

# Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt      Roei Gilron      Roy Mukamel

August 6, 2016

## Abstract

[TODO]

## 1 Introduction

A common workflow in 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 neuroscientific literature include Golland and Fischl [2003], Pereira et al. [2009], Varoquaux et al. [2016], and especially the recently popularized *multivariate pattern analysis* (MVPA) framework of Kriegeskorte et al. [2006]. This practice is also observed in the genetics literature, but to a lesser extent [Radmacher et al., 2002, Jiang et al., 2008].

To fix ideas, we will adhere to a concrete example. In Gilron et al. [2016], the authors seek to detect brain regions which encode differences between vocal and non-vocal stimuli. Following the MVPA 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].

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 is different given a vocal/non-vocal stimulus. As put in Pereira et al. [2009]:

... the problem of deciding whether the classifier learned to discriminate the classes can be subsumed into the more general question as to whether there is evidence that the underlying distributions of each class are equal or not.

A practitioner may then call upon a two-group location test such as Hotelling’s  $T^2$  [Anderson, 2003]. Alternatively, if the size of a brain region is too large compared to the number of observations, so that the spatial covariance cannot be fully estimated, then a high dimensional version of Hotelling’s test can be called upon, such as in Schäfer and Strimmer [2005] or Srivastava [2013]. For brevity, and in contrast to *accuracy tests*, we will call these two-sample multivariate tests simply *location tests*, also termed *class comparisons* in Simon et al. [2003].

At this point, it becomes unclear which is preferable: a location test or an accuracy test? The former with a heritage dating back to Hotelling [1931], and the latter being extremely popular, as the 959 citations<sup>1</sup> of Kriegeskorte et al. [2006] suggest.

The comparison between location and accuracy tests was precisely the goal of Ramdas et al. [2016], who compared the  $T^2$  location test to the accuracy of *Fisher’s linear discriminant analysis* classifier (LDA). By comparing the rates of convergence of the powers to 1, Ramdas et al. [2016] concluded that accuracy and location tests are rate equivalent. Judging by convergence rates alone, not much is (asymptotically) lost by using an accuracy test.

Asymptotic relative efficiency measures (ARE) are typically used by statisticians to compare between test statistics with similar rates [van der Vaart, 1998]. The ARE between Hotelling’s  $T^2$  (location) test and Fisher’s LDA (accuracy) test in Ramdas et al. [2016] is lower bounded by  $\sqrt{2\pi} \approx 2.5$ . This means that Fisher’s LDA requires at least 2.5 more samples to achieve the same (asymptotic) power than the  $T^2$  test. In this light, the accuracy test is remarkably inefficient compared to the location test. For comparison, the t-test is only 1.04 more (asymptotically) efficient than Wilcoxon’s rank-sum test [Lehmann, 2009], so that an ARE of 2.5 is strong evidence in favor of the location test.

Before discarding accuracy tests as inefficient, we recall that Ramdas et al. [2016] analyzed a *half-sample* holdout. The authors thus conjecture that a leave-one-out approach, which makes more efficient use of the data, may have better performance. Also, the analysis in Ramdas et al. [2016] is asymptotic. This eschews the discrete nature of the accuracy statistic, which will be shown to have crucial impact. Since typical sample sizes in neuroscience are not large, we seek to study which test is to be preferred

---

<sup>1</sup>GoogleScholar. Accessed on Aug 4, 2016.

in finite samples? Our conclusion will be quite simple: *location tests almost always have more power than accuracy tests.*

The main argument for our statement rests upon the observation that with typical sample sizes, the accuracy test statistic is highly discrete. Discrete test statistics are known to be conservative [Hemerik and Goeman, 2014], since they are insensitive to mild perturbations of the data, and they cannot exhaust the permissible false positive rate. The degree of discretization is governed by the number of samples. In our neuroscience example from [Gilron et al., 2016], the classification is performed based on 40 trials, so that the test statistic may assume only 40 possible values. This number of examples is not unusual if considering this is the number of subjects, or the number of trial-repeats in an neuroimaging study.

The discretization effect is aggravated if the test statistic is highly concentrated. For an intuition consider the usage of a the *resubstitution accuracy* as a test statistic. This statistic simply means that the accuracy is not cross validated. If the data is high dimensional, the resubstitution accuracy will be very high due to over fitting [McLACHLAN, 1976, Theorem 1]. In an extreme case, the resubstitution accuracy will be 1 for the observed data, but also for any permutation. The concentration of resubstitution accuracy near 1, and its discreteness, render this test completely useless, with a power of 0.

To compare the power of accuracy tests and location tests in finite samples, we perform a simulation study of a battery of test statistics. The main findings are reported in Section 4, and the intuition for our findings is provided in Section 6, but first, the problem’s setup.

## 2 Problem setup

Let  $y \in \mathcal{Y}$  be a class encoding. Let  $x \in \mathcal{X}$  be a  $p$  dimensional feature vector. In our vocal/non-vocal example we have  $\mathcal{Y} = \{-1, 1\}$  and  $p$ , the number of voxels in a brain region so that  $\mathcal{X} = \mathbb{R}^{27}$ .

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

101 that the classes are distinct.

## 102 2.1 Candidate Tests

103 The design of a permutation test using the prediction accuracy, requires the  
104 following design choices:

- 105 1. How to estimate accuracy?
- 106 2. Is the statistic cross validated or not?
- 107 3. For a K-fold cross validated test statistic: should the data be refolded  
108 in each permutation?
- 109 4. Permute labels of features?
- 110 5. For a K-fold cross validated test statistic: should the data folding be bal-  
111 anced (a.k.a. stratified)?
- 112 6. How many folds?

113 We will now address these questions while bearing in mind that unlike the  
114 typical supervised learning setup, we are not interested in an unbiased esti-  
115 mate of the prediction error, but rather in the mere detection of a difference  
116 between two groups.

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

129 **Cross validate or not?** Were we interested in an unbiased estimator of  
130 the prediction error, there is no question that some independent validation  
131 is in order. Since we are merely interested in detecting a difference between  
132 classes, a biased error estimate is not an issue provided that bias is consistent  
133 over all permutations. The underlying intuition is that if the exact same  
134 computation is performed over all permutations, then a permutation test  
135 will be “fair”, i.e., will not inflate the false positive rate. We will thus be  
136 considering both cross validated accuracies, and resubstitution accuracies as  
137 our test statistics, a.k.a. *resubstitution classification*.

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

143 **Permute labels of features?** While seemingly identical, the compound-  
144 ing of permutations with data foldings renders these two approaches distinct.  
145 As an example, consider balanced (stratified) K-fold cross validation where  
146 the initial data folding is balanced. After a label permutation, the original  
147 folds will probably not be balanced. If the *features* are permuted, then the  
148 labels conserve their original fold assignments, and the original folds are bal-  
149 anced after each permutation. Since we only report results while refolding  
150 the data in each permutation, then the only difference between permuting  
151 labels and permuting features seems to be a computational one. We thus  
152 adhere to the more common, albeit computationally less efficient practice of  
153 permuting labels.

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

162 **How many folds?** Different authors suggest different rules for the num-  
163 ber of folds. We will be varying the number of folds. This will affect the  
164 concentration of permutation distribution of the estimated accuracy, which

will have a crucial effect on the conservativeness of the accuracy test. Our intuition suggests that since more folds imply a less concentrated estimate, then leave-one-out should be the less conservative, and 2-fold should be the most conservative.

The of tests we will be comparing is collected for convenience in Table 1.

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

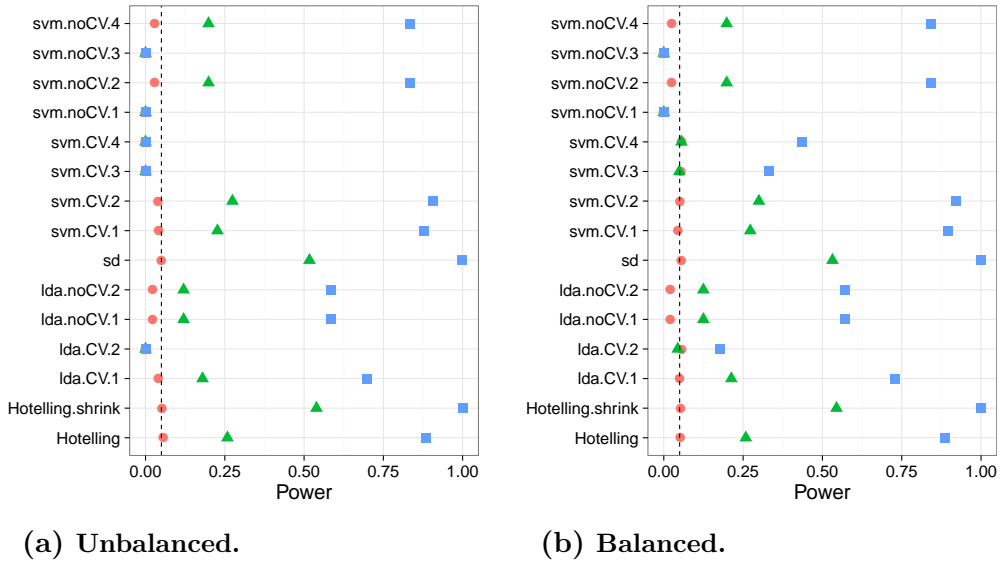
Table 1: This table enumerates the various test statistics we will be studying. Three are location tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group  $T^2$  statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer and Strimmer [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 ( $|T_f^{acc} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$ , see Section 2.1). Another example is *lda.noCV.1*, which is Fisher’s LDA, returning the resubstitution accuracy, without cross validation, and without z-scoring.

### 3 Controlling the False Positive Rate

Figure 1 demonstrates that all of the tests considered conserve the desired 0.05 false positive rate, up to varying levels of conservatism. This can be seen from the fact that the probability of rejection is no larger than 0.05 in the absence of any effect, encoded by a red circle. This is true, in particular if: (a) the folds are balanced or not, (b) the tuning parameters of some test

176 statistic are varied, (d) the number of folds is varied. We also observe that the  
 177 most conservative tests are the accuracy tests that are not cross validated.  
 178 We return to this matter in the Discussion.

*Figure 1: The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in Table 1. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix B. Cross-validation was performed with balanced (stratified) and unbalanced data folding. See sub-captions.*



## 179 4 Power

180 Having established that all of the tests in our battery control the false pos-  
 181 itive rate, it remains to be seen if they have similar power– especially when  
 182 comparing the power of location tests to accuracy tests. From the simulation  
 183 results reported in Appendix C we collect the following insights:

- 184 1. Location tests have more power than accuracy tests in all our configu-  
 185 rations.
- 186 2. The conservativeness decays as the sample grows (Figure 6), supporting  
 187 the statement that discretization is responsible for power loss.
- 188 3. The power is may increase or decrease with the number of folds (Fig-  
 189 ure 3). [TODO:effect of n.folds.]

- 190 4. The z-scoring of the accuracies was introduced to deal with unbalanced  
191 foldings. If the z-scoring has any effect at all, it merely kills power.  
192 There is really no reason to use it.
- 193 5. Both accuracy and location tests are inappropriate for scale alternatives  
194 (Figure 5a). This was to be expected and is reported mostly as a sanity  
195 check.
- 196 6. The presence of heavy tails (Figure 5b) may reduce power, but does  
197 not quantitatively change results.
- 198 7. Balancing typically has no effect. It increased power only for the z-  
199 scored statistics (Figure 1). This is surprising given they were precisely  
200 designed to deal with the presence of imbalance.
- 201 8. Varying the accuracy test’s tuning parameter, such as the cost (i.e.  
202 margins) has no effect on the power of the test.
- 203 9. ... [TODO: correlation between voxels].

204 The major insight from simulations is that the use of accuracy tests for  
205 signal detection is underpowered compared to location tests. We now verify  
206 this finding on a neuroimaging dataset.

## 207 5 Neuroimaging Example

208 Figure 2 is an application of both a location and an accuracy test to the data  
209 of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI  
210 data while subjects were exposed to the sounds of human speech (vocal),  
211 and other non-vocal sounds. Each subject was exposed to 20 sounds of each  
212 type, totaling in  $n = 40$  trials in each scan. The study was rather large and  
213 consisted of about 200 subjects. The data was kindly made available by the  
214 authors at the OpenfMRI website<sup>2</sup>.

215 We perform group inference using within-subject permutations using the  
216 pipeline of Stelzer et al. [2013], which was also reported in Gilron et al. [2016].  
217 For completeness, the pipeline is described in Appendix A. To demonstrate  
218 our point, we compare the *sd* location test with the *svm.cv.1* accuracy test  
219 (see Table 1 for the definition of these statistics).

220 In agreement with our simulation results, the location test (*sd*) discovers  
221 more brain regions when compared to an accuracy test (*svm.cv.1*). The  
222 former discovers 1,232 regions, while the latter only 441, as depicted in

---

<sup>2</sup><https://openfmri.org/>



Figure 2. We emphasize that both test statistics were compared with the same permutation scheme, and the same error controls, so that any difference in detections is due to their different power.

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

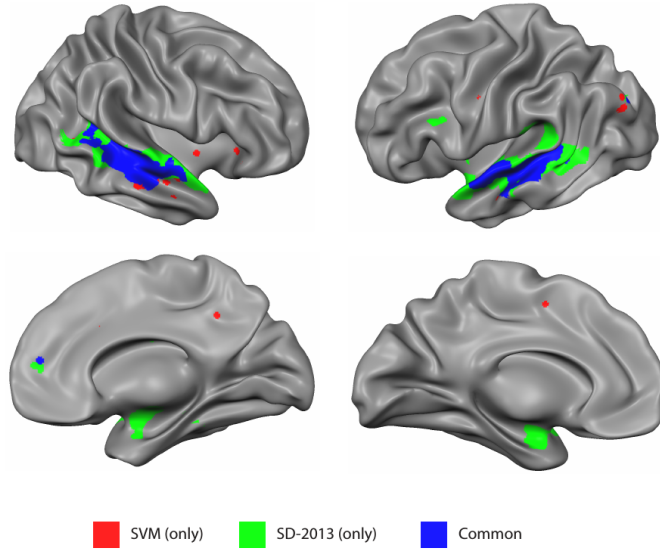


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (svm.cv.1), and a location test (sd). svm.cv.1 was computed using 5-fold cross validation, and a cost parameter 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].

## 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 several phenomena: (a) Discretization introduced in finite samples by the accuracy test statistic.

(b) Inefficient use of the data for the validation holdout set. In our high dimensional setup, we also confirmed that high-dimensional versions of the  $T^2$  test, such as Srivastava [2013] or Schäfer and Strimmer [2005] are preferable over the original  $T^2$ .

The sensitivity of the power to the number of folds suggests that most of the power is lost due to the discretization and not to the holdout. The degree of discretization is governed by the sample size. For this reason, an asymptotic analysis such as Ramdas et al. [2016] may uncover the holdout inefficiency, but will not uncover the discretization effect. The practical advice for the practitioner, is that for the purpose of signal detection, there is typically a multivariate test (be it a location test or other), that is more powerful than an accuracy test. There is also a good chance that it would be easier to implement, since no validation will be involved.

## 6.1 Neyman-Pearson Classification

[TODO]

## 6.2 A good accuracy test

In Section 6.5 we discuss cases where an accuracy test cannot replace a location test. For such cases we collect some conclusions from our simulations on the best practices for accuracy tests.

1. The conservativeness due to discretization decreases with sample size.
2. Cross-validate. For moderate sample sizes, the power loss due to the holdout inefficiency is smaller than the power loss due to the concentration of the resubstitution accuracy.
3. Permuting features is easier than permuting labels. It allows to preserve balanced folds after a permutation without refolding.
4. There is no gain in z-scoring the accuracy scores.
5. Cross validated accuracy with balanced folds has more power than unbalanced folds. We currently have no intuition to offer for this phenomenon.
6. It is unclear what is the effect of the number of folds. More folds increase power by reducing the number of holdout samples. On the other hand, it increases the concentration of the accuracy statistic. Compounded with the discreteness of the accuracy statistic, this decreases power.

268 7. The value of the tuning parameters of a classifier do not matter.

## 269 6.3 Related Literature

270 Olivetti et al. [2012] and Olivetti et al. [2014] also looked into a similar  
271 problem as we do, namely, what is the preferred accuracy test? They propose  
272 a new test they call an *independence test*, and demonstrate by simulation that  
273 it has more power than other accuracy tests, and can deal with non-balanced  
274 data sets. We did not include this test in the battery we compared, but we  
275 note the following: (a) The independence test of Olivetti et al. [2012] relies on  
276 a discrete test statistic. This means that in the cases that the accuracy test is  
277 called upon for discriminating populations, it will probably be underpowered  
278 compared to location tests. (b) In contrast with the underlying motivation  
279 of Olivetti et al. [2012]’s independence test, we did not find that balancing  
280 the data folds is crucial for an accuracy test.

## 281 6.4 Non-linear predictors

282 It may be argued that accuracy tests permits the separation between classes  
283 in high dimensions, such as in *reproducing kernel Hilbert spaces* (RKHS) by  
284 using non-linear predictors. This is immaterial since group tests can also be  
285 performed in higher dimensions (see Gretton et al. [2012]).

## 286 6.5 Reservations

287 Some reservations to the generality of our findings are in order. Firstly, not  
288 all accuracy tests are concerned with signal detection. Indeed, it is possible  
289 that the purpose of the test is not to detect a difference between classes,  
290 but to actually test the performance of a particular classifier. Examples  
291 include brain decoding for machine interfaces, and clinical diagnosis, where  
292 the presence of a medical condition is “predicted” from imaging data. [e.g.  
293 Olivetti et al., 2012, Wager et al., 2013]

294 Secondly, not all signals are manifested in a shift of the null distribution  
295 Our focus on location tests is misleading. Perhaps Simon et al. [2003]’s *class*  
296 *comparison* is a more appropriate name, in that it does not only imply a  
297 shift alternative. Indeed, one may consider signal, i.e. effects, as a change in  
298 scale, such as the *spiked covariance* model. In this case, other-than-Hotelling  
299 type tests are appropriate [e.g. Nadler, 2008]. Tests have been proposed even  
300 when the nature of the difference between populations is left unspecified [e.g.  
301 Gretton et al., 2012]. The fact that in our neuroimaging example (Section 5)

302 some brain regions were detected with the accuracy test, and not the location  
303 test, is consistent with this observation.

304 The reservation to the reservation is that the far greater power of the  
305 location test, certainly in our example, does serve as an empirical evidence  
306 that changes in location are a prevalent phenomenon.

## 307 6.6 Ease of implementation

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

## 315 6.7 Epilogue

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

322       The recent popularity of machine learning has resulted in the ex-  
323       tensive teaching and use of prediction in theoretical and applied  
324       communities and the relative lack of awareness or popularity of  
325       the topic of Neyman-Pearson style hypothesis testing in the com-  
326       puter science and related “data science” communities.

## 327 References

- 328 T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-  
329       Interscience, Hoboken, NJ, 3 edition edition, July 2003. ISBN 978-0-471-  
330       36091-9.
- 331 Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a prac-  
332       tical and powerful approach to multiple testing. *JOURNAL-ROYAL STA-*  
333       *TISTICAL SOCIETY SERIES B*, 57:289–289, 1995.

- 334 R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. Quan-  
 335 tifying spatial pattern similarity in multivariate analysis using functional  
 336 anisotropy. *arXiv:1605.03482 [q-bio]*, May 2016.
- 337 P. Golland and B. Fischl. Permutation tests for classification: towards statis-  
 338 tical significance in image-based studies. In *IPMI*, volume 3, pages 330–341.  
 339 Springer, 2003.
- 340 A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A  
 341 Kernel Two-sample Test. *J. Mach. Learn. Res.*, 13:723–773, Mar. 2012.  
 342 ISSN 1532-4435.
- 343 J. Hemerik and J. Goeman. Exact testing with random permutations.  
 344 *arXiv:1411.7565 [math, stat]*, Nov. 2014.
- 345 H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Math-*  
 346 *ematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990.  
 347 doi: 10.1214/aoms/1177732979.
- 348 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for  
 349 prediction error in microarray classification using resampling. *Statistical*  
 350 *Applications in Genetics and Molecular Biology*, 7(1), 2008.
- 351 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional  
 352 brain mapping. *Proceedings of the National Academy of Sciences of the*  
 353 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,  
 354 1091-6490. doi: 10.1073/pnas.0600244103.
- 355 E. L. Lehmann. Parametric versus nonparametrics: two alternative method-  
 356 ologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN  
 357 1048-5252. doi: 10.1080/10485250902842727.
- 358 G. J. McLACHLAN. The bias of the apparent error rate in discriminant  
 359 analysis. *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-  
 360 3510. doi: 10.1093/biomet/63.2.239.
- 361 B. Nadler. Finite sample approximation results for principal component  
 362 analysis: A matrix perturbation approach. *The Annals of Statistics*, 36  
 363 (6):2791–2817, Dec. 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/  
 364 08-AOS618.
- 365 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with  
 366 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-  
 367 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*

- 368 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,  
369 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2  
370 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9\_6.
- 371 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the  
372 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.  
373 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.
- 374 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and  
375 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,  
376 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 377 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.  
378 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and  
379 P. Belin. The human voice areas: Spatial organization and inter-individual  
380 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–  
381 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 382 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for  
383 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*  
384 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/  
385 106652702760138592.
- 386 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy  
387 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 388 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance  
389 Matrix Estimation and Implications for Functional Genomics. *Statistical*  
390 *Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN  
391 1544-6115. doi: 10.2202/1544-6115.1175.
- 392 C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning.  
393 *IEEE Transactions on Information Theory*, 51(11):3806–3819, Nov. 2005.  
394 ISSN 0018-9448. doi: 10.1109/TIT.2005.856955.
- 395 R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the  
396 Use of DNA Microarray Data for Diagnostic and Prognostic Classification.  
397 *Journal of the National Cancer Institute*, 95(1):14–18, Jan. 2003. ISSN  
398 0027-8874, 1460-2105. doi: 10.1093/jnci/95.1.14.
- 399 M. S. Srivastava. On testing the equality of mean vectors in high dimension.  
400 *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1):  
401 31–56, June 2013. ISSN 2228-4699. doi: 10.12697/ACUTM.2013.17.03.

- 402 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high  
403 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.  
404 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 405 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-  
406 ing correction in classification-based multi-voxel pattern analysis (MVPA):  
407 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.  
408 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.
- 409 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,  
410 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-  
411 2.
- 412 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,  
413 and B. Thirion. Assessing and tuning brain decoders: cross-validation,  
414 caveats, and guidelines. working paper or preprint, June 2016.
- 415 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.  
416 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-  
417 nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:  
418 10.1056/NEJMoa1204471.

## 419 A Analysis pipeline

420 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in  
 421 Gilron et al. [2016]. Denoting by  $i = 1, \dots, I$  the subject index,  $v = 1, \dots, V$   
 422 the voxel index, and  $s = 1, \dots, S$  the permutation index. Since regions<sup>3</sup> are  
 423 centered around a unique voxel, the voxel index  $v$  also serves as a unique  
 424 region index. Algorithm 1 computes a region-wise test statistic, which is  
 425 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
```

---

<sup>3</sup>*searchlight* or *sphere* in the MVPA parlance



## 428 B Simulation Details

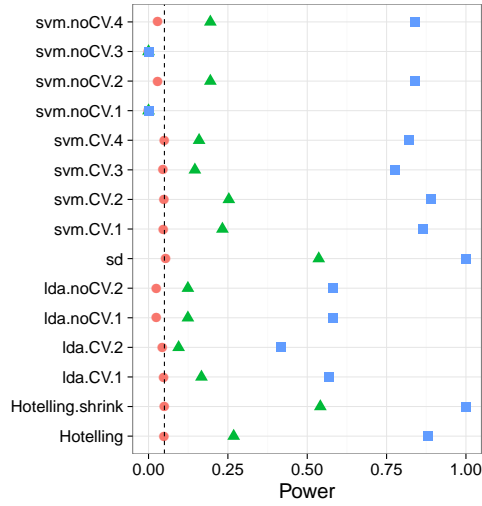
429 The following details are common to all the reported simulations, unless stated  
430 otherwise in a figure’s caption. The R code for the simulations can be found  
431 in [TODO].

432 Each simulation is based on 4,000 replications. In each replication, we  
433 generate  $n$  i.i.d. samples from a shift model  $\mathbf{x}_i = \mu \mathbf{y}_i^* + \eta_i$ . Where  $y_i^* = \{0, 1\}$   
434 is the class of subject  $i$  in dummy coding. Recalling that  $y_i = \{-1, 1\}$  is the  
435 class in effect coding, then clearly  $y_i = 2y_i^* - 1$ . The noise is distributed as  
436  $\eta_i \sim \mathcal{N}_p(0, \Sigma)$ . The sample size  $n = 40$ . The dimension of the data is  $p = 23$ .  
437 The covariance  $\Sigma = I$ . Effects, i.e. shifts  $\mu$ , are equal coordinate  $p$ -vectors  
438 with coordinates that vary over  $\mu \in \{0, 1/4, 1/2\}$ .

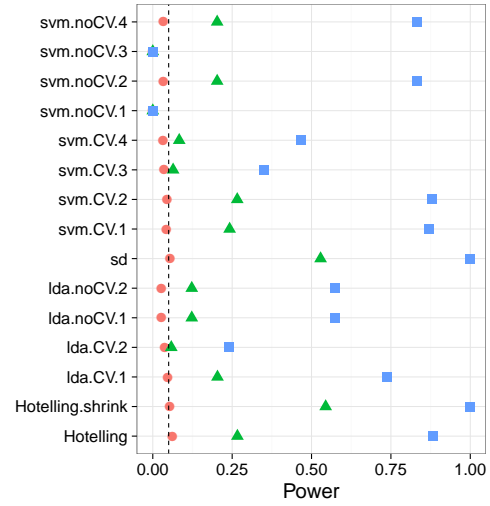
439 Having generated the data, we compute each of the test statistics in Ta-  
440 ble 1. For test statistics that require data folding, we used 8 folds. We then  
441 compute a permutation p-value by permuting the class labels, and recomput-  
442 ing each test statistic. We perform 400 such permutations. We then reject  
443 the  $\mu_i = 0$  null hypothesis if the permutation p-value is smaller than 0.05.  
444 The reported power is the proportion of replication where the permutation  
445 p-value falls below 0.05.

## C Simulation Results

Figure 3: *Simulation details in Appendix B except the changes in the sub-captions.*

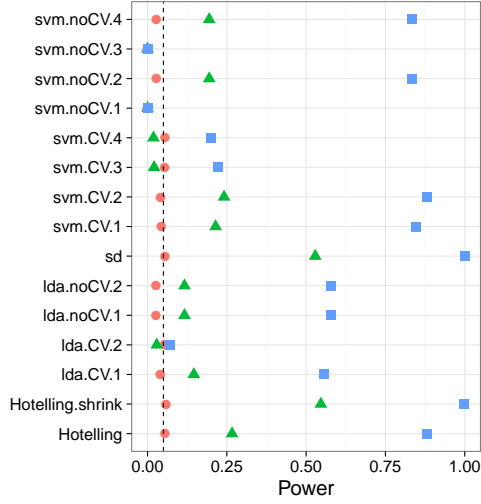


(a) 2-fold cross validation.  
Balanced folding.

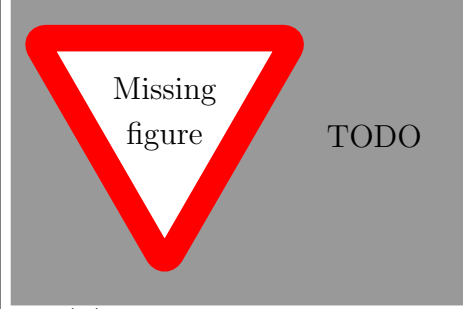


(b) 20-fold cross validation.  
Balanced folding

Figure 4: *Simulation details in Appendix B except the changes in the sub-captions.*

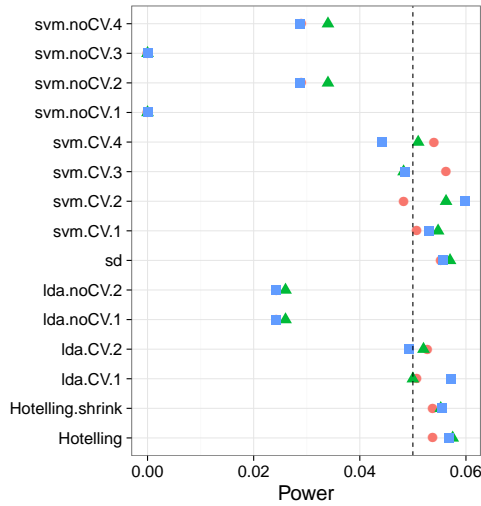


(a) **2-fold** cross validation.  
Unbalanced folding.

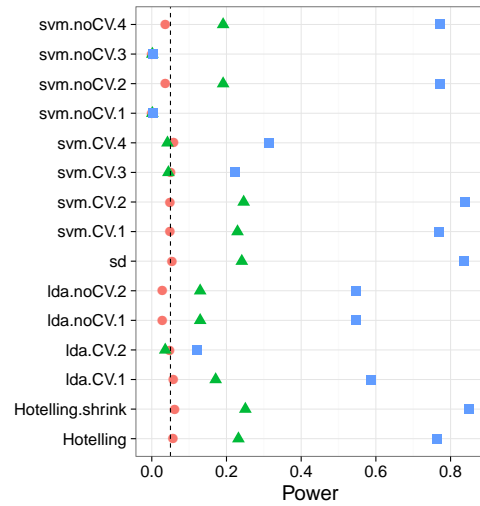


(b) **20-fold** cross validation.  
Unbalanced folding.

Figure 5: *Simulation details in Appendix B except the changes in the sub-captions.*

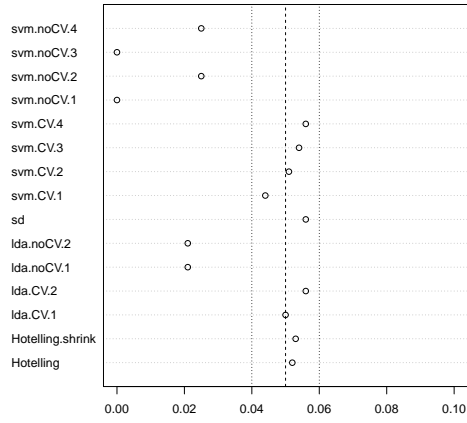


(a) **Scale Change:**  $\mathbf{x}_i = \eta_i * \mu^{\mathbf{y}_i^*}$   
so that the effect are a scale  
change.

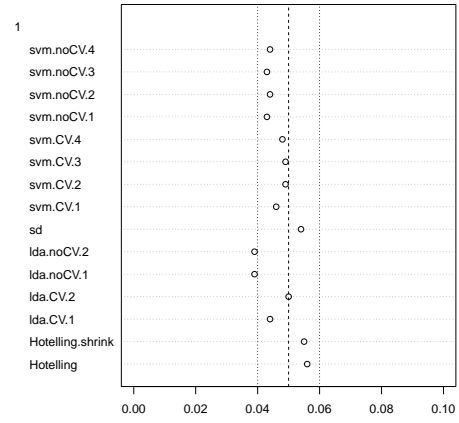


(b) **Heavytailed:**  $\eta_i$  is not  
 $p$ -variate Gaussian, but rather  
 $p$ -variate  $t$ , with  $df = 3$ .

Figure 6: *Simulation details in Appendix B except the changes in the sub-captions.*

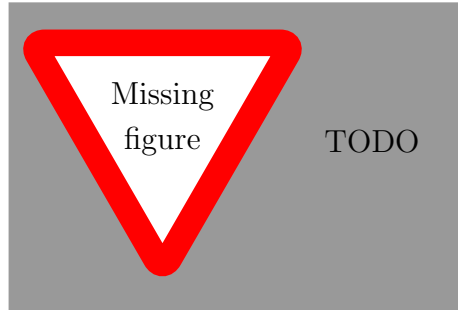


(a) **Low-Dimension:** False positive rates for  $n = 40$ .



(b) **High-Dimension:** False positive rates for  $n = 400$ .

*Figure 7: Simulation details in Appendix B except the changes in the sub-captions.*



**(a) High-Dimension, local alternative:**  $n = 400$ ,  
 $\mu \in \frac{\sqrt{40}}{\sqrt{400}} \times \{0, 1/4, 1/2\}$ .



**(b) AR(1) dependence:**  
 $\Sigma_{k,l} = \rho^{|k-l|}; \rho = 0.8$ .