

Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt Roei Gilron Roy Mukamel

August 14, 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 very 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 then call upon a two-group population test such as
31 Hotelling’s T^2 [Anderson, 2003]. Alternatively, if the size of a brain re-
32 gion is large compared to the number of observations, so that the spatial
33 covariance cannot be fully estimated, then a high dimensional version of
34 Hotelling’s test can be called upon, such as in Schäfer and Strimmer [2005]
35 or Srivastava [2007]. For brevity, and in contrast to *accuracy tests*, we will
36 call any two-sample multivariate tests simply *population tests*, also termed
37 *class comparisons*. [TODO: rename to parameter test?]

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

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

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

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

¹GoogleScholar. Accessed on Aug 4, 2016.

65 shown to have crucial impact. Since typical sample sizes in neuroscience are
 66 not large, we seek to study which test is to be preferred in finite samples?
 67 Our conclusion will be quite simple: *population tests almost always have more*
 68 *power than accuracy tests.*

69 Our statement rests upon the observation that with typical sample sizes,
 70 the accuracy test statistic is highly discrete. Permutation testing with dis-
 71 crete test statistics are known to be conservative [Hemerik and Goeman,
 72 2014], since they are insensitive to mild perturbations of the data, and they
 73 cannot exhaust the permissible false positive rate. The degree of discretiza-
 74 tion is governed by the number of samples. In our neuroscience example
 75 from Gilron et al. [2016], the classification is performed based on 40 trials,
 76 so that the test statistic may assume only 40 possible values. This number
 77 of examples is not unusual if considering this is the number of trial-repeats,
 78 or the number of subjects in an neuroimaging study.

79 The discretization effect is aggravated if the test statistic is highly concen-
 80 trated. For an intuition consider the usage of a the *resubstitution accuracy*
 81 as a test statistic. This statistic simply means that the accuracy is not cross
 82 validated. If the data is high dimensional, the resubstitution accuracy will be
 83 very high due to over fitting. In a very high dimensional model, the resubsti-
 84 tution accuracy will be 1 for the observed data [McLachlan, 1976, Theorem
 85 1], but also for any permutation. The concentration of resubstitution accu-
 86 racy near 1, and its discreteness, render this test completely useless, with a
 87 power tending to 0 for any (fixed) effect size, as the dimension of the model
 88 grows.

89 To compare the power of accuracy tests and population tests in finite sam-
 90 ples, we perform a simulation study of a battery of test statistics. We start
 91 with formalizing the problem in Section 2. The main findings are reported
 92 in Sections 4 and 5. A discussion follows in Section 6.

93 2 Problem setup

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

97 Given n pairs of (x_i, y_i) , typically assumed i.i.d., a population test amounts
 98 to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$. I.e.,
 99 we test if the multivariate voxel activation pattern has the same distribution
 100 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, $\mathcal{A}_S(x) : \mathcal{X} \rightarrow \mathcal{Y}$, is the output of a learning algorithm \mathcal{A} when applied to the dataset, $\mathcal{A} : \mathcal{S} \rightarrow \mathcal{A}_S(x)$. The accuracy of predictor $\mathcal{A}_S(x)$ is defined as the probability of $\mathcal{A}_S(x)$ 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

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

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

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

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

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

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

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

107 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, 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

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

108 2.1 Candidate Tests

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

- 111 1. Is $\hat{\mathcal{E}}_{\mathcal{A}}$ cross validated or not?

112 2. For a V-fold cross validated test statistic:

113 (a) Should the data be refolded in each permutation?

114 (b) Should the data folding be balanced (a.k.a. stratified)?

115 (c) How many folds?

116 3. How to estimate $\hat{\mathcal{E}}_{\mathcal{A}}$?

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

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

126 **Balanced folding?** The standard practice when cross validating is to con-
127 strain the data folds to be balanced, i.e. stratified [e.g. Ojala and Garriga,
128 2010]. This means that each fold has the same number of examples from
129 each class. We will report results with both balanced and unbalanced data
130 foldings, only to discover, it does not really matter.

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

136 **How many folds?** Different authors suggest different rules for the number
137 of folds. We will be varying the number of folds, and ultimately discover that
138 the power *decreases with the number of folds*.

How to estimate accuracy? Given a predictor for the i 'th data example², $\mathcal{A}_S(x)$, a natural test statistic its empirical accuracy

$$\hat{\mathcal{E}}_{\mathcal{A}_S(x)} := \sum_i I(\mathcal{A}_S(x)(x_i) = y_i). \quad (5)$$

Since low accuracies, even 0, are evidence that the classes are separated, we can consider the departure from chance level $|\hat{\mathcal{E}}_{\mathcal{A}_S(x)} - 0.5|$, as another 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 $|\hat{\mathcal{E}}_{\mathcal{A}_S(x)} - \hat{\pi}|$. 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 propose the z-scored accuracy statistic, $\hat{\mathcal{Z}}_{\mathcal{A}_S(x)}$:

$$\hat{\mathcal{Z}}_{\mathcal{A}_S(x)} := \frac{|\hat{\mathcal{E}}_{\mathcal{A}_S(x)} - \hat{\pi}|}{\sqrt{\hat{\pi}(1 - \hat{\pi})}}. \quad (6)$$

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

²We add the unorthodox observation index i to $\mathcal{A}_S(x)$ to accommodate for cross validation methods, where each prediction is made with a different classifier, fit to a slightly different dataset.

Name	Basis	CV	Accuracy	Parameters
Hotelling	Hotelling	—	—	—
Hotelling.shrink	Hotelling	—	—	—
lda.CV.1	LDA	V-fold	$\hat{\mathcal{E}}$	—
lda.CV.2	LDA	V-fold	$\hat{\mathcal{Z}}$	—
lda.noCV.1	LDA	—	accuracy	—
lda.noCV.2	LDA	—	z-accuracy	—
sd	SD	—	—	—
svm.CV.1	SVM	V-fold	accuracy	cost=1e1
svm.CV.2	SVM	V-fold	accuracy	cost=1e-1
svm.CV.3	SVM	V-fold	z-accuracy	cost=1e1
svm.CV.4	SVM	V-fold	z-accuracy	cost=1e-1
svm.noCV.1	SVM	—	accuracy	cost=1e1
svm.noCV.2	SVM	—	accuracy	cost=1e-1
svm.noCV.3	SVM	—	z-accuracy	cost=1e1
svm.noCV.4	SVM	—	z-accuracy	cost=1e-1

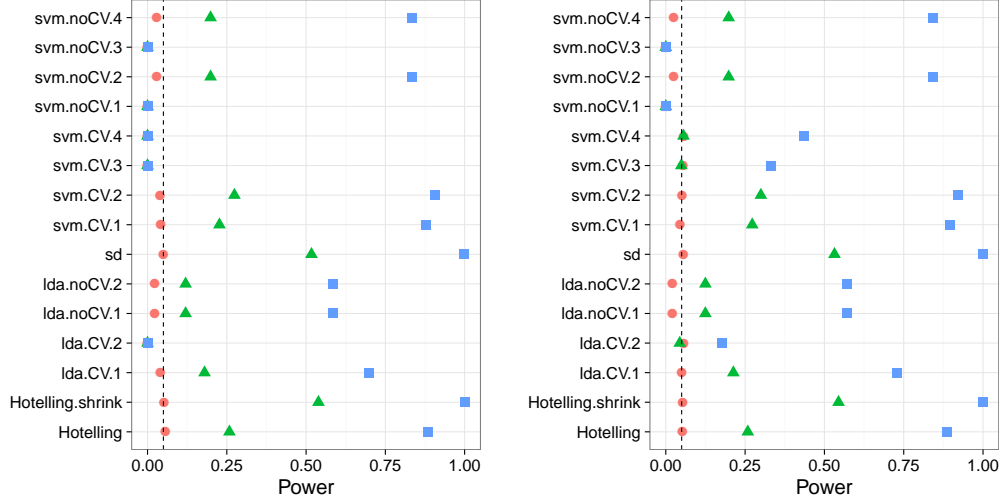
Table 1: This table collects the various test statistics we will be studying. Three are population tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group T^2 statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer and Strimmer [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 implemented with the *svm* R function, the cost parameter set at 0.1, and using the cross validated z-scored accuracy in Eq. 6. Another example is *lda.noCV.1*, which is Fisher’s LDA, returning the resubstitution accuracy.

140

141 3 Controlling the False Positive Rate

142 Figure 1 demonstrates that all of the tests considered conserve the desired
143 0.05 false positive rate, up to varying levels of conservatism. This can be
144 seen from the fact that the probability of rejection is no larger than 0.05 in
145 the absence of any effect, encoded by a red circle. This is true, in particular
146 if: (a) the folds are balanced or not, (b) the tuning parameters of some test
147 statistic are varied, (d) the number of folds is varied. We also observe that
148 the most conservative tests are the resubstitution accuracy statistics. We
149 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.



(a) Unbalanced.

(b) Balanced.

4 Power

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

1. population tests have more power than accuracy tests in all our configurations.
2. The conservativeness decays as the sample grows (Figures 9a, 9b and 10a)
3. For heavy tailed distributions (Figure 8b), the extra power of the location test vanishes.
4. The presence of correlations between coordinates reduces the signal to noise ratio (SNR), thus reduces power. More importantly, in the presence of correlations the effect of regularization is amplified, increasing the power difference between regularized and non-regularized test

165 statistics. Put differently- in low SNR regimes, regularization proves
166 crucial (Figure 10b).

167 5. The z-scoring of the accuracies was introduced to deal with unbalanced
168 foldings. If the z-scoring has any effect at all, it merely kills power.

169 6. Both accuracy and population tests are inappropriate for scale alter-
170 natives (Figure 8a). This was to be expected and is reported mostly as
171 a sanity check.

172 7. Balanced folding only affects the z-scored accuracy, in the opposite
173 direction than we anticipated.

174 8. Increasing the SVM’s cost parameter, which reduces the number of
175 support vectors entering the classifier, reduces power.

176 The major insight from simulations is that the use of accuracy tests for
177 signal detection is underpowered compared to population tests. We now
178 verify this finding on a neuroimaging dataset.

179 5 Neuroimaging Example

180 Figure 2 is an application of both a location and an accuracy test to the data
181 of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI
182 data while subjects were exposed to the sounds of human speech (vocal),
183 and other non-vocal sounds. Each subject was exposed to 20 sounds of each
184 type, totaling in $n = 40$ trials in each scan. The study was rather large and
185 consisted of about 200 subjects. The data was kindly made available by the
186 authors at the OpenfMRI website³.

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

192 In agreement with our simulation results, the population test (*sd*) dis-
193 covers more brain regions of interest when compared to an accuracy test
194 (*svm.cv.1*). The former discovers 1,232 regions, while the latter only 441, as
195 depicted in Figure 2. We emphasize that both test statistics were compared
196 with the same permutation scheme, and the same error controls, so that any
197 difference in detections is due to their different power.

³<https://openfmri.org/>

198 Having established that accuracy tests are typically underpowered for sig-
 199 nal detection compared to population tests, we wish to identify the conditions
 200 under which this will occur, and discuss practical implications.

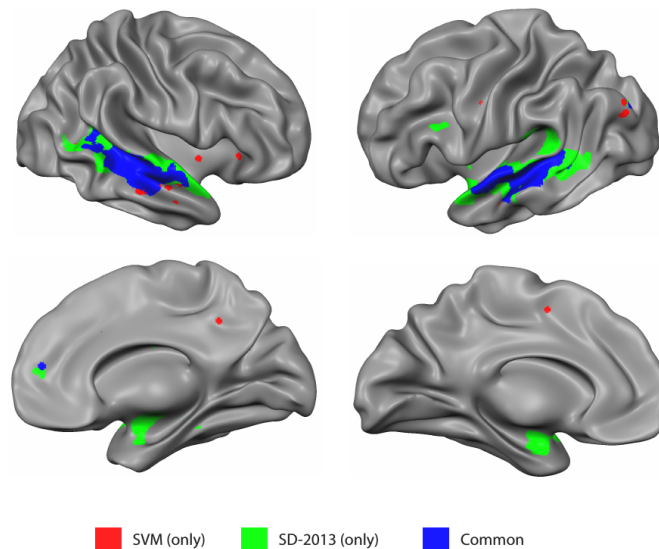


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a population test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost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

201 6 Discussion

202 We have set out to understand which of the tests is more powerful: the
 203 accuracy test or the population test. No amount of simulations can replace
 204 the insight provided by a good closed-form analytic result. The finite sample
 205 power of permutation tests is a formidable mathematical problem, so we
 206 currently content ourselves with simulations We have concluded that the
 207 population tests are typically preferable. Their high dimensional versions,
 208 such as Srivastava [2007] and Schäfer and Strimmer [2005], are particularly
 209 well suited for neuroimaging problems such as MVPA. We attribute this

210 to several phenomena: (a) Discretization introduced in finite samples by
 211 the accuracy test statistic. (b) Inefficient use of the data for the validation
 212 holdout set. (c) Regularization crucial in high dimensional problems.

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

217 The degree of discretization is governed by the sample size. For this
 218 reason, an asymptotic analysis such as Ramdas et al. [2016] may uncover the
 219 holdout inefficiency, but will not uncover the discretization effect.

220 The practical advice for the practitioner, is that for the purpose of signal
 221 detection, there is typically a population test that is more powerful than
 222 an accuracy test. There is also a good chance that it would be easier to
 223 implement, and faster to run, since no cross validation will be involved.

224 6.1 Ease of implementation

225 A very important consideration is the ease of implementation. The need for
 226 cross validation of the accuracy test greatly increases its computational com-
 227 plexity. Moreover, anyone who has actually implemented tests with discrete
 228 statistics, will attest they are more prone to programming errors. This is
 229 because their unforgiveness to the type of inequalities used. Indeed, mistak-
 230 enly replacing a weak inequality with a strong inequality in one’s program
 231 may considerably change the results. This is not the case for continuous test
 232 statistics.

233 6.2 Reservations

234 Some reservations to the generality of our findings are in order. Firstly,
 235 not all accuracy tests are concerned with signal detection. Consider brain
 236 decoding for machine interfaces, or clinical diagnosis, where the presence of
 237 a medical condition is predicted from imaging data [e.g. Olivetti et al., 2012,
 238 Wager et al., 2013]. In those examples, the purpose of the test is not to
 239 detect a difference between classes, but to actually test the performance of a
 240 particular classifier.

241 Secondly, it may be argued that accuracy tests permits the separation
 242 between classes in high dimensions, such as in *reproducing kernel Hilbert*
 243 *spaces* (RKHS) by using non-linear predictors. This is a false argument—
 244 accuracy test do not have any more flexibility than population tests. Indeed,
 245 it is possible to test for location in the same dimension the classifier is learned.
 246 Gretton et al. [2012] is an example where the test for location is performed

in the RKHS of the data. It is also possible to test for the equality of two multivariate distributions [TODO: cite vogelstein]. On the other hand, based on our reported neuroimaging example, and others, we find that a population test in the original feature space is indeed 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 is easier than permuting labels. It allows to preserve balanced folds after a permutation without refolding. Although we did not find a power difference between balanced and unbalanced foldings.

Use less folds. For V-fold CV, power decreases as the number of folds increases. This is quite interesting since two phenomena compete as the number of folds increase: (a) the train set is larger so that better accuracies are achievable. (b) The test set is smaller so that the accuracy estimate is more variable. The decrease in power with increase fold number suggests that the latter dominates the former. Put differently: it is easier to detect a small stable departure from chance level, than a large but unstable one.

Resubstitution accuracy in low dimension. Resubstitution accuracy is useful in low dimension. 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 computational burden of cross validation can be avoided by using the resubstitution accuracy. There is a fundamental difference between V-folding and resubstitution. The latter should not be thought of as the limit of the former.

281 **Regularize** Regularizing the accuracy test proves very useful in high di-
 282 mensional problems. Put differently: reducing variance by adding some bias
 283 is very useful to detect better-than-chance classification.

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

286 6.4 Smoothing accuracy estimates

287 It may be possible to alleviate the effect of discretization by appropriate
 288 cross-validation. The discreteness of the accuracy statistic is governed by
 289 the number of examples in the union (over all validation iterations) of test
 290 sets. For V-fold CV, for instance, this number is simply the sample size. This
 291 suggests that the accuracy can be “smoothed” by allowing the test sample to
 292 be drawn with replacement. The *bootstrap* may seem like a good candidate
 293 approach since it samples examples with replacement. It does so, however,
 294 for the train set, and not the test set. An algorithm that samples test sets
 295 with replacement is the *leave-one-out bootstrap estimator* (bLOO) and its
 296 derivation– the *0.632 bootstrap estimator* (b0.632) [Hastie et al., 2003, Sec
 297 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:

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

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,

$$\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} = \frac{1}{B} \sum_{b=1}^B \frac{1}{|S^{(b)}|} \sum_{i \in S^{(b)}} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i) = y_i\}. \quad (8)$$

Definition 4 (b0.632). Denoting by \mathcal{E}_{resub} the resubstitution accuracy estimate, the b0.632 accuracy estimator, $\mathcal{E}_{0.632}$, is defined as

$$\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} := 0.368 \hat{\mathcal{E}}_{\mathcal{A}}^{Resub} + 0.632 \hat{\mathcal{E}}_{\mathcal{A}}^{bLOO}. \quad (9)$$

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, itself reducing the level of discretization. The type of bootstrap, bLOO versus b0.632, does not change the power.

Name	Basis	Type	B	Accuracy	Parameters
lda.Boot.1	LDA	b0.632	10	accuracy	—
lda.Boot.2	LDA	bLOO	10	accuracy	—
svm.Boot.1	SVM	b0.632	10	accuracy	cost=1e1
svm.Boot.2	SVM	bLOO	10	accuracy	cost=1e1
svm.Boot.3	SVM	b0.632	50	accuracy	cost=1e1
svm.Boot.4	SVM	bLOO	50	accuracy	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 to 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 (*sd* and *Hotelling.shrink*), 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], which can be thought of a method intersecting Fisher’s LDA and 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 was to be expected, since we already saw in (e.g. Figure 1a) that the unregularized parameter test, *Hotelling*, is slightly more powerful than the

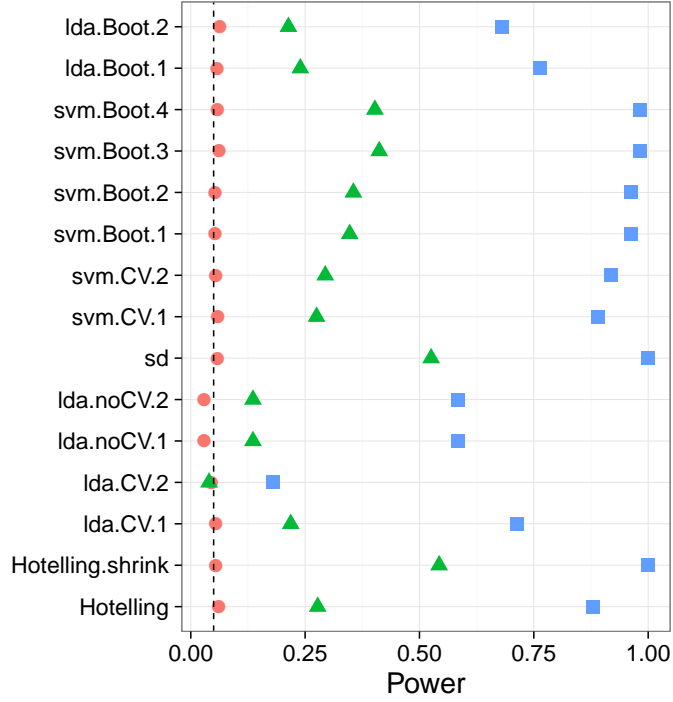


Figure 3: Bootstrap— The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square).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

324 regularized accuracy test, *svm.CV.1* for instance.

325 We can compound regularization in this section with the bootstrapping
326 from Section 6.4, to improve finite sample power of the accuracy tests. This
327 is done in the *svm.highdim.2* test, which still falls short from the power of the
328 location tests, but is a much more powerful accuracy test than the original
329 non-regularized, V-fold validated, version of *svm.CV.1*.

Name	Basis	CV	Accuracy	Parameters
svm.highdim.1	SVM	V-fold	accuracy	cost=1e-1
svm.highdim.2	SVM	B=50	accuracy	cost=1e-1
lda.highdim.1	LDA	V-fold	accuracy	—
lda.highdim.2	LDA	V-fold	accuracy	—
lda.highdim.3	LDA	V-fold	accuracy	—

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].

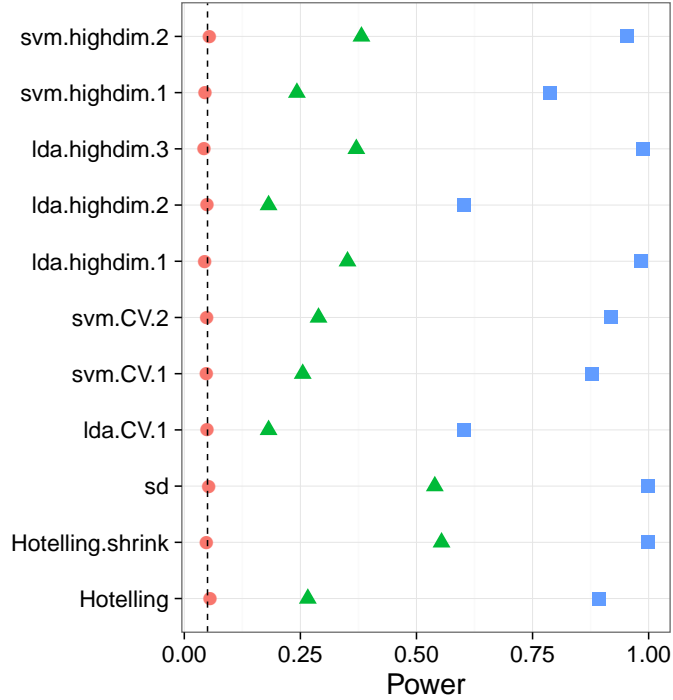


Figure 4: **HighDim Classifier**— The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 3. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5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

6.6 Related Literature

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

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

Golland et al. [2005] study accuracy tests using simulation, neuroimaging data, genetic data, and analytically. Their analytic results formalize our intuition from Section 1 on the effect of concentration of the accuracy statistic: The finite Vapnik–Chervonenkis (VC) dimension requirement [Golland and Fischl, 2003, 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 (Figure 4, middle). This is seen in their Figure 4, 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 shift model ($p = 0$). We now perform the same simulation as Golland et al. [2005], after parameterizing p so that $p = 0$ corresponds to the null model, and in the same dimensionality as our previous simulations. We find that also in this mixture class of models a population test has more power than an accuracy test (Figure 5).

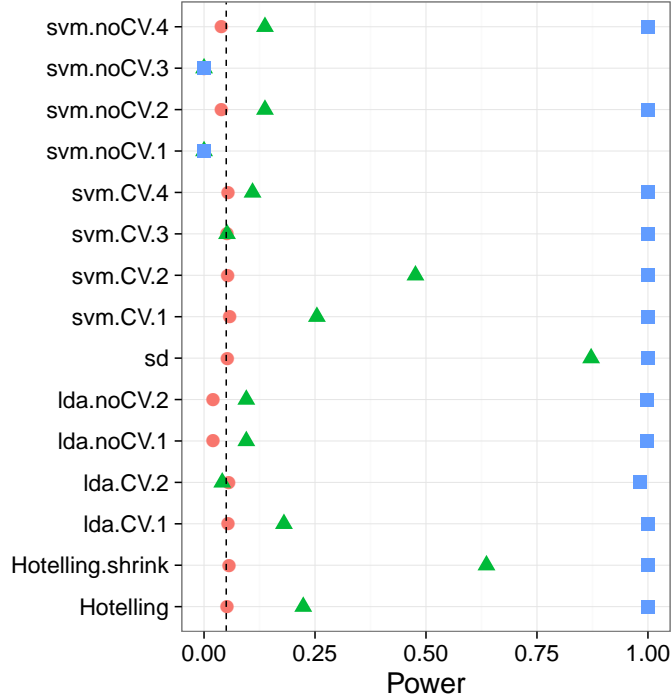


Figure 5: **Mixture**— $\mathbf{x}_i = \chi_i \mu + \eta_i$; $\chi_i = \{-1, 1\}$ and $Prob(\chi_i = 1) = (1/2 - p)^{y_i^*} (1/2 + p)^{1-y_i^*}$. μ is a p -vector with $3/\sqrt{p}$ in all coordinates. The effect, p , is color and shape coded and varies over 0 (red circle), $1/4$ (green triangle) and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

369 6.7 Epilogue

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

376 The recent popularity of machine learning has resulted in the ex-
 377 tensive teaching and use of prediction in theoretical and applied
 378 communities and the relative lack of awareness or popularity of
 379 the topic of Neyman-Pearson style hypothesis testing in the com-
 380 puter science and related “data science” communities.

381 And more simply by Frank Harrell in the CrossValidated Q&A site⁴:

⁴[http://stats.stackexchange.com/questions/17408/
how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier](http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier).

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

385 **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.

- 420 T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, July 2003. ISBN 0-387-95284-5.
421
- 422 J. Hemerik and J. Goeman. Exact testing with random permutations.
423 *arXiv:1411.7565 [math, stat]*, Nov. 2014.
- 424 H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Math-*
425 *ematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990.
426 doi: 10.1214/aoms/1177732979.
- 427 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for
428 prediction error in microarray classification using resampling. *Statistical*
429 *Applications in Genetics and Molecular Biology*, 7(1), 2008.
- 430 L. Juan and H. Iba. Prediction of tumor outcome based on gene expression
431 data. *Wuhan University Journal of Natural Sciences*, 9(2):177–182, Mar.
432 2004. ISSN 1007-1202, 1993-4998. doi: 10.1007/BF02830598.
- 433 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional
434 brain mapping. *Proceedings of the National Academy of Sciences of the*
435 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,
436 1091-6490. doi: 10.1073/pnas.0600244103.
- 437 E. L. Lehmann. Parametric versus nonparametrics: two alternative method-
438 ologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN
439 1048-5252. doi: 10.1080/10485250902842727.
- 440 G. J. McLachlan. The bias of the apparent error rate in discriminant analysis.
441 *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi:
442 10.1093/biomet/63.2.239.
- 443 S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell,
444 T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements
445 for classifying DNA microarray data. *Journal of Computational Biology:*
446 *A Journal of Computational Molecular Cell Biology*, 10(2):119–142, 2003.
447 ISSN 1066-5277. doi: 10.1089/106652703321825928.
- 448 M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Perfor-
449 mance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.
450 ISSN ISSN 1533-7928.
- 451 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
452 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
453 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*

- 454 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
 455 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
 456 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.
- 457 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the
 458 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.
 459 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.
- 460 H. Pang, T. Tong, and H. Zhao. Shrinkage-based Diagonal Discriminant
 461 Analysis and Its Applications in High-Dimensional Data. *Biometrics*, 65
 462 (4):1021–1029, Dec. 2009. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2009.
 463 01200.x.
- 464 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
 465 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
 466 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 467 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
 468 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
 469 P. Belin. The human voice areas: Spatial organization and inter-individual
 470 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
 471 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 472 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
 473 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*
 474 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
 475 106652702760138592.
- 476 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
 477 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 478 J. A. Ramey, C. K. Stein, P. D. Young, and D. M. Young. High-Dimensional
 479 Regularized Discriminant Analysis. *arXiv preprint arXiv:1602.01182*,
 480 2016.
- 481 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance
 482 Matrix Estimation and Implications for Functional Genomics. *Statistical*
 483 *Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN
 484 1544-6115. doi: 10.2202/1544-6115.1175.
- 485 D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class
 486 Prediction and Discovery Using Gene Expression Data. In *Proceedings of*
 487 *the Fourth Annual International Conference on Computational Molecular*

- 488 *Biology*, RECOMB '00, pages 263–272, New York, NY, USA, 2000. ACM.
489 ISBN 978-1-58113-186-4. doi: 10.1145/332306.332564.
- 490 M. S. Srivastava. Multivariate Theory for Analyzing High Dimensional Data.
491 *Journal of the Japan Statistical Society*, 37(1):53–86, 2007. doi: 10.14490/
492 jjss.37.53.
- 493 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
494 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
495 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 496 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
497 ing correction in classification-based multi-voxel pattern analysis (MVPA):
498 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
499 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.
- 500 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,
501 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-
502 2.
- 503 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
504 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
505 caveats, and guidelines. working paper or preprint, June 2016.
- 506 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.
507 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-
508 nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:
509 10.1056/NEJMoa1204471.

510 A Analysis pipeline

511 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 512 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 513 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions⁵ are
 514 centered around a unique voxel, the voxel index v also serves as a unique
 515 region index. Algorithm 1 computes a region-wise test statistic, which is
 516 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
```

⁵*searchlight* or *sphere* in the MVPA parlance

519 B Simulation Details

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

523 Each simulation is based on 4,000 replications. In each replication, we
524 generate n i.i.d. samples from a shift model $\mathbf{x}_i = \mu \mathbf{y}_i^* + \eta_i$. Where $y_i^* = \{0, 1\}$
525 is the class of subject i in dummy coding. Recalling that $y_i = \{-1, 1\}$ is the
526 class in effect coding, then clearly $y_i = 2y_i^* - 1$. The noise is distributed as
527 $\eta_i \sim \mathcal{N}_p(0, \Sigma)$. The sample size $n = 40$. The dimension of the data is $p = 23$.
528 The covariance $\Sigma = I$. Effects, i.e. shifts μ , are equal coordinate p -vectors
529 with coordinates that vary over $\mu \in \{0, 1/4, 1/2\}$.

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

C Simulation Results

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

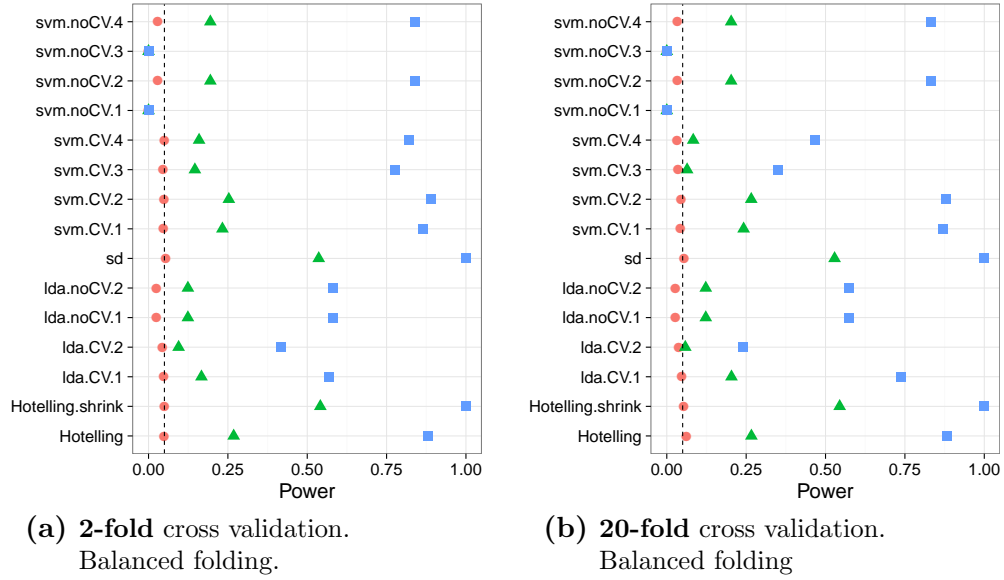


Figure 7: Simulation details in Appendix B except the changes in the sub-captions.

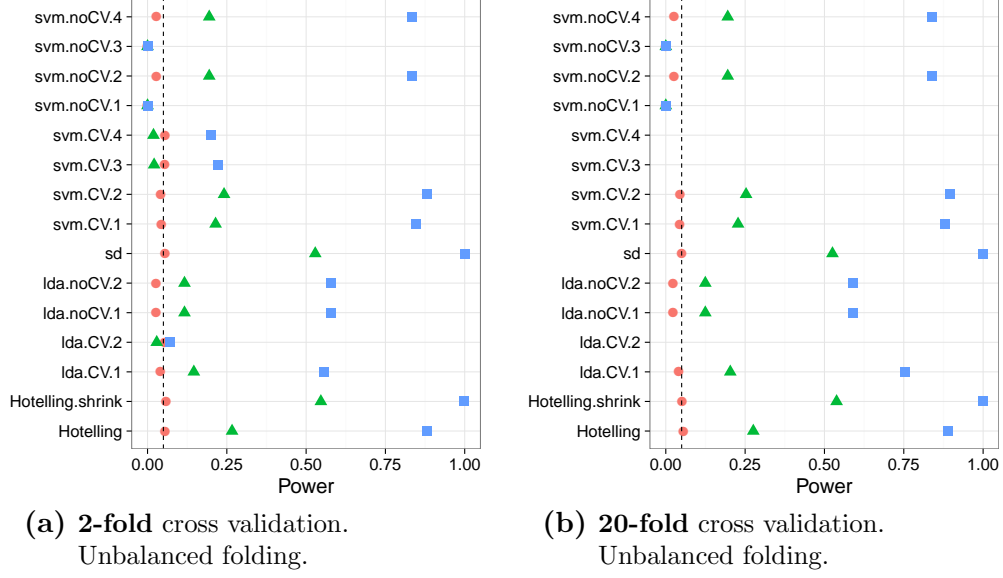


Figure 8: Simulation details in Appendix B except the changes in the sub-captions.

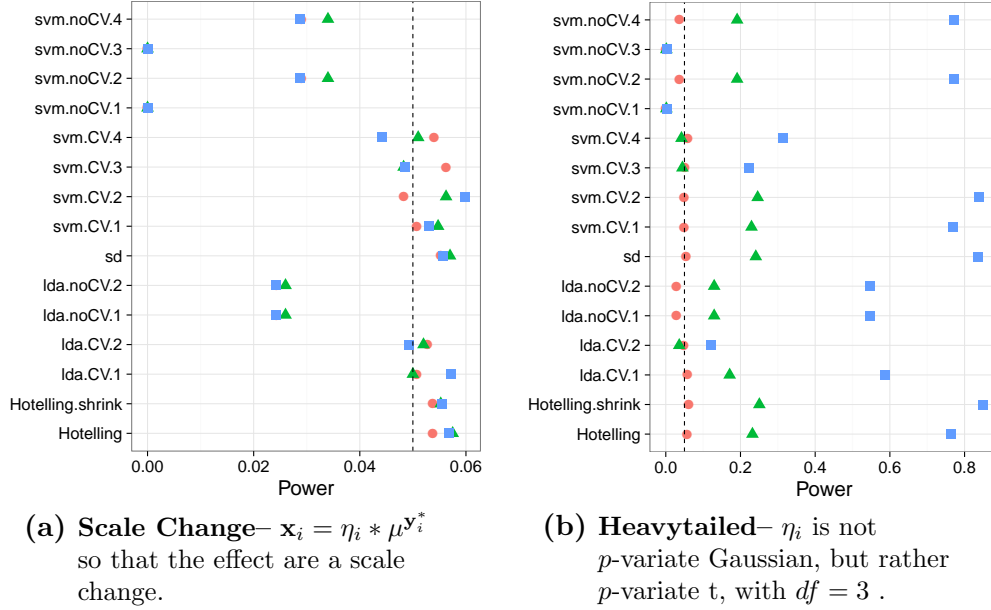
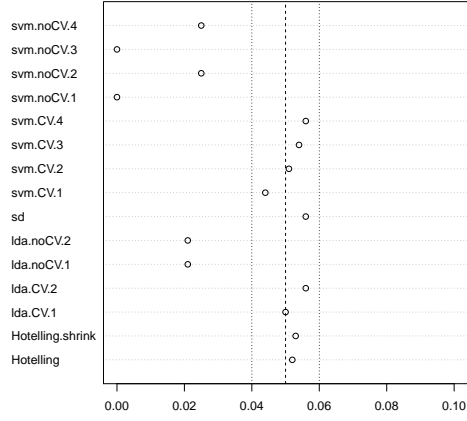
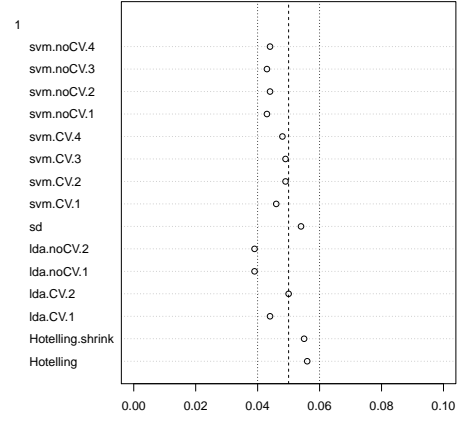


Figure 9: Simulation details in Appendix B except the changes in the sub-captions.

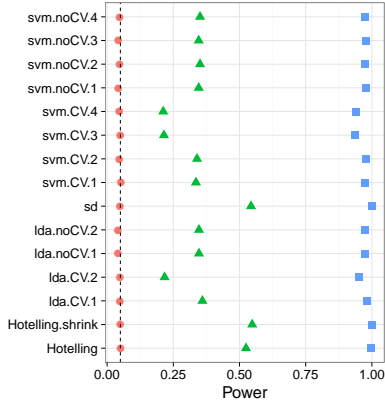


(a) **Low-Dimension**— False positive rates for $n = 40$.

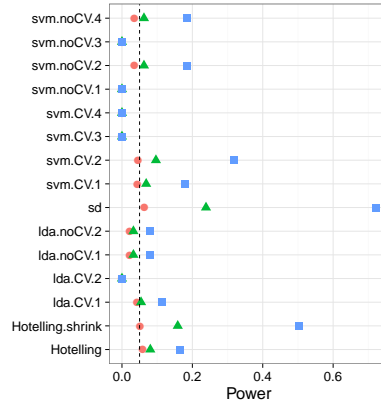


(b) **High-Dimension**— False positive rates for $n = 400$.

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