

Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt Roei Gilron Roy Mukamel

August 16, 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 some high profile publications in the genetics literature: Golub et al. [1999], Slonim et al. [2000], Radmacher et al. [2002], Mukherjee et al. [2003], Juan and Iba [2004], 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*, or *pattern discrimination*.

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]:

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

30 A practitioner may thus approach the signal detection problem with a two-
 31 group population test such as Hotelling’s T^2 [Anderson, 2003]. Alternatively,
 32 if the size of brain region of interest is large compared to the number of
 33 observations, so that the spatial covariance cannot be fully estimated, then
 34 a high dimensional version of Hotelling’s test can be called upon, such as
 35 in Schäfer and Strimmer [2005] or Srivastava [2007]. For brevity, and in
 36 contrast to *accuracy tests*, we will call any two-sample multivariate tests
 37 simply *population tests*, a.k.a. *class comparisons*. [TODO: rename population
 38 test to parameter test?]

39 At this point, it becomes unclear which is preferable: a population test or
 40 an accuracy test? The former with a heritage dating back to Hotelling [1931],
 41 and the latter being extremely popular, as the 959 citations¹ of Kriegeskorte
 42 et al. [2006] suggest.

43 The comparison between population and accuracy tests was precisely the
 44 goal of Ramdas et al. [2016], who compared the T^2 population test to the
 45 accuracy of *Fisher’s linear discriminant analysis* classifier (LDA). By com-
 46 paring the rates of convergence of the powers to 1, Ramdas et al. [2016]
 47 concluded that accuracy and population tests are rate equivalent.

48 Asymptotic relative efficiency measures (ARE) are typically used by statis-
 49 ticians to compare between rate-equivalent test statistics [van der Vaart,
 50 1998]. Ramdas et al. [2016] derive the asymptotic power functions of the
 51 two test statistics, which allows to compute the ARE between Hotelling’s T^2
 52 (population) test and Fisher’s LDA (accuracy) test. Theorem 14.7 of van der
 53 Vaart [1998] relates asymptotic power functions to ARE. Using this theorem
 54 and the results of Ramdas et al. [2016] we deduce that the ARE is lower
 55 bounded by $2\pi \approx 6.3$. This means that Fisher’s LDA requires at least 6.3
 56 more samples to achieve the same (asymptotic) power than the T^2 test. In
 57 this light, the accuracy test is remarkably inefficient compared to the pop-
 58 ulation test. For comparison, the t-test is only 1.04 more (asymptotically)
 59 efficient than Wilcoxon’s rank-sum test [Lehmann, 2009], so that an ARE of
 60 6.3 is strong evidence in favor of the population test.

61 Before discarding accuracy tests as inefficient, we recall that Ramdas
 62 et al. [2016] analyzed a *half-sample* holdout. The authors conjectured that a
 63 leave-one-out approach, which makes more efficient use of the data, may have
 64 better performance. Also, the analysis in Ramdas et al. [2016] is asymptotic.

¹GoogleScholar. Accessed on Aug 4, 2016.

65 This eschews the discrete nature of the accuracy statistic, which will be
66 shown to have crucial impact. Since typical sample sizes in neuroscience are
67 not large, we seek to study which test is to be preferred in finite samples?
68 Our conclusion will be quite simple: *population tests typically have more*
69 *power than accuracy tests, and are easier to implement.*

70 Our statement rests upon the observation that with typical sample sizes,
71 the accuracy test statistic is highly discrete. Permutation testing with dis-
72 crete test statistics are known to be conservative [Hemerik and Goeman,
73 2014], since they are insensitive to mild perturbations of the data, and they
74 cannot exhaust the permissible false positive rate. As simply put by Frank
75 Harrell in `CrossValidated`² post back in 2011:

76 ... your use of proportion classified correctly as your accuracy
77 score. This is a discontinuous improper scoring rule that can be
78 easily manipulated because it is arbitrary and insensitive.

79 The degree of discretization is governed by the number of samples. In our
80 neuroscience example from Gilron et al. [2016], the classification is performed
81 based on 40 trials, so that the test statistic may assume only 40 possible
82 values. This number of examples is not unusual if considering this is the
83 number of trial-repeats, or the number of subjects, in an neuroimaging study.

84 The discretization effect is aggravated if the test statistic is highly concen-
85 trated. For an intuition consider the usage of a the *resubstitution accuracy*
86 as a test statistic. This statistic simply means that the accuracy is not cross
87 validated, but rather evaluated on the training data. If the data is high
88 dimensional, the resubstitution accuracy will be very high due to over fit-
89 ting. In a very high dimensional regime, the resubstitution accuracy will
90 be 1 for the observed data [McLachlan, 1976, Theorem 1], but also for any
91 permutation. The concentration of resubstitution accuracy near 1, and its
92 discreteness, render this test completely useless, with power tending to 0 for
93 any (fixed) effect size, as the dimension of the model grows.

94 To compare the power of accuracy tests and population tests in finite
95 samples, we study a battery of test statistics by means of simulation. We start
96 with formalizing the problem in Section 2. The main findings are reported
97 in Sections 4, 5 and Appendix C. A discussion follows in Section 6.

²A Q&A website for statistical questions: <http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier>

98 2 Problem setup

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

102 Given n pairs of (x_i, y_i) , typically assumed i.i.d., a population test amounts
 103 to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$. I.e.,
 104 we test if the multivariate voxel activation pattern has the same distribution
 105 when given a vocal stimulus, as when given a non-vocal stimulus.

An accuracy test amounts to learning a predictive model and testing if its
 predictions $y|x$ are better than chance. Denoting a dataset by $\mathcal{S} := (x_i, y_i)_{i=1}^n$,
 the a predictor, $\mathcal{A}_{\mathcal{S}}(x) : \mathcal{X} \rightarrow \mathcal{Y}$, is the output of a learning algorithm \mathcal{A}
 when applied to the dataset \mathcal{S} , so that $\mathcal{A} : \mathcal{S} \rightarrow \mathcal{A}_{\mathcal{S}}(x)$. The accuracy
 of predictor $\mathcal{A}_{\mathcal{S}}(x)$ is defined as the probability of $\mathcal{A}_{\mathcal{S}}(x)$ making a correct
 prediction. Denoting by \mathcal{P} the probability measure of (x, y) , and by \mathcal{P}^n the
 same for the i.i.d sample \mathcal{S} , then

$$\mathcal{E}_{\mathcal{A}_{\mathcal{S}}(x)} := \mathcal{P}(\mathcal{A}_{\mathcal{S}}(x) = y). \quad (1)$$

The accuracy of an algorithm \mathcal{A} is defined as the average accuracy, over all
 possible data sets

$$\mathcal{E}_{\mathcal{A}} := \int_{\mathcal{S}} \mathcal{E}_{\mathcal{A}_{\mathcal{S}}} d\mathcal{P}^n(\mathcal{S}). \quad (2)$$

106 Denoting an estimate of $\mathcal{E}_{\mathcal{A}_{\mathcal{S}}(x)}$ by $\hat{\mathcal{E}}_{\mathcal{A}_{\mathcal{S}}(x)}$, and $\mathcal{E}_{\mathcal{A}}$ by $\hat{\mathcal{E}}_{\mathcal{A}}$, a statistically sig-
 107 nificant “better than chance” estimate of either, is evidence that the classes
 108 are distinct. In a typical application, the predictor is not fixed, so that $\hat{\mathcal{E}}_{\mathcal{A}}$,
 109 and not $\hat{\mathcal{E}}_{\mathcal{A}_{\mathcal{S}}(x)}$, will be used for the testing.

110 Two popular estimates of $\hat{\mathcal{E}}_{\mathcal{A}}$ are the *resubstitution estimate*, and the
 111 V-fold cross validation (CV) estimate [Hastie et al., 2003].

Definition 1 (Resubstitution accuracy). The resubstitution accuracy esti-
 mator, $\hat{\mathcal{E}}_{\mathcal{A}}^{resub}$, is defined as

$$\hat{\mathcal{E}}_{\mathcal{A}}^{Resub} := \frac{1}{n} \sum_{i=1}^n \mathcal{I}\{\mathcal{A}_{\mathcal{S}}(x_i) = y_i\}, \quad (3)$$

112 where $\mathcal{I}\{A\}$ is the indicator function of event A .

Definition 2 (V-fold CV). Denoting by \mathcal{S}^v the v 'th partition of the dataset,
 and by $\mathcal{S}^{(v)}$ its complement, so that $\mathcal{S}^v \cup \mathcal{S}^{(v)} = \cup_{v=1}^V \mathcal{S}^v = \mathcal{S}$, the V-fold CV

accuracy estimator, $\hat{\mathcal{E}}_{\mathcal{A}}^{Vfold}$, is defined as

$$\hat{\mathcal{E}}_{\mathcal{A}}^{Vfold} := \frac{1}{V} \sum_{v=1}^V \frac{1}{|\mathcal{S}^v|} \sum_{i \in \mathcal{S}^v} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^v}(x_i) = y_i\}, \quad (4)$$

113 2.1 Candidate Tests

114 The design of a permutation test using $\hat{\mathcal{E}}_{\mathcal{A}}$, requires the following design
115 choices:

- 116 1. Is $\hat{\mathcal{E}}_{\mathcal{A}}$ cross validated or not?
- 117 2. For a V-fold cross validated test statistic:
 - 118 (a) Should the data be refolded in each permutation?
 - 119 (b) Should the data folding be balanced (a.k.a. stratified)?
 - 120 (c) How many folds?
- 121 3. How to estimate $\hat{\mathcal{E}}_{\mathcal{A}}$?

122 We will now address these questions while bearing in mind that unlike
123 the typical supervised learning setup, we are not interested in an unbiased
124 estimate of $\mathcal{E}_{\mathcal{A}}$, but rather in its mere departure from chance level.

125 **Cross validate or not?** Given our goal, a biased estimate of $\hat{\mathcal{E}}_{\mathcal{A}}$ is not a
126 problem provided that bias is consistent over all permutations. The under-
127 lying intuition is that a permutation test will be unbiased, provided that the
128 exact same computation is performed over all permutations. We will thus
129 be considering both cross validated accuracies, and *resubstitution accuracies*,
130 where the accuracy is evaluated on the training set and not on a holdout.

131 **Balanced folding?** The standard practice when cross validating is to con-
132 strain the data folds to be balanced, i.e. stratified [e.g. Ojala and Garriga,
133 2010]. This means that each fold has the same number of examples from
134 each class. We will report results with both balanced and unbalanced data
135 foldings, only to discover, it does not really matter.

136 **Refolding?** The standard practice in neuroimaging is to permute labels
137 and refold the data after each permutation, so that the balance of the classes
138 in each fold is preserved. We will adhere to this practice due to its popularity,
139 even though it can be simplified by permuting features instead of labels, as
140 done by Golland et al. [2005].

141 **How many folds?** Different authors suggest different rules for the number
142 of folds. We will look into the effect of the number of folds.

How to estimate accuracy? Low accuracies, even 0, are evidence that the classes are separated so that for our purposes, we should consider the departure from chance level $|\hat{\mathcal{E}}_{\mathcal{A}} - 0.5|$ as candidate test statistic. For unbalanced classes, chance level is not 0.5, but rather the probability of the majority class, we denote by $\hat{\pi}$. This suggests the following test statistic $|\hat{\mathcal{E}}_{\mathcal{A}} - \hat{\pi}|$. Since we will be aggregating these statistics over random data sets where $\hat{\pi}$ may vary, it seems appropriate to standardize the scale. We thus study, along with the naive accuracy estimate, $\hat{\mathcal{E}}_{\mathcal{A}}$, also the *z-scored accuracy* of algorithm \mathcal{A} :

$$\hat{\mathcal{Z}}_{\mathcal{A}} := \frac{|\hat{\mathcal{E}}_{\mathcal{A}} - \hat{\pi}|}{\sqrt{\hat{\pi}(1 - \hat{\pi})}}. \quad (5)$$

143 Table 1 collects an initial battery of tests we will be comparing.

Name	Algorithm	Accuracy	Z-scored	Parameters
Hotelling	Hotelling	—	—	—
Hotelling.shrink	Hotelling	—	—	—
sd	SD	—	—	—
lda.CV.1	LDA	V-fold	FALSE	—
lda.CV.2	LDA	V-fold	TRUE	—
lda.noCV.1	LDA	Resubstitution	FALSE	—
lda.noCV.2	LDA	Resubstitution	TRUE	—
svm.CV.1	SVM	V-fold	FALSE	cost=10
svm.CV.2	SVM	V-fold	FALSE	cost=0.1
svm.CV.3	SVM	V-fold	TRUE	cost=10
svm.CV.4	SVM	V-fold	TRUE	cost=0.1
svm.noCV.1	SVM	Resubstitution	FALSE	cost=10
svm.noCV.2	SVM	Resubstitution	FALSE	cost=0.1
svm.noCV.3	SVM	Resubstitution	TRUE	cost=10
svm.noCV.4	SVM	Resubstitution	TRUE	cost=0.1

Table 1: This table collects the various test statistics we will be studying. Three are population 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 from 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 implemented with the *svm* R function [Meyer et al., 2015], the cost parameter set at 0.1, and using the cross validated z-scored accuracy in Eq. 5. Another example is *lda.noCV.1*, which is Fisher’s LDA, returning the resubstitution accuracy.

144

145 3 Controlling the False Positive Rate

146 Our simulation show that all of the tests considered conserve the desired
147 0.05 false positive rate, up to varying levels of conservatism. This can be
148 seen from the fact that the probability of rejection is no larger than 0.05 in
149 the absence of any effect, encoded by a red circle. This is true, in particular
150 if:

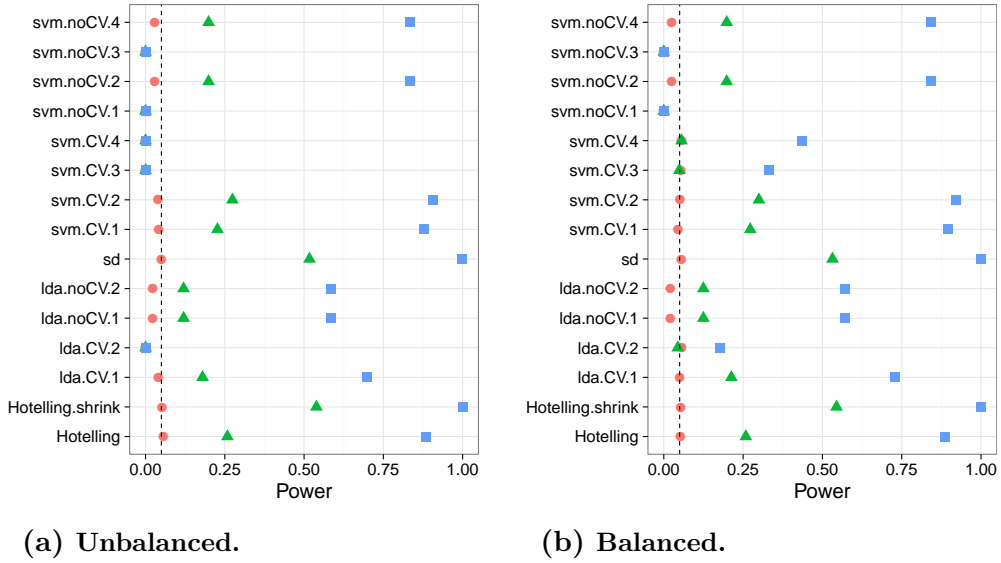
- 151 (a) The folds are balanced or not (Figures 1,6 and 7)
- 152 (b) The tuning parameters are varied (cost=10 versus cost=0.1).
- 153 (c) The number of folds is varied (Figures 6 and 7).
- 154 (d) The noise is heavytailed (Figure 8b).

155 (e) The problem is high or low dimensional (Figure 9.)

156 (f) The noise is correlated (Figure 10b).

157 We also observe that the most conservative tests are the resubstitution ac-
 158 curacy statistics. 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 and unbalanced data folding. See sub-captions.



159 4 Power

160 Having established that all of the tests in our battery control the false pos-
 161 itive rate, it remains to be seen if they have similar power— especially when
 162 comparing population tests to accuracy tests. From the simulation results
 163 reported in Appendix C we collect the following insights:

- 164 1. Population tests have more power than accuracy tests in all our con-
 165 figurations.
- 166 2. The conservativeness decays as the sample grows (Figures 9a, 9b and
 167 10a)
- 168 3. For heavy tailed distributions (Figure 8b), the extra power of the pop-
 169 ulation test vanishes.

- 170 4. Regularization is most beneficial to power in low signal to noise (SNR)
171 regimes. Low SNR may be the result of a high-dimensional problem,
172 or due to correlations. Indeed, the presence of positive correlations
173 amplifies the contribution of regularization to power ((Figure 10b)).
- 174 5. The z-scoring of the accuracies was introduced to deal with unbalanced
175 foldings. If the z-scoring has any effect at all, it merely kills power.
- 176 6. Both accuracy and population tests are inappropriate for scale alter-
177 natives (Figure 8a). This was to be expected and is reported mostly as
178 a sanity check.
- 179 7. Balanced folding only affects the z-scored accuracy, in the opposite
180 direction than we anticipated.
- 181 8. Increasing the SVM’s cost parameter, which reduces the number of
182 support vectors entering the classifier, reduces power.

183 The major insight from simulations is that the use of accuracy tests for
184 signal detection is underpowered compared to population tests. We have not
185 established, however, that the dominance of the population tests is not due to
186 their regularization. Indeed, the unregularized *Hotelling* test, is only slightly
187 superior to the accuracy tests. We return to this matter in Section 6.5. We
188 now verify our finding on a neuroimaging dataset.

189 5 Neuroimaging Example

190 Figure 2 is an application of both a population and an accuracy test to the
191 data of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI
192 data while subjects were exposed to the sounds of human speech (vocal), and
193 other non-vocal sounds. Each subject was exposed to 20 sounds of each type,
194 totaling in $n = 40$ trials. The study was rather large and consisted of about
195 200 subjects. The data was kindly made available by the authors at the
196 OpenfMRI website³.

197 We perform group inference using within-subject permutations along the
198 analysis pipeline of Stelzer et al. [2013], which was also reported in Gilron
199 et al. [2016]. For completeness, the pipeline is described in Appendix A. To
200 demonstrate our point, we compare the *sd* population test with the *svm.cv.1*
201 accuracy test.

202 In agreement with our simulation results, the population test (*sd*) dis-
203 covers more brain regions of interest when compared to an accuracy test

³<https://openfmri.org/>

204 (*svm.cv.1*). The former discovers 1,232 regions, while the latter only 441, as
 205 depicted in Figure 2. We emphasize that both test statistics were compared
 206 with the same permutation scheme, and the same error controls, so that any
 207 difference in detections is due to their different power.

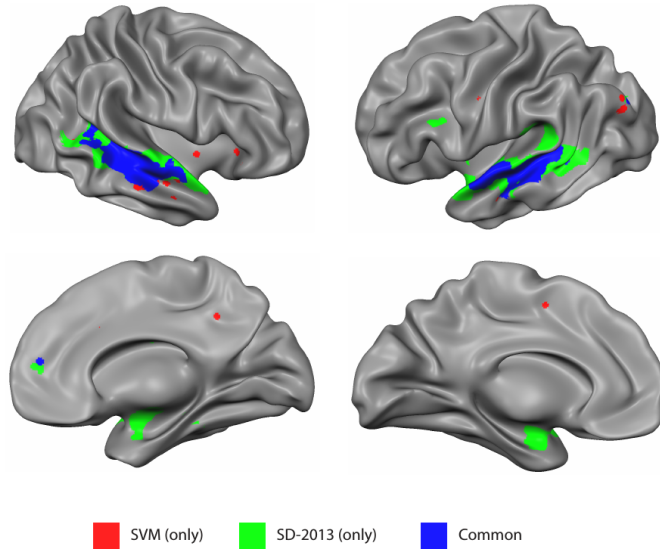


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a population test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost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

208 6 Discussion

209 We have set out to understand which of the tests is more powerful: the ac-
 210 curacy test or the population test. No amount of simulations can replace the
 211 insight provided by a closed-form analytic result. The finite sample power
 212 of permutation tests is a formidable mathematical problem, so we currently
 213 content ourselves with simulations. We have concluded that the population
 214 tests are typically preferable. Their high dimensional versions, such as Sri-
 215 vastava [2007] and Schäfer and Strimmer [2005], are particularly well suited

216 for neuroimaging problems such as MVPA. We attribute this to several ef-
217 fects:

- 218 (a) The discrete nature of the accuracy test in finite samples.
- 219 (b) Inefficient use of the data when validating with a holdout set.
- 220 (c) The lack of regularization in high SNR regimes (high dimension or cor-
221 relations).

222
223 The degree of discretization is governed by the sample size. For this
224 reason, an asymptotic analysis such as Ramdas et al. [2016] may uncover
225 the holdout inefficiency, but will not uncover the discretization effect. An
226 asymptotic analysis of a finite complexity model would also fail to reveal the
227 effect of the concentration of the resubstitution accuracy near 1. This effect
228 would render the resubstitution estimates a legitimate asymptotic test, and
229 a terrible finite sample test.

230 The presence of heavy tails shrinks the power advantage of the population
231 tests over accuracy tests. Our empirical example suggests that even if the
232 population test does not necessarily dominate the accuracy test in power,
233 empirically, it does have an advantage.

234 The practical advice for the practitioner, is that for the purpose of signal
235 detection, there is typically a population test that is more powerful than an
236 accuracy test. The class of population tests we examined, in particular their
237 regularized versions, are good performers in a wide range of simulation setups
238 and empirically. They are also typically easier to implement, and faster to
239 run, since no cross validation will be involved.

240 6.1 Ease of implementation

241 A very important consideration is the ease of implementation. The need
242 for cross validation of the accuracy test greatly increases its computational
243 complexity. Moreover, programming with discrete statistics is more prone to
244 errors. This is because their unforgiveness to the type of inequalities used.
245 Indeed, mistakenly replacing a weak inequality with a strong inequality in
246 one’s program may considerably change the results. This is not the case for
247 continuous test statistics.

248 6.2 Reservations

249 Some reservations to the generality of our findings are in order. Firstly,
250 not all accuracy tests are concerned with signal detection. Consider brain
251 decoding for machine interfaces, or clinical diagnosis, where the presence of
252 a medical condition is predicted from imaging data [e.g. Olivetti et al., 2012,

Wager et al., 2013]. In those examples, the purpose of the test is not to detect a difference between classes, but to actually test the performance of a particular classifier.

Secondly, it may be argued that accuracy tests permits the separation between classes in high dimensions, such as in *reproducing kernel Hilbert spaces* (RKHS) by using non-linear predictors while population tests do not. This is a false argument— accuracy test do not have any more flexibility than population tests. Indeed, it is possible to test for location in the same space the classifier is learned. Gretton et al. [2012] is an example where the test for location is performed in RKHS. It is also possible to test for the equality of two multivariate distributions on an arbitrary, unspecified manifold [e.g. Heller et al., 2013][TODO: verify]. On the other hand, based on our experience, and the reported neuroimaging example, we find that a population test in the original feature space is a simple and powerful approach to signal detection.

6.3 A good accuracy test

For the cases a population test cannot replace an accuracy test, we collect some conclusions and best practices from our simulations. We give particular emphasis in this section to V-fold cross validation due to its popularity, but note that sampling the test set with replacement is actually preferable, as we discuss in Section 6.4.

Sample size. The conservativeness of accuracy tests decrease with sample size.

Permute features. Permuting features, such as in Golland et al. [2005], is easier than permuting labels. It allows to preserve the balance of folds after a permutation, without refolding.

Resubstitution accuracy in low dimension. Resubstitution accuracy is useful in low SNR regimes, such as low dimensional problems, because it avoids cross validation without compromising power. In high dimension, the power loss is considerable compared to a cross validated approach. We attribute this to the compounding of discretization and concentration effects: the difference between the sampling distribution of the resubstitution accuracy is simply indistinguishable under the null and under the alternative. In low dimensional problems, the discretization is less impactful, and the

287 computational burden of cross validation can be avoided by using the resub-
 288 stitution accuracy. There is a fundamental difference between V-folding and
 289 resubstitution. The latter should not be thought of as the limit of the former.

290 **Regularize** Regularizing a classifier proves crucial to detection power in
 291 low SNR regimes; in high dimension in particular. We also conjecture that
 292 the power-maximizing regularization is larger than the error-minimizing reg-
 293 ularization.

294 **Don't z-score.** There is no gain in z-scoring the accuracy scores. Our
 295 motivating rational was clearly flawed. [TODO: why?]

296 6.4 Smoothing accuracy estimates

297 It may be possible to alleviate the effect of discretization by appropriate
 298 cross-validation. The discreteness of the accuracy statistic is governed by the
 299 number of examples in the union of test sets, over all validation iterations.
 300 For V-fold CV, for instance, the accuracy may assume as many values as the
 301 sample size. This suggests that the accuracy can be “smoothed” by allowing
 302 the test sample to be drawn with replacement. An algorithm that samples
 303 test sets with replacement is the *leave-one-out bootstrap estimator*, and its
 304 derivatives, such as the *0.632 bootstrap*, and *0.632+ bootstrap* [Hastie et al.,
 305 2003, Sec 7.11].

Definition 3 (bLOO). The *leave-one-out bootstrap* estimate is the average
 accuracy of the holdout observations, over all bootstrap samples. Denoting
 by \mathcal{S}^b , a bootstrap sample b , sampled with replacement from \mathcal{S} . Also denote
 by $C^{(i)}$ the index set of bootstrap samples, b , not containing observation i .
 The leave-one-out bootstrap estimate, $\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO}$, is defined as:

$$\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} := \frac{1}{n} \sum_{i=1}^n \frac{1}{|C^{(i)}|} \sum_{b \in C^{(i)}} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i) = y_i\}. \quad (6)$$

where $|A|$ is the cardinality of set A . Equivalently [TODO: verify], denoting
 by $S^{(b)}$ the indexes of observations, i , that are *not* in the bootstrap sample b
 and are not empty,

$$\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} = \frac{1}{B} \sum_{b=1}^B \frac{1}{|S^{(b)}|} \sum_{i \in S^{(b)}} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i) = y_i\}. \quad (7)$$

Definition 4 (b0.632). The *0.632 bootstrap* estimator, $\hat{\mathcal{E}}_{\mathcal{A}}^{0.632}$, is defined as

$$\hat{\mathcal{E}}_{\mathcal{A}}^{0.632} := 0.368 \hat{\mathcal{E}}_{\mathcal{A}}^{Resub} + 0.632 \hat{\mathcal{E}}_{\mathcal{A}}^{bLOO}. \quad (8)$$

Simulation results reported in Figure 3 with naming conventions in Table 2. It can be seen that selecting test sets with replacement does increase the power, when compared to V-fold cross validation, but still falls short from the power of population tests. It can also be seen that power increases with the number of bootstrap replications, as was to be expected, since more replications reduce the level of discretization. The type of bootstrap, bLOO versus b0.632, does not change the power.

Name	Algorithm	Accuracy	B	Z-scored	Parameters
lda.Boot.1	LDA	b0.632	10	FALSE	–
lda.Boot.2	LDA	bLOO	10	FALSE	–
svm.Boot.1	SVM	b0.632	10	FALSE	cost=1e1
svm.Boot.2	SVM	bLOO	10	FALSE	cost=1e1
svm.Boot.3	SVM	b0.632	50	FALSE	cost=1e1
svm.Boot.4	SVM	bLOO	50	FALSE	cost=1e1

Table 2: The same as Table 1 for bootstrapped accuracy estimates. bLOO and b0.632 are defined in definitions 3 and 4 respectively. B denotes the number of Bootstrap samples.

6.5 High dimensional classifiers

Inspecting Figure 1a (for instance), it can be seen that Hotelling’s T^2 test has similar power as accuracy tests. It should thus be argued that the real advantage of the population tests is due to their adaptation to high dimension by regularization, and not only to discretization. To study this, we call upon several *regularized classifiers*, designed for high dimensional problems. In the spirit of the regularized covariance of *Hotelling.shrink*, we try an l_2 regularized SVM Friedman et al. [2010], and shrinkage based LDA [Pang et al., 2009, Ramey et al., 2016]. In the spirit of the diagonalized covariance of *sd*, we try a diagonalized LDA [Dudoit et al., 2002], a.k.a. *Gaussian naive Bayes*.

Simulation results reported in Figure 4 with naming conventions in Table 3. It can be seen that regularizing a classifier in high dimension, just like a parameter test, improves power. It can also be seen that (regularized) parameter tests are still more powerful than (regularized) accuracy tests. This

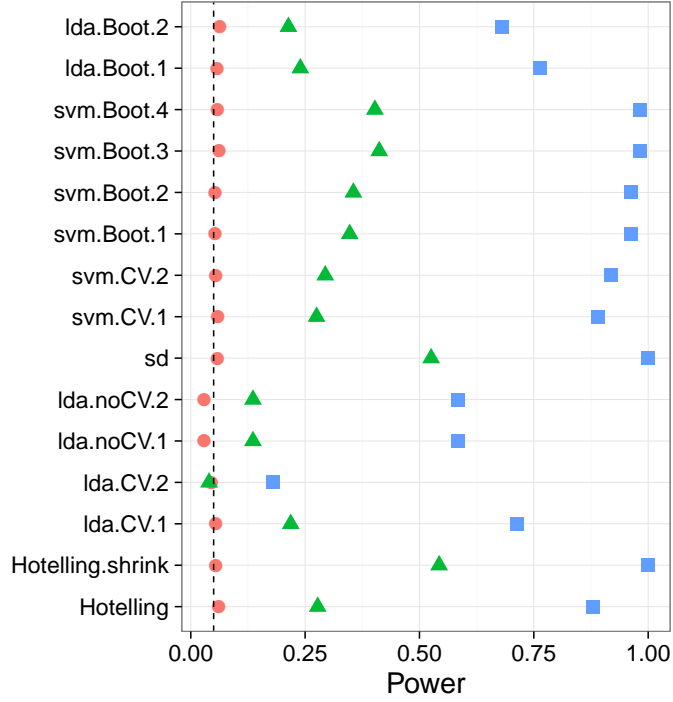


Figure 3: Bootstrap— The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix B.

was to be expected, since we already saw in (e.g. Figure 1a) that the unregularized parameter test, *Hotelling*, is slightly more powerful than unregularized accuracy tests— *svm.CV.1* for instance.

We can compound the regularization with the bootstrapping from Section 6.4, to improve finite sample power of the accuracy tests. This is done in the *svm.highdim.2* and *lda.highdim.4* tests. The latter being the first accuracy test that achieves the same power as the best of population tests. This is exciting news for the cases a population test cannot replace an accuracy test. The implication for practitioners is that powerful accuracy tests can be constructed by sampling test sets with replacement, and regularizing the classifier.

Name	Algorithm	Accuracy	Z-scored	Parameters
svm.highdim.1	SVM	V-fold	FALSE	cost=10, V=4
svm.highdim.2	SVM	b0.632	FALSE	cost=10, B=50
lda.highdim.1	LDA	V-fold	FALSE	V=4
lda.highdim.2	LDA	V-fold	FALSE	V=4
lda.highdim.3	LDA	V-fold	FALSE	V=4
lda.highdim.4	LDA	b0.632	FALSE	B=50

Table 3: The same as Table 1 for regularized (high dimensional) predictors. *svm.highdim.1* is an l_2 regularized SVM [Friedman et al., 2010]. *svm.highdim.2* is the same with b0.632 instead of V-fold cross validation. *lda.highdim.1* is the Diagonal Linear Discriminant Analysis of Dudoit et al. [2002]. *lda.highdim.2* is the High-Dimensional Regularized Discriminant Analysis of Ramey et al. [2016]. *lda.highdim.3* is the Shrinkage-based Diagonal Linear Discriminant Analysis of Pang et al. [2009]. *lda.highdim.4* is the same with b0.632.

340

341 6.6 Related Literature

342 Ojala and Garriga [2010] study the power of two accuracy tests differing in
343 the permutation scheme: One testing the “no signal” null hypothesis, and
344 the other testing the “independent features” null hypothesis. They perform
345 an asymptotic analysis, and a simulation study. They also apply various
346 classifiers to various data sets. Their emphasis is the effect of the underlying
347 classifier on the power, and the potential of the “independent features” test
348 for feature selection. This is a very different emphasis from our own.

349 Olivetti et al. [2012] and Olivetti et al. [2014] looked into the problem
350 of choosing a good accuracy test. They propose a new test they call an
351 *independence test*, and demonstrate by simulation that it has more power
352 than other accuracy tests, and can deal with non-balanced data sets. We did
353 not include this test in the battery we compared, but we note the following:
354 (a) The independence test of Olivetti et al. [2012] relies on a discrete test
355 statistic. It may probably be improved with the methods discussed in this
356 section, before the application of Olivetti et al. [2012]’s independence test.
357 (b) In contrast with the underlying motivation of Olivetti et al. [2012]’s
358 independence test, we did not find that balancing the data folds affects the
359 power of the test.

360 Golland and Fischl [2003] and Golland et al. [2005] study accuracy tests
361 using simulation, neuroimaging data, genetic data, and analytically. Their

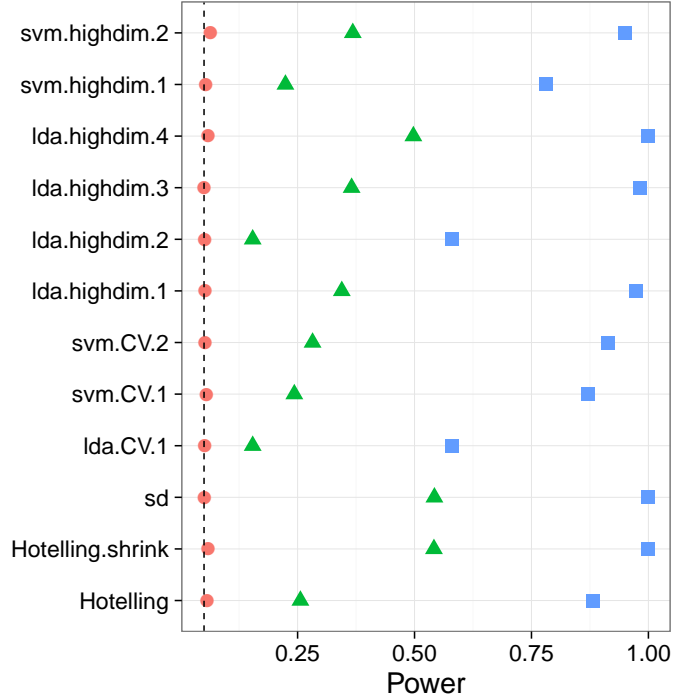


Figure 4: HighDim Classifier— The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 3. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix B.

analytic results formalize our intuition from Section 1 on the effect of concentration of the accuracy statistic: The finite Vapnik–Chervonenkis dimension requirement [Golland et al., 2005, Sec 4.3] prevents the permutation p-value from (asymptotically) concentrating near 1. Like ourselves, they also find that the power increases with the size of the test set. This is seen in Fig. 4 of Golland et al. [2005], where the size of the test-set, K , governs the discretization. Since they permute features, not labels, then all their permutation samples are balanced, and there is no issue of refolding.

Golland et al. [2005] simulate the power of accuracy tests by sampling from a Gaussian mixture family of models, and not from a location family as our own simulations. Under their model

$$(x_i|y_i = 1) \sim p\mathcal{N}(\mu_1, I) + (1 - p)\mathcal{N}(\mu_2, I)$$

and

$$(x_i|y_i = -1) \sim (1 - p)\mathcal{N}(\mu_1, I) + p\mathcal{N}(\mu_2, I).$$

Varying p interpolates between the null distribution ($p = 0.5$) and a location

371 shift model ($p = 0$). We now perform the same simulation as Golland et al.
 372 [2005], after parameterizing p so that $p = 0$ corresponds to the null model,
 373 and in the same dimensionality as our previous simulations We find that also
 374 in this mixture class of models a population test has more power than an
 375 accuracy test (Figure 5).

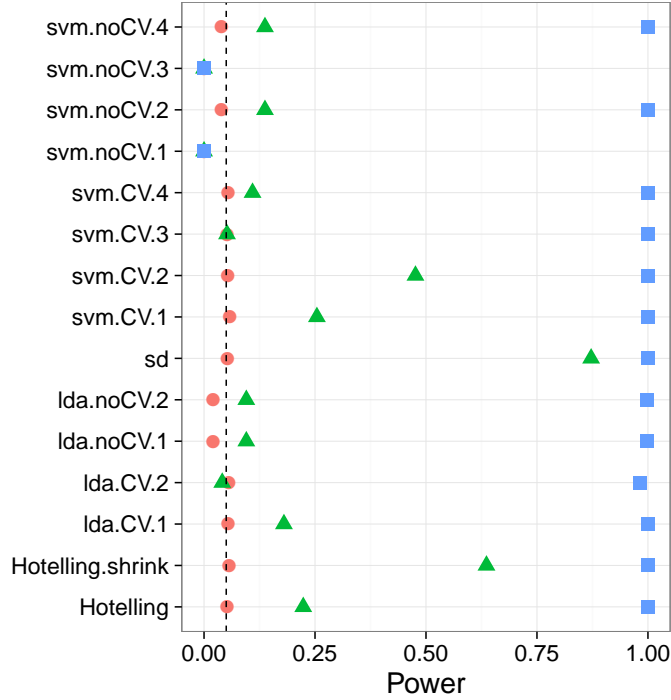


Figure 5: **Mixture**— $\mathbf{x}_i = \chi_i \mu + \eta_i$; $\chi_i = \{-1, 1\}$ and $\text{Prob}(\chi_i = 1) = (1/2 - p)^{\mathbf{y}_i^*} (1/2 + p)^{1 - \mathbf{y}_i^*}$. μ is a p -vector with $3/\sqrt{p}$ in all coordinates. The effect, p , is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

376 6.7 Epilogue

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

383 The recent popularity of machine learning has resulted in the ex-
 384 tensive teaching and use of prediction in theoretical and applied

385 communities and the relative lack of awareness or popularity of
386 the topic of Neyman-Pearson style hypothesis testing in the com-
387 puter science and related “data science” communities.

388 **7 Acknowledgments**

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.
- S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, Mar. 2002. ISSN 0162-1459. doi: 10.1198/016214502753479248.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- 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.
- P. Golland, F. Liang, S. Mukherjee, and D. Panchenko. Permutation Tests for Classification. In P. Auer and R. Meir, editors, *Learning Theory*, number 3559 in Lecture Notes in Computer Science, pages 501–515. Springer Berlin Heidelberg, June 2005. ISBN 978-3-540-26556-6 978-3-540-31892-7. doi: 10.1007/11503415_34.
- T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, Oct. 1999. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.286.5439.531.
- 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.

- 423 T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learn-*
424 *ing*. Springer, July 2003. ISBN 0-387-95284-5.
- 425 R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of associ-
426 ation based on ranks of distances. *Biometrika*, 100(2):503–510, Jan. 2013.
427 ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/ass070.
- 428 J. Hemerik and J. Goeman. Exact testing with random permutations.
429 *arXiv:1411.7565 [math, stat]*, Nov. 2014.
- 430 H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Math-*
431 *ematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990.
432 doi: 10.1214/aoms/1177732979.
- 433 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for
434 prediction error in microarray classification using resampling. *Statistical*
435 *Applications in Genetics and Molecular Biology*, 7(1), 2008.
- 436 L. Juan and H. Iba. Prediction of tumor outcome based on gene expression
437 data. *Wuhan University Journal of Natural Sciences*, 9(2):177–182, Mar.
438 2004. ISSN 1007-1202, 1993-4998. doi: 10.1007/BF02830598.
- 439 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional
440 brain mapping. *Proceedings of the National Academy of Sciences of the*
441 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,
442 1091-6490. doi: 10.1073/pnas.0600244103.
- 443 E. L. Lehmann. Parametric versus nonparametrics: two alternative method-
444 ologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN
445 1048-5252. doi: 10.1080/10485250902842727.
- 446 G. J. McLachlan. The bias of the apparent error rate in discriminant analysis.
447 *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi:
448 10.1093/biomet/63.2.239.
- 449 D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071:*
450 *Misc Functions of the Department of Statistics, Probability Theory Group*
451 *(Formerly: E1071), TU Wien*. 2015. R package version 1.6-7.
- 452 S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell,
453 T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements
454 for classifying DNA microarray data. *Journal of Computational Biology:*
455 *A Journal of Computational Molecular Cell Biology*, 10(2):119–142, 2003.
456 ISSN 1066-5277. doi: 10.1089/106652703321825928.

- 457 M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Perfor-
458 mance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.
459 ISSN ISSN 1533-7928.
- 460 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
461 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
462 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*
463 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
464 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
465 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.
- 466 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the
467 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.
468 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.
- 469 H. Pang, T. Tong, and H. Zhao. Shrinkage-based Diagonal Discriminant
470 Analysis and Its Applications in High-Dimensional Data. *Biometrics*, 65
471 (4):1021–1029, Dec. 2009. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2009.
472 01200.x.
- 473 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
474 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
475 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 476 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
477 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
478 P. Belin. The human voice areas: Spatial organization and inter-individual
479 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
480 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 481 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
482 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*
483 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
484 106652702760138592.
- 485 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
486 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 487 J. A. Ramey, C. K. Stein, P. D. Young, and D. M. Young. High-Dimensional
488 Regularized Discriminant Analysis. *arXiv preprint arXiv:1602.01182*,
489 2016.
- 490 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance
491 Matrix Estimation and Implications for Functional Genomics. *Statistical*

492 *Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN
493 1544-6115. doi: 10.2202/1544-6115.1175.

494 D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class
495 Prediction and Discovery Using Gene Expression Data. In *Proceedings of*
496 *the Fourth Annual International Conference on Computational Molecular*
497 *Biology*, RECOMB '00, pages 263–272, New York, NY, USA, 2000. ACM.
498 ISBN 978-1-58113-186-4. doi: 10.1145/332306.332564.

499 M. S. Srivastava. Multivariate Theory for Analyzing High Dimensional Data.
500 *Journal of the Japan Statistical Society*, 37(1):53–86, 2007. doi: 10.14490/
501 jjss.37.53.

502 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
503 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
504 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.

505 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
506 ing correction in classification-based multi-voxel pattern analysis (MVPA):
507 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
508 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.

509 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,
510 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-
511 2.

512 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
513 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
514 caveats, and guidelines. working paper or preprint, June 2016.

515 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.
516 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-*
517 *nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:
518 10.1056/NEJMoa1204471.

519 A Analysis pipeline

520 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 521 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 522 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions⁴ are
 523 centered around a unique voxel, the voxel index v also serves as a unique
 524 region index. Algorithm 1 computes a region-wise test statistic, which is
 525 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
```

⁴*searchlight* or *sphere* in the MVPA parlance

528 B Simulation Details

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

532 Each simulation is based on 4,000 replications. In each replication, we
533 generate n i.i.d. samples from a shift model $\mathbf{x}_i = \mu \mathbf{y}_i^* + \eta_i$. Where $y_i^* = \{0, 1\}$
534 is the class of subject i in dummy coding. Recalling that $y_i = \{-1, 1\}$ is the
535 class in effect coding, then clearly $y_i = 2y_i^* - 1$. The noise is distributed as
536 $\eta_i \sim \mathcal{N}_p(0, \Sigma)$. The sample size $n = 40$. The dimension of the data is $p = 23$.
537 The covariance $\Sigma = I$. Effects, i.e. shifts μ , are equal coordinate p -vectors
538 with coordinates that vary over $\mu \in \{0, 1/4, 1/2\}$.

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

C Simulation Results

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

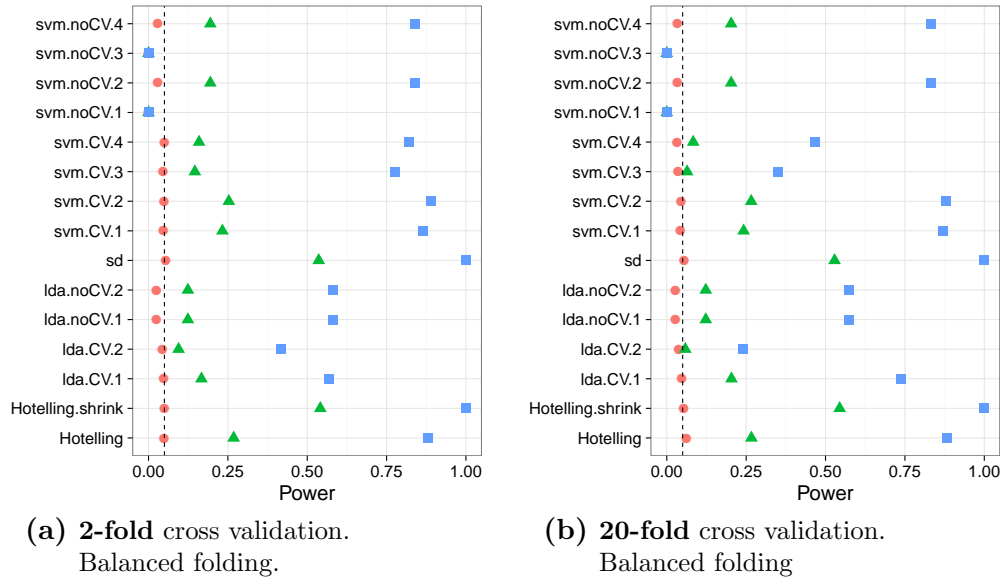


Figure 7: Simulation details in Appendix B except the changes in the sub-captions.

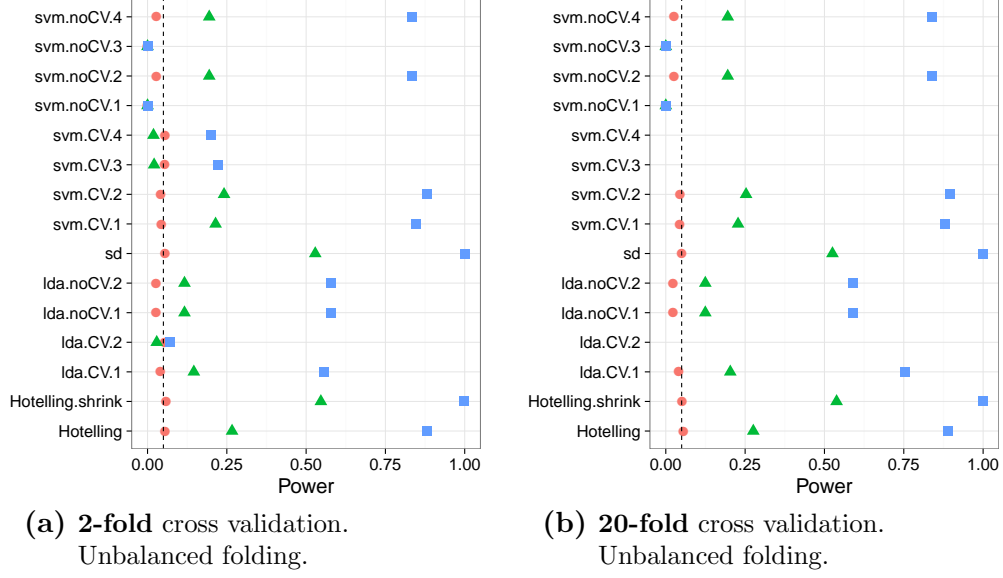


Figure 8: Simulation details in Appendix B except the changes in the sub-captions.

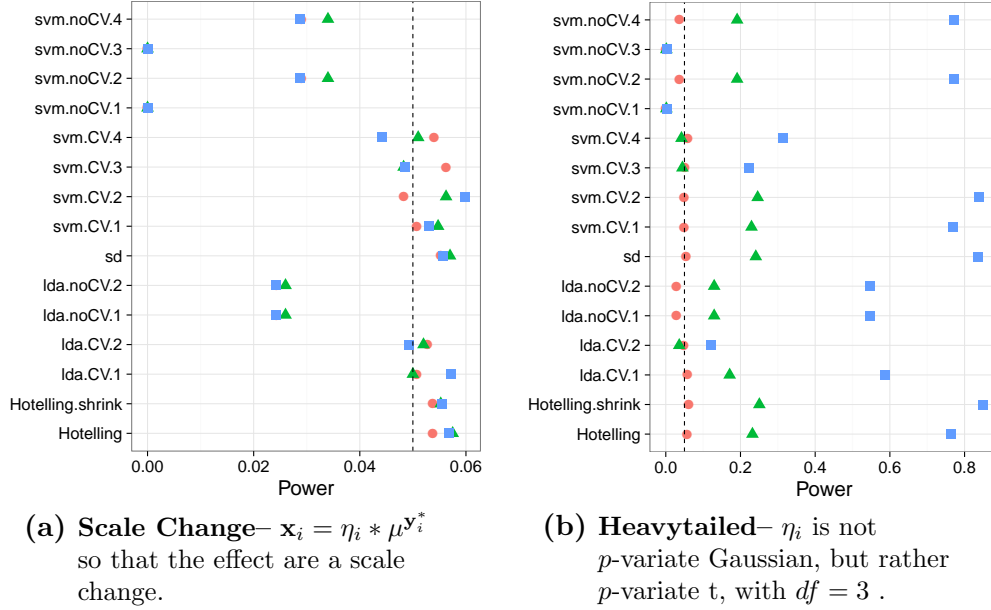
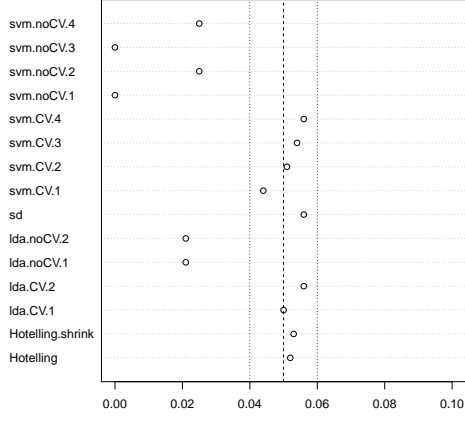
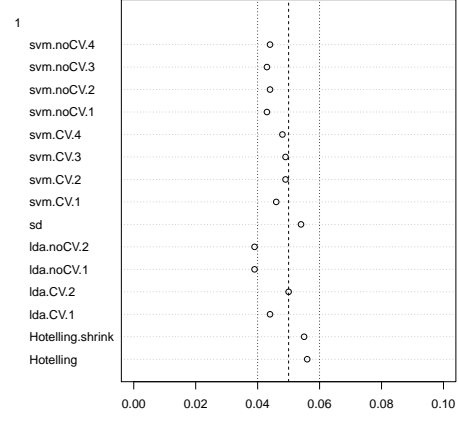


Figure 9: Simulation details in Appendix B except the changes in the sub-captions.

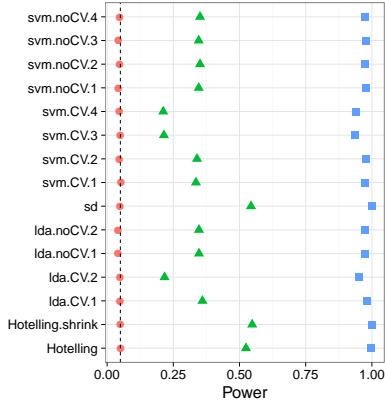


(a) **Low-Dimension**– False positive rates for $n = 40$.

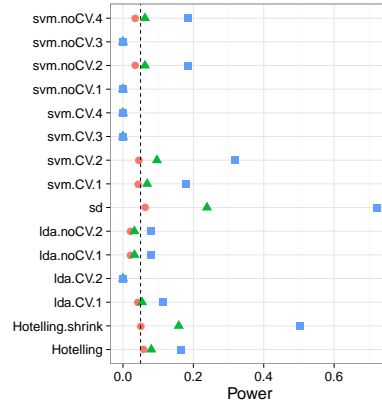


(b) **High-Dimension**– False positive rates for $n = 400$.

Figure 10: Simulation details in Appendix B except the changes in the sub-captions.



(a) **High-Dimension, local alternative**–
 $n = 400$,
 $\mu \in \frac{1}{\sqrt{10}} \times \{0, 1/4, 1/2\}$.



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