

Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt Roei Gilron Roy Mukamel

August 5, 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 citations¹ 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, 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. On the other hand, the analysis in Ramdas et al. [2016] is asymptotic. This eschews the discrete nature of the accuracy statistic, which will shown to have crucial impact. Since typical sample sizes in neuroscience

¹GoogleScholar. Accessed on Aug 4, 2016.

are not large, we seek to study which test is to be preferred 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) .

101 A statistically significant “better than chance” estimate of $T_{\hat{f}}^{acc}$ is evidence
 102 that the classes are distinct.

103 2.1 Candidate Tests

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

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

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

118 **How to estimate accuracy?** Given a predictor \hat{f} , a natural test statis-
 119 tic is some estimate of its accuracy $T_{\hat{f}}^{acc}$. Complicating matters: very low
 120 accuracies, even 0, is evidence that the classes are separated, and we only
 121 need to invert the predictions. We can thus consider $|T_{\hat{f}}^{acc} - 0.5|$ as the test
 122 statistic. This, however, implies that if the classes are identical, random
 123 guessing has 0.5 accuracy. This is not true if the classes are not balanced.
 124 The chance level in which case is the prevalence of the dominant class, we
 125 denote by \hat{p}_{max} . This suggests the following test statistic $|T_{\hat{f}}^{acc} - \hat{p}_{max}|$. Since
 126 we will be aggregating these statistics over random data sets where the dom-
 127 inant class may have varying frequencies, it seems appropriate to standard-
 128 ize the scale of this statistic. We thus also consider the z-scored accuracy:
 129 $|T_{\hat{f}}^{acc} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$.

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

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

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

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

163 **How many folds?** Different authors suggest different rules for the num-
164 ber of folds. We will be varying the number of folds. This will affect the
165 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

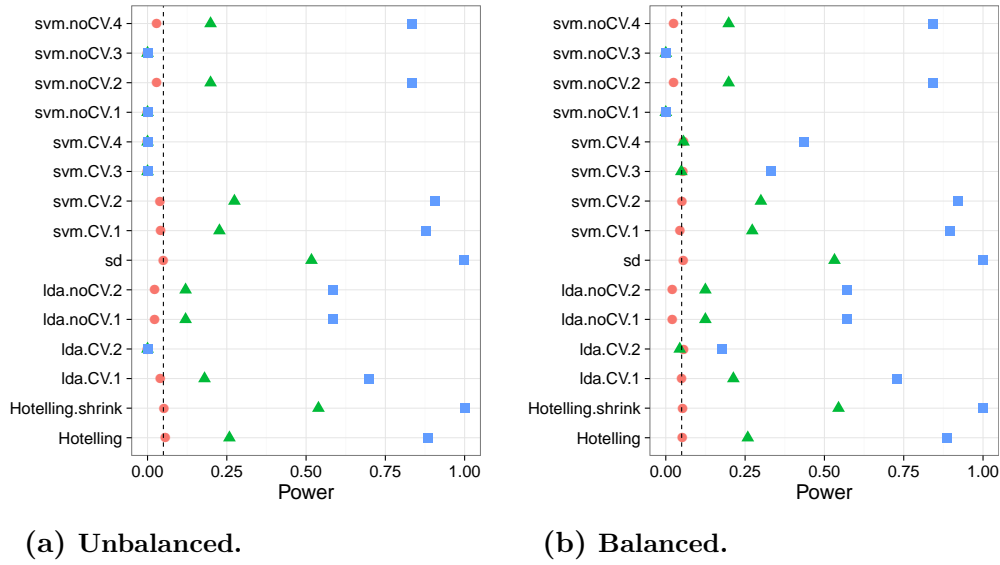
Table 1: This table enumerates the various test statistics we will be studying. Three are location tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group T^2 statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer and Strimmer [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

177 statistic are varied, (d) the number of folds is varied. We also observe that the
 178 most conservative tests are the accuracy tests that are not cross validated.
 179 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.



180 4 Power

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

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

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

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

208 5 Neuroimaging Example

209 Figure 2 is an application of both a location and an accuracy test to the data
210 of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI
211 data while subjects were exposed to the sounds of human speech (vocal),
212 and other non-vocal sounds. Each subject was exposed to 20 sounds of each
213 type, totaling in $n = 40$ trials in each scan. The study was rather large and
214 consisted of about 200 subjects. The data was kindly made available by the
215 authors at the OpenfMRI website².

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

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

²<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.

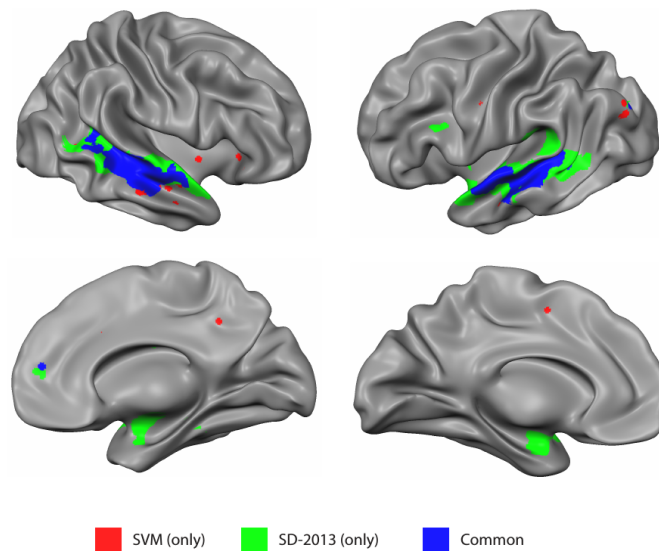


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (svm.cv.1), and a location test (sd). svm.cv.1 was computed using 5-fold cross validation, and a cost parameter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

Classifiers are designed to solve $\operatorname{argmin}\{FP + FN\}$ where FP and FN are the false positive and false negative error rates, or *type I* and *type II* errors, respectively. One might thus argue that our simulations are unfair since our simulation benchmark is $\operatorname{argmax}\{FN \text{ s.t. } FP \leq \alpha\}$, as in the classical Neyman-Pearson framework for which the location tests are designed. Our view is that this is not the case because, and the comparison is indeed “fair”. This is because the permutation test ensures that the constrain

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.

- 267 5. Cross validated accuracy with balanced folds has more power than un-
268 balanced folds. We currently have no intuition to offer for this phe-
269 nomenon.
- 270 6. It is unclear what is the effect of the number of folds. More folds in-
271 crease power by reducing the number of holdout samples. On the other
272 hand, it increases the concentration of the accuracy statistic. Com-
273 pounded with the discreteness of the accuracy statistic, this decreases
274 power.
- 275 7. The value of the tuning parameters of a classifier do not matter.

276 6.3 Related Literature

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

288 6.4 Non-linear predictors

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

293 6.5 Reservations

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

301 Secondly, not all signals are manifested in a shift of the null distribution
 302 Our focus on location tests is misleading. Perhaps Simon et al. [2003]’s *class*
 303 *comparison* is a more appropriate name, in that it does not only imply a
 304 shift alternative. Indeed, one may consider signal, i.e. effects, as a change in
 305 scale, such as the *spiked covariance* model. In this case, other-than-Hotelling
 306 type tests are appropriate [e.g. Nadler, 2008]. Tests have been proposed even
 307 when the nature of the difference between populations is left unspecified [e.g.
 308 Gretton et al., 2012]. The fact that in our neuroimaging example (Section 5)
 309 some brain regions were detected with the accuracy test, and not the location
 310 test, is consistent with this observation.

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

314 6.6 Ease of implementation

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

322 6.7 Epilogue

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

329 The recent popularity of machine learning has resulted in the ex-
 330 tensive teaching and use of prediction in theoretical and applied
 331 communities and the relative lack of awareness or popularity of
 332 the topic of Neyman-Pearson style hypothesis testing in the com-
 333 puter science and related “data science” communities.

References

- T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, Hoboken, NJ, 3 edition edition, July 2003. ISBN 978-0-471-36091-9.
- 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.
- R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. Quantifying spatial pattern similarity in multivariate analysis using functional anisotropy. *arXiv:1605.03482 [q-bio]*, May 2016.
- P. Golland and B. Fischl. Permutation tests for classification: towards statistical significance in image-based studies. In *IPMI*, volume 3, pages 330–341. Springer, 2003.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-sample Test. *J. Mach. Learn. Res.*, 13:723–773, Mar. 2012. ISSN 1532-4435.
- J. Hemerik and J. Goeman. Exact testing with random permutations. *arXiv:1411.7565 [math, stat]*, Nov. 2014.
- H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177732979.
- W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.
- N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0600244103.
- E. L. Lehmann. Parametric versus nonparametrics: two alternative methodologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN 1048-5252. doi: 10.1080/10485250902842727.
- G. J. McLACHLAN. The bias of the apparent error rate in discriminant analysis. *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/63.2.239.

- 368 B. Nadler. Finite sample approximation results for principal component
369 analysis: A matrix perturbation approach. *The Annals of Statistics*, 36
370 (6):2791–2817, Dec. 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/
371 08-AOS618.
- 372 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
373 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
374 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*
375 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
376 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
377 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.
- 378 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the
379 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.
380 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.
- 381 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
382 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
383 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 384 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
385 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
386 P. Belin. The human voice areas: Spatial organization and inter-individual
387 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
388 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 389 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
390 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*
391 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
392 106652702760138592.
- 393 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
394 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 395 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance
396 Matrix Estimation and Implications for Functional Genomics. *Statistical*
397 *Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN
398 1544-6115. doi: 10.2202/1544-6115.1175.
- 399 C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning.
400 *IEEE Transactions on Information Theory*, 51(11):3806–3819, Nov. 2005.
401 ISSN 0018-9448. doi: 10.1109/TIT.2005.856955.

- 402 R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the
403 Use of DNA Microarray Data for Diagnostic and Prognostic Classification.
404 *Journal of the National Cancer Institute*, 95(1):14–18, Jan. 2003. ISSN
405 0027-8874, 1460-2105. doi: 10.1093/jnci/95.1.14.
- 406 M. S. Srivastava. On testing the equality of mean vectors in high dimension.
407 *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1):
408 31–56, June 2013. ISSN 2228-4699. doi: 10.12697/ACUTM.2013.17.03.
- 409 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
410 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
411 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 412 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
413 ing correction in classification-based multi-voxel pattern analysis (MVPA):
414 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
415 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.
- 416 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,
417 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-
418 2.
- 419 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
420 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
421 caveats, and guidelines. working paper or preprint, June 2016.
- 422 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.
423 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-
424 nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:
425 10.1056/NEJMoA1204471.

426 A Analysis pipeline

427 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 428 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 429 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions³ are
 430 centered around a unique voxel, the voxel index v also serves as a unique
 431 region index. Algorithm 1 computes a region-wise test statistic, which is
 432 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

435 B Simulation Details

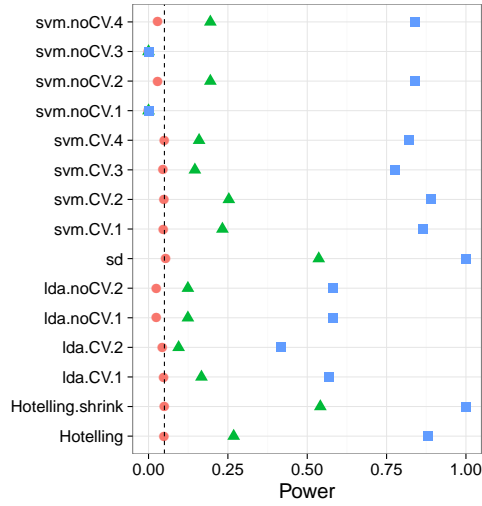
436 The following details are common to all the reported simulations, unless stated
437 otherwise in a figure’s caption. The R code for the simulations can be found
438 in [TODO].

439 Each simulation is based on 4,000 replications. In each replication, we
440 generate n i.i.d. samples from a shift model $\mathbf{x}_i = \mu \mathbf{y}_i^* + \eta_i$. Where $y_i^* = \{0, 1\}$
441 is the class of subject i in dummy coding. Recalling that $y_i = \{-1, 1\}$ is the
442 class in effect coding, then clearly $y_i = 2y_i^* - 1$. The noise is distributed as
443 $\eta_i \sim \mathcal{N}_p(0, \Sigma)$. The sample size $n = 40$. The dimension of the data is $p = 23$.
444 The covariance $\Sigma = I$. Effects, i.e. shifts μ , are equal coordinate p -vectors
445 with coordinates that vary over $\mu \in \{0, 1/4, 1/2\}$.

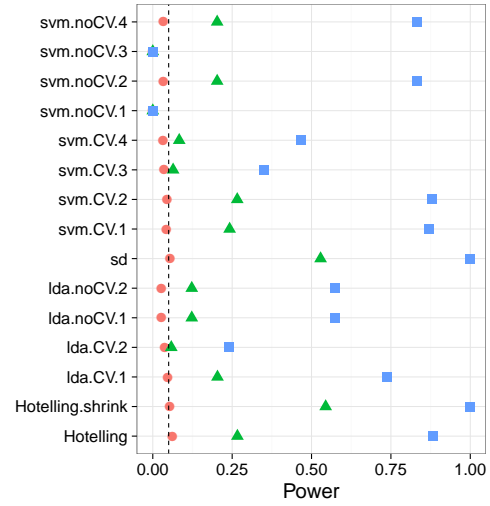
446 Having generated the data, we compute each of the test statistics in Ta-
447 ble 1. For test statistics that require data folding, we used 8 folds. We then
448 compute a permutation p-value by permuting the class labels, and recomput-
449 ing each test statistic. We perform 400 such permutations.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

C Simulation Results

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

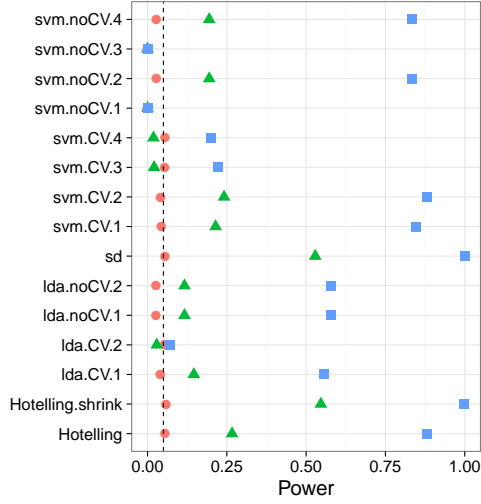


(a) 2-fold cross validation.
Balanced folding.

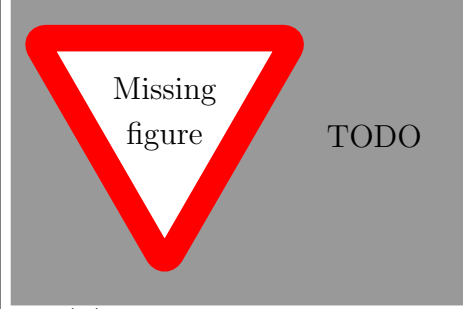


(b) 20-fold cross validation.
Balanced folding

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

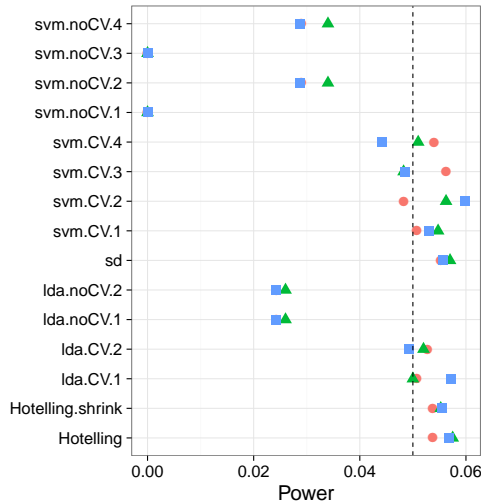


(a) **2-fold** cross validation.
Unbalanced folding.

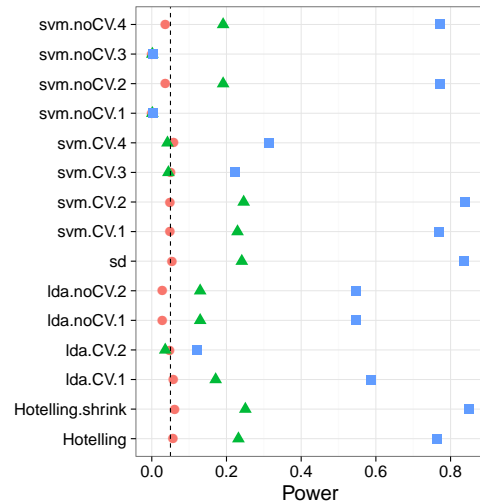


(b) **20-fold** cross validation.
Unbalanced folding.

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

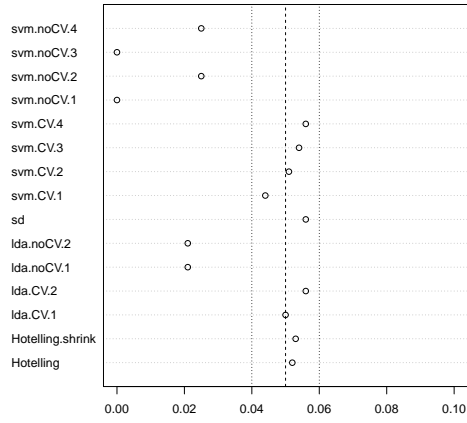


(a) **Scale Change:** $\mathbf{x}_i = \eta_i * \mu^{\mathbf{y}_i^*}$
so that the effect are a scale
change.

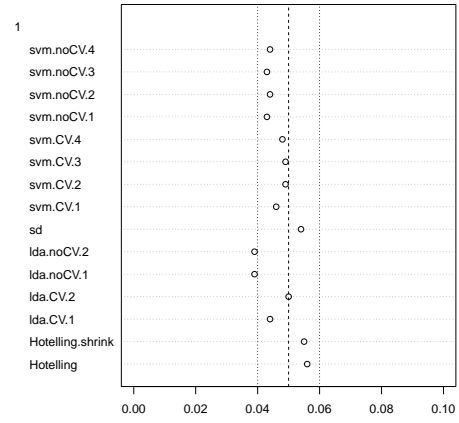


(b) **Heavytailed:** η_i is not
 p -variate Gaussian, but rather
 p -variate t , with $df = 3$.

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

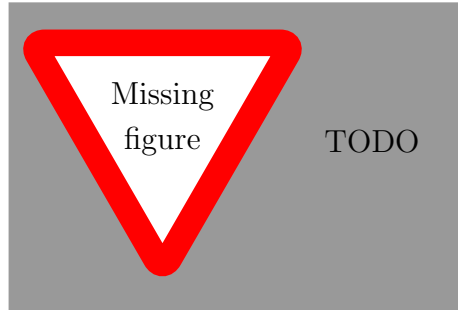


(a) **Low-Dimension:** False positive rates for $n = 40$.



(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:
 $\Sigma_{k,l} = \rho^{|k-l|}; \rho = 0.8$.