

Better-than-chance classification for signal detection

Jonathan Rosenblatt Roei Gilron Roy Mukamel

August 2, 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]. Examples in 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], or *pattern discrimination* in Pereira et al. [2009].

¹Based on GoogleScholar. Accesses on 26.7.2016.

23 This same signal detection task can be also approached as a two-group
 24 multivariate test: Inferring that a region encodes vocal/non-vocal informa-
 25 tion, is essentially inferring that the spatial distribution of brain activations
 26 is different given a vocal/non-vocal stimulus. As put in Pereira et al. [2009]:

27 ... the problem of deciding whether the classifier learned to dis-
 28 criminate the classes can be subsumed into the more general ques-
 29 tion as to whether there is evidence that the underlying distribu-
 30 tions of each class are equal or not.

31 A practitioner may then call upon a two-group location test such as Hotelling’s
 32 T^2 [Fujikoshi et al., 2011]. Alternatively, if the size of the brain region is too
 33 large compared to the number of observations, so that the spatial covariance
 34 cannot be fully estimated, then a high dimensional version of Hotelling’s test
 35 can be called upon, such as in Srivastava [2013] or Schäfer et al. [2005]. In
 36 contrast to *accuracy tests*, we call these *location tests*, a.k.a. *class comparison*
 37 in Simon et al. [2003].

38 At this point, it becomes unclear which is the preferred test. The compar-
 39 ison between location and accuracy tests was precisely the topic of Ramdas
 40 et al. [2016], who compared the Hotelling location test to the accuracy of
 41 *Fisher’s linear discriminant analysis* classifier (LDA) [Hastie et al., 2003].
 42 Using an asymptotic analysis, Ramdas et al. [2016] concluded that accuracy
 43 and location tests are equivalent with respect to their order of convergence
 44 to a consistent test, while they differ in constants. Judging by rate of conver-
 45 gence alone, this result may suggest that not much is (asymptotically) lost by
 46 using an accuracy test. On the other hand, asymptotic relative efficiency mea-
 47 sures (ARE) such as *Pitman’s*, *Bahadur’s*, *Lehman-Sheffe’s*, always assume
 48 equivalent convergence rates [van der Vaart, 1998].

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

54 The main argument for our statement rests upon the observation that
 55 with typical sample sizes, the accuracy test statistic is highly discrete. Dis-
 56 crete test statistics are known to be conservative [?], since they cannot ex-
 57 haust the permissible false positive rate. For accuracy tests, the degree of
 58 discretization is governed by the number of samples. In our running neu-
 59 roscience example [Gilron et al., 2016], the classification is performed based
 60 on 40 trials, so that the test statistic may assume only 40 possible values.
 61 This number of examples is not unusual if considering this is the number of

62 subject in a genetic study, or the number of trial-repeats in an fMRI brain
63 scan.

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

70 2 Problem setup

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

76 Given n pairs of (x_i, y_i) , typically assumed i.i.d., a location test amounts
77 to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$ (or
78 at least the same location). I.e., the multivariate voxel activation pattern
79 has the same distribution when given a vocal stimulus, as when given a non-
80 vocal stimulus. An accuracy test amounts to learning a predictive model $\hat{f}(x)$
81 from some assumed model class $\hat{f} \in \mathcal{F}$. The prediction accuracy, denoted
82 $T_{\hat{f}}^{acc}$, is defined as the probability of a given classifier \hat{f} of making a correct
83 prediction $T_{\hat{f}}^{acc} := Prob(\hat{f}(x) = y)$ when given a new, randomly drawn data
84 point, (x, y) . A statistically significant “better than chance” estimate of $T_{\hat{f}}^{acc}$
85 is evidence that the classes are distinct.

86 2.1 Candidate Tests

87 The design of a permutation test using the prediction accuracy, requires the
88 following design choices:

- 89 1. How to estimate accuracy?
- 90 2. Is the statistic cross validated or not?
- 91 3. For a K-fold cross validated test statistic: should the data be refolded
92 in each permutation?
- 93 4. Permute labels of features?

94 5. For a K-fold cross validated test statistic: should the data folding bal-
95 anced? (a.k.a. stratified).

96 6. How many folds?

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

101 **How to estimate accuracy?** Given a predictor \hat{f} , a natural test statis-
102 tic is some estimate of its accuracy $T_{\hat{f}}^{acc}$. Complicating matters: very low
103 accuracies, even 0, is evidence that the classes are separated, and we only
104 need to invert the predictions. We can thus consider $|T_{\hat{f}}^{acc} - 0.5|$ as the test
105 statistic. This, however, implies that if the classes are identical, random
106 guessing has a 0.5 accuracy. This is not true if the classes are not balanced.
107 The chance level in which case is the prevalence of the dominant class, we
108 denote by \hat{p}_{max} . This suggests the following test statistic $|T_{\hat{f}}^{acc} - \hat{p}_{max}|$. Since
109 we will be aggregating these statistic over random data sets where the dom-
110 inant class may have varying frequencies, it seems appropriate to standard-
111 ize the scale of this statistic. We thus also consider the z-scored accuracy:
112 $|T_{\hat{f}}^{acc} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$.

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

122 **Refolding?** The standard practice in neuroimaging is to refold the data
123 after each permutation [Pereira et al., 2009]. This is imperative if permuting
124 labels while aiming at balanced data folds. This is not, however, imperative
125 in general. For simplicity, we will adhere to the standard practice of refolding
126 the data within each permutation.

127 **Permute labels of features?** While seemingly identical, the compound-
 128 ing of permutations with data foldings renders these two approaches distinct.
 129 As an example, consider balanced (stratified) K-fold cross validation where
 130 the initial data folding is balanced. After a label permutation, the original
 131 folds will probably not be balanced. If the *features* are permuted, then the
 132 labels conserve their original fold assignments, and the original folds are bal-
 133 anced after each permutation. Since we only report results while refolding the
 134 data in each permutation, then the only difference between permuting labels
 135 and permuting features seems to be a computational one. We thus adhere
 136 to the more common, albeit less efficient practice, of permuting labels.

137 **Balanced folding?** As already implied, a standard practice when cross
 138 validating is to constrain the data folds to be balanced (i.e. stratified). This
 139 is well justified when aiming at unbiased accuracy estimation. This also
 140 simplifies matter when aiming at signal detection, as can be seen from the
 141 above discussion of the appropriate test statistic. On the other hand, it
 142 may complicate matters, as can be seen from the above discussion on label
 143 versus feature permutation. We will report results with both balanced and
 144 unbalanced data foldings, only to discover, it does not really matter.

145 **How many folds?** Different authors suggest different rules for the num-
 146 ber of folds. We will be varying the number of folds. This will affect the
 147 concentration of permutation distribution of the estimated accuracy, which
 148 will have a crucial effect on the conservativeness of the accuracy test. Our
 149 intuition suggests that since more folds imply a less concentrated estimate,
 150 then leave-one-out should be the less conservative, and 2-fold should be the
 151 most conservative.

152 There are indeed many design choices when performing a permutation test
 153 using a cross validated statistic. The subset of tests we will be comparing is
 154 collected for convenience in Table 1.

155 3 Controlling the False Positive Rate

156 We start by verifying that the battery of tests in Table 1 control the false
 157 positive rate at the desired 0.05 level, with varying conservativeness levels.
 158 Figure 1 demonstrates that this is indeed the case. All our candidate tests
 159 control the type I error, with varying degrees of conservativeness. In particu-
 160 lar: (a) if the folds are balanced or not, (b) if the tuning parameters of some
 161 test statistic are varied, (d) if the number of folds is varied.

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.

162 4 Power

163 Having established that all of the tests in our battery control the false positive rate, it remains to be seen if they have similar power; Especially, when
164 comparing the power of the location tests to the accuracy tests. The theoretical results of Ramdas et al. [2016] suggest that power should be of the
165 same order. On the other hand, the results of our previous sections suggest
166 that the conservativeness of some of the considered tests can be considerable,
167 rendering them underpowered.

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

171 We see by now that the use of accuracy tests for signal detection is underpowered compared to location tests. Simulations alone cannot, however,
172 support such a universal statement. We will thus verify on a neuroimaging
173 dataset, and discuss the causes for this phenomenon with implications on the
174

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].



scope of our statement.

5 Neuroimaging Example

Figure 2 is an application of both a location and an accuracy test 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 $n = 40$ trials in each scan. The study was rather large and consisted of about 200 subjects. The data was kindly made available by the authors at the OpenfMRI website².

We perform permutation inference using the pipeline of Stelzer et al. [2013], which was also used in Gilron et al. [2016]. For completeness, the pipeline is described in Appendix A. To demonstrate our point, we compare the *sd* location test with the *svm.cv.1* accuracy test (see Table 1 for the definition of these statistics).

In agreement with our simulation results, the location test (*sd*) discovers

²<https://openfmri.org/>

190 more brain regions when compared to an accuracy test (*svm.cv.1*). The
 191 former discovers 1,232 regions, while the latter only 441, as reported in
 192 Figure 2. We emphasize that both test statistics were compared with the
 193 same permutation scheme, and the same error controls, so that any difference
 194 in detections is due to their different power.

195 Having established that accuracy tests are underpowered both in simula-
 196 tion and in application, we wish to identify the conditions under which this
 197 will occur, and discuss implications on the practice of accuracy tests.

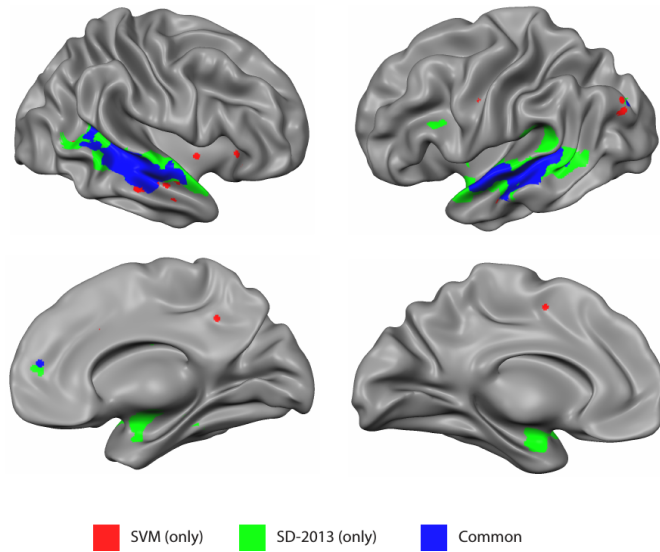


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centres of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a location test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost parameter 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, 399 of which are common to both. For the details of the analysis see Appendix A and Gilron et al. [2016].

198 6 Discussion

199 We have set out to understand which of the tests is more powerful: the
 200 accuracy test or the location test. Using simulations, we have concluded
 201 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. [TODO: relate to large sample simulation].

Olivetti et al. [2012] and Olivetti et al. [2014] also looked into a similar problem as we do, namely, what is the preferred accuracy test? They propose a new test they call an *independence test*, and demonstrate by simulation that it has more power than other accuracy tests, and can deal with non-balanced data sets. We did not include this test in the battery we compared, but we note the following: (a) The independence test of Olivetti et al. [2012] relies on a discrete test statistic. This means that in the cases that the accuracy test is called upon for discriminating populations, it will probably be under-powered compared to location tests. (b) The problem of the accuracy test with unbalanced data-sets, which motivates Olivetti et al. [2012]’s independence test, can also be remedied by replacing the accuracy statistic with its z-score, as suggested in Section 2.1.

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 in decoding applications, like brain-machine interfaces, where the localization of a signal is not enough. Clinical diagnosis is another application, where the presence of a medical condition is “predicted” from imaging data. [e.g. Olivetti et al., 2012, Wager et al., 2013]

Secondly, not all signals are manifested in a shift of the null distribution. Put differently, the preferred alternative to an accuracy test is not always a location test. Indeed, one may consider signal, i.e. effects, as a change in scale, such as the *spiked covariance* model. In this case, other-than-Hotelling type tests are appropriate [TODO: cite change in covariance alternative]. Tests have been proposed even when the nature of the difference between populations is left unspecified [e.g. Gretton et al., 2012]. The fact that in our neuroimaging example (Section 5) some brain regions were detected with the accuracy test, and not the location test, is consistent with this observation. On the other hand, the far greater power of the location test, certainly in our example, does serve as empirical evidence that changes in location are a prevalent phenomenon. [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, mistakenly replacing a weak inequality with a strong inequality in one’s program may considerably change

243 the results. This is not the case for continuous test statistics.

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

250 The recent popularity of machine learning has resulted in the ex-
251 tensive teaching and use of prediction in theoretical and applied
252 communities and the relative lack of awareness or popularity of
253 the topic of Neyman-Pearson style hypothesis testing in the com-
254 puter science and related “data science” communities.

255 References

- 256 Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a prac-
257 tical and powerful approach to multiple testing. *JOURNAL-ROYAL STA-*
258 *TISTICAL SOCIETY SERIES B*, 57:289–289, 1995.
- 259 Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu. *Multivariate Statistics: High-*
260 *Dimensional and Large-Sample Approximations*. John Wiley & Sons, Aug.
261 2011. ISBN 978-0-470-53986-6.
- 262 R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. Quan-
263 tifying spatial pattern similarity in multivariate analysis using functional
264 anisotropy. *arXiv:1605.03482 [q-bio]*, May 2016.
- 265 P. Golland and B. Fischl. Permutation tests for classification: towards statis-
266 tical significance in image-based studies. In *IPMI*, volume 3, pages 330–341.
267 Springer, 2003.
- 268 A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola.
269 A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13
270 (Mar):723–773, 2012. ISSN ISSN 1533-7928.
- 271 T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learn-*
272 *ing*. Springer, July 2003. ISBN 0-387-95284-5.
- 273 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for
274 prediction error in microarray classification using resampling. *Statistical*
275 *Applications in Genetics and Molecular Biology*, 7(1), 2008.

- 276 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional
277 brain mapping. *Proceedings of the National Academy of Sciences of the*
278 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,
279 1091-6490. doi: 10.1073/pnas.0600244103.
- 280 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
281 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
282 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*
283 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
284 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
285 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.
- 286 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the
287 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.
288 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.
- 289 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
290 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
291 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 292 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
293 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
294 P. Belin. The human voice areas: Spatial organization and inter-individual
295 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
296 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 297 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
298 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*
299 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
300 106652702760138592.
- 301 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
302 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 303 J. Schäfer, K. Strimmer, and others. A shrinkage approach to large-scale co-
304 variance matrix estimation and implications for functional genomics. *Sta-*
305 *tistical applications in genetics and molecular biology*, 4(1):32, 2005.
- 306 R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the
307 Use of DNA Microarray Data for Diagnostic and Prognostic Classification.
308 *Journal of the National Cancer Institute*, 95(1):14–18, Jan. 2003. ISSN
309 0027-8874, 1460-2105. doi: 10.1093/jnci/95.1.14.

- 310 M. S. Srivastava. On testing the equality of mean vectors in high dimension.
 311 *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1):
 312 31–56, June 2013. ISSN 2228-4699. doi: 10.12697/ACUTM.2013.17.03.
- 313 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
 314 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
 315 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 316 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
 317 ing correction in classification-based multi-voxel pattern analysis (MVPA):
 318 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
 319 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.
- 320 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,
 321 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-
 322 2.
- 323 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
 324 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
 325 caveats, and guidelines. working paper or preprint, June 2016.
- 326 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.
 327 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-*
 328 *nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:
 329 10.1056/NEJMoA1204471.

330 A Analysis pipeline

331 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 332 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 333 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions³ are
 334 centred around a unique voxel, the voxel index v also serves as a unique
 335 region index. Algorithm 1 computes a region-wise test statistic, which is
 336 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].

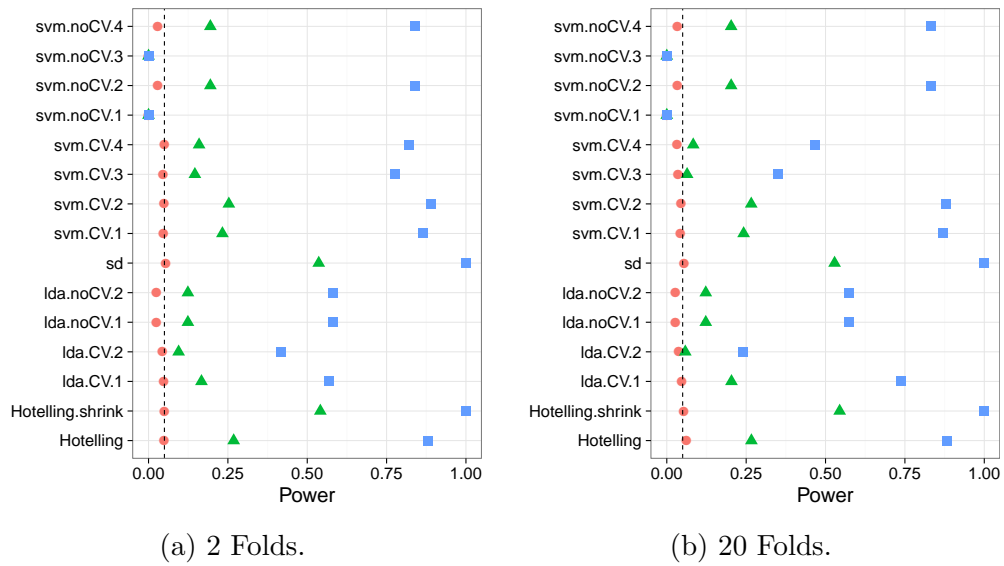


Figure 4: [TODO].

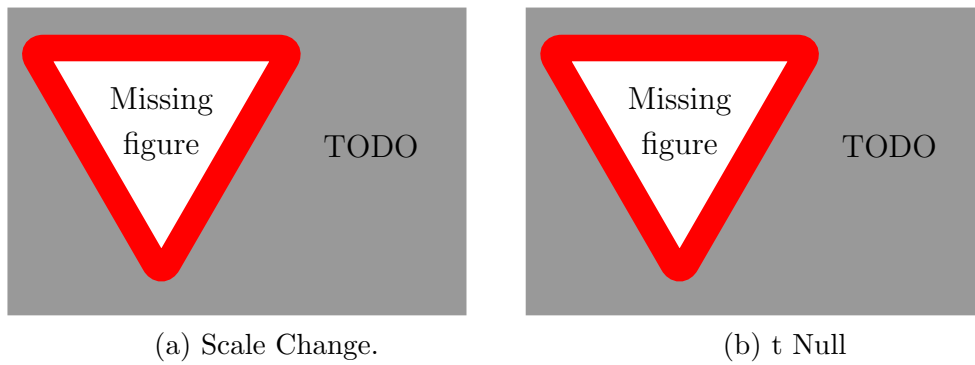
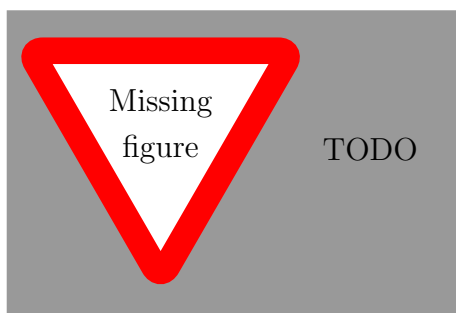


Figure 5: [TODO].

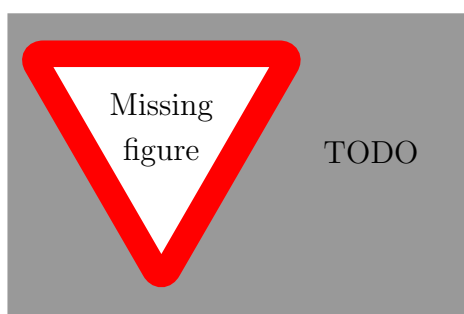


(a) Compound symmetry

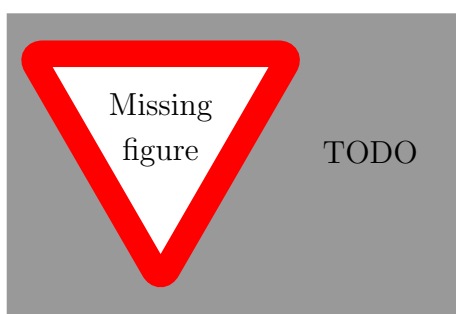


(b) AR(1)

Figure 6: [TODO].



(a) $n=400$



(b) ?