

Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt Roei Gilron Roy Mukamel

August 11, 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]:

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

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

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

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

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

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

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 in
 92 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 $\hat{f}(x)$ from some
 assumed model class $\hat{f} \in \mathcal{F}$. The prediction accuracy, denoted $\mathcal{E}_{\hat{f}}$, is de-

defined as the probability of a given classifier \hat{f} of making a correct prediction. Denoting by $I(A)$ the indicator function of the event A , we get

$$\mathcal{E}_{\hat{f}} := \mathbf{E} \left[I(\hat{f}(x) = y) \right] \quad (1)$$

when given a randomly drawn data point, (x, y) . A statistically significant “better than chance” estimate of $\mathcal{E}_{\hat{f}}$ 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. Is the statistic cross validated or not?
2. For a V-fold cross validated test statistic:
 - (a) Should the data be refolded in each permutation?
 - (b) Should the data folding be balanced (a.k.a. stratified)?
 - (c) How many folds?
3. How to estimate accuracy?

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.

Cross validate or not? Since we are merely interested in detecting a difference between classes, a biased error estimate is not an issue provided that bias is consistent over all permutations. The underlying intuition is that if the exact same computation is performed over all permutations, then a permutation test will be “fair”, i.e., will not inflate the false positive rate. We will thus be considering both cross validated accuracies, and resubstitution accuracies as our test statistics.

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

128 **Refolding?** The standard practice in neuroimaging is to refold the data
 129 after each permutation, so that data folds are balanced after each label per-
 130 mutation. We will adhere, even though it can be circumvented by permuting
 131 features instead of labels, as done by Golland et al. [2005].

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

How to estimate accuracy? Given a predictor \hat{f} , a natural accuracy test
 statistic is its accuracy $\mathcal{E}_{\hat{f}}$. Since low accuracies, even 0, are evidence that the
 classes are separated, can consider the departure from chance level, $|\mathcal{E}_{\hat{f}} - 0.5|$,
 as the test statistic. For unbalanced classes, chance level is not 0.5, but rather
 the probability of the majority class, we denote by \hat{p}_{max} . This suggests
 the following test statistic $|\mathcal{E}_{\hat{f}} - \hat{p}_{max}|$. Since we will be aggregating these
 statistics over random data sets where \hat{p}_{max} may vary, it seems appropriate to
 standardize the scale of this statistic. We thus propose the z-scored accuracy
 statistic:

$$|\mathcal{E}_{\hat{f}} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}. \quad (2)$$

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

Name	Basis	CV	Accuracy	Parameters
Hotelling	Hotelling	—	—	—
Hotelling.shrink	Hotelling	—	—	—
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

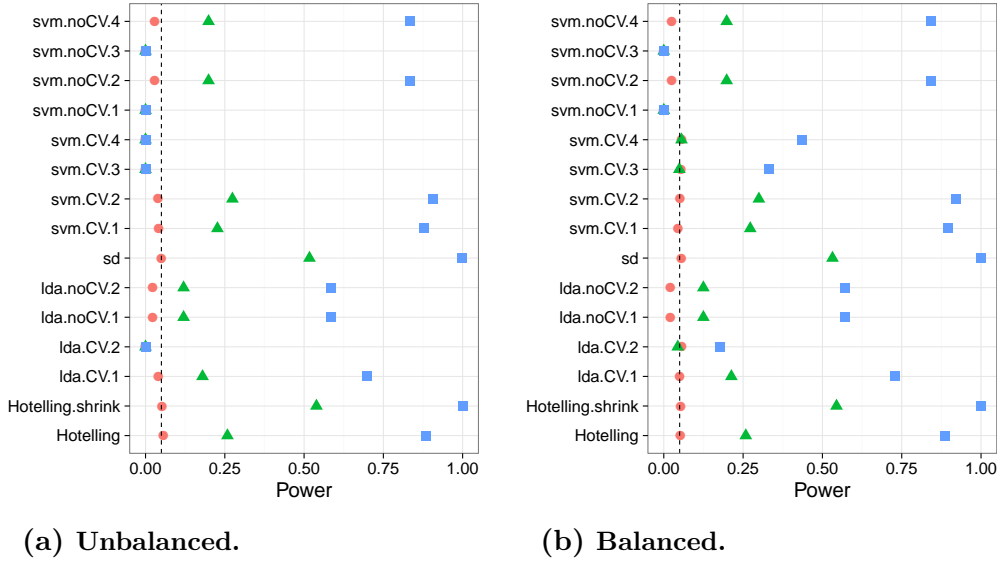
Table 1: This table collects the various test statistics we will be studying. Three are population tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group T^2 statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer and Strimmer [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, with *libsvm*’s cost parameter set at 0.1, using the cross validated z-scored accuracy ($|\mathcal{E}_{\hat{f}} - \hat{p}_{max}|/\sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$, 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.

136

137 3 Controlling the False Positive Rate

138 Figure 1 demonstrates that all of the tests considered conserve the desired
139 0.05 false positive rate, up to varying levels of conservatism. This can be
140 seen from the fact that the probability of rejection is no larger than 0.05 in
141 the absence of any effect, encoded by a red circle. This is true, in particular
142 if: (a) the folds are balanced or not, (b) the tuning parameters of some test
143 statistic are varied, (d) the number of folds is varied. We also observe that
144 the most conservative tests are the resubstitution accuracy measures. We
145 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.



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 the power of 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 8a, 8b and 9a), suggesting that concentration and/or discretization is responsible for power loss.
3. The power may increase or decrease with the number of folds (Figure 5).
4. The z-scoring of the accuracies was introduced to deal with unbalanced foldings. If the z-scoring has any effect at all, it merely kills power. There is really no reason to use it.

- 160 5. Both accuracy and population tests are inappropriate for scale alter-
161 natives (Figure 7a). This was to be expected and is reported mostly as
162 a sanity check.
- 163 6. The presence of heavy tails (Figure 7b) may reduce power, but does
164 not quantitatively change results.
- 165 7. Balanced folding typically has no effect. It increased power only for
166 the z-scored statistics (Figure 1). This is surprising given they were
167 precisely designed to deal with the presence of imbalance.
- 168 8. Varying the accuracy test’s tuning parameter, such as the cost (i.e.
169 margins) has no effect on the power of the test.
- 170 9. Correlation between coordinates, mimicking temporal correlation in
171 fMRI data, has no effect on conclusions, since all test statistics account
172 for this correlation (Figure 9b).

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

176 5 Neuroimaging Example

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

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

189 In agreement with our simulation results, the population test (*sd*) discov-
190 ers more brain regions when compared to an accuracy test (*svm.cv.1*). The
191 former discovers 1,232 regions, while the latter only 441, as depicted in Fig-
192 ure 2. We emphasize that both test statistics were compared with the same

²<https://openfmri.org/>

193 permutation scheme, and the same error controls, so that any difference in
 194 detections is due to their different power.

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

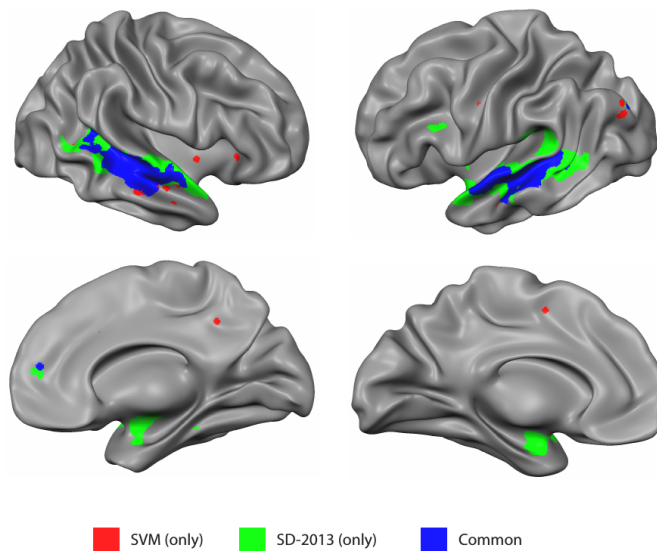


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (svm.cv.1), and a population test (sd). svm.cv.1 was computed using 5-fold cross validation, and a cost parameter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

198 6 Discussion

199 We have set out to understand which of the tests is more powerful: the
 200 accuracy test or the population test. Using simulations, we have concluded
 201 that the population tests are preferable. Their high dimensional versions
 202 such as Srivastava [2007] and Schäfer and Strimmer [2005] are preferable for
 203 typical neuroimaging problems such as MVPA. We attribute this to several
 204 phenomena: (a) Discretization introduced in finite samples by the accuracy

205 test statistic. (b) Inefficient use of the data for the validation holdout set.
 206 The presence of heavy tails shrinks the power advantage of the population
 207 tests over accuracy tests.

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

217 **6.1 Ease of implementation**

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

226 **6.2 A good accuracy test**

227 In Section 6.6 we discuss cases where an accuracy test cannot replace a pop-
 228 ulation test. For such cases we collect some conclusions from our simulations
 229 on the best practices for accuracy tests.

- 230 1. The conservativeness of accuracy tests decrease with sample size.
- 231 2. Permuting features is easier than permuting labels. It allows to preserve
 232 balanced folds after a permutation without refolding, thus reducing
 233 computational complexity.
- 234 3. For V-fold CV, it is unclear what is the effect of the number of folds.
 235 More folds increase power by reducing the number of holdout samples.
 236 On the other hand, it increases the concentration of the accuracy statis-
 237 tic. Compounded with the discreteness of the accuracy statistic, this
 238 decreases power. This suggests that the optimal number of folds may
 239 be problem specific.

- 240 4. Cross validating has no less power than resubstitution. The power loss
 241 due to the training sub-samples when cross validating, is smaller than
 242 the power loss due to the concentration of the resubstitution statistic
 243 (Figure 8). For large sample sizes, discretization and concentration
 244 have weaker effects, so that the cross validated accuracy may be re-
 245 placed with the computationally more efficiency resubstitution accu-
 246 racy (Figure 9a). This also implies that there is a fundamental differ-
 247 ence between V-folding and resubstitution, so that latter should not be
 248 thought of as the limit of the former.
- 249 5. There is no gain in z-scoring the accuracy scores. Our motivating
 250 rational was clearly flawed. [TODO: why?]
- 251 6. Cross validated accuracy with balanced folds has more power than
 252 unbalanced folds. [TODO: Why?].
- 253 7. The value of the tuning parameters of a classifier have little to no
 254 effect.

255 6.3 Smoothing accuracy estimates

256 It may be possible to alleviate the effect of discretization by appropriate cross-
 257 validation. The discreteness of the accuracy statistic can be “smoothed” by
 258 allowing the test sample to be drawn with replacement. The *bootstrap* may
 259 seem like a candidate approach, but since the original data always serves as
 260 a test set, the accuracy can still only assume $1/n$ values. This is not the case,
 261 however, for the *leave-one-out bootstrap estimator* (B-LOO) and the *0.632*
 262 *bootstrap estimator* (B-0.632) [Hastie et al., 2003, Sec 7.11], which we define
 263 below for completeness. By the same rational, the degree of conservatism
 264 should decrease with the number of bootstrap samples.

Definition 1 (B-LOO). Denoting by $C^{(i)}$ the index set of bootstrap samples,
 b , where observation i is not in the train set, *leave-one-out bootstrap* estimate
 is defined as:

$$\mathcal{E}_{BLOO} := \frac{1}{n} \sum_{i=1}^n \frac{1}{|C^{(i)}|} \sum_{b \in C^{(i)}} I(\hat{f}^b(x_i) = y_i).$$

Equivalently, denoting by $S^{(b)}$ the indexes of observations, i , that are not in
 the bootstrap train sample b ,

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

Definition 2 (B-0.632). Denoting by \mathcal{E}_{resub} the resubstitution accuracy estimate, the B-0.632 accuracy estimator, $\mathcal{E}_{0.632}$, is defined as

$$\mathcal{E}_{0.632} := 0.368 \mathcal{E}_{resub} + 0.632 \mathcal{E}_{BLOO}.$$

265 The simulation results reported in Figure 3, with naming conventions in
 266 Table 2. It can be seen that selecting test sets with replacement does increase
 267 the power, when compared to V-fold cross validation, but still falls short from
 268 the power of population tests. It can also be seen that power increases with
 269 the number of Bootstrap replications, itself reducing the level of discretiza-
 270 tion. The type of Bootstrap, B-LOO versus B-0.632, does not change the
 271 power. Again, consistent with the observation that it is discretization that
 272 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=1e1
svm.Boot.3	SVM	B-0.632	50	accuracy	cost=1e1
svm.Boot.4	SVM	B-LOO	50	accuracy	cost=1e1

Table 2: The same as Table 1 for bootstrapped accuracy estimates. B-LOO and B-0.632 are defined in definitions 1 and 2 respectively. B denotes the number of Bootstrap samples.

273

274 6.4 High dimensional classifiers

275 It is known that when $p > n$ Hotelling’s T^2 , and Fisher’s LDA are not
 276 computable. In our simulations, in which $p = 23$ and $n = 40$ is “almost”
 277 high dimensional, but still allows to compute both tests. We have simulated
 278 two high dimensional versions of Hotelling’s T^2 : *sd* [Srivastava, 2007] and
 279 *Hotelling.shrink* [Schäfer and Strimmer, 2005]. The former solves the dimen-
 280 sionality problem by assuming independence over coordinates, and the latter
 281 by Tikhonov regularization of the covariance, a-la ridge regression. The cor-
 282 responding high dimensional accuracy tests would be a *naive Bayes* classifier,
 283 and l_2 regularized SVM [Ramdas et al., 2016]. We conjecture that they would
 284 not alter our conclusions, since the main force driving the conservatism is
 285 discretization, which they do not solve.

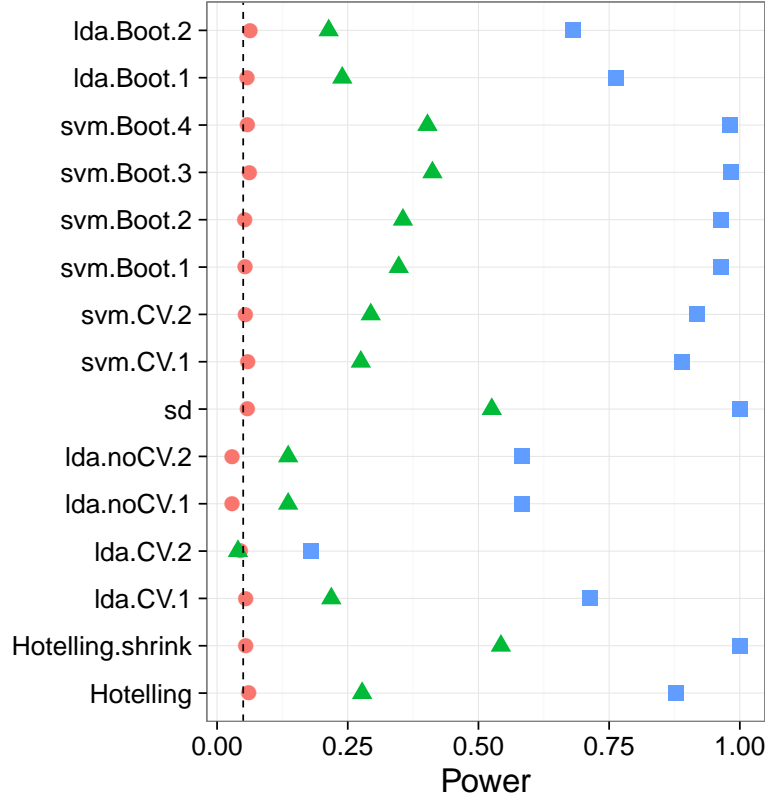


Figure 3: **Bootstrap:** The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square).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

286 6.5 Related Literature

287 Olivetti et al. [2012] and Olivetti et al. [2014] looked into the problem of
 288 choosing a good accuracy test. They propose a new test they call an *inde-*
 289 *pendence test*, and demonstrate by simulation that it has more power than
 290 other accuracy tests, and can deal with non-balanced data sets. We did not
 291 include this test in the battery we compared, but we note the following: (a)
 292 The independence test of Olivetti et al. [2012] relies on a discrete test statis-
 293 tic. This means that in the cases that the accuracy test is called upon for
 294 discriminating populations, it will probably be underpowered compared to
 295 population tests. (b) In contrast with the underlying motivation of Olivetti
 296 et al. [2012]’s independence test, we did not find that balancing the data

297 folds is crucial for an accuracy test.

298 Golland et al. [2005] study accuracy tests using simulation, neuroimaging
 299 data, genetic data, and analytically. Their analytic results formalize our in-
 300 tuition from Section 1 on the effect of concentration of the accuracy statistic:
 301 The finite Vapnik–Chervonenkis (VC) dimension requirement [Golland and
 302 Fischl, 2003, Sec 4.3] prevents the permutation p-value from (asymptotically)
 303 concentrating. They also find that the power decreases with the level of dis-
 304 cretization of the statistic. This is seen in their Figure 4, where the size of
 305 the test-set, K , governs the discretization. Since they permute features, and
 306 not labels, then all their permutation samples are balanced, and there is no
 307 issue of refolding.

308 Golland et al. [2005] simulate the power of an accuracy test using a mul-
 309 tivariate Gaussian mixture, with a parameter p governing the separation be-
 310 tween classes. Under their model $(x_i|y_i = 1) \sim p\mathcal{N}(\mu_1, I) + (1 - p)\mathcal{N}(\mu_2, I)$
 311 and $(x_i|y_i = -1) \sim (1 - p)\mathcal{N}(\mu_1, I) + p\mathcal{N}(\mu_2, I)$. Varying p interpolates be-
 312 tween the null distribution ($p = 0.5$) and a location shift model ($p = 0$). We
 313 perform the same simulation as Golland et al. [2005], after reparametrizing p
 314 so that $p = 0$ corresponds to the null model, and $p = 23$ to be comparable to
 315 our other simulations. We find that in this mixture class of models, like the
 316 location class of models, a population test has more power than an accuracy
 317 test (Figure 4).

318 6.6 Reservations

319 Some reservations to the generality of our findings are in order. Firstly,
 320 not all accuracy tests are concerned with signal detection. Consider brain
 321 decoding for machine interfaces, and clinical diagnosis, where the presence
 322 of a medical condition is predicted from imaging data [e.g. Olivetti et al.,
 323 2012, Wager et al., 2013]. In those examples, the purpose of the test is not
 324 to detect a difference between classes, but to actually test the performance
 325 of a particular classifier. As put by Ojala and Garriga [2010]:

326 ...these tests study whether the classifier is using the described
 327 properties and not whether the plain data contain such properties.
 328 For studying the characteristics of a population represented by
 329 the data, standard statistical test could be used.

330 This is because classification is harder than detection. We may be able
 331 to detect a difference between classes, but not be able to classify examples
 332 significantly better than chance.

333 Secondly, it may be argued that accuracy tests permits the separation
 334 between classes in high dimensions, such as in *reproducing kernel Hilbert*

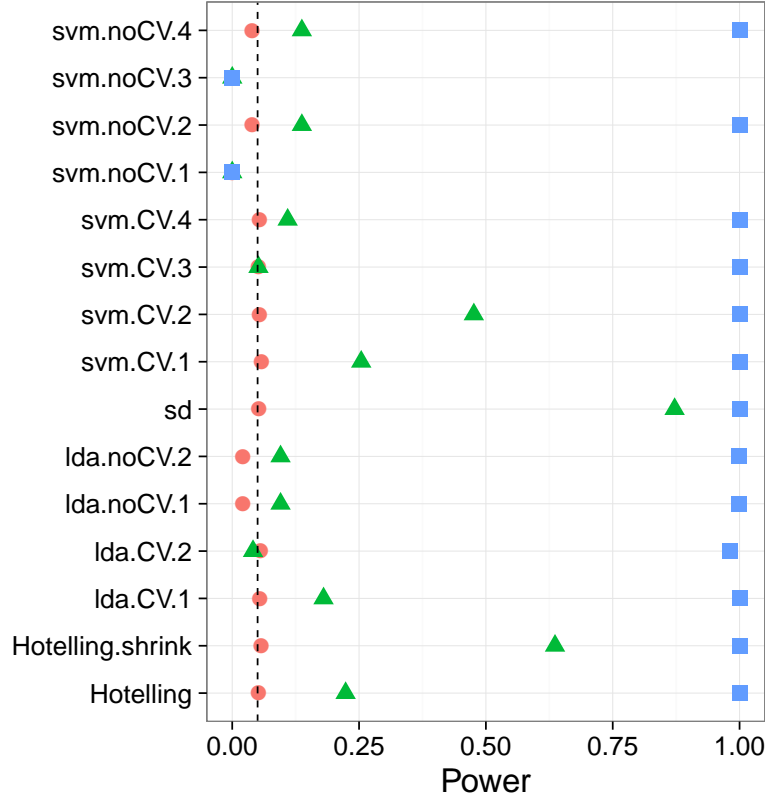


Figure 4: **Mixture:** $\mathbf{x}_i = \chi_i \mu + \eta_i$; $\chi_i \in \{-1, 1\}$ and $\text{Prob}(\chi_i = 1) = (1/2 - p)^{y_i^*} (1/2 + p)^{1-y_i^*}$. μ is a p -vector with $3/\sqrt{p}$ in all coordinates. The effect, p , is color and shape coded and varies over 0 (red circle), 1/4 (green triangle)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

spaces (RKHS) by using non-linear predictors. This is a false argument—
accuracy test do not have any more flexibility than population tests. Indeed,
it is possible to test for location in the same 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. ?].
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.

344 6.7 Epilogue

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

351 The recent popularity of machine learning has resulted in the ex-
352 tensive teaching and use of prediction in theoretical and applied
353 communities and the relative lack of awareness or popularity of
354 the topic of Neyman-Pearson style hypothesis testing in the com-
355 puter science and related “data science” communities.

356 And more simply by Frank Harrell in the `CrossValidated` Q&A site³:

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

360 7 Acknowledgments

³[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).

References

- T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, Hoboken, NJ, 3 edition edition, July 2003. ISBN 978-0-471-36091-9.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES B*, 57:289–289, 1995.
- R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. Quantifying spatial pattern similarity in multivariate analysis using functional anisotropy. *arXiv:1605.03482 [q-bio]*, May 2016.
- P. Golland and B. Fischl. Permutation tests for classification: towards statistical significance in image-based studies. In *IPMI*, volume 3, pages 330–341. Springer, 2003.
- 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.
- J. Hemerik and J. Goeman. Exact testing with random permutations. *arXiv:1411.7565 [math, stat]*, Nov. 2014.
- H. Hotelling. The Generalization of Student’s Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177732979.

- 395 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for
396 prediction error in microarray classification using resampling. *Statistical*
397 *Applications in Genetics and Molecular Biology*, 7(1), 2008.
- 398 L. Juan and H. Iba. Prediction of tumor outcome based on gene expression
399 data. *Wuhan University Journal of Natural Sciences*, 9(2):177–182, Mar.
400 2004. ISSN 1007-1202, 1993-4998. doi: 10.1007/BF02830598.
- 401 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional
402 brain mapping. *Proceedings of the National Academy of Sciences of the*
403 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,
404 1091-6490. doi: 10.1073/pnas.0600244103.
- 405 E. L. Lehmann. Parametric versus nonparametrics: two alternative method-
406 ologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN
407 1048-5252. doi: 10.1080/10485250902842727.
- 408 G. J. McLachlan. The bias of the apparent error rate in discriminant analysis.
409 *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi:
410 10.1093/biomet/63.2.239.
- 411 S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell,
412 T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements
413 for classifying DNA microarray data. *Journal of Computational Biology:*
414 *A Journal of Computational Molecular Cell Biology*, 10(2):119–142, 2003.
415 ISSN 1066-5277. doi: 10.1089/106652703321825928.
- 416 M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Perfor-
417 mance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.
418 ISSN 1533-7928.
- 419 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
420 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
421 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*
422 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
423 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
424 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.
- 425 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the
426 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.
427 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.

- 428 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
429 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
430 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 431 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
432 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
433 P. Belin. The human voice areas: Spatial organization and inter-individual
434 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
435 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 436 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
437 Class Prediction Using Gene Expression Profiles. *Journal of Computa-*
438 *tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
439 106652702760138592.
- 440 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
441 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 442 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance
443 Matrix Estimation and Implications for Functional Genomics. *Statistical*
444 *Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN
445 1544-6115. doi: 10.2202/1544-6115.1175.
- 446 D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class
447 Prediction and Discovery Using Gene Expression Data. In *Proceedings of*
448 *the Fourth Annual International Conference on Computational Molecular*
449 *Biology*, RECOMB ’00, pages 263–272, New York, NY, USA, 2000. ACM.
450 ISBN 978-1-58113-186-4. doi: 10.1145/332306.332564.
- 451 M. S. Srivastava. Multivariate Theory for Analyzing High Dimensional Data.
452 *Journal of the Japan Statistical Society*, 37(1):53–86, 2007. doi: 10.14490/
453 jjss.37.53.
- 454 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
455 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
456 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 457 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
458 ing correction in classification-based multi-voxel pattern analysis (MVPA):
459 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
460 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.

- 461 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,
462 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-
463 2.
- 464 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
465 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
466 caveats, and guidelines. working paper or preprint, June 2016.
- 467 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.
468 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-*
469 *nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:
470 10.1056/NEJMoa1204471.

471 A Analysis pipeline

472 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 473 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 474 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions⁴ are
 475 centered around a unique voxel, the voxel index v also serves as a unique
 476 region index. Algorithm 1 computes a region-wise test statistic, which is
 477 compared to its permutation null distribution computed by Algorithm 2.

Algorithm 1: Compute a group parametric map.

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

```

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

Algorithm 2: Compute a permutation p-value map.

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

```

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

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

480 B Simulation Details

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

484 Each simulation is based on 4,000 replications. In each replication, we
485 generate n i.i.d. samples from a shift model $\mathbf{x}_i = \mu \mathbf{y}_i^* + \eta_i$. Where $y_i^* = \{0, 1\}$
486 is the class of subject i in dummy coding. Recalling that $y_i = \{-1, 1\}$ is the
487 class in effect coding, then clearly $y_i = 2y_i^* - 1$. The noise is distributed as
488 $\eta_i \sim \mathcal{N}_p(0, \Sigma)$. The sample size $n = 40$. The dimension of the data is $p = 23$.
489 The covariance $\Sigma = I$. Effects, i.e. shifts μ , are equal coordinate p -vectors
490 with coordinates that vary over $\mu \in \{0, 1/4, 1/2\}$.

491 Having generated the data, we compute each of the test statistics in Ta-
492 ble 1. For test statistics that require data folding, we used 8 folds. We then
493 compute a permutation p-value by permuting the class labels, and recomput-
494 ing each test statistic. We perform 400 such permutations. We then reject
495 the $\mu_i = 0$ null hypothesis if the permutation p-value is smaller than 0.05.
496 The reported power is the proportion of replication where the permutation
497 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.

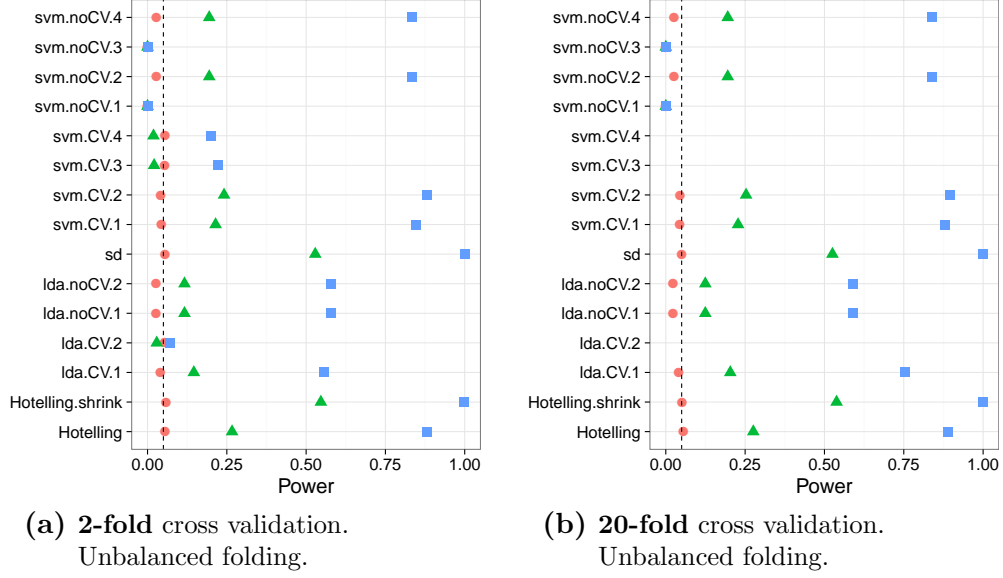
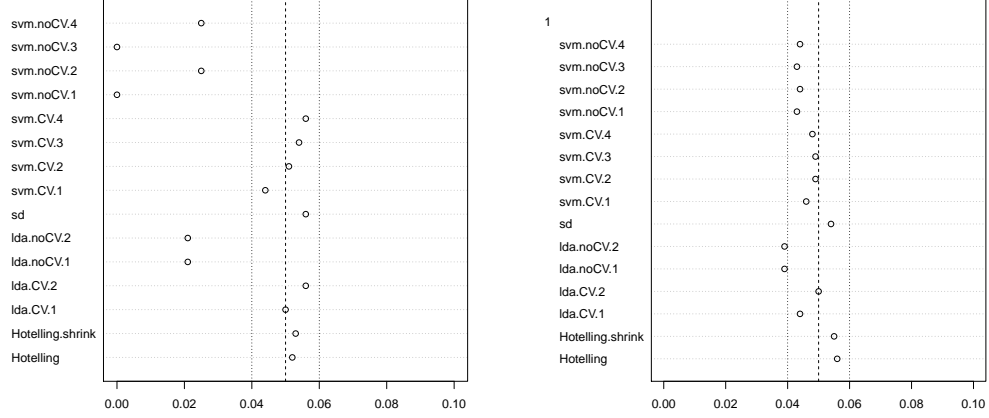


Figure 7: Simulation details in Appendix B except the changes in the sub-captions.



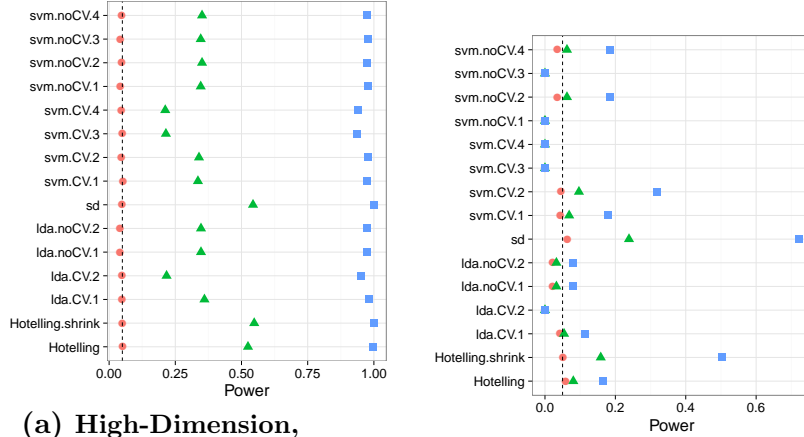
Figure 8: Simulation details in Appendix B except the changes in the sub-captions.



(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:**
 $n = 400$,
 $\mu \in \frac{1}{\sqrt{10}} \times \{0, 1/4, 1/2\}$.

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