 to the more common, albeit less efficient practice, of permuting labels.

128 **Balanced folding?** A standard practice when cross validating is to con-
 129 strain the data folds to be balanced (i.e. stratified). This is well justified
 130 when aiming at unbiased accuracy estimation. This also simplifies matter
 131 when aiming at signal detection, as can be seen from the above discussion
 132 of the appropriate test statistic. Then again, it may complicate matters, as
 133 can be seen from the above discussion on label versus feature permutation.
 134 In general , it is not imperative in general, and we will indeed be comparing
 135 the effect of balanced foldings versus unbalanced. We will thus report results
 136 with both balanced and unbalanced data foldings.

137 **How many folds?** Different authors suggest different rules for the number
 138 of folds. We will be varying the number of folds, since it will affect the
 139 concentration of the estimated accuracy, which will have a crucial effect on
 140 the conservativeness of the permutation test. Our intuition suggests that
 141 since more folds imply a less concentrated estimate, then leave-one-out should
 142 be the less conservative, and 2-fold should be the most conservative.

143 By now, the reader will have observed that there are indeed many ways
 144 to perform a permutation test using a cross validated statistic. The subset
 145 of tests we will be comparing is collected for convenience in Table 1.

146 3 Controlling the False Positive Rate

147 In the first of our battery of simulations we verify that various test statistics
 148 and permutation schemes control the type I error. Figure ?? demonstrates
 149 that this is indeed the case. All our candidate tests control the type I error,
 150 with varying degrees of conservativeness. In particular: (a) if the folds are
 151 balanced or not, (b) if the labels are permuted or the features, (c) if the test
 152 statistic is varied, (d) if the regularization level of the support vector machine
 153 classifier (SVM) is varied, (e) if the number of folds is varied.

154 4 Power

155 Having established that all of the tests in our battery control the false pos-
 156 itive rate, it remains to be seen if they have similar power– at least when
 157 comparing the power of the various classifiers and multivariate tests. The re-
 158 sults of Ramdas et al. [2016] suggest that power should be of the same order.
 159 On the other hand, the results of our previous sections suggest that the con-
 160 servativeness of some of the considered tests can be considerable, rendering
 161 them underpowered.

Name	Basis	CV	Accuracy	Parameters
Hotelling	Hotelling	–	–	shrink=FALSE
Hotelling.shrink	Hotelling	–	–	shrink=TRUE
lda.CV.1	LDA	TRUE	accuracy	–
lda.CV.2	LDA	TRUE	z-accuracy	–
lda.noCV.1	LDA	FALSE	accuracy	–
lda.noCV.2	LDA	FALSE	z-accuracy	–
sd	SD	–	–	–
svm.CV.1	SVM	TRUE	accuracy	cost=1e1
svm.CV.2	SVM	TRUE	accuracy	cost=1e-1
svm.CV.3	SVM	TRUE	z-accuracy	cost=1e1
svm.CV.4	SVM	TRUE	z-accuracy	cost=1e-1
svm.noCV.1	SVM	FALSE	accuracy	cost=1e1
svm.noCV.2	SVM	FALSE	accuracy	cost=1e-1
svm.noCV.3	SVM	FALSE	z-accuracy	cost=1e1
svm.noCV.4	SVM	FALSE	z-accuracy	cost=1e-1

Table 1: This table enumerates the various test statistics we will be studying. Three are location tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group T^2 statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer et al. [2005]. *sd* is another high dimensional version of the T^2 , from Srivastava et al. [2013]. The rest of the tests are variations of the linear SVM, and Fisher’s LDA, with varying accuracy measures, cross validated or not, and varying tuning parameters. For example, *svm.CV.4* is a linear SVM, with *libsvm*’s cost parameter set at 0.1, using the cross validated z-scored accuracy ($|T_{\hat{f}}^{acc} - \hat{p}_{max}|/\sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$, see Section 2.1). Another example is *lda.noCV.1*, which is Fisher’s LDA, returning the train accuracy, without cross validation, and without z-scoring.

[TODO: discuss power of various tests after finishing simulations]

We see by now that the use of accuracy tests for signal detection is underpowered compared to location tests. The above simulations can hardly support such a universal statement. We will thus verify on a neuroimaging dataset, and discuss the causes for this phenomenon thus the scope of the statement.

5 Neuroimaging Example

Figure 2 is an application of our battery of tests to the data of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI data while subjects were exposed to the sounds of human speech (vocal), and other non-vocal sounds. Each subject was exposed to 20 sounds of each type, totalling in

Figure 1: The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. They are assumed to be equal in all the 23 dimensions, and vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). The various statistics on the y axis. Their details are given in Table 1. Simulation code available at [TODO].



173 $n = 40$ trials in each scan. The study was rather large and consisted of
 174 about 200 subjects. The data was kindly made available by the authors at
 175 the OpenfMRI website².

176 To verify the observation that location tests have more power than accu-
 177 racy tests, we perform permutation inference using the pipeline of Stelzer
 178 et al. [2013], which was also used in Gilron et al. [2016]. For completeness,
 179 the pipeline is described in Appendix A. To demonstrate our point, we com-
 180 pare the *sd* location test with the *svm.cv.1* accuracy test (see Table 1 for the
 181 definition of these statistics).

182 In agreement with our simulation results, the location test (*sd*) discovers
 183 more brain regions that encode information discriminating between vocal and
 184 non-vocal stimuli when compared to an accuracy test (*svm.cv.1*). The former
 185 discovers 1,232 regions, while the latter only 441, 399 of which are common
 186 to both as reported in Figure 2. We emphasize that both test statistics were
 187 compared with the same permutation scheme, and the same error controls,
 188 so that any difference in detections is due to their different power.

189 Having established that accuracy tests are underpowered both in simula-
 190 tion and in application, we wish to identify the conditions under which this

²<https://openfmri.org/>

will occur, and discuss implications on the practice of accuracy tests.

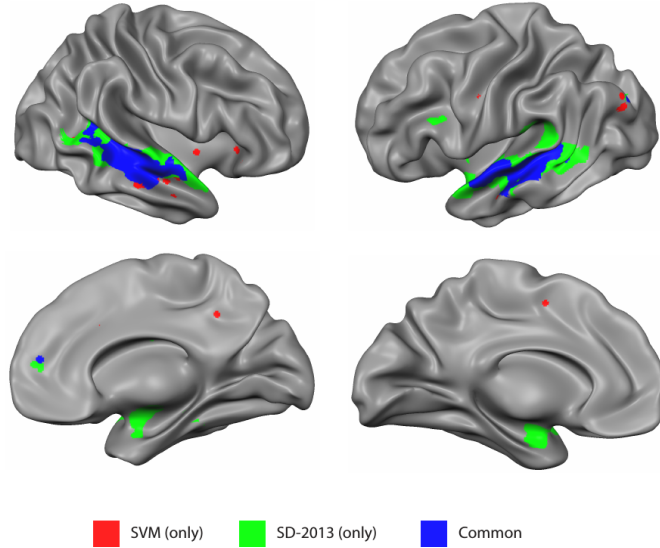


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centres of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a location test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of Stelzer et al. [2013], followed by region-wise $FDR \leq 0.05$ control using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. Number of permutations equals 400. The location test detect 1,232 regions, and the accuracy test 441. The overlap is such that 90% of the accuracy test regions, are also detected by the location test. For the details of the analysis see Appendix A and Gilron et al. [2016].

6 Discussion

We have set out to understand which of the tests is more powerful: the accuracy test or the location test. Using simulations, we have concluded that the location tests are preferable. We attribute this to the discretization introduced in finite samples by the accuracy test statistic. This also explains why an asymptotic analysis, such as Ramdas et al. [2016], did not find a qualitative difference.

At this point some reservations to the generality of our findings are in order. Firstly, not all accuracy tests are concerned with signal detection. Indeed, it is possible that the purpose of the test is not to detect a difference between classes, but to actually test if a particular classifier is better

than chance. This would be the case, for instance, with brain-machine interfaces, where the detection of a signal is not enough [Olivetti et al., 2012]. In such cases, the performance of a particular classifier is the object of study, rendering the accuracy test the appropriate choice.

Secondly, there may be cases where the accuracy test does have more power than the location test. Our simulations were unable to point out such a scenario, but the fact that in our neuroimaging example (Section 5) some brain regions were detected with the accuracy test, and not the location test, suggest that the accuracy test does have more power for particular types of signal. [TODO: signal in scale? heavy tails?]

A very important point is the ease of implementation. The need for cross validation of the accuracy test greatly increases its computational complexity. Moreover, anyone who has actually implemented tests with discrete statistics, will attest they are considerably harder to implement. This is because their unforgiveness to the type of inequality. Indeed, replacing a weak inequality with a strong inequality may considerably change the results. This is not the case for continuous test statistics.

Given all the above, we find the popularity of accuracy tests quite puzzling. We believe this is due to a reversal of the inference cascade. Researchers first fit a classifier, and then ask if the classes are any different. Were they to start by asking if classes are any different, and only then try to classify, then location tests would naturally arise as the preferred method. As put by Ramdas et al. [2016]:

The recent popularity of machine learning has resulted in the extensive teaching and use of prediction in theoretical and applied communities and the relative lack of awareness or popularity of the topic of Neyman-Pearson style hypothesis testing in the computer science and related “data science” communities.

References

- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES B*, 57:289–289, 1995.
- Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu. *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. John Wiley & Sons, Aug. 2011. ISBN 978-0-470-53986-6.
- R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. Quan-

239 tifying spatial pattern similarity in multivariate analysis using functional
240 anisotropy. *arXiv:1605.03482 [q-bio]*, May 2016.

241 P. Golland and B. Fischl. Permutation tests for classification: towards statis-
242 tical significance in image-based studies. In *IPMI*, volume 3, pages 330–341.
243 Springer, 2003.

244 T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learn-*
245 *ing*. Springer, July 2003. ISBN 0-387-95284-5.

246 J. Hemerik and J. Goeman. Exact testing with random permutations.
247 *arXiv:1411.7565 [math, stat]*, Nov. 2014.

248 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for
249 prediction error in microarray classification using resampling. *Statistical*
250 *Applications in Genetics and Molecular Biology*, 7(1), 2008.

251 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional
252 brain mapping. *Proceedings of the National Academy of Sciences of the*
253 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,
254 1091-6490. doi: 10.1073/pnas.0600244103.

255 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
256 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
257 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*
258 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
259 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
260 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.

261 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
262 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
263 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.

264 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
265 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
266 P. Belin. The human voice areas: Spatial organization and inter-individual
267 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
268 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.

269 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
270 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*
271 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
272 106652702760138592.

- 273 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
274 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 275 J. Schäfer, K. Strimmer, and others. A shrinkage approach to large-scale co-
276 variance matrix estimation and implications for functional genomics. *Sta-*
277 *tistical applications in genetics and molecular biology*, 4(1):32, 2005.
- 278 R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the
279 Use of DNA Microarray Data for Diagnostic and Prognostic Classification.
280 *Journal of the National Cancer Institute*, 95(1):14–18, Jan. 2003. ISSN
281 0027-8874, 1460-2105. doi: 10.1093/jnci/95.1.14.
- 282 M. S. Srivastava. On testing the equality of mean vectors in high dimension.
283 *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1):
284 31–56, June 2013. ISSN 2228-4699. doi: 10.12697/ACUTM.2013.17.03.
- 285 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
286 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
287 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 288 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
289 ing correction in classification-based multi-voxel pattern analysis (MVPA):
290 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
291 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.
- 292 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
293 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
294 caveats, and guidelines. working paper or preprint, June 2016.

295 A Analysis pipeline

296 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 297 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 298 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions³ are
 299 centred around a unique voxel, the voxel index v also serves as a unique
 300 region index. Algorithm 1 computes a region-wise test statistic, which is
 301 compared to its permutation null distribution computed by Algorithm 2.

Algorithm 1: Compute a group parametric map.

Data: fMRI scans, and experimental design.
Result: Brain map of group statistics: $\{\bar{T}_v\}_{v=1}^V$

```

1 for  $v \in 1, \dots, V$  do
2   for  $i \in 1, \dots, I$  do
3      $T_{i,v} \leftarrow$  test statistic for subject  $i$  in a region centered at  $v$ .
4    $\bar{T}_v \leftarrow \frac{1}{I} \sum_{i=1}^I T_{i,v}$ .
```

Algorithm 2: Compute a permutation p-value map.

Data: fMRI scans of 20 subjects, experimental design.
Result: Brain map of permutation p-values: $\{p_v\}_{v=1}^V$

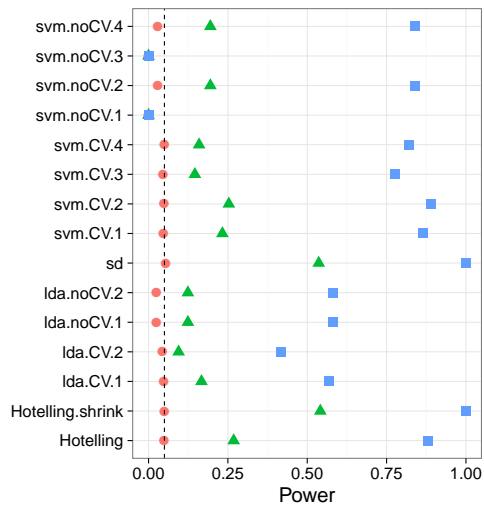
```

1 for  $s \in 1, \dots, S$  do
2   permute labels;
3    $\bar{T}_v^s \leftarrow$  parametric map
```

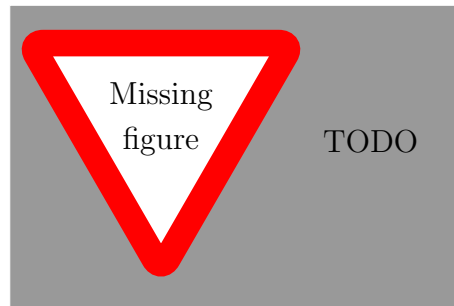
³*searchlight* or *sphere* in the MVPA parlance

B More Simulations

Figure 3: [TODO].



(a) 2 Folds.

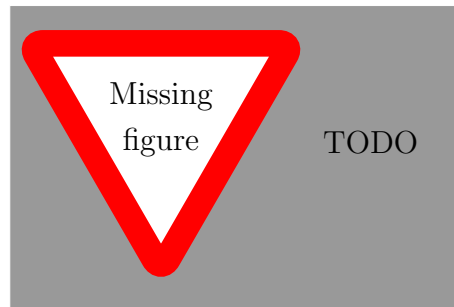


(b) 20 Folds.

Figure 4: [TODO].

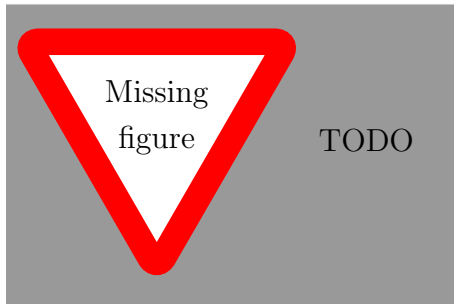


(a) Scale Change.

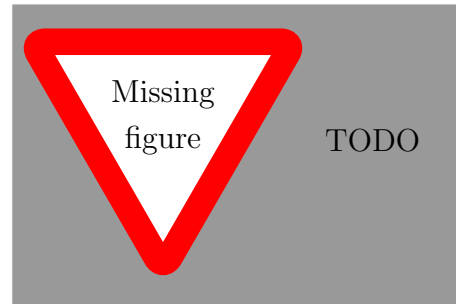


(b) t Null

Figure 5: [TODO].

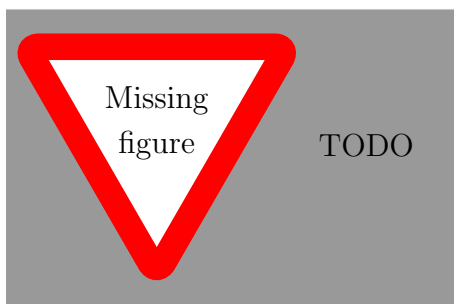


(a) Compound symmetry

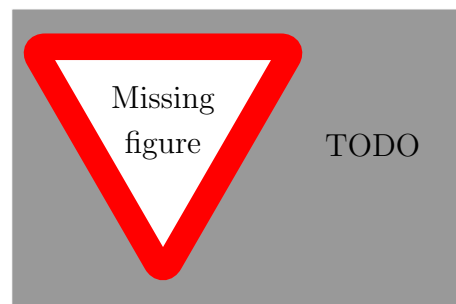


(b) AR(1)

Figure 6: [TODO].



(a) $n=400$



(b) ?