

# Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt      Roei Gilron      Roy Mukamel

August 9, 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 location test such as Hotelling’s  
 31  $T^2$  [Anderson, 2003]. Alternatively, if the size of a brain region is too large  
 32 compared to the number of observations, so that the spatial covariance can-  
 33 not be fully estimated, then a high dimensional version of Hotelling’s test can  
 34 be called upon, such as in Schäfer and Strimmer [2005] or Srivastava [2013].  
 35 For brevity, and in contrast to *accuracy tests*, we will call any two-sample  
 36 multivariate tests simply *location tests*, also termed *class comparisons*.

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

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

46 Asymptotic relative efficiency measures (ARE) are typically used by statis-  
 47 ticians to compare between test statistics with similar rates [van der Vaart,  
 48 1998]. Ramdas et al. [2016] derive the asymptotic power functions of the  
 49 two test statistics, which allow to extract the ARE between Hotelling’s  $T^2$   
 50 (location) test and Fisher’s LDA (accuracy) test. Using the Theorem 14.7 in  
 51 van der Vaart [1998], we deduce that the ARE is lower bounded by  $2\pi \approx 6.3$ .  
 52 This means that Fisher’s LDA requires at least 6.3 more samples to achieve  
 53 the same (asymptotic) power than the  $T^2$  test. In this light, the accuracy test  
 54 is remarkably inefficient compared to the location test. For comparison, the  
 55 t-test is only 1.04 more (asymptotically) efficient than Wilcoxon’s rank-sum  
 56 test [Lehmann, 2009], so that an ARE of 2.5 is strong evidence in favor of  
 57 the location test.

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

---

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

65 Our conclusion will be quite simple: *location tests almost always have more*  
 66 *power than accuracy tests.*

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

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

86 To compare the power of accuracy tests and location tests in finite sam-  
 87 ples, we perform a simulation study of a battery of test statistics. The main  
 88 findings are reported in Sections 4 and 5, and the intuition for our findings  
 89 is provided in Section 6, but first, the problem’s setup.

## 90 2 Problem setup

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

94 Given  $n$  pairs of  $(x_i, y_i)$ , typically assumed i.i.d., a location test amounts  
 95 to testing whether  $x|y = 1$  has the the same distribution as  $x|y = -1$ . I.e.,  
 96 we test if the multivariate voxel activation pattern has the same distribution  
 97 when given a vocal stimulus, as when given a non-vocal stimulus. An accu-  
 98 racy test amounts to learning a predictive model  $\hat{f}(x)$  from some assumed  
 99 model class  $\hat{f} \in \mathcal{F}$ . The prediction accuracy, denoted  $T_{\hat{f}}^{acc}$ , is defined as the  
 100 probability of a given classifier  $\hat{f}$  of making a correct prediction. Denoting by  
 101  $I(A)$  the indicator function of the event  $A$ , we have  $T_{\hat{f}}^{acc} := \mathbf{E} \left[ I(\hat{f}(x) = y) \right]$

when given a randomly drawn data point,  $(x, y)$ . A statistically significant “better than chance” estimate of  $T_{\hat{f}}^{acc}$  is evidence that the classes are distinct.

## 2.1 Candidate Tests

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

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

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

**How to estimate accuracy?** Given a predictor  $\hat{f}$ , a natural test statistic is some estimate of its accuracy  $T_{\hat{f}}^{acc}$ . Complicating matters: very low accuracies, even 0, is evidence that the classes are separated, and we only need to invert the predictions. We can thus consider  $|T_{\hat{f}}^{acc} - 0.5|$  as the test statistic. This, however, implies that if the classes are identical, random guessing has 0.5 accuracy. This is not true if the classes are not balanced. For unbalanced data the chance level is the probability of the minority class, we denote by  $\hat{p}_{min}$  [Golland et al., 2005, Sec 4.1]. This suggests the following test statistic  $|T_{\hat{f}}^{acc} - \hat{p}_{min}|$ . Since we will be aggregating these statistics over random data sets where  $\hat{p}_{min}$  may vary, it seems appropriate to standardize the scale of this statistic. We thus also consider the z-scored accuracy:  $|T_{\hat{f}}^{acc} - \hat{p}_{min}| / \sqrt{\hat{p}_{min}(1 - \hat{p}_{min})}$ .

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

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

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

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

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

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

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

Name	Basis	CV	Accuracy	Parameters
Hotelling	Hotelling	—	—	shrink=FALSE
Hotelling.shrink	Hotelling	—	—	shrink=TRUE
lda.CV.1	LDA	TRUE	accuracy	—
lda.CV.2	LDA	TRUE	z-accuracy	—
lda.noCV.1	LDA	FALSE	accuracy	—
lda.noCV.2	LDA	FALSE	z-accuracy	—
sd	SD	—	—	—
svm.CV.1	SVM	TRUE	accuracy	cost=1e1
svm.CV.2	SVM	TRUE	accuracy	cost=1e-1
svm.CV.3	SVM	TRUE	z-accuracy	cost=1e1
svm.CV.4	SVM	TRUE	z-accuracy	cost=1e-1
svm.noCV.1	SVM	FALSE	accuracy	cost=1e1
svm.noCV.2	SVM	FALSE	accuracy	cost=1e-1
svm.noCV.3	SVM	FALSE	z-accuracy	cost=1e1
svm.noCV.4	SVM	FALSE	z-accuracy	cost=1e-1

Table 1: This table enumerates the various test statistics we will be studying. Three are location tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group  $T^2$  statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer and Strimmer [2005]. *sd* is another high dimensional version of the  $T^2$ , from Srivastava et al. [2013]. The rest of the tests are variations of the linear SVM, and Fisher’s LDA, with varying accuracy measures, cross validated or not, and varying tuning parameters. For example, *svm.CV.4* is a linear SVM, with *libsvm*’s cost parameter set at 0.1, using the cross validated z-scored accuracy ( $|T_f^{acc} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$ , see Section 2.1). Another example is *lda.noCV.1*, which is Fisher’s LDA, returning the resubstitution accuracy, without cross validation, and without z-scoring.

172

### 3 Controlling the False Positive Rate

173

Figure 1 demonstrates that all of the tests considered conserve the desired 0.05 false positive rate, up to varying levels of conservatism. This can be seen from the fact that the probability of rejection is no larger than 0.05 in the absence of any effect, encoded by a red circle. This is true, in particular

177

178 if: (a) the folds are balanced or not, (b) the tuning parameters of some test  
 179 statistic are varied, (d) the number of folds is varied. We also observe that  
 180 the most conservative tests are the resubstitution accuracy measures. We  
 181 return to this matter in the Discussion.

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



## 182 4 Power

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

- 187 1. Location tests have more power than accuracy tests in all our configurations.
- 188
- 189 2. The conservativeness decays as the sample grows (Figures 8a, 8b and  
 190 9a), suggesting that either concentration or discretization is responsible  
 191 for power loss.

- 192 3. The power may increase or decrease with the number of folds (Figure 5).
- 193 4. The z-scoring of the accuracies was introduced to deal with unbalanced  
194 foldings. If the z-scoring has any effect at all, it merely kills power.  
195 There is really no reason to use it.
- 196 5. Both accuracy and location tests are inappropriate for scale alternatives  
197 (Figure 7a). This was to be expected and is reported mostly as a sanity  
198 check.
- 199 6. The presence of heavy tails (Figure 7b) may reduce power, but does  
200 not quantitatively change results.
- 201 7. Balanced folding typically has no effect. It increased power only for  
202 the z-scored statistics (Figure 1). This is surprising given they were  
203 precisely designed to deal with the presence of imbalance.
- 204 8. Varying the accuracy test’s tuning parameter, such as the cost (i.e.  
205 margins) has no effect on the power of the test.
- 206 9. Correlation between coordinates, mimicking temporal correlation in  
207 fMRI data, has no effect on conclusions, since all test statistics account  
208 for this correlation (Figure 9b).

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

## 212 5 Neuroimaging Example

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

220 We perform group inference using within-subject permutations using the  
221 pipeline of Stelzer et al. [2013], which was also reported in Gilron et al. [2016].  
222 For completeness, the pipeline is described in Appendix A. To demonstrate

---

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



our point, we compare the *sd* location test with the *svm.cv.1* accuracy test (see Table 1 for the definition of these statistics).

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

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



*Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a location test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of Stelzer et al. [2013], followed by region-wise  $FDR \leq 0.05$  control using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. Number of permutations equals 400. The location test detect 1,232 regions, and the accuracy test 441, 399 of which are common to both. For the details of the analysis see Appendix A and Gilron et al. [2016].*

## 234 6 Discussion

235 We have set out to understand which of the tests is more powerful: the ac-  
236 curacy test or the location test. Using simulations, we have concluded that  
237 the location tests are preferable. Their high dimensional versions such as  
238 Srivastava [2013] and Schäfer and Strimmer [2005] are preferable for typical  
239 neuroimaging problems such as MVPA. We attribute this to several phe-  
240 nomena: (a) Discretization introduced in finite samples by the accuracy test  
241 statistic. (b) Inefficient use of the data for the validation holdout set. In our  
242 high dimensional setup, we also confirmed that high-dimensional versions of  
243 the  $T^2$  test, such as Srivastava [2013] or Schäfer and Strimmer [2005] are  
244 preferable over the original  $T^2$ .

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

### 254 6.1 Ease of implementation

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

### 262 6.2 A good accuracy test

263 In Section 6.6 we discuss cases where an accuracy test cannot replace a  
264 location test. For such cases we collect some conclusions from our simulations  
265 on the best practices for accuracy tests.

- 266 1. The conservativeness due to discretization decreases with sample size.
- 267 2. Cross validating the accuracy statistic increases power in moderate  
268 sample sizes. The power loss due to the holdout inefficiency is smaller

- 269 than the power loss due to the concentration of the resubstitution ac-  
270 curacy. For large sample sizes, discretization and concentration have  
271 weaker effects, and the cross validated accuracy may be replaced with  
272 the computationally more efficiency resubstitution accuracy.
- 273 3. Permuting features is easier than permuting labels. It allows to preserve  
274 balanced folds after a permutation without refolding, thus reducing  
275 computational complexity.
  - 276 4. There is no gain in z-scoring the accuracy scores.
  - 277 5. Cross validated accuracy with balanced folds has more power than un-  
278 balanced folds. We currently have no intuition to offer for this phe-  
279 nomenon.
  - 280 6. It is unclear what is the effect of the number of folds. More folds in-  
281 crease power by reducing the number of holdout samples. On the other  
282 hand, it increases the concentration of the accuracy statistic. Com-  
283 pounded with the discreteness of the accuracy statistic, this decreases  
284 power.
  - 285 7. The value of the tuning parameters of a classifier have little to no  
286 effect.

### 287 6.3 Smoothing accuracy estimates

288 It may be possible to alleviate the effect of discretization by appropriate cross-  
289 validation. The discreteness of the accuracy statistic can be “smoothed” by  
290 allowing the test sample to be drawn with replacement. The *bootstrap* may  
291 seem like a candidate approach, but since the original data always serves as  
292 a test set, the accuracy can still only assume  $1/n$  values. This is not the  
293 case, however, for the *leave-one-out bootstrap estimator* (B-LOO) and the  
294 *0.632 bootstrap estimator* (B-0.632) [Hastie et al., 2003, Sec 7.11], which we  
295 define below for completeness. By the same rationale, the degree of conserva-  
296 tivism should decrease with the number of bootstrap samples. The naming  
297 conventions of the bootstrapped estimates are detailed in Table 2.

**Definition 1** (B-LOO). Denoting by  $S^b$  the indexes of the test set of Boot-  
strap sample  $b$ , the *leave-one-out* bootstrap estimate is defined as

$$T_{BLOO}^{acc} := \frac{1}{B} \sum_{b=1}^B \frac{1}{|S^b|} \sum_{i \in S^b} I(\hat{f}^b(x_i) = y_i)$$

**Definition 2** (B-0.632). Denoting by  $T_{resub}^{acc}$  the resubstitution accuracy estimate, the B-0.632 accuracy estimator,  $T_{0.632}^{acc}$ , is defined as

$$T_{0.632}^{acc} := 0.368 T_{resub}^{acc} + 0.632 T_{BLOO}^{acc}$$

298 The simulation results are reported in Figure 3. It can be seen that select-  
 299 ing test sets with replacement does increase the power, when compared to  
 300 V-fold cross validation, but still falls short from the power of location tests.  
 301 It can also be seen that power increases with the number of Bootstrap repli-  
 302 cations, itself reducing the level of discretization. The type of Bootstrap,  
 303 B-LOO versus B-0.632, does not change the power. Again, consistent with  
 304 the observation that it is discretization that drives the power loss.

Name	Basis	Boot Type	B	Accuracy	Parameters
lda.Boot.1	LDA	B-0.632	10	accuracy	—
lda.Boot.2	LDA	B-LOO	10	accuracy	—
svm.Boot.1	SVM	B-0.632	10	accuracy	cost=1e1
svm.Boot.2	SVM	B-LOO	10	accuracy	cost=1e-1
svm.Boot.3	SVM	B-0.632	50	accuracy	cost=1e1
svm.Boot.4	SVM	B-LOO	50	accuracy	cost=1e-1

Table 2: The same as Table 1 for bootstrapped accuracy estimates.

305

## 306 6.4 High dimensional classifiers

307 It is known that when  $p > n$  Hotelling’s  $T^2$ , and Fisher’s LDA are not  
 308 computable. In our simulations, in which  $p = 23$  and  $n = 40$  is “almost”  
 309 high dimensional, but still allows to compute both tests. We have simulated  
 310 two high dimensional versions of Hotelling’s  $T^2$ : *sd* [Srivastava, 2013] and  
 311 *Hotelling.shrink* [Schäfer and Strimmer, 2005]. The former solves the dimen-  
 312 sionality problem by assuming independence over coordinates, and the latter  
 313 by Tikhonov regularization of the covariance, a-la ridge regression. The cor-  
 314 responding high dimensional accuracy tests would be a *naive Bayes* classifier,  
 315 and  $l_2$  regularized SVM [Ramdas et al., 2016]. We conjecture that they would  
 316 not alter our conclusions, since the main force driving the conservatism  
 317 is discretization, which they do not solve.

## 318 6.5 Related Literature

319 Olivetti et al. [2012] and Olivetti et al. [2014] looked into the problem of  
 320 choosing a good accuracy test. They propose a new test they call an *inde-*

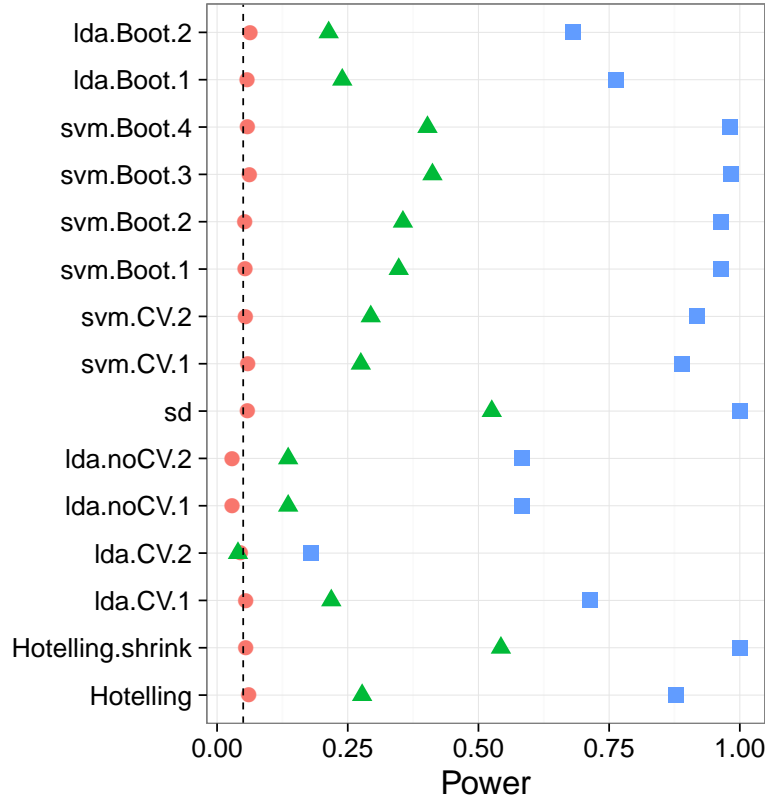


Figure 3: **Bootstrap:**

pendence test, and demonstrate by simulation that it has more power than other accuracy tests, and can deal with non-balanced data sets. We did not include this test in the battery we compared, but we note the following: (a) The independence test of Olivetti et al. [2012] relies on a discrete test statistic. This means that in the cases that the accuracy test is called upon for discriminating populations, it will probably be underpowered compared to location tests. (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. They also find that the power decreases with the level of discretization of the statistic. This is seen in their Figure 4, where the size of

the test-set,  $K$ , governs the discretization. Since they permute features, and not labels, then all their permutation samples are balanced, and there is no issue of refolding.

Golland et al. [2005] simulate the power of an accuracy test using a multivariate Gaussian mixture, with a parameter  $p$  governing the separation between classes. 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 perform the same simulation as Golland et al. [2005], after reparametrizing  $p$  so that  $p = 0$  corresponds to the null model, and  $p = 23$  to be comparable to our other simulations. We find that in this mixture class of models, like the location class of models, a location test has more power than an accuracy test (Figure 4).

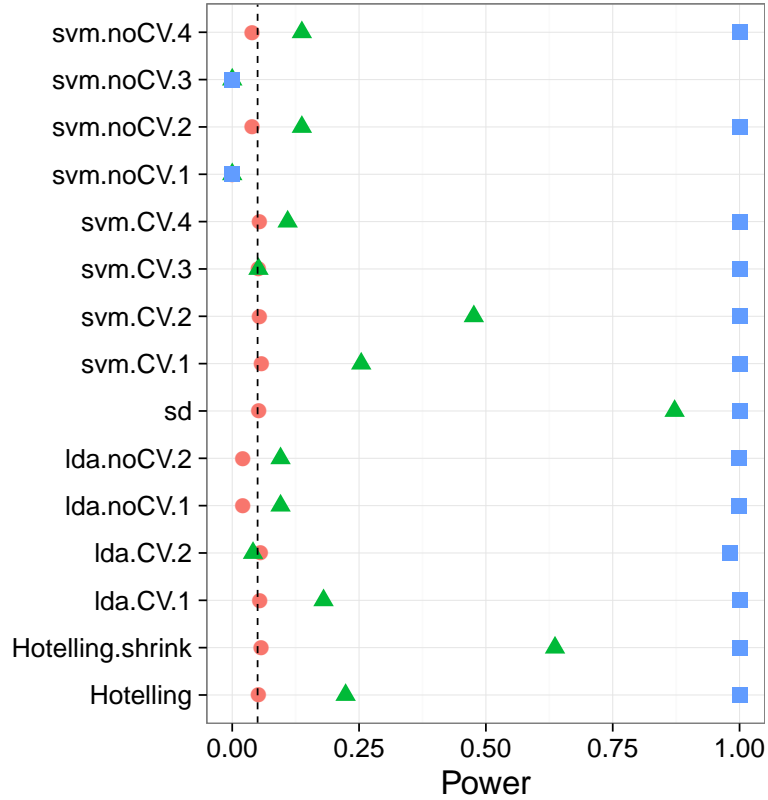


Figure 4: **Mixture:**  $\mathbf{x}_i = \chi_i \mu + \eta_i; \chi_i = \{-1, 1\}$  and  $\text{Prob}(\chi_i = 1) = (1/2 - p)^{y_i^*} (1/2 + p)^{1 - y_i^*}$ .  $\mu$  is a  $p$ -vector with  $3/\sqrt{p}$  in all coordinates. The effect,  $p$ , is color and shape coded and varies over 0 (red circle),  $1/4$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

## 6.6 Reservations

Some reservations to the generality of our findings are in order. Firstly, not all accuracy tests are concerned with signal detection. Indeed, it is possible that the purpose of the test is not to detect a difference between classes, but to actually test the performance of a particular classifier. Put differently—classification is harder than detection, so that we may be able to detect a difference between classes, but not be able to classify examples significantly better than chance. Examples of such problems include brain decoding for machine interfaces, and clinical diagnosis, where the presence of a medical condition is predicted from imaging data. [e.g. Olivetti et al., 2012, Wager et al., 2013]

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. This is a false argument—accuracy test do not have any more flexibility than location tests. Indeed, it is possible to test for location in the same dimension the classifier is learned. 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 without specifying any a-priori alternative [e.g. Heller et al., 2012]). On the other hand, based on our reported neuroimaging example, and others, we find that a location test in the original feature space is indeed a simple and powerful approach to signal detection.

## 6.7 Epilogue

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

The recent popularity of machine learning has resulted in the extensive teaching and use of prediction in theoretical and applied communities and the relative lack of awareness or popularity of the topic of Neyman-Pearson style hypothesis testing in the computer science and related “data science” communities.

And more simply by Frank Harrell in the CrossValidated Q&A site<sup>3</sup>:

---

<sup>3</sup><http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier>.

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



## 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.
- B. Efron and R. Tibshirani. Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association*, 92(438):548–560, June 1997. ISSN 0162-1459. doi: 10.2307/2965703.
- 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.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, July 2003. ISBN 0-387-95284-5.
- R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, page ass070, Dec. 2012. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/ass070.

- 423 J. Hemerik and J. Goeman. Exact testing with random permutations.  
424 *arXiv:1411.7565 [math, stat]*, Nov. 2014.
- 425 H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Math-*  
426 *ematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990.  
427 doi: 10.1214/aoms/1177732979.
- 428 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for  
429 prediction error in microarray classification using resampling. *Statistical*  
430 *Applications in Genetics and Molecular Biology*, 7(1), 2008.
- 431 L. Juan and H. Iba. Prediction of tumor outcome based on gene expression  
432 data. *Wuhan University Journal of Natural Sciences*, 9(2):177–182, Mar.  
433 2004. ISSN 1007-1202, 1993-4998. doi: 10.1007/BF02830598.
- 434 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional  
435 brain mapping. *Proceedings of the National Academy of Sciences of the*  
436 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,  
437 1091-6490. doi: 10.1073/pnas.0600244103.
- 438 E. L. Lehmann. Parametric versus nonparametrics: two alternative method-  
439 ologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN  
440 1048-5252. doi: 10.1080/10485250902842727.
- 441 G. J. McLachlan. The bias of the apparent error rate in discriminant analysis.  
442 *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi:  
443 10.1093/biomet/63.2.239.
- 444 S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell,  
445 T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements  
446 for classifying DNA microarray data. *Journal of Computational Biology:*  
447 *A Journal of Computational Molecular Cell Biology*, 10(2):119–142, 2003.  
448 ISSN 1066-5277. doi: 10.1089/10665270321825928.
- 449 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with  
450 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-  
451 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*  
452 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,  
453 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2  
454 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9\_6.
- 455 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the  
456 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.  
457 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.

- 458 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and  
459 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,  
460 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 461 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.  
462 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and  
463 P. Belin. The human voice areas: Spatial organization and inter-individual  
464 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–  
465 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 466 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for  
467 Class Prediction Using Gene Expression Profiles. *Journal of Computa-  
468 tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/  
469 106652702760138592.
- 470 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy  
471 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 472 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance  
473 Matrix Estimation and Implications for Functional Genomics. *Statistical  
474 Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN  
475 1544-6115. doi: 10.2202/1544-6115.1175.
- 476 D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class  
477 Prediction and Discovery Using Gene Expression Data. In *Proceedings of  
478 the Fourth Annual International Conference on Computational Molecular  
479 Biology*, RECOMB ’00, pages 263–272, New York, NY, USA, 2000. ACM.  
480 ISBN 978-1-58113-186-4. doi: 10.1145/332306.332564.
- 481 M. S. Srivastava. On testing the equality of mean vectors in high dimension.  
482 *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1):  
483 31–56, June 2013. ISSN 2228-4699. doi: 10.12697/ACUTM.2013.17.03.
- 484 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high  
485 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.  
486 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 487 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-  
488 ing correction in classification-based multi-voxel pattern analysis (MVPA):  
489 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.  
490 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.

- 491 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,  
492 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-  
493 2.
- 494 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,  
495 and B. Thirion. Assessing and tuning brain decoders: cross-validation,  
496 caveats, and guidelines. working paper or preprint, June 2016.
- 497 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.  
498 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-*  
499 *nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:  
500 10.1056/NEJMoa1204471.

## 501 A Analysis pipeline

502 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in  
 503 Gilron et al. [2016]. Denoting by  $i = 1, \dots, I$  the subject index,  $v = 1, \dots, V$   
 504 the voxel index, and  $s = 1, \dots, S$  the permutation index. Since regions<sup>4</sup> are  
 505 centered around a unique voxel, the voxel index  $v$  also serves as a unique  
 506 region index. Algorithm 1 computes a region-wise test statistic, which is  
 507 compared to its permutation null distribution computed by Algorithm 2.

**Algorithm 1:** Compute a group parametric map.

**Data:** fMRI scans, and experimental design.  
**Result:** Brain map of group statistics:  $\{\bar{T}_v\}_{v=1}^V$

```

1 for  $v \in 1, \dots, V$  do
2   for  $i \in 1, \dots, I$  do
3      $T_{i,v} \leftarrow$  test statistic for subject  $i$  in a region centered at  $v$ .
4    $\bar{T}_v \leftarrow \frac{1}{I} \sum_{i=1}^I T_{i,v}$ .
```

**Algorithm 2:** Compute a permutation p-value map.

**Data:** fMRI scans of 20 subjects, experimental design.  
**Result:** Brain map of permutation p-values:  $\{p_v\}_{v=1}^V$

```

1 for  $s \in 1, \dots, S$  do
2   permute labels;
3    $\bar{T}_v^s \leftarrow$  parametric map
```

---

<sup>4</sup>*searchlight* or *sphere* in the MVPA parlance

## 510 B Simulation Details

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

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

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

## C Simulation Results

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



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



(b) 20-fold cross validation.  
Balanced folding

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

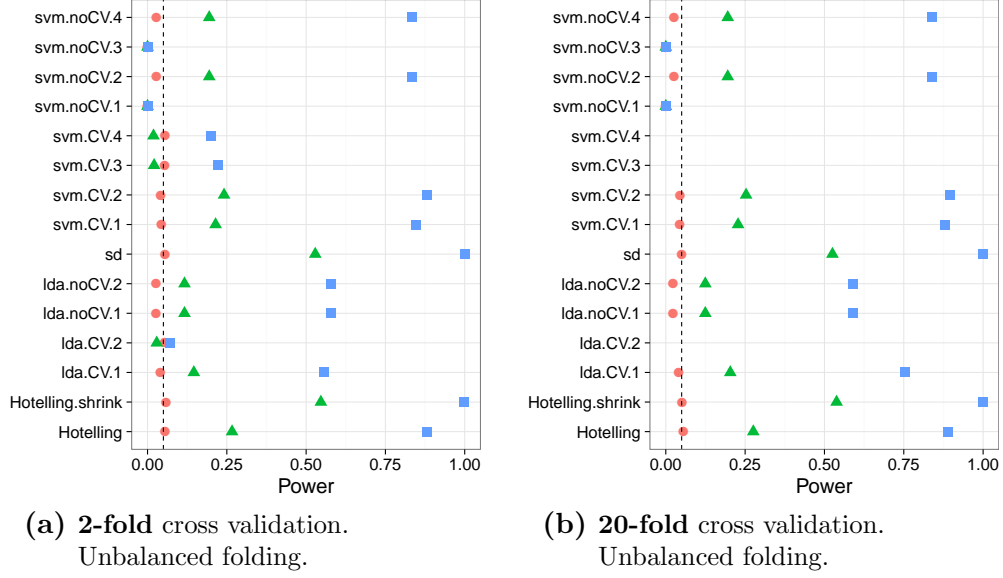
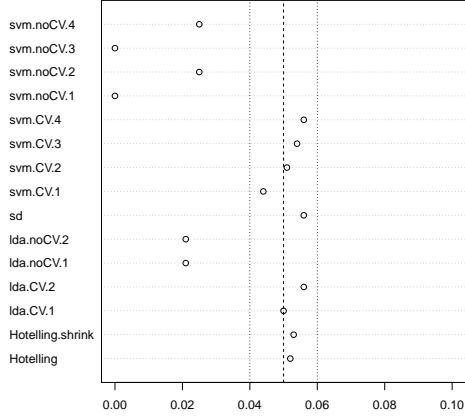


Figure 7: *Simulation details in Appendix B except the changes in the sub-captions.*

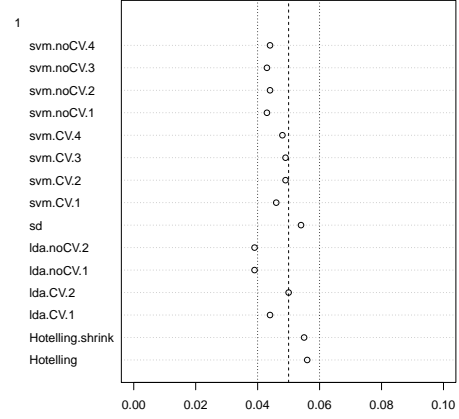




Figure 8: *Simulation details in Appendix B except the changes in the sub-captions.*



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

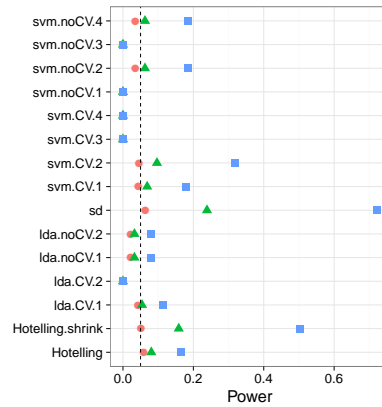


(b) **High-Dimension:** False positive rates for  $n = 400$ .

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



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



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