Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt
Ben Gurion University,

Roee Gilron
Tel Aviv University,

Roy Mukamel
Tel Aviv University.

September 12, 2017

Abstract

We show that using a classifier's accuracy as a test statistic, is an underpowered strategy for the purpose of finding a difference between populations, compared to a bona-fide statistical test. It is also more complicated to implement. For the cases that the purposes of the analysis is not the mere existence of a difference between populations, but rather the performance of a particular classifier, we suggest several improvements to increase power.

1 Introduction

A common workflow in neuroimaging consists of fitting a classifier, and estimating its predictive accuracy using cross validation. Given that the cross validated accuracy is a random quantity, it is then common to test if the cross validated accuracy is significantly better than chance using a permutation test. Examples in the neuroscientific literature include Golland and Fischl [2003], Pereira et al. [2009], Varoquaux et al. [2016], and especially the recently popularized *multivariate pattern analysis* (MVPA) framework of Kriegeskorte et al. [2006]. This practice is also observed in some high profile publications in the genetics literature: Golub et al. [1999], Slonim et al. [2000], Radmacher et al. [2002], Mukherjee et al. [2003], Juan and Iba [2004], Jiang et al. [2008].

To fix ideas, we will adhere to a concrete example. In Gilron et al. [2016], the authors seek to detect brain regions which encode differences between

vocal and non-vocal stimuli. Following the MVPA workflow, the localization problem is cast as a supervised learning problem: if the type of the stimulus can be predicted from the spatial activation pattern significantly better than chance, then a region is declared to encode vocal/non-vocal information. We call this an *accuracy test*, because it uses the prediction accuracy as a test statistic [1].

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 thus approach the signal detection problem with a two-group location test such as Hotelling's T^2 [Anderson, 2003]. Alternatively, if the size of brain region of interest 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], Goeman et al. [2006], or Srivastava [2007] . For brevity, and in contrast to *accuracy tests*, we will call these *location tests*[2], because they test for the equality of location of a multivariate distribution.

At this point, it becomes unclear which is preferable: a location test or an accuracy test? The former with a heritage dating back to Hotelling [1931], and the latter being extremely popular, as the $1,170$ citations[3] of Kriegeskorte et al. [2006] suggest.

The comparison between location and accuracy tests was precisely the goal of Ramdas et al. [2016], who compared Hotelling's T^2 location test to accuracy tests using *Fisher's linear discriminant analysis* classifier (LDA). By comparing the rates of convergence of the power of each statistic to 1, Ramdas et al. [2016] concluded that accuracy and location tests are rate equivalent. Rates, however, are only a first stage when comparing test statistics.

Asymptotic relative efficiency measures (ARE) are typically used by statisticians to compare between rate-equivalent test statistics [van der Vaart, 1998]. ARE is the limiting ratio of the samples sizes required by two statistics to achieve similar power. Ramdas et al. [2016] derive the asymptotic

[1] Known as *class prediction*, or *pattern discrimination*.

[2] Known as *class comparisons*.

[3] GoogleScholar. Accessed Aug 2017.

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 this theorem and 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 as the T^2 test. In this light, the accuracy test is remarkably inefficient compared to the location 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 location 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 we will show to have crucial impact. Since typical sample sizes in neuroscience are not large, we seek to study which test is to be preferred in finite samples, and not only asymptotically. Our conclusion will be quite simple: *location tests typically have more power than accuracy tests, and are easier to implement.*

Our statement rests upon the observation that with typical sample sizes, the accuracy test statistic is highly discrete. Permutation testing with discrete test statistics are known to be conservative [Hemerik and Goeman, 2014], since they are insensitive to mild perturbations of the data, and they cannot exhaust the permissible false positive rate. As put by Prof. Frank Harrell in CrossValidated[4] post back in 2011:

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

The degree of discretization is governed by the number of samples. In our example from Gilron et al. [2016], the classification accuracy is computed using 40 examples, so that the test statistic may assume only 40 possible values. This number of examples is not unusual if considering this is the number of trial-repeats, or the number of subjects, in an neuroimaging study.

The discretization effect is aggravated if the test statistic is highly concentrated. For an intuition consider the usage of the *resubstitution accuracy* as a test statistic. This statistic simply means that the accuracy is not cross validated, but rather evaluated on the training data. If the data is high

[4]A Q&A website for statistical questions: `http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier`

dimensional, the resubstitution accuracy will be very high due to over fitting. In a very high dimensional regime, the resubstitution accuracy will be 1 for the observed data [McLachlan, 1976, Theorem 1], but also for any permutation. The concentration of resubstitution accuracy near 1, and its discreteness, render this test completely useless, with power tending to 0 for any (fixed) effect size, as the dimension of the model grows.

To compare the power of accuracy tests and location tests in finite samples, we study a battery of test statistics by means of simulation. We start with formalizing the problem in Section 2. The main findings are reported in Sections 3, and 4. A discussion follows.

2 Problem setup

2.1 Multivariate Testing

Let $y \in y$ be a class encoding. Let $x \in x$ be a p dimensional feature vector. In our vocal/non-vocal example we have $y = \{0, 1\}$ and p, the number of voxels in a brain region so that $x = \mathbb{R}^{27}$.

Given n pairs of (x_i, y_i), typically assumed i.i.d., a multivariate test amounts to testing whether $x|y = 1$ has the the same distribution as $x|y = 0$. I.e., we test if the multivariate voxel activation pattern has the same distribution when given a vocal stimulus, as when given a non-vocal stimulus. The comparison metric between statistics is their power, i.e., the probability to infer that $x|y = 1$ is not distributed like $x|y = 0$.

2.2 Prediction Accuracy as a Test Statistic

An accuracy test amounts to setting a test statistic $\mathcal{T}$, to be some measure prediction accuracy.

Denoting a dataset by $\mathcal{S} := \{x_i, y_i\}_{i=1}^n$, a predictor[5], $\mathcal{A}_\mathcal{S} : x \to y$, is the output of a learning algorithm $\mathcal{A}$ when applied to the dataset $\mathcal{S}$. The accuracy of predictor[6], $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$, is defined as the probability of $\mathcal{A}_\mathcal{S}$ making a correct prediction. The accuracy of an algorithm[7], $\mathcal{E}_\mathcal{A}$, is defined as the expected accuracy over all possible data sets $\mathcal{S}$. Formalizing, we denote by $\mathcal{P}$ the probability measure of (x, y), and by $\mathcal{P}_\mathcal{S}$ the joint probability measure

[5]Known as a *hypothesis* in the machine learning literature.

[6]Known as the complement of the *test error* in Friedman et al. [2001]

[7]Known as the complement of the *expected test error* in Friedman et al. [2001]

of the sample $\mathcal{S}$. We can then write

$$\mathcal{E}_{\mathcal{A}_\mathcal{S}} := \int_{(x,y)} \mathcal{I}\{\mathcal{A}_\mathcal{S}(x) = y\}\, d\mathcal{P}(x,y), \tag{1}$$

and

$$\mathcal{E}_\mathcal{A} := \int_\mathcal{S} \mathcal{E}_{\mathcal{A}_\mathcal{S}}\, d\mathcal{P}_\mathcal{S}, \tag{2}$$

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

Denoting an estimate of $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$ by $\hat{\mathcal{E}}_{\mathcal{A}_\mathcal{S}}$, and $\mathcal{E}_\mathcal{A}$ by $\hat{\mathcal{E}}_\mathcal{A}$, a statistically significant "better than chance" estimate of either, is evidence that the classes are distinct.

Two popular estimates of $\hat{\mathcal{E}}_\mathcal{A}$ are the *resubstitution estimate*[8], and the V-fold Cross Validation (CV) estimate.

Definition 1 (Resubstitution estimate)**.** The resubstitution accuracy estimator of a learning algorithm $\mathcal{A}$, denoted $\hat{\mathcal{E}}^{Resub}_\mathcal{A}$, is defined as

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

Definition 2 (V-fold CV estimate)**.** Denoting by $\mathcal{S}^v$ the v'th partition, or *fold*, of the dataset, and by $\mathcal{S}^{(v)}$ its complement, so that $\mathcal{S}^v \cup \mathcal{S}^{(v)} = \cup_{v=1}^{V} \mathcal{S}^v = \mathcal{S}$, the V-fold CV accuracy estimator, denoted $\hat{\mathcal{E}}^{Vfold}_\mathcal{A}$, is defined as

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

where $|A|$ denotes the cardinality of a set A.

2.3 A Battery of Candidate Tests

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

1. How to estimate $\hat{\mathcal{E}}_\mathcal{A}$? In particular, should it be resampled and how?
2. If resampling using V-fold cross validation: How many folds? Should the data be refolded after each permutation? Should the folding be balanced?

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 $\mathcal{E}_\mathcal{A}$, but rather in the detection of its departure from chance level.

[8]Known as the *train error* in Friedman et al. [2001].

Cross validate or not? Given our goal, a biased estimate of $\hat{\mathcal{E}}_{\mathcal{A}}$ is not a problem provided that bias is consistent over all permutations. The underlying intuition is that a permutation test will be unbiased, provided that the exact same computation is performed over all permutations. We will thus be considering both cross validated accuracies, and resubstitution accuracies.

Balanced folding? The standard practice when V-fold 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.

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

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

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

Name	Algorithm	Resampling	Parameters
Hotelling	Hotelling	Resubstitution	–
Oracle	Hotelling	Resubstitution	–
Hotelling.shrink	Hotelling	Resubstitution	–
SD	Hotelling	Resubstitution	–
LDA.CV.1	LDA	V-fold	–
LDA.noCV.1	LDA	Resubstitution	–
LDA.CV.1	SVM	V-fold	cost=10
SVM.CV.2	SVM	V-fold	cost=0.1
SVM.noCV.1	SVM	Resubstitution	cost=10
SVM.noCV.2	SVM	Resubstitution	cost=0.1

Table 1: This table collects the various test statistics we will be studying. Location tests include: *Oracle*, *Hotelling*, *Hotelling.shrink*, and *SD*. *Hotelling* is the classical two-group T^2 statistic [Anderson, 2003]. *Oracle* is the same as Hotelling's T^2, only using the generative covariance, and not an estimated one. *Hotelling.shrink* is a high dimensional version of T^2, with the regularized covariance from Schäfer and Strimmer [2005]. *SD* is another high dimensional version of the T^2, from Srivastava et al. [2013]. The rest of the tests are accuracy tests, with details given in the table. For example, *SVM.CV.2* is a linear SVM, with V-fold cross validated accuracy, and cost parameter set at 0.1 [Meyer et al., 2015]. Another example is *LDA.noCV.1*, which is Fisher's LDA, with a resubstituted accuracy estimate.

2.4 Collecting the Pieces Into a Permutation Test

The various test statistics in Table 1 will be compared using their power. Because our problems of interest are typically high-dimensional, i.e. $n \gg p$ does not hold, them central limit laws will not apply and we recur to permutation tests. Because we focus on two-group testing under an independent sampling assumption, we know that a label-switching permutation test is valid in that it has a false positive rate no larger than the desired level.
The sketch of our test is the following:
(a) Fix a test statistic $\mathcal{T}$.
(b) Permute labels and recompute $\mathcal{T}$ to recover its null.
(c) Declare classes differ if the observed $\mathcal{T}$ is beyond its 95%'th permutation percentile.

3 Results

Now that we can construct various accuracy tests, we can compare them to location tests. We do so via simulation, which details are now described.

3.1 Simulation Details

The following details are common to all the reported simulations, unless stated otherwise in a figure's caption. The R code for the simulations can be found in `http://www.john-ros.com/permuting_accuracy/`.

Each simulation is based on $1,000$ replications. In each replication, we generate n i.i.d. samples from a shift class $\mathbf{x}_i = \mu \mathbf{y}_i^* + \eta_i$, where $y_i^* = \{0, 1\}$ encodes the class of subject i, the noise is distributed as $\eta_i \sim \mathcal{N}_p(0, \Sigma)$, the sample size $n = 40$, and the dimension of the data is $p = 23$. The covariance $\Sigma = I$. Effects, i.e. shifts μ, are equal equal in all p coordinates, and vary over $\mu \in \{0, 1/4, 1/2\}$.

Having generated the data, we compute each of the test statistics in Table 1. For test statistics that require data folding, we used 4 folds. We then compute a permutation p-value by permuting the class labels, and recomputing each test statistic. We perform 300 such permutations. We then reject the $\mu_i = 0$ null hypothesis against a two sided alternative if the permutation p-value is smaller than 0.05. The reported power is thus the proportion of replication where the permutation p-value falls below 0.05.

3.2 False Positive Rate

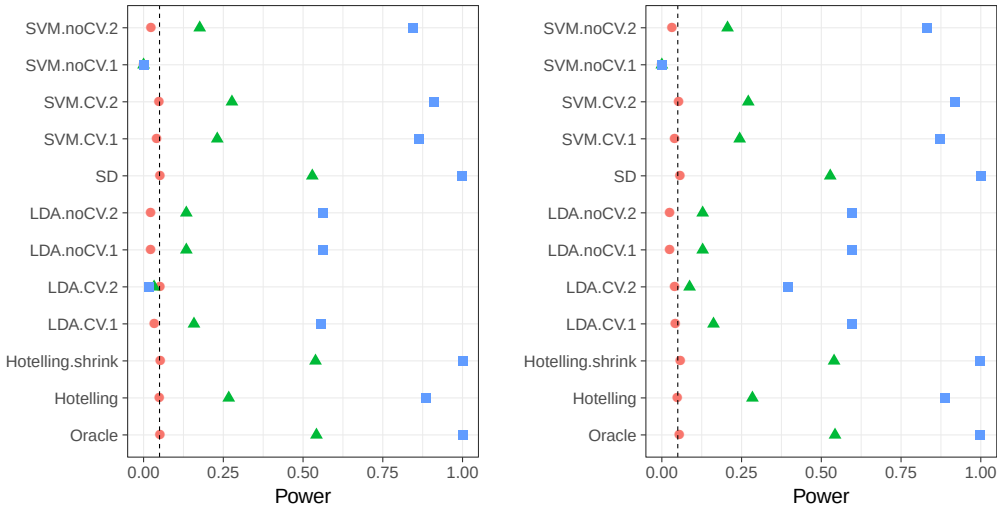
We start with a sanity check of the simulation. Theory suggests that all permutation tests should control the false positive rates. Our simulations confirm this. In all our results, such as Figure 1, we encode the null case, where no signal is present and $x|y = 1$ has the same distribution as $x|y = 0$, by a red circle. Since the red circles are always below the desired 0.05 error rate, then the false positive rate of all test statistics, in all simulations is controlled. We may thus proceed and compare the power of each test statistic.

3.3 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 location tests to accuracy tests.

From Figure 1 we learn that location tests are more powerful than accuracy tests. This is particularly visible for intermediate signal strength (green triangle), and the high-dimensional location tests: *SD* and *Hotelling.shrink* (see Table 1).

Figure 1: The power of the permutation test with various test statistics. The power on the x axis. Effects 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 Section 3.1. Cross-validation was performed with balanced and unbalanced data folding. See sub-captions.



(a) Unbalanced. **(b) Balanced.**

Clearly, the statement that location tests have more power than accuracy tests is currently limited to the simulation setup of Figure 1. Are location tests always preferable to accuracy tests? No. There exist scenarios where accuracy tests have no less, or even more power, than location tests. These are discussed in Section 5.

3.4 Scale Alternatives

Is it possible for an accuracy test to have more power than a location test? Yes. *Scale alternatives*, where signal is carried in the covariance and not in means, provide one such example.

In Figure 2a we report a simulation designed to show the possibility of an accuracy test being more powerful then a location test. This result is merely a proof of existence, and is not to be interpreted as "accuracy tests have more power than location tests against scale alternatives". If the practitioner were

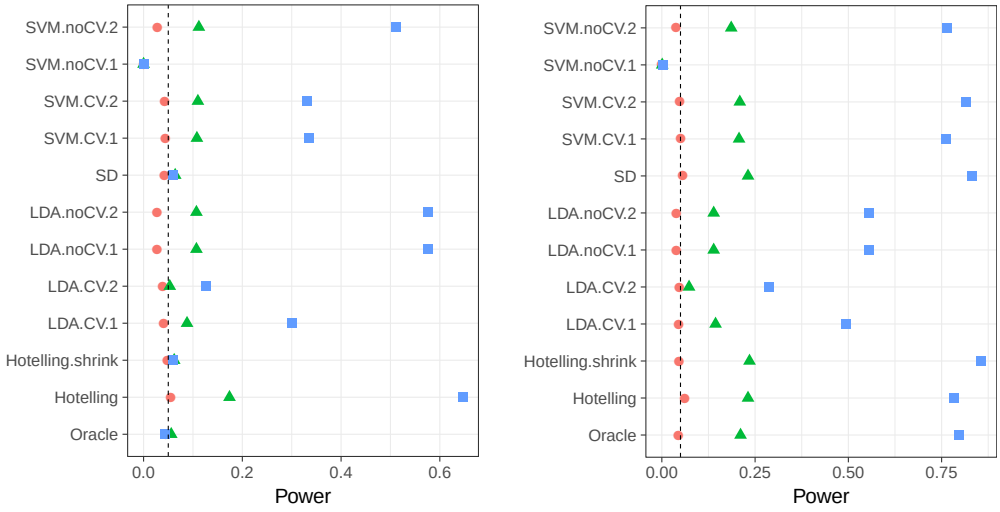
expecting a change in scatter, then a Neyman-Pearson type reasoning would lead them to abandon both location and accuracy tests, and adopt a bona-fide covariance test [e.g. Ley et al., 2015]. Our empirical example in Section 4, and experience, actually support the opposite conclusion: that location tests have more power than accuracy tests.

3.5 Departure From Gaussianity

The Neyman-Pearson type reasoning that favors the location test over accuracy tests may fail when the data is not multivariate Guassian, thus Hotelling's T^2 statistic no longer a generalized-likelihood-ratio test.

To check this, we replaced the multivariate Gaussian distribution with a heavy-tailed multivariate-t distribution. The dominance of the location tests was preserved even under a multivariate heavy-tailed distribution, but this does reduce the advantage of the location tests over accuracy tests (Figure 2b).

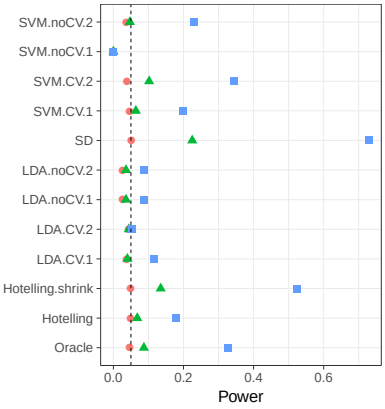
Figure 2: Simulation details in Appendix 3.1 except the changes in the sub-captions.



(a) Scale Change– $\mathbf{x}_i = \eta_i * \mu^{\mathbf{y}_i^*}$ so that the effects are a scale change.

(b) Heavytailed– η_i is not p-variate Gaussian, but rather p-variate t, with $df = 3$.

Figure 3: Simulation details in Appendix 3.1 except the changes in the sub-captions.



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

(b) Long-memory Brownian motion dependence: $\Sigma_{k,l} = \rho \min\{k, l\}; \rho = 0.8$. .

3.6 Dependence

3.7 The Effect of High Dimension

Inspecting Figure 1a (for instance), it can be seen that Hotelling's unregularized T^2 test has similar power as accuracy tests. It could thus be argued that the real advantage of *SD* or *Hotelling.shrink* is not in their sensitivity to location, but rather, in their adaptation to high dimension[9] by regularization. To study this, we call upon several *covariance regularized classifiers*, designed for high dimensional problems. We try an l_2 regularized SVM [Friedman et al., 2010], and a shrinkage based LDA [Pang et al., 2009, Ramey et al., 2016], which are similar in spirit to the *Hotelling.shrink* test. We also try we try a diagonalized LDA[10] [Dudoit et al., 2002], which is similar in spirit to the *SD* location test.

Simulation results are reported in Figure 4 with naming conventions in Table 2. It can be seen that regularizing a classifier in high dimension, just like a location test, improves power. It can also be seen that (regularized) location tests are still more powerful than (regularized) accuracy tests. This was to be expected, since we already saw in (e.g.) Figure 1a that the unregularized location test, *Hotelling*, is slightly more powerful than unregularized accuracy tests such as (e.g.) *SVM.CV.1*; a phenomenon we attribute to the

[9]By *high-dimension* it is typically meant that p/n is not too small, regardless of p and n themselves.

[10]Known as *Gaussian Naïve Bayes*.

discretization effect.

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

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

4 Neuroimaging Example

Figure 5 is an application of both a location and an accuracy test to the data of Pernet et al. [2015]. The authors of Pernet et al. [2015] collected fMRI data while subjects were exposed to the sounds of human speech (vocal), and other non-vocal sounds. Each subject was exposed to 20 sounds of each type, totaling in $n = 40$ trials. The study was rather large and consisted of about 200 subjects. The data was kindly made available by the authors at the OpenfMRI website[11].

We perform group inference using within-subject permutations along the analysis pipeline of Stelzer et al. [2013], which was also reported in Gilron et al. [2016]. To demonstrate our point, we compare the *SD* location test with the *SVM.cv.1* accuracy test.

In agreement with our simulation results, the location test (*sd*) discovers more brain regions of interest when compared to an accuracy test (*svm.cv.1*). The former discovers $1,232$ regions, while the latter only 441, as depicted in Figure 5. We emphasize that both test statistics were compared with the

[11] `https://openfmri.org/`

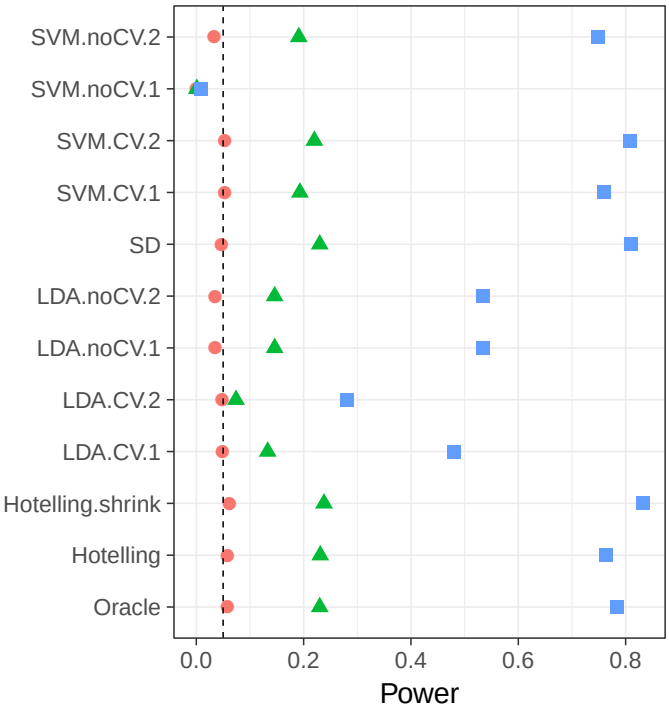


Figure 4: **HighDim Classifier**– The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix 3.1.

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

5 Discussion

We have set out to understand which of the tests is more powerful: accuracy tests or location tests. Our practical advice for the practitioner, is that accuracy tests are never optimal. There is always a multivariate test, possibly a location test, that dominates in power. The class of location tests we examined, in particular their regularized versions, are good performers in a wide range of simulation setups and empirically. They are also typically easier to implement, and faster to run, since no cross validation will be involved. Their high-dimensional versions, such as Schäfer and Strimmer [2005], Goeman et al. [2006], and Srivastava [2007], are particularly well suited for empirical problems such as neuroimaging and genetics.

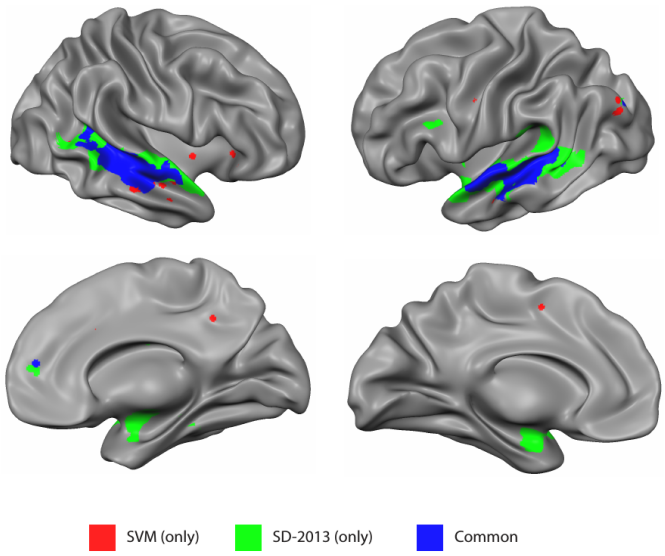


Figure 5: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a location test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost parameter 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 Gilron et al. [2016].

5.1 Where do Accuracy Tests Lose Power?

The low power of the accuracy tests compared to location tests can be attributed to the following causes:
(a) A Neyman-Pearson type argument, favoring location tests for the shift alternatives we simulated (even if not maximal power).
(b) The discrete nature of accuracy tests.
(c) Inefficient use of the data when validating with a holdout set.
(d) Inappropriate regularization in high SNR regimes: testing requires less regularization than predicting.

Shift Alternatives We focused on shift alternatives. One may argue that empirical alternatives are rarely of shift type, thus limiting our conclusions. We reply that our empirical evidence in Section 4, also favors location tests.

More generally: discretization alone suffice to state that there will always be a test statistic with more power than an accuracy test. When looking for a "one-size fits all" strategy, it is our experience that location tests are a better candidate than accuracy tests.

Discreteness The degree of discretization is governed by the sample size. For this reason, an asymptotic analysis such as Ramdas et al. [2016], or Golland et al. [2005], will not capture power loss due to discretization[12]. These studies have other limitations as well. Ramdas et al. [2016] assume a split-sample holdout, without cross-validation. This hurts the power of accuracy tests. The asymptotic analysis of the resubstitution accuracy statistics in Golland et al. [2005] not only conceals the discreteness of the test statistic, but also the effects of its concentration. This analysis renders resubstitution accuracy estimates a legitimate asymptotic test, while they are a terrible finite sample test.

5.2 Interpretation

Multivariate tests, and location tests in particular, are easier to interpret. To do so we typically use a Neyman-Pearson type argument, and think of the type of signal a test is sensitive to. Accuracy tests are seen as "black boxes", even though they can be analyzed in the same way. Gilron et al. [2017] for instance, demonstrate that the type of signal captured by accuracy tests is less obvious to neuroimaging practitioners than location tests.

5.3 Testing in Augmented Spaces

It may be argued that accuracy tests permits the separation between classes in high dimensions, such as in *reproducing kernel Hilbert spaces* (RKHS) by using non-linear predictors while location tests do not. This is a false argument– accuracy tests do not have any more flexibility than location tests. Indeed, it is possible to test for location in the same space the classifier is learned. For independence tests in high dimensional spaces see for example Székely and Rizzo [2009] or Gretton et al. [2012].

5.4 Smoothing Accuracy Estimates

It may be possible to alleviate the effect of discretization via the cross-validation scheme. The discreteness of the accuracy statistic is governed

[12]This actually holds for all power analyses relying on a *contiguity* argument [van der Vaart, 1998, Ch.6].

by the number of examples in the union of holdout test sets. For V-fold CV, for instance, the accuracy may assume as many values as the sample size. This suggests that the accuracy can be "smoothed" by allowing the test sample to be drawn with replacement. An algorithm that samples test sets with replacement is the *leave-one-out bootstrap estimator*, and its derivatives, such as the *0.632 bootstrap*, and *0.632+ bootstrap* [Friedman et al., 2001, Sec 7.11].

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

$$\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{(i)}|}\sum_{b\in C^{(i)}}\mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i)=y_i\}. \tag{5}$$

where $|A|$ is the cardinality of set A. An equivalent formulation, which stresses the Bootstrap nature of the algorithm is the following. Denoting by $S^{(b)}$ the indexes of observations, i, that are *not* in the bootstrap sample b and are not empty,

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

Definition 4 (b0.632)**.** The *0.632 bootstrap* accuracy estimate, b0.632, is a weighted average of the resubstitution error and the bLOO. Formally:

$$\hat{\mathcal{E}}_{\mathcal{A}}^{0.632} := 0.368\ \hat{\mathcal{E}}_{\mathcal{A}}^{Resub} + 0.632\ \hat{\mathcal{E}}_{\mathcal{A}}^{bLOO}. \tag{7}$$

Simulation results are reported in Figure 6 with naming conventions in Table 3. It can be seen that selecting test sets with replacement does increase the power, when compared to V-fold cross validation, but still falls short from the power of location tests. It can also be seen that power increases with the number of bootstrap replications, as was to be expected, since more replications reduce the level of discretization. The type of bootstrap, bLOO versus b0.632, does not change the power.

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

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

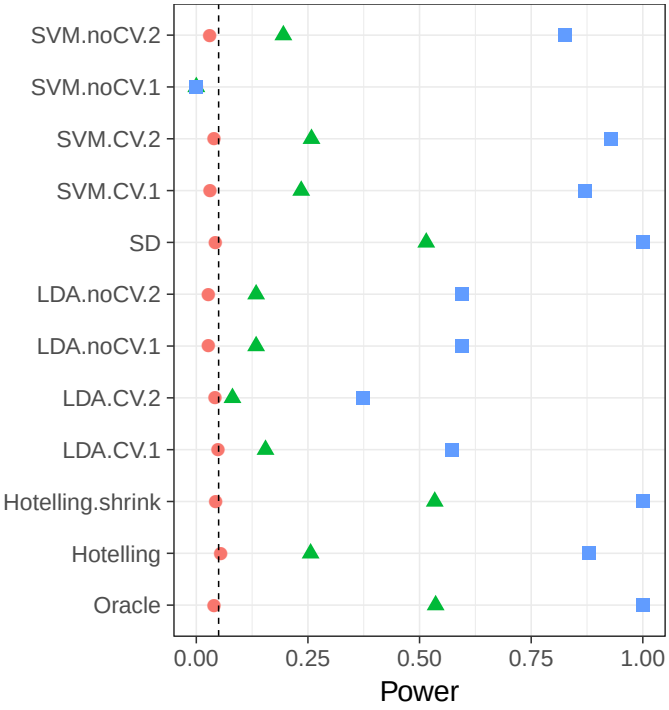


Figure 6: **Bootstrap**– The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 3. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix 3.1.

5.5 A Good Accuracy Test

There are cases where an accuracy test cannot be replaced by some location, or other, statistical test. Brain-computer interfaces and clinical diagnostics [e.g. Olivetti et al., 2012, Wager et al., 2013] are examples where we want to know not only if information is encoded in a region, but rather, that a particular decoder can extract it. For these cases, we collect some conclusions and best practices.

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

Regularize. Regularization proves crucial to detection power in low signal to noise regimes: in high dimension and/or in the presence of strong correlations. We find that the Shrinkage-based Diagonal Linear Discriminant Analysis of Pang et al. [2009] is a particularly good performer, but more research is required on this matter. We also conjecture that the power-maximizing regularization is larger than the error-minimizing regularization.

Smooth accuracy. Smooth accuracy estimate by cross validating with replacement. The bLOO estimator, in particular, is preferable over V-fold.

We can compound the regularization with the bootstrapping from Section 5.4, to improve finite sample power of the accuracy tests. This is done in the *SVM.highdim.2* and *LDA.highdim.4* tests. The latter being one of the very few accuracy tests that achieve the same power as location tests. This is exciting news since it shows how to design powerful new high-powered accuracy tests: by sampling test sets with replacement, and by regularizing the classifiers.

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

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

In low dimensional problems, the discretization is less impactful, and the computational burden of cross validation can be avoided by using the resubstitution accuracy. There is a fundamental difference between V-folding and resubstitution. The latter should not be thought of as the limit of the former.

Don't z-score. There is no gain in z-scoring the accuracy scores. Our motivating rational was clearly flawed.

5.6 Related Literature

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

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

Golland and Fischl [2003] and 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 dimension requirement [Golland et al., 2005, Sec 4.3] prevents the permutation p-value from (asymptotically) concentrating near 1. Like ourselves, they also find that the power increases with the size of the test set. This is seen in Fig.4 of Golland et al. [2005], where the size of the test-set, K, governs the discretization. Since they permute features, not labels, then all their permutation samples are balanced, and there is no issue of refolding.

Golland et al. [2005] simulate the power of accuracy tests by sampling from a Gaussian mixture family of models, and not from a location family

as our own simulations. Under their model

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

and

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

Varying p interpolates between the null distribution ($p = 0.5$) and a location shift model ($p = 0$). We now perform the same simulation as Golland et al. [2005], after parameterizing p so that $p = 0$ corresponds to the null model, and in the same dimensionality as our previous simulations. We find that also in this mixture class of models a location test has more power than an accuracy test (Figure 7).

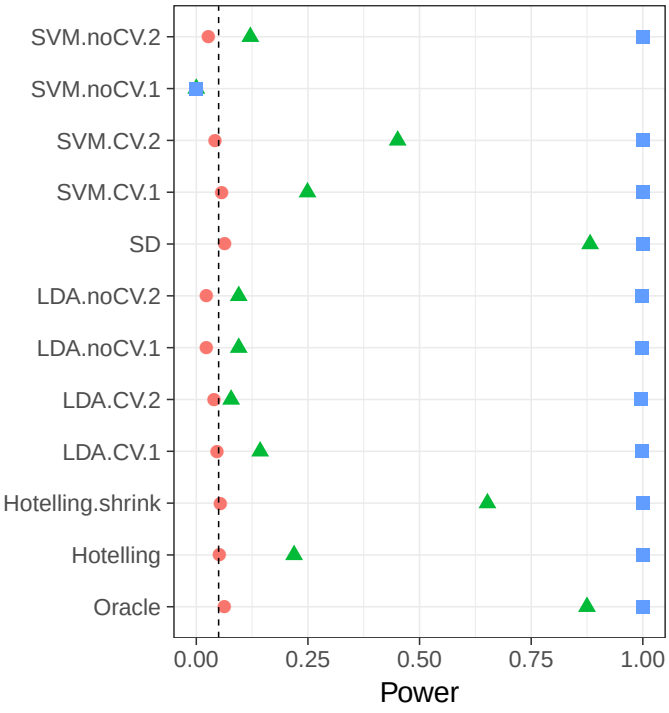


Figure 7: **Mixture**– $\mathbf{x}_i = \chi_i\mu + \eta_i; \chi_i = \{-1, 1\}$ and $Prob(\chi_i = 1) = (1/2 - p)^{\mathbf{y}_i^*}(1/2 + p)^{1-\mathbf{y}_i^*}$. μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, p, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

5.7 Epilogue

Given all the above, we find the popularity of accuracy tests for signal detection 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.

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, Hoboken, NJ, 3 edition edition, July 2003. ISBN 978-0-471-36091-9.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES B*, 57:289–289, 1995.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, Mar. 2002. ISSN 0162-1459. doi: 10.1198/016214502753479248.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. Quantifying spatial pattern similarity in multivariate analysis using functional anisotropy. *arXiv:1605.03482 [q-bio]*, May 2016.

R. Gilron, J. Rosenblatt, O. Koyejo, R. A. Poldrack, and R. Mukamel. What's in a pattern? examining the type of signal multivariate analysis uncovers at the group level. *NeuroImage*, 146:113–120, 2017.

J. J. Goeman, S. A. Van De Geer, and H. C. Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

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.

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.

W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

L. Juan and H. Iba. Prediction of tumor outcome based on gene expression data. *Wuhan University Journal of Natural Sciences*, 9(2):177–182, Mar. 2004. ISSN 1007-1202, 1993-4998. doi: 10.1007/BF02830598.

N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868, July 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0600244103.

E. L. Lehmann. Parametric versus nonparametrics: two alternative methodologies. *Journal of Nonparametric Statistics*, 21(4):397–405, 2009. ISSN 1048-5252. doi: 10.1080/10485250902842727.

C. Ley, D. Paindaveine, and T. Verdebout. High-dimensional tests for spherical location and spiked covariance. *Journal of Multivariate Analysis*, 139: 79–91, 2015.

G. J. McLachlan. The bias of the apparent error rate in discriminant analysis. *Biometrika*, 63(2):239–244, Jan. 1976. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/63.2.239.

D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. 2015. R package version 1.6-7.

S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 10(2):119–142, 2003. ISSN 1066-5277. doi: 10.1089/106652703321825928.

M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010. ISSN ISSN 1533-7928.

E. Olivetti, S. Greiner, and P. Avesani. Induction in Neuroscience with Classification: Issues and Solutions. In G. Langs, I. Rish, M. Grosse-Wentrup, and B. Murphy, editors, *Machine Learning and Interpretation in Neuroimaging*, number 7263 in Lecture Notes in Computer Science, pages 42–50. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34712-2 978-3-642-34713-9. doi: 10.1007/978-3-642-34713-9_6.

E. Olivetti, S. Greiner, and P. Avesani. Statistical independence for the evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec. 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.

H. Pang, T. Tong, and H. Zhao. Shrinkage-based Diagonal Discriminant Analysis and Its Applications in High-Dimensional Data. *Biometrics*, 65 (4):1021–1029, Dec. 2009. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2009.01200.x.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209, Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.

C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. Charest, P. E. G. Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and P. Belin. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–174, Oct. 2015. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2015.06.050.

M. D. Radmacher, L. M. McShane, and R. Simon. A Paradigm for Class Prediction Using Gene Expression Profiles. *Journal of Computational Biology*, 9(3):505–511, June 2002. ISSN 1066-5277. doi: 10.1089/106652702760138592.

A. Ramdas, A. Singh, and L. Wasserman. Classification Accuracy as a Proxy for Two Sample Testing. *arXiv:1602.02210 [cs, math, stat]*, Feb. 2016.

J. A. Ramey, C. K. Stein, P. D. Young, and D. M. Young. High-Dimensional Regularized Discriminant Analysis. *arXiv preprint arXiv:1602.01182*, 2016.

J. Schäfer and K. Strimmer. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), Jan. 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1175.

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

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

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

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

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, Dec. 2009. ISSN 1932-6157, 1941-7330. doi: 10.1214/09-AOAS312.

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

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

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