

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]:

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 large com-
 32 pared to the number of observations, so that the spatial covariance cannot
 33 be fully estimated, then a high dimensional version of Hotelling’s test can be
 34 called upon, such as in Schäfer and Strimmer [2005] or Srivastava [2007]. For
 35 brevity, and in contrast to *accuracy tests*, we will call any two-sample mul-
 36 tivariate tests simply *location tests*, also termed *class comparisons*. [TODO:
 37 rename to parameter test?]

38 At this point, it becomes unclear which is preferable: a location test or an
 39 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 location test to the accu-
 44 racy of *Fisher’s linear discriminant analysis* classifier (LDA). By comparing
 45 the rates of convergence of the powers to 1, Ramdas et al. [2016] concluded
 46 that accuracy and location 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 test
 56 is remarkably inefficient compared to the location test. For comparison, the
 57 t-test is only 1.04 more (asymptotically) efficient than Wilcoxon’s rank-sum
 58 test [Lehmann, 2009], so that an ARE of 6.3 is strong evidence in favor of
 59 the location 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: *location 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 location 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 location 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 accu-
 101 racy test amounts to learning a predictive model $\hat{f}(x)$ from some assumed
 102 model class $\hat{f} \in \mathcal{F}$. The prediction accuracy, denoted $\mathcal{E}_{\hat{f}}$, is 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 have $\mathcal{E}_{\hat{f}} := \mathbf{E} \left[I(\hat{f}(x) = y) \right]$ when given a randomly drawn data point, (x, y) . A statistically significant “better than chance” estimate of $\mathcal{E}_{\hat{f}}$ 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. 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 $\mathcal{E}_{\hat{f}}$. 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 $|\mathcal{E}_{\hat{f}} - 0.5|$ 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 accuracy chance level is the probability of the majority class, we denote by \hat{p}_{max} [Golland et al., 2005, Sec 4.1]. 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 also consider the z-scored accuracy: $|\mathcal{E}_{\hat{f}} - \hat{p}_{max}| / \sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$.

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

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

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

159 **Balanced folding?** As already implied, a standard practice when cross
160 validating is to constrain the data folds to be balanced (i.e. stratified) [e.g.
161 Ojala and Garriga, 2010]. This is well justified when aiming at unbiased accu-
162 racy estimation. This also simplifies matter when aiming at signal detection,
163 as can be seen from the above discussion of the appropriate test statistic. On
164 the other hand, it may complicate matters, as can be seen from the above
165 discussion on label versus feature permutation. We will report results with
166 both balanced and unbalanced data foldings, only to discover, it does not
167 really matter.

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

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

176

3 Controlling the False Positive Rate

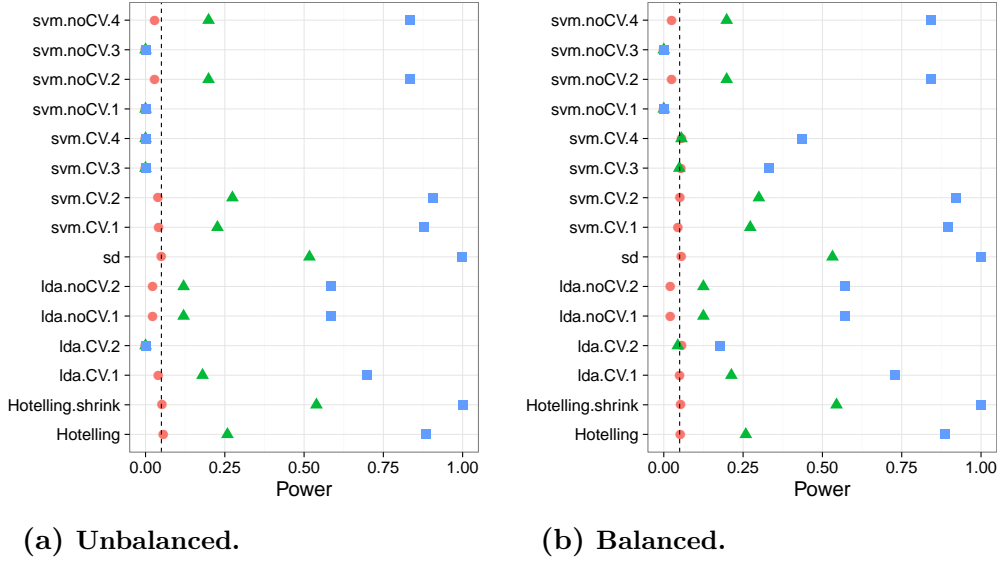
177

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

180

the absence of any effect, encoded by a red circle. This is true, in particular if: (a) the folds are balanced or not, (b) the tuning parameters of some test statistic are varied, (d) the number of folds is varied. We also observe that the most conservative tests are the resubstitution accuracy measures. We return to this matter in the Discussion.

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



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 location tests to accuracy tests. From the simulation results reported in Appendix C we collect the following insights:

1. Location tests have more power than accuracy tests in all our configurations.

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

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

216 5 Neuroimaging Example

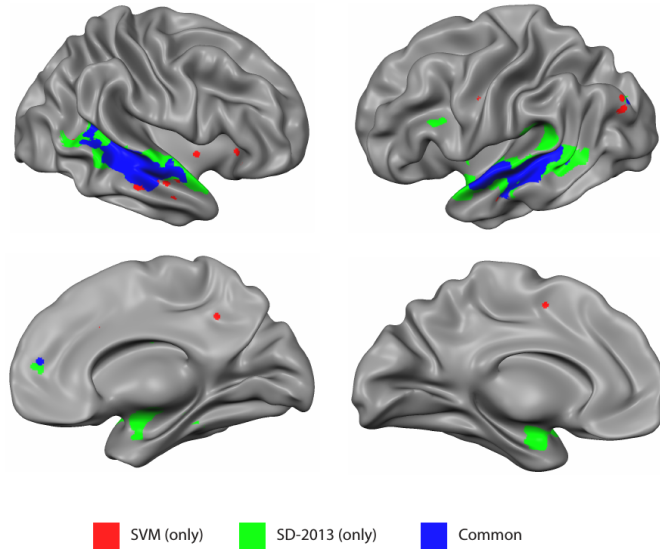
217 Figure 2 is an application of both a location and an accuracy test to the data
218 of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI
219 data while subjects were exposed to the sounds of human speech (vocal),
220 and other non-vocal sounds. Each subject was exposed to 20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

²<https://openfmri.org/>

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

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

235 Having established that accuracy tests are underpowered both in simula-
 236 tion and in application, we wish to identify the conditions under which this
 237 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].*

238 6 Discussion

239 We have set out to understand which of the tests is more powerful: the ac-
240 curacy test or the location test. Using simulations, we have concluded that
241 the location tests are preferable. Their high dimensional versions such as
242 Srivastava [2007] and Schäfer and Strimmer [2005] are preferable for typical
243 neuroimaging problems such as MVPA. We attribute this to several phe-
244 nomena: (a) Discretization introduced in finite samples by the accuracy test
245 statistic. (b) Inefficient use of the data for the validation holdout set. The
246 presence of heavy tails shrinks the power advantage of the location tests over
247 accuracy tests.

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

257 6.1 Ease of implementation

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

266 6.2 A good accuracy test

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

- 270 1. The conservativeness of accuracy tests decrease with sample size.
- 271 2. Permuting features is easier than permuting labels. It allows to preserve
272 balanced folds after a permutation without refolding, thus reducing

- 273 computational complexity.
- 274 3. For V-fold CV, it is unclear what is the effect of the number of folds.
 275 More folds increase power by reducing the number of holdout samples.
 276 On the other hand, it increases the concentration of the accuracy statis-
 277 tic. Compounded with the discreteness of the accuracy statistic, this
 278 decreases power. This suggests that the optimal number of folds may
 279 be problem specific.
- 280 4. Cross validating has no less power than resubstitution. The power loss
 281 due to the training sub-samples when cross validating, is smaller than
 282 the power loss due to the concentration of the resubstitution statistic
 283 (Figure 8). For large sample sizes, discretization and concentration
 284 have weaker effects, so that the cross validated accuracy may be re-
 285 placed with the computationally more efficiency resubstitution accu-
 286 racy (Figure 9a). This also implies that there is a fundamental differ-
 287 ence between V-folding and resubstitution, so that latter should not be
 288 thought of as the limit of the former.
- 289 5. There is no gain in z-scoring the accuracy scores. Our motivating
 290 rational was clearly flawed. [TODO: why?]
- 291 6. Cross validated accuracy with balanced folds has more power than
 292 unbalanced folds. [TODO: Why?].
- 293 7. The value of the tuning parameters of a classifier have little to no
 294 effect.

295 6.3 Smoothing accuracy estimates

296 It may be possible to alleviate the effect of discretization by appropriate cross-
 297 validation. The discreteness of the accuracy statistic can be “smoothed” by
 298 allowing the test sample to be drawn with replacement. The *bootstrap* may
 299 seem like a candidate approach, but since the original data always serves as
 300 a test set, the accuracy can still only assume $1/n$ values. This is not the case,
 301 however, for the *leave-one-out bootstrap estimator* (B-LOO) and the *0.632*
 302 *bootstrap estimator* (B-0.632) [Hastie et al., 2003, Sec 7.11], which we define
 303 below for completeness. By the same rational, the degree of conservatism
 304 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

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

305 The simulation results reported in Figure 3, with naming conventions in
 306 Table 2. It can be seen that selecting test sets with replacement does increase
 307 the power, when compared to V-fold cross validation, but still falls short from
 308 the power of location tests. It can also be seen that power increases with the
 309 number of Bootstrap replications, itself reducing the level of discretization.
 310 The type of Bootstrap, B-LOO versus B-0.632, does not change the power.
 311 Again, consistent with the observation that it is discretization that drives
 312 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.

313

314 6.4 High dimensional classifiers

315 It is known that when $p > n$ Hotelling's T^2 , and Fisher's LDA are not
 316 computable. In our simulations, in which $p = 23$ and $n = 40$ is “almost”

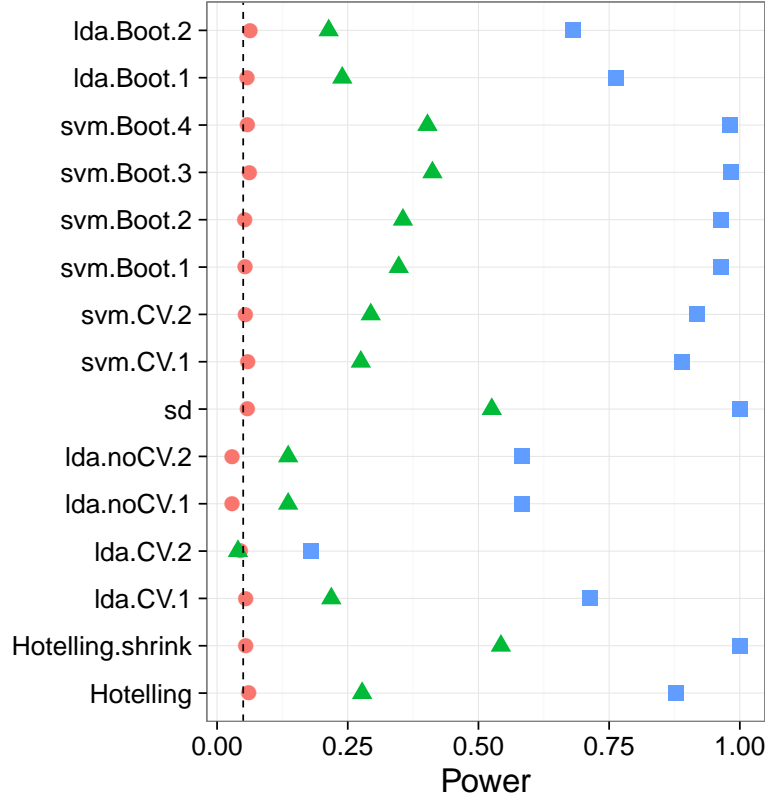


Figure 3: Bootstrap: The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square).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

high dimensional, but still allows to compute both tests. We have simulated two high dimensional versions of Hotelling's T^2 : *sd* [Srivastava, 2007] and *Hotelling.shrink* [Schäfer and Strimmer, 2005]. The former solves the dimensionality problem by assuming independence over coordinates, and the latter by Tikhonov regularization of the covariance, a-la ridge regression. The corresponding high dimensional accuracy tests would be a *naive Bayes* classifier, and l_2 regularized SVM [Ramdas et al., 2016]. We conjecture that they would not alter our conclusions, since the main force driving the conservatism is discretization, which they do not solve.

6.5 Related Literature

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. 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).

6.6 Reservations

Some reservations to the generality of our findings are in order. Firstly, not all accuracy tests are concerned with signal detection. Consider 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]. In those examples, the purpose of the test is not

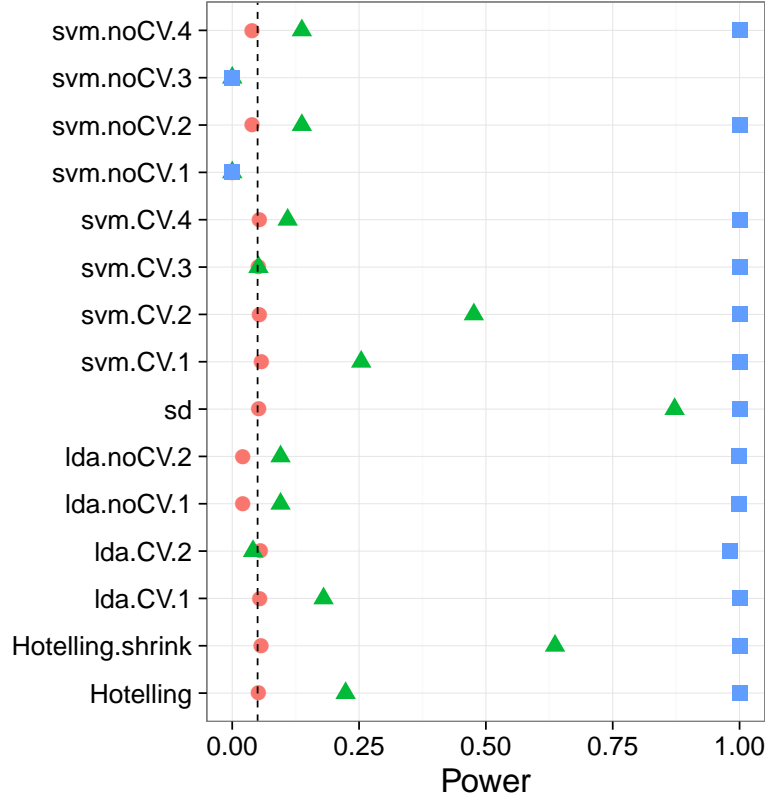


Figure 4: **Mixture:** $\mathbf{x}_i = \chi_i \mu + \eta_i$; $\chi_i = \{-1, 1\}$ and $\text{Prob}(\chi_i = 1) = (1/2 - p)^{y_i^*} (1/2 + p)^{1-y_i^*}$. μ is a p -vector with $3/\sqrt{p}$ in all coordinates. The effect, p , is color and shape coded and varies over 0 (red circle), $1/4$ (green triangle)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

364 to detect a difference between classes, but to actually test the performance
 365 of a particular classifier. As put by Ojala and Garriga [2010]:

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

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

373 Secondly, it may be argued that accuracy tests permits the separation
 374 between classes in high dimensions, such as in *reproducing kernel Hilbert*
 375 *spaces* (RKHS) by using non-linear predictors. This is a false argument–

376 accuracy test do not have any more flexibility than location tests. Indeed, it
377 is possible to test for location in the same dimension the classifier is learned.
378 Gretton et al. [2012] is an example where the test for location is performed
379 in the RKHS of the data. It is also possible to test for the equality of two
380 multivariate distributions without specifying any a-priori alternative [e.g. ?]).
381 On the other hand, based on our reported neuroimaging example, and others,
382 we find that a location test in the original feature space is indeed a simple
383 and powerful approach to signal detection.

384 6.7 Epilogue

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

391 The recent popularity of machine learning has resulted in the ex-
392 tensive teaching and use of prediction in theoretical and applied
393 communities and the relative lack of awareness or popularity of
394 the topic of Neyman-Pearson style hypothesis testing in the com-
395 puter science and related “data science” communities.

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

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

400 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.

- 435 W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for
436 prediction error in microarray classification using resampling. *Statistical*
437 *Applications in Genetics and Molecular Biology*, 7(1), 2008.
- 438 L. Juan and H. Iba. Prediction of tumor outcome based on gene expression
439 data. *Wuhan University Journal of Natural Sciences*, 9(2):177–182, Mar.
440 2004. ISSN 1007-1202, 1993-4998. doi: 10.1007/BF02830598.
- 441 N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional
442 brain mapping. *Proceedings of the National Academy of Sciences of the*
443 *United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424,
444 1091-6490. doi: 10.1073/pnas.0600244103.
- 445 E. L. Lehmann. Parametric versus nonparametrics: two alternative method-
446 ologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN
447 1048-5252. doi: 10.1080/10485250902842727.
- 448 G. J. McLachlan. The bias of the apparent error rate in discriminant analysis.
449 *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi:
450 10.1093/biomet/63.2.239.
- 451 S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell,
452 T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements
453 for classifying DNA microarray data. *Journal of Computational Biology:*
454 *A Journal of Computational Molecular Cell Biology*, 10(2):119–142, 2003.
455 ISSN 1066-5277. doi: 10.1089/106652703321825928.
- 456 M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Perfor-
457 mance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.
458 ISSN 1533-7928.
- 459 E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with
460 Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-
461 Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation*
462 *in Neuroimaging*, number 7263 in Lecture Notes in Computer Science,
463 pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2
464 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.
- 465 E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the
466 evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec.
467 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.

- 468 F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and
469 fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209,
470 Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.
- 471 C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G.
472 Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and
473 P. Belin. The human voice areas: Spatial organization and inter-individual
474 variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–
475 174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.
- 476 M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for
477 Class Prediction Using Gene Expression Profiles. *Journal of Computa-
478 tional Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/
479 106652702760138592.
- 480 A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy
481 for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.
- 482 J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance
483 Matrix Estimation and Implications for Functional Genomics. *Statistical
484 Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN
485 1544-6115. doi: 10.2202/1544-6115.1175.
- 486 D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class
487 Prediction and Discovery Using Gene Expression Data. In *Proceedings of
488 the Fourth Annual International Conference on Computational Molecular
489 Biology*, RECOMB ’00, pages 263–272, New York, NY, USA, 2000. ACM.
490 ISBN 978-1-58113-186-4. doi: 10.1145/332306.332564.
- 491 M. S. Srivastava. Multivariate Theory for Analyzing High Dimensional Data.
492 *Journal of the Japan Statistical Society*, 37(1):53–86, 2007. doi: 10.14490/
493 jjss.37.53.
- 494 M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high
495 dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb.
496 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.
- 497 J. Stelzer, Y. Chen, and R. Turner. Statistical inference and multiple test-
498 ing correction in classification-based multi-voxel pattern analysis (MVPA):
499 Random permutations and cluster size control. *NeuroImage*, 65:69–82, Jan.
500 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.09.063.

- 501 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press,
502 Cambridge, UK ; New York, NY, USA, Oct. 1998. ISBN 978-0-521-49603-
503 2.
- 504 G. Varoquaux, P. R. Raamana, D. Engemann, A. Hoyos-Idrobo, Y. Schwartz,
505 and B. Thirion. Assessing and tuning brain decoders: cross-validation,
506 caveats, and guidelines. working paper or preprint, June 2016.
- 507 T. D. Wager, L. Y. Atlas, M. A. Lindquist, M. Roy, C.-W. Woo, and E. Kross.
508 An fMRI-Based Neurologic Signature of Physical Pain. *New England Jour-*
509 *nal of Medicine*, 368(15):1388–1397, Apr. 2013. ISSN 0028-4793. doi:
510 10.1056/NEJMoa1204471.

511 A Analysis pipeline

512 Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in
 513 Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$
 514 the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions⁴ are
 515 centered around a unique voxel, the voxel index v also serves as a unique
 516 region index. Algorithm 1 computes a region-wise test statistic, which is
 517 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

520 B Simulation Details

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

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

531 Having generated the data, we compute each of the test statistics in Ta-
532 ble 1. For test statistics that require data folding, we used 8 folds. We then
533 compute a permutation p-value by permuting the class labels, and recomput-
534 ing each test statistic. We perform 400 such permutations. We then reject
535 the $\mu_i = 0$ null hypothesis if the permutation p-value is smaller than 0.05.
536 The reported power is the proportion of replication where the permutation
537 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.*

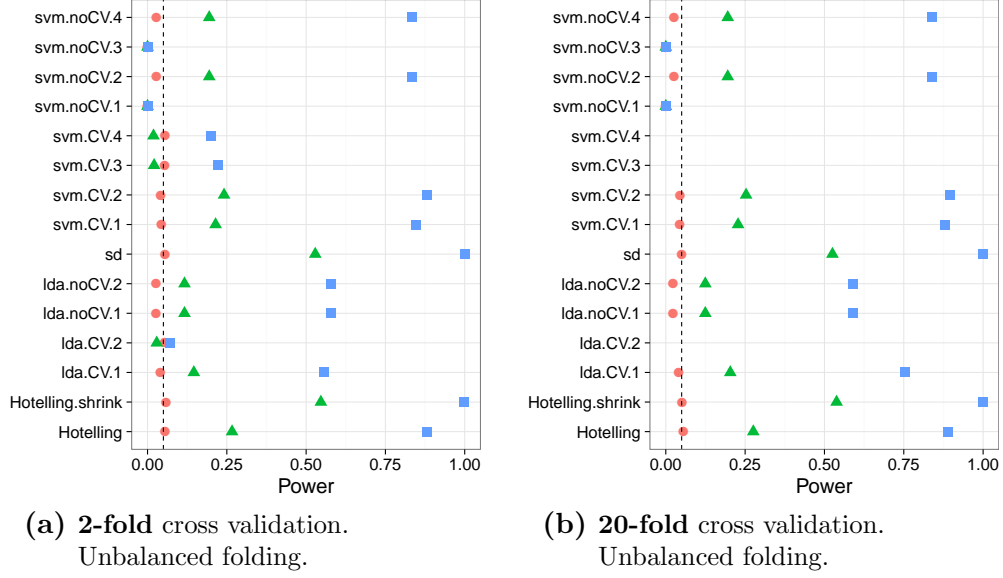
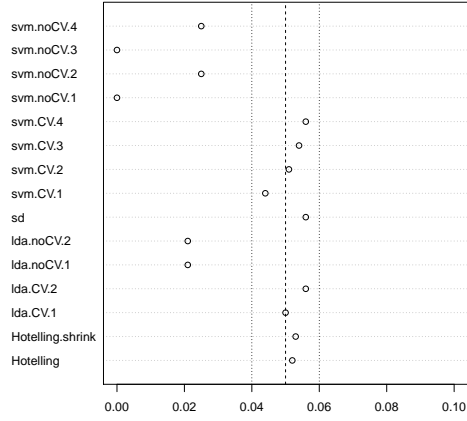


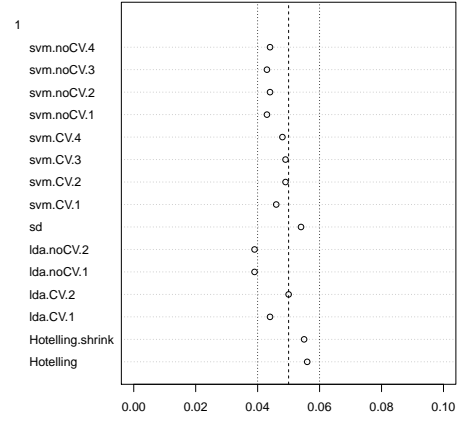
Figure 7: *Simulation details in Appendix B except the changes in the sub-captions.*



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

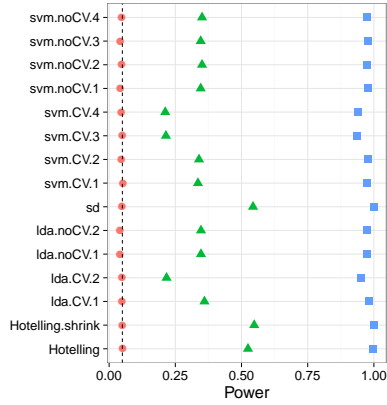


(a) **Low-Dimension:** False positive rates for $n = 40$.

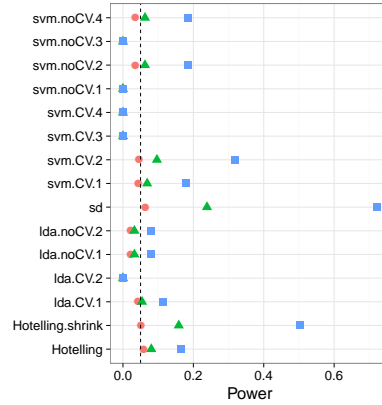


(b) **High-Dimension:** False positive rates for $n = 400$.

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