Better-Than-Chance Classification for Signal Detection

Jonathan D. Rosenblatt
Department of IE&M and
Zlotowsky Center for Neuroscience,
Ben Gurion University of the Negev, Israel.

Yuval Benjamini
Department of Statistics,
Hebrew University, Israel

Roee Gilron
Movement Disorders and Neuromodulation Center,
University of California, San Francisco.

Roy Mukamel
School of Psychological Science
Tel Aviv University, Israel.

Jelle Goeman
Department of Medical Statistics and Bioinformatics,
Leiden University Medical Center, The Netherlands.

April 9, 2018

Abstract

The estimated accuracy of a classifier is a random quantity with variability. A common practice in supervised machine learning, is thus to test if the estimated accuracy is significantly better than chance level. This method of signal detection is particularly popular in neuroimaging and genetics. We provide evidence that using a classifier's accuracy as a test statistic can be an underpowered strategy for finding differences between populations, compared to a bona-fide statistical test. It is also computationally more demanding than a statistical

test. Via simulation, we compare test statistics that are based on classification accuracy, to others based on multivariate test statistics. We find that probability of detecting differences between two distributions is lower for accuracy based statistics. We examine several candidate causes for the low power of accuracy tests. These causes include: the discrete nature of the accuracy test statistic, the type of signal accuracy tests are designed to detect, their inefficient use of the data, and their regularization. When the purposes of the analysis is not signal detection, but rather, the evaluation of a particular classifier, we suggest several improvements to increase power. In particular, to replace V-fold cross validation with the Leave-One-Out Bootstrap.

Keywords: signal-detection; multivariate-testing; supervised-learning; hypothesis-testing; high-dimension

1 Introduction

Many neuroscientists and geneticists detect signal by fitting a classifier and testing whether it's prediction accuracy is better than chance. The workflow 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. This general idea has been promoted by Friedman [2003], Eric et al. [2008], Lopez-Paz and Oquab [2016]. Examples in the genetics literature include Golub et al. [1999], Slonim et al. [2000], Radmacher et al. [2002], Mukherjee et al. [2003], Juan and Iba [2004], Jiang et al. [2008], Yu et al. [2007]. Other examples include speaker verification [Gretton et al., 2012], text classification [Dhillon et al., 2003, Lopez-Paz and Oquab, 2016], distinguishing between facial expressions [Lopez-Paz and Oquab, 2016], data integration [Gretton et al., 2012], attribute matching in databases systems (a.k.a. record linkage) [Gretton et al., 2012, Hall and Tajvidi, 2002], optical character recognition [Pérez-Cruz, 2009], multimedia [Moreno et al., 2004], and functional data analysis [Hall and Tajvidi, 2002].

Examples in the neuroscientific literature, which is our motivating use-case, include Golland and Fischl [2003], Pereira et al. [2009], Schreiber and Krekelberg [2013], Olivetti et al. [2013], Varoquaux et al. [2016], and especially the recently popularized *multivariate pattern analysis* (MVPA) framework of Kriegeskorte et al. [2006].

To fix ideas, we will adhere to a concrete example. In Gilron et al. [2016], the authors seek to detect brain regions that 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 brain's 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.

This same signal detection task can also be approached as a 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 hypothesis test. Two-groups hypothesis tests may be divided into tests for equality of location (i.e., equality of means), and two-sample goodness of fit tests (i.e., equality of the two whole distribution). The t-test or a Kolmogorov-Smirnov test, are univariate examples of such tests, respectively. The canonical multivariate location test is Hotelling's T^2 test [Hotelling, 1931]. Multivariate goodness-of-fit tests date back to [Bickel, 1969], but see Gretton et al. [2012], and references therein, for some recent advances.

Crucially for our applications, we will assume that the number of samples is similar to the dimension of each sample. In the statistical literature this is known as a *high-dimensional* problem. We emphasize that by high-dimension it is not necessarily implied that the sample is large, even if it is often the case. In our running example it means that the size of the brain's region of interest is large compared to the number of subjects in the experiment.

In a seminal contribution, Bai and Saranadasa [1996] noted that in-high dimension, multivariate tests tend to be low powered unless some regularization is involved. Since then, many high-dimensional tests have been proposed. Examples of high-dimensional goodness-of-fit tests include Hall and Tajvidi [2002], Székely and Rizzo [2009], Gretton et al. [2012]. Examples of high-dimensional locations tests include Schäfer and Strimmer [2005], Goeman et al. [2006], Srivastava [2007], Lopes et al. [2011], Nishiyama et al. [2013], Thulin [2014], Shen and Lin [2015], Xu et al. [2016], Zhang and Pan [2016], and certainly others.

At this point, it becomes unclear which is preferable, in particular for genetics and neuroimaging? Various authors have addressed this matter, often with contradicting conclusions. Yu et al. [2007], for instance, finds that an accuracy test based on a tree predictor is preferable over testing. Their

simulated signal is sparse, so that it is no surprise that a tree-type predictors outperforms linear predictors and tests. Olivetti et al. [2013] compares the Kernel test of Gretton et al. [2012] to a logistic-regression based accuracy-test. Their results are inconclusive with a slight advantage to the logistic regression. Lopez-Paz and Oquab [2016] compare several accuracy tests to several location and GOF tests and conclude that a neural-net based accuracy test is preferable. Their argument is that the neural-net is able to learn the features that best separate the samples. Their examples, however, are low-dimensional. This feature learning is typically impossible in high-dimension. Ramdas et al. [2016], is currently the only analytic analysis. Comparing Hotelling's T^2 location test to *Fisher's linear discriminant analysis* (LDA) accuracy test. By comparing the rates of convergence of the power of each statistic, 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 sample sizes required by two statistics to achieve similar power. Ramdas et al. [2016] derive the asymptotic power functions of the two test statistics, which allows to compute the ARE between Hotelling's T^2 (location) test and Fisher's LDA (accuracy) test. Theorem 14.7 of van der Vaart [1998] relates asymptotic power functions to ARE. Using 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.

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 can be summarized as follows:

(1) There is always a location test that dominates an accuracy test.

(2) In high-dimension, performing a location test in the original sample/feature space is a practical and good-powered strategy.

Our statement rests upon the following observations, which rely on a battery of simulations and the discussion that follows.

1. In our typical sample sizes, the accuracy test statistic is highly discrete, and needlessly so.

2. In high-dimension, learning a feature map may introduce more variance than it reduces bias, so that performing the test in the original space is a competitive strategy.

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. The sample sizes in the simulations were typical for a neuroimaging study, which can be thought of as a high-dimension–small-sample setup, but we also show that our conclusions apply to a high-dimension–large-sample setup. We conclude with a discussion.

2 Problem setup

2.1 Multivariate Testing

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

Denoting a dataset by $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n$, a multivariate test amounts to testing whether the distribution of x given $y = 1$ is the same as x given $y = 0$. For example, we can test whether multivariate voxel activation patterns (x) are similarly distributed when given a vocal stimulus ($y = 1$) or a non-vocal one ($y = 0$). The tests are calibrated to have a fixed false positive rate ($\alpha = 0.05$). The comparison metric between statistics is power, i.e., the probability to infer that $x|y = 1$ is not distributed like $x|y = 0$.

2.2 From a Test Statistic to a Permutation Test

The multivariate tests we will be considering rely on fixing some test statistic, and comparing it to it's permutation distribution. The tests differ in the statistic they employ. Our comparison metric is their power, i.e., their true positive rate. We adhere to permutation tests and not parametric inference because our problems of interest are typically high-dimensional. This means that $n \gg p$ does not hold, and central limit laws do not apply. Because we focus on two-group testing under an independent sampling assumption, we know that a label-switching permutation test is valid even if possibly conservative. The sketch of our permutation test is the following:

(a) Fix a test statistic $\mathcal{T}$ with a right tailed rejection region.
(b) Sample a random permutation of the class labels, $\pi(y)$.
(c) Permute labels and recompute the statistic $\mathcal{T}_\pi$.
(d) Repeat (a)-(c) R times.
(e) The permutation p-value is the proportion of $\mathcal{T}_\pi$ larger than the observed $\mathcal{T}$. Formally: $\mathbb{P}\{\mathcal{T}_\pi \geq \mathcal{T}\} := \frac{1}{R}\sum_\pi I\{\mathcal{T}_\pi \geq \mathcal{T}\}$.
(f) Declare classes differ if the permutation p-value is smaller than α, which we set to $\alpha = 0.05$.

We now detail the various test statistics that will be compared.

2.3 Location Tests and Hotelling's T^2

The most prevalent interpretation of "$x|y = 1$ is not distributed like $x|y = 0$" is to assume they differ in means. In his seminal work, Hotelling [1931] has proposed the T^2 test statistic for testing the equality in means of two multivariate distributions. Using our notations this statistic is proportional to the difference between group means, measured with the Mahalanobis norm:

$$T^2 \propto (\bar{x}_{y=1} - \bar{x}_{y=0})' \hat{\Sigma}^{-1} (\bar{x}_{y=1} - \bar{x}_{y=0}), \tag{1}$$

where $\bar{x}_{y=j}$ is the p-vector of means in the $y = j$ group, and $\hat{\Sigma}$ is a pooled covariance estimator. Perhaps more intuitively, T^2 is Euclidean norm of the mean difference vector, but after transforming to decorrelated scales. For more background see, for example, Anderson [2003].

The major difficulty with these multivariate tests is that Σ has $p(p+1)/2$ free parameters, so that n has to be very large to apply these tests. If n is not much larger than p, or in low signal-to-noise (SNR), the test is very low powered, as shown by Bai and Saranadasa [1996]. In these cases, high dimensional versions of the T^2 should be applied, which essentially regularize the estimator of Σ, thus reducing the dimensionality of the problem and improving the SNR and power.

2.4 Prediction Accuracy as a Test Statistic

An accuracy test amounts to using a predictor's accuracy as a test statistic.

A predictor[1], $\mathcal{A}_\mathcal{S} : \mathcal{X} \to \mathcal{Y}$, is the output of a learning algorithm $\mathcal{A}$ when applied to the dataset $\mathcal{S}$. The accuracy of predictor[2], $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$, is defined as the probability of $\mathcal{A}_\mathcal{S}$ making a correct prediction. The accuracy of an

[1]Known as a *hypothesis* in the machine learning literature.
[2]Known as (the complement of) the *test error* in Friedman et al. [2001]

algorithm[3], $\mathcal{E}_{\mathcal{A}}$, 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 of the sample $\mathcal{S}$. We can then write

$$\mathcal{E}_{\mathcal{A}_{\mathcal{S}}} := \int_{(x,y)} \mathcal{I}\{\mathcal{A}_{\mathcal{S}}(x) = y\}\, d\mathcal{P}, \tag{2}$$

and

$$\mathcal{E}_{\mathcal{A}} := \int_{\mathcal{S}} \mathcal{E}_{\mathcal{A}_{\mathcal{S}}}\, d\mathcal{P}_{\mathcal{S}}, \tag{3}$$

where $\mathcal{I}\{A\}$ is the indicator function[4] of the set 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*, and the V-fold Cross Validation (CV) estimate.

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

$$\hat{\mathcal{E}}_{\mathcal{A}}^{Resub} := \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\{\mathcal{A}_{\mathcal{S}}(x_i) = y_i\}. \tag{4}$$

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 $\hat{\mathcal{E}}_{\mathcal{A}}^{Vfold}$, is defined as

$$\hat{\mathcal{E}}_{\mathcal{A}}^{Vfold} := \frac{1}{V} \sum_{v=1}^{V} \frac{1}{|\mathcal{S}^v|} \sum_{i \in \mathcal{S}^v} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^{(v)}}(x_i) = y_i\}, \tag{5}$$

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

2.5 How to Estimate Accuracies?

Estimating $\hat{\mathcal{E}}_{\mathcal{A}}$ requires the following design choices: Should it be cross-validated and how? If cross validating using V-fold CV then how many folds? Should the folding be balanced? If estimation is part of a permutation test: should the data be refolded after each permutation?

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.

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

[4]Mutatis mutandis for continous y.

Cross validate or not? For the purpose of statistical testing, bias in $\hat{\mathcal{E}}_{\mathcal{A}}$ is not a problem, as long as it does not invalidate the error rate guarantees. The underlying intuition is that if the same bias is introduced in all permutations, it will not affect the properties of the permutation test. We will thus be considering both cross validated accuracies, and resubstitution accuracies.

Balanced folding? The standard practice in V-fold CV 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? In V-fold CV, *folding* the data means assigning each observation to one of the V data folds. The standard practice in neuroimaging is to permute labels and refold the data after each permutation. This is done because permuting labels will unbalance the original balanced folding. We will adhere to this practice due to its popularity, even though it is computationally more efficient to permute features[5] instead of labels, as done by Golland et al. [2005].

How many folds? Different authors suggest different rules for the number of folds. We fix the number of folds to $V = 4$, and do dot discuss the effect of V because we will ultimately show that V-fold CV is dominated by other cross-validation procedures, and thus, never recommended.

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

[5]The difference between permuting labels or features is in the mapping to folds. When permuting features, the *label* assignment to folds is fixed. When permuting labels, the *feature* assignment to folds is fixed.

Name	Algorithm	Resampling	Parameters
Oracle	Hotelling	Resubstitution	–
Hotelling	Hotelling	Resubstitution	–
Hotelling.shrink	Hotelling	Resubstitution	–
Goeman	Hotelling	Resubstitution	–
sd	Hotelling	Resubstitution	–
lda.CV.1	LDA	V-fold	–
lda.noCV.1	LDA	Resubstitution	–
svm.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*, *Goeman*, and *sd*. *Oracle* is the same as Hotelling's T^2, only using the generative covariance, and not an estimated one. *Hotelling* is the classical two-group T^2 statistic [Anderson, 2003]. *Hotelling.shrink* is a high dimensional version of T^2, with the regularized covariance from Schäfer and Strimmer [2005]. *Goeman* and *sd* are other high dimensional versions of the T^2, from Goeman et al. [2006] and Srivastava [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.

3 Results

We now compare the power of our various statistics in various configurations. We do so via simulation. The basic simulation setup is presented in Section 3.1. Following sections present variations on the basic setup. The R code for the simulations can be found in `http://www.john-ros.com/permuting_accuracy/`.

3.1 Basic Simulation Setup

Each simulation is based on $1,000$ replications. In each replication, we generate n i.i.d. samples from a shift class

$$\mathbf{x}_i = \mu \mathbf{y}_i + \eta_i, \tag{6}$$

where $\mathbf{y}_i \in \mathcal{Y} = \{0, 1\}$ encodes the class of subject i, μ is a p-dimensional shift vector, the noise η_i is distributed as $\mathcal{N}_p(0, \Sigma)$, the sample size $n = 40$,

and the dimension of the data is $p = 23$. The covariance $\Sigma = I$. In this basic setup, reported in Figure 1, the shift effect is captured by μ. Shifts are equal in all p coordinates of μ. With e being a p-vector of ones, then $\mu := c\,e$. We will use c to index the signal's strength, and vary it over $c \in \{0, 1/4, 1/2\}$. The (squared) Euclidean and Mahalanobis norms of the signal are $\|\mu\|_2^2 = \|\mu\|_\Sigma^2 = c^2 p \approx \{0, 1.4, 5.7\}$. These can be thought as the effect's size.

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 "$x|y = 0$ distributed like $x|y = 1$" null hypothesis if the permutation p-value is smaller than 0.05. The reported power is the proportion of replication where the permutation p-value fell below 0.05.

3.2 False Positive Rate

We start with a sanity check. Theory suggests that all test statistics should control their false positive rate. 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

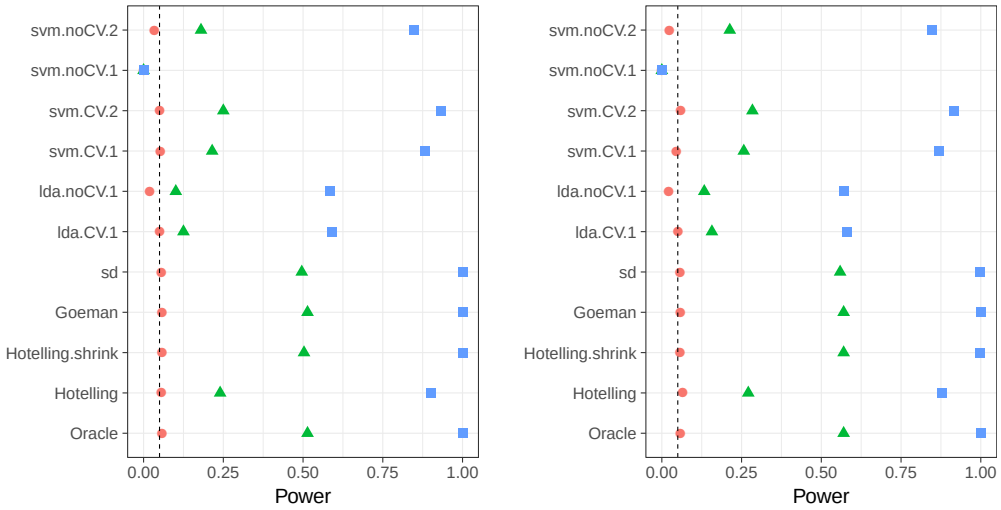
Having established that all of the tests in our battery control the false positive rate, it remains to be seen if they have similar power– especially when comparing location tests to accuracy tests.

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 location tests *Goeman*, *sd* and *Hotelling.shrink* defined in Table 1.

3.4 Tie Breaking

As already stated in the introduction, the accuracy statistic is highly discrete. Especially the resubstitution accuracy tests. Discrete test statistics lose power by not exhausting the permissible false positive rate. A common remedy is a *randomized test*, in which the rejection of the null is decided at

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 $c = 0$ (red circle), $c = 1/4$ (green triangle), and $c = 1/2$ (blue square). Simulation details in Section 3.1. Cross-validation was performed with balanced and unbalanced data folding; see sub-captions.



(a) Unbalanced V-fold CV. **(b) Balanced V-fold CV.**

random in a manner that exhausts the false positive rate. Formally, denoting by $\mathcal{T}$ the observed test statistic, by $\mathcal{T}_\pi$, its value after under permutation π, and by $\mathbb{P}\{A\}$ the proportion of permutations satisfying A then the randomized version of our tests imply that if the permutation p-value, $\mathbb{P}\{\mathcal{T}_\pi \geq \mathcal{T}\}$, is greater than α then we reject the null with probability

$$max\left\{\frac{\alpha - \mathbb{P}\{\mathcal{T}_\pi > \mathcal{T}\}}{\mathbb{P}\{\mathcal{T}_\pi = \mathcal{T}\}}, 0\right\}.$$

Figure 2 reports the same analysis as in Figure 1b, after allowing for random tie breaking. It demonstrates that the power disadvantage of accuracy tests, cannot be remedied by random tie breaking.

3.5 Departure From Gaussianity

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

To check this, we replaced the multivariate Gaussian distribution of η in Eq.(6) with a heavy-tailed multivariate-t distribution. In this heavytailed

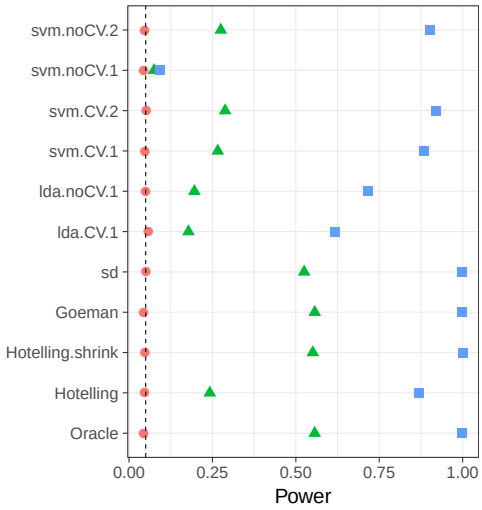


Figure 2: The same as Figure 1b, with random tie breaking.

setup, the dominance of the location tests was preserved, even if less evident than in the Gaussian case (Figure 3).

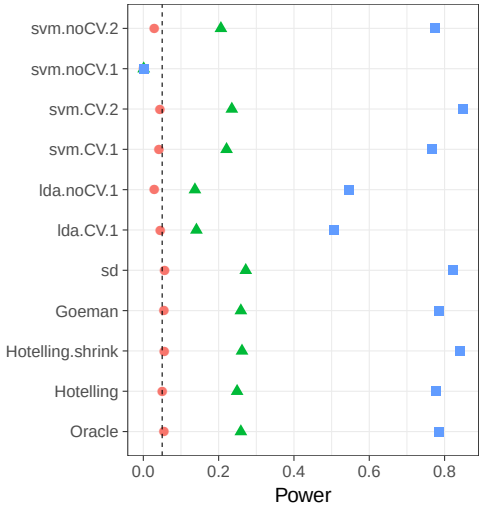


Figure 3: **Heavytailed.** η_i is p-variate t, with $df = 3$.

3.6 Departure from Sphericity

We now test the robustness of our results to the correlations in x. In terms of Eq.(6), Σ will no longer be the identity matrix. Intuitively- both location tests and accuracy tests include the estimation of Σ, so that correlations

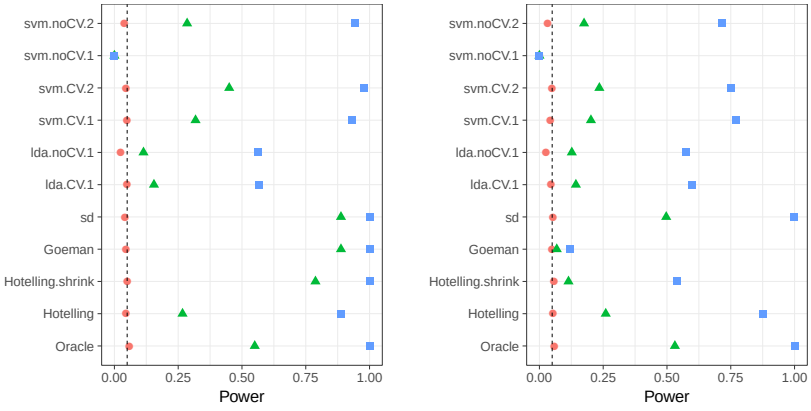
should be accounted for. To keep the comparisons "fair" as the correlations vary, we kept $\|\mu\|_\Sigma := \sqrt{\mu'\Sigma^{-1}\mu}$ fixed.

Which test has more power: accuracy or location? We address this question using various correlation structures. We also vary the direction of the signal, μ, and distinguish between signal in high variance principal component (PC) of Σ, and in the low variance PC.

The simulation results reveal some non trivial phenomena. First, when the signal is in the direction of the high variance PC, the high dimensional location tests are far superior than accuracy tests. This holds true for various correlation structures: the short memory correlations of $AR(1)$ in Figure 4a, the long memory correlations of a Brownian motion in Figure 5a, and the arbitrary correlation in Figure 6a.

When the signal is in the direction of the low variance PC, a different phenomenon appears. There is no clear preference between location or accuracy tests. Instead the non-regularized tests are the clear victors. This holds true for various correlation structures: the short memory correlations of $AR(1)$ in Figure 4b, the long memory correlations of a Brownian motion in Figure 5b, and the arbitrary correlation in Figure 6b. We attribute this phenomenon to the bias introduced by the regularization, which masks the signal. This matter is further discussed in Section 5.3.

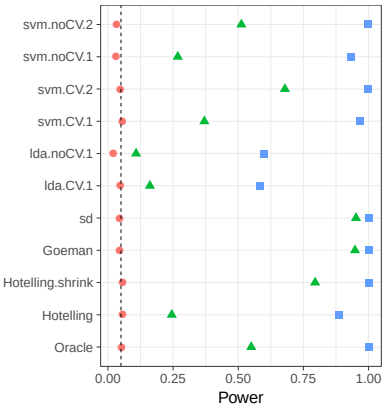
Figure 4: Short memory, AR(1) correlation. $\Sigma_{k,l} = \rho^{|k-l|}; \rho = 0.6$



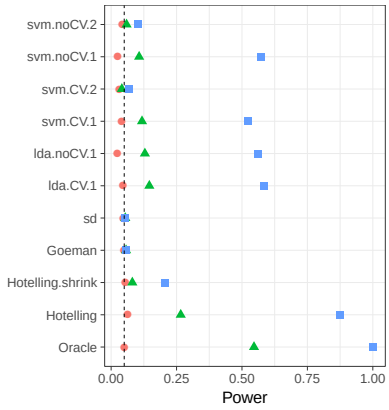
(a) Signal in direction of highest variance PC of Σ.

(b) Signal in direction of lowest variance PC of Σ.

Figure 5: Long-memory Brownian motion correlation: $\Sigma = D^{-1}RD^{-1}$ where D is diagonal with $D_{jj} = \sqrt{R_{jj}}$, and $R_{k,l} = \min\{k, l\}$.



(a) Signal in direction of highest variance PC of Σ.



(b) Signal in direction of lowest lowest variance PC of Σ.

3.7 Departure from Homoskedasticity

Our previous simulations assume variables a have unit variance. The heteroskedastic case, where difference coordinates have different variance, is of lesser importance, since we can typically normalize the variable-wise variance. Some test statistics have built-in variance normalization, and are known as *scalar invariant*. The *sd* test statistic is scalar invariant. Statistics that are not scalar-invariant such as the *Goeman* statistic, will give less importance to high-variance directions than to low-variance directions.

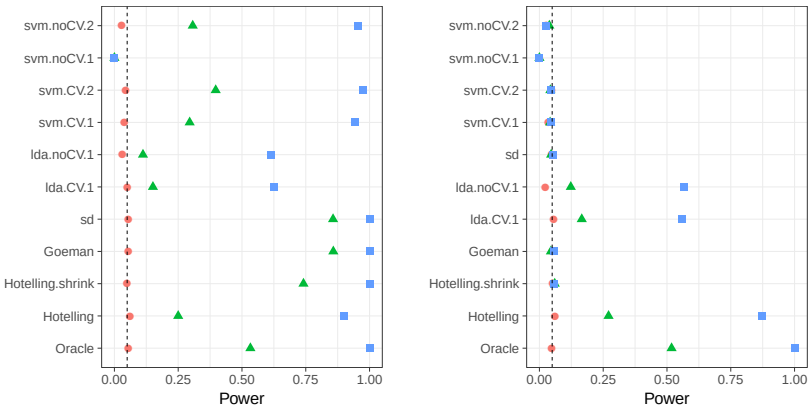
In Figure 7a we see that as before, location tests dominate accuracy tests. For the first time, we can see the difference between the scalar-invariant *sd* and *Goeman*: the latter gaining power by focusing on low variance coordinates. Since the signal's magnitude is the same in all coordinates, *Goeman* gains power by putting emphasis where it is needed.

When the signal is in the low variance PC, *Goeman* puts emphasis on variables which carry little signal. For this reason it has less power than *sd*, as seen in Figure 7b.

3.8 Departure from V-fold CV

Intuition suggests we may alleviate the discretization of the accuracy test statistic by replacing the V-fold CV, and resampling *with replacement*. The discretization of the accuracy statistic is governed by the number of samples in the union of test sets. For V-fold CV, for instance, the accuracy may

Figure 6: Arbitrary Correlation. $\Sigma = D^{-1}RD^{-1}$ where D is diagonal with $D_{jj} = \sqrt{R_{jj}}$, and $R = A'A$ where A is a Gaussian $p \times p$ random matrix with independent $\mathcal{N}(0,1)$ entries.



(a) Signal in direction of highest variance PC of Σ.

(b) Signal in direction of lowest variance PC of Σ.

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 not containing observation i. The leave-one-out bootstrap estimate, $\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO}$, is defined as:

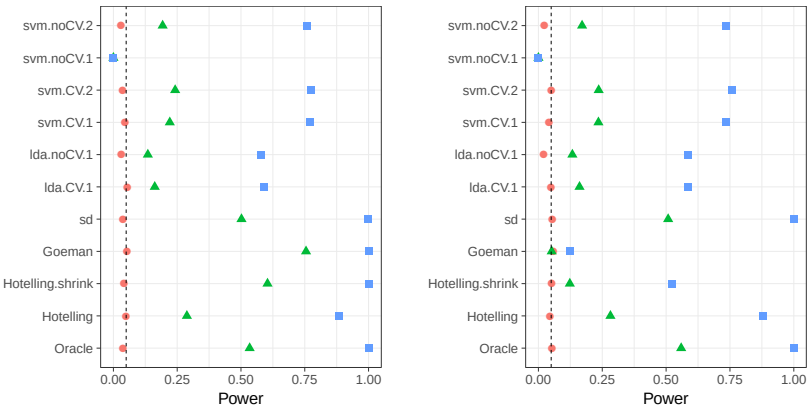
$$\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{(i)}|}\sum_{b\in C^{(i)}}\mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i) = y_i\}. \tag{7}$$

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

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

Simulation results are reported in Figure 8 with naming conventions in Table 2. As expected, selecting test sets with replacement does increase the

Figure 7: Heteroskedasticity: Σ is diagonal with $\Sigma_{jj} = j$.



(a) μ in the high variance PC of Σ.

(b) μ in the low variance PC of Σ.

power of accuracy tests, 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, since more replications reduce the level of discretization.

Name	Algorithm	Resampling	B	Parameters
LDA.Boot.1	LDA	bLOO	10	–
SVM.Boot.1	SVM	bLOO	10	cost=10
SVM.Boot.2	SVM	bLOO	10	cost=0.1
SVM.Boot.3	SVM	bLOO	50	cost=10
SVM.Boot.4	SVM	bLOO	50	cost=0.1

Table 2: The same as Table 1 for bootstraped accuracy estimates. bLOO is defined in 3. B denotes the number of Bootstrap samples.

3.9 The Effect of High Dimension

Our setup of $n = 40$ and $p = 23$ is high dimensional in that p/n is not too small. This surfaces finite samples effects, not manifested in classical $p/n \to 0$ asymptotic analysis. Our best performing tests, *sd*, *Goeman*, and *Hotelling.shrink*, alleviate the dimensionality of the problem by regularizing the estimation of Σ, thus reducing variance at the cost of some bias. It may thus be argued that the power advantages of the location tests are driven by the regularization of the covariance, and not the statistic itself. We would

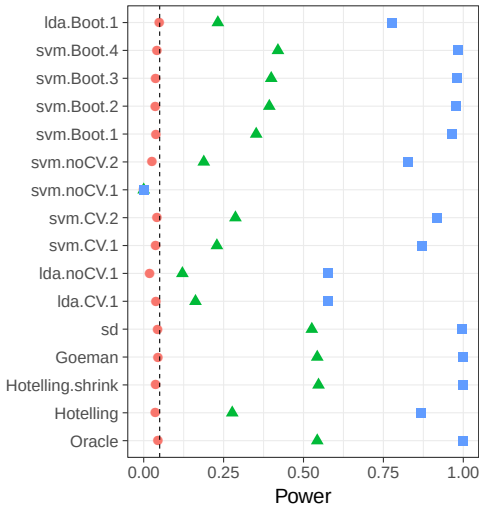


Figure 8: **Bootstrap.** The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix 3.1.

thus augment the comparison with various covariance-regularized accuracy tests. The l_2 regularization in our SVM accuracy test, already regularizes the covariance, but it is certainly not the only way to do so. We thus add some covariance-regularized accuracy tests such as a shrinkage based LDA [Pang et al., 2009, Ramey et al., 2016], where similarly to *Hotteling.shrink*, Tikhonov regularization of $\hat{\Sigma}$ is employed. We also try we try a diagonalized LDA[6] [Dudoit et al., 2002], which regularizes $\hat{\Sigma}$ similarly to *sd* and *Goeman*.

Simulation results are reported in Figure 9 with naming conventions in Table 3. The proper regularization of the covariance of a classifier, just like a location test, can improve power. See, for instance, *svm.CV.6* which is clearly the best regularized SVM for testing. Replacing the V-fold with a bootstrap allows us to further increase the power, as done with *lda.highdim.4*. Even so, the out-of-the-box location tests outperform the accuracy tests.

[6]Known as *Gaussian Naïve Bayes*.

Name	Algorithm	Resampling	Parameters
svm.CV.5	SVM	V-fold	cost=100
svm.CV.6	SVM	V-fold	cost=0.01
lda.highdim.1	LDA	V-fold	–
lda.highdim.2	LDA	V-fold	–
lda.highdim.3	LDA	V-fold	–
lda.highdim.4	LDA	bLOO	B=50

Table 3: The same as Table 1 for regularized (high dimensional) predictors. *svm.CV.5* and *svm.CV.6* are l_2 regularized SVM, with varying regularization penalty. *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 bLOO.

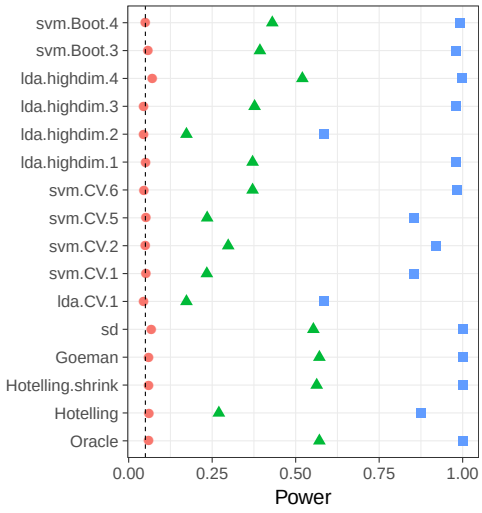


Figure 9: **HighDim Classifier.** The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 3. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Section 3.1.

4 Neuroimaging Example

Figure 10 is an application of both a location and an accuracy test to the neuroimaging 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[7].

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 10. We emphasize that both test statistics were compared with the same permutation scheme, and the same error controls, so that any difference in detections is due to their different power.

5 Discussion

We have set out to understand which of the tests is more powerful: accuracy tests or location tests. Our current observation is that accuracy tests are never optimal. There is always a multivariate test, possibly a location test, that dominates in power. Our advice to the practitioner is that location tests, 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 resampling is required. 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.

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) **Shift Alternatives**: We focused on shift alternative so that location tests are expectedly superior via an NPL type argument.

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

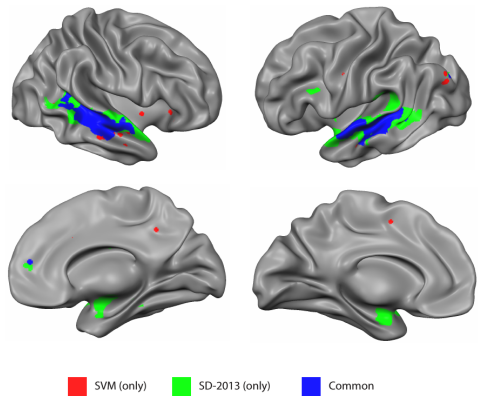


Figure 10: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.CV.1*), and a location test (*sd*). *svm.CV.1* was computed using 5-fold cross validation, and a cost 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].

(b) **Inefficient** use of the data when validating with a holdout set: the variability of the test statistic is governed by the size of the test set, which may be several orders smaller than the whole data.
(c) Inappropriate **regularization** in high SNR regimes: testing requires less regularization than predicting. Intuitively, when the signal is in the location μ, then predicting requires a regularizing bias in the order of $\|\mu\|_2$, while testing requires a bias in the order of $\|\mu\|_\infty$. The gap may be considerable in high-dimension. This phenomenon was also observed by Cheng et al. [2017], which observe that recovering the support of a function requires less regularization than recovering the whole function.
(d) **Discretization**: discussed in the next section.

5.1.1 Discretization

Permutation testing with discrete test statistics are known to be conservative. Firstly, because a Monte-Carlo sample of permutations will always be conservative compared to a full enumeration of permutations [Hemerik and Goeman, 2017]. Secondly, because of the presence of ties. And thirdly, because a highly discrete test-statistic, is insensitive to mild perturbations of the data. For an intuition consider the usage of the *resubstitution accuracy*,

a.k.a. the *train error*, or *empirical risk*, as a test statistic. Resubstitution accuracy is the accuracy of the classifier evaluated on the training set. If data is high dimensional, the resubstitution accuracy will be very high due to over fitting. In a very high dimensional regime, the resubstitution accuracy may be as high as 1 for the observed data [McLachlan, 1976, Theorem 1], but also for any permutation. The concentration of resubstitution accuracy near 1, and its discretization, render this test completely useless, with power tending to 0 for any (fixed) effect size, as the dimension of the model grows.

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[8]. An asymptotic analysis may suggest resubstitution accuracy estimates are good test statistics, while they suffer from very low finite-sample power. The canonical remedy for ties— random tie breaking — showed only a minor improvement (Sec. 3.4).

The matter of discretization was summarized in a 2011 post by by Prof. Frank Harrell in CrossValidated[9]:

> ... 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.

5.2 Interpretation

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

Some authors prefer accuracy tests because they can be seen as effect-size estimates, invariant to the sample size. This is true, but the multivariate-statistics literature provides many multivariate effect-size estimators, that generalize Cohen's d. Examples can be found, for instance, in Stevens [2012] and references therein.

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

[9] A Q&A website for statistical questions: `http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier`. But also in "Problems Caused by Categorizing Continuous Variables": `http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous`

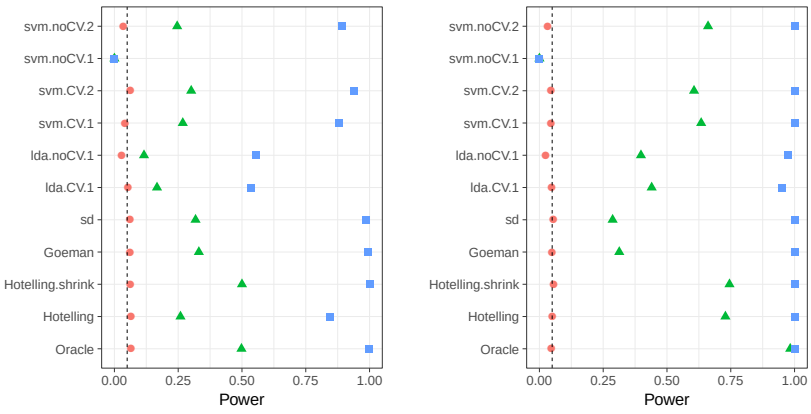
5.3 Fixed SNR

For a fair comparison between simulations, in particular between those with different Σ, we needed to fix the difficulty of the problem. We defined "a fair comparison" to be such that a maximal power test would have the same power, justifying our choice of fixing the Mahalanobis norm of μ. Formally, in all our simulations we set $\|\mu\|_\Sigma^2 = c^2 p$.

Our choice implies that the Euclidean norm of μ varies with the covariance, and with the direction of the signal. An initial intuition may suggest that detecting signal in the low variance PCs is easier than in the high variance PCs. This is true when fixing $\|\mu\|_2$, but not when fixing $\|\mu\|_\Sigma$.

For completeness, Figure 11 reports the power analysis under $AR(1)$ correlations, but with $\|\mu\|_2$ fixed instead of $\|\mu\|_\Sigma$. We compare the power of a shift in the direction of some high variance PC (Figure 11a), versus a shift in the direction of a low variance PC (Figure 11b). The intuition that it is easier to detect signal in the low variance directions is confirmed. It is also consistent with Figure 4, in the following aspects: (i)*Hotelling.shrink* is a good performed "on average", (ii) *sd* and *Goeman* have the best power to detect signal in the noisiest directions, but low power for signal in the noiseless directions.

Figure 11: Short memory, AR(1) correlation. $\|\mu\|_2$ fixed.



(a) μ in PC7 of Σ. **(b)** μ in PC15 of Σ.

5.4 Detecting Signal in Different Directions

Figures 4, 5, 6 and 11, demonstrate that detecting signal in the direction of the high variance PCs is very different than detecting in the low variance PCs. Why is that?

We attribute this phenomenon to regularization. While the signal, μ varies in direction, the regularization of $\hat{\Sigma}$ does not. The various regularization methods deflate the high variance directions, thus, relatively inflate the low variance directions. If the signal is in the low variance directions, the regularization may mask it. This is what we see in figures 4b, 5b, and 6b: the unregularized tests have more power than the regularized.

5.5 Implications to Other Problems

Our work studies signal detection in the two-group multivariate testing framework, i.e., MANOVA framework. The same problem can be cast in the univariate generalized linear models framework, and in particular, as a Brenoulli Regression problem. If any of the predictors, x, carries any signal, then $x|y = 0$ has a different distribution than $x|y = 1$. This view is the one adopted Goeman et al. [2006].

Another related problem is that of multinomial-regression, i.e., multi-class classification. We conjecture that power differences in favor of location tests versus accuracy tests will increase as the number of classes increases.

5.6 Testing in Augmented Spaces

It may be argued that only accuracy tests permits the separation between classes in augmented spaces, such as in *reproducing kernel Hilbert spaces* (RKHS) by using non-linear predictors. 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 with kernels see for example Székely and Rizzo [2009] or Gretton et al. [2012].

5.7 A Good Accuracy 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 predictor can extract it. In these cases an accuracy test cannot be replaced by a location, or other, statistical test. For the cases an accuracy test cannot be replaced with other tests, we collect the following observations.

Sample size. The conservativeness of accuracy tests, due to discretization, decrease with sample size.

Regularize. Regularization proves crucial to detection power in low SNR regimes, such as when n is in the order of p, or under 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. Particularly, in the possibility of regularizing in directions orthogonal to μ.

Smooth accuracy. Smooth accuracy estimate by cross validating with replacement. The bLOO estimator, in particular, is preferable over V-fold. This was also observed by Yu et al. [2007], albeit attributed to the stability of the accuracy estimate, and not to its smoothness.

Resubstitution accuracy in high SNR. Resubstitution accuracy is useful in high SNR regimes, such as $n \gg p$, because it avoids cross validation without compromising power. In low SNR, the power loss is considerable. 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 high SNR, the concentration is less impactful, and the computational burden of cross validation can be avoided by using the resubstitution accuracy.

5.8 Related Literature

We now review some related accuracy-testing literature, with an emphasis on neuroimaging applications. Ojala and Garriga [2010] study the power of two accuracy tests differing in their 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 that the independence test of Olivetti et al. [2012] relies on a discrete test statistic. It may thus be improved by regularizing and resampling with replacement.

Schreiber and Krekelberg [2013] used null simulations to study the statistical properties of linear SVM's for signal detection, and in particular, false positive rates. They did not study the matter of power. They recommended to test the significance of accuracy estimates using permutation testing instead of parametric t-tests, or binomial tests. They recommend so due to the correlations between data folds in V-fold CV. The authors were also concerned with temporal correlations, which biases accuracy estimates even if cross validated. Bias in accuracy estimates is of great concern when studying a classifier, but it is of lesser concern when using the accuracy merely for localization. Their recommendations differ from ours: they recommend to ensure independent data foldings in V-fold CV, whereas we claim discretization is the real concern, and thus recommend bLOO.

Golland and Fischl [2003] and Golland et al. [2005] study accuracy tests using simulation, neuroimaging data, genetic data, and analytically. The finite Vapnik–Chervonenkis dimension requirement [Golland et al., 2005, Sec 4.3] implies a the problem is low dimensional and prevents the permutation p-value from (asymptotically) concentrating near 1. They 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. We attribute this to the reduced discretization of the accuracy statistic.

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 (with some abuse of notation)

$$\begin{aligned}(x_i|y_i=1) &\sim \pi\mathcal{N}(\mu_1, I) + (1-\pi)\mathcal{N}(\mu_2, I)\,,\\ (x_i|y_i=0) &\sim (1-\pi)\mathcal{N}(\mu_1, I) + \pi\mathcal{N}(\mu_2, I)\,.\end{aligned}$$

Varying π interpolates between the null distribution ($\pi = 0.5$) and a location shift model ($\pi = 0$). We now perform the same simulation as Golland et al. [2005], and in the same dimensionality as our previous simulations. We reparameterize so that $\pi = 0$ corresponds to the null model:

$$\begin{aligned}(x_i|y_i=1) &\sim (1/2-\pi)\mathcal{N}(\mu_1, I) + (1/2+\pi)\mathcal{N}(\mu_2, I)\,,\\ (x_i|y_i=0) &\sim (1/2+\pi)\mathcal{N}(\mu_1, I) + (1/2-\pi)\mathcal{N}(\mu_2, I)\,.\end{aligned} \tag{9}$$

From Figure 12, we see that also for the mixture class of Golland et al. [2005] locations tests are to be preferred over accuracy tests.

5.9 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

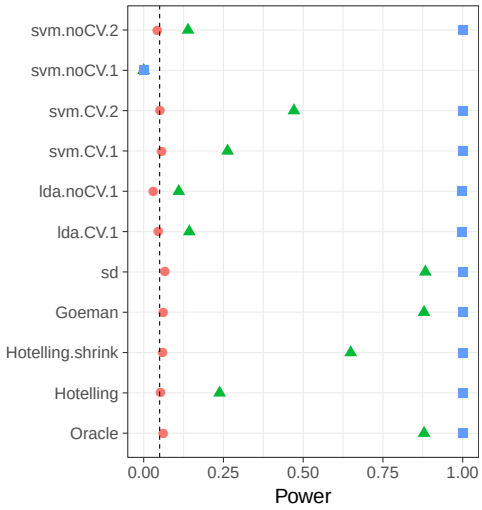


Figure 12: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(9). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

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.

Acknowledgments

JDR was supported by the ISF 900/60 research grant. JDR also wishes to thank, Jesse B.A. Hemerik, Yakir Brechenko, Omer Shamir, Joshua Vogelstein, Gilles Blanchard, and Jason Stein for their valuable inputs.

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.

Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329, 1996.

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.

P. J. Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

D. Cheng, A. Schwartzman, et al. Multiple testing of local maxima for detection of peaks in random fields. *The Annals of Statistics*, 45(2):529–556, 2017.

I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of machine learning research*, 3(Mar):1265–1287, 2003.

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.

M. Eric, F. R. Bach, and Z. Harchaoui. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2008.

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

J. H. Friedman. On multivariate goodness of fit and two sample testing. *eConf*, 30908(SLAC-PUB-10325):311–313, 2003.

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.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

J. Hemerik and J. Goeman. Exact testing with random permutations. *TEST*, Nov 2017. ISSN 1863-8260. doi: 10.1007/s11749-017-0571-1. URL `https://doi.org/10.1007/s11749-017-0571-1`.

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.

M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

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.

P. J. Moreno, P. P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, pages 1385–1392, 2004.

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.

T. Nishiyama, M. Hyodo, T. Seo, and T. Pavlenko. Testing linear hypotheses of mean vectors for high-dimension data with unequal covariance matrices. *Journal of Statistical Planning and Inference*, 143(11):1898–1911, 2013.

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, D. Benozzo, S. M. Kia, M. Ellero, and T. Hartmann. The kernel two-sample test vs. brain decoding. In *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*, pages 128–131. IEEE, 2013.

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.

F. Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *Advances in neural information processing systems*, pages 1257–1264, 2009.

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.

K. Schreiber and B. Krekelberg. The statistical analysis of multi-voxel patterns in functional imaging. *PLoS One*, 8(7):e69328, 2013.

Y. Shen and Z. Lin. An adaptive test for the mean vector in large-p-small-n problems. *Computational Statistics & Data Analysis*, 89:25–38, 2015.

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. On testing the equality of mean vectors in high dimension. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1): 31–56, June 2013. ISSN 2228-4699. doi: 10.12697/ACUTM.2013.17.03.

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.

J. P. Stevens. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

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.

M. Thulin. A high-dimensional two-sample test for the mean using random subspaces. *Computational Statistics & Data Analysis*, 74:26–38, 2014.

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.

G. Xu, L. Lin, P. Wei, and W. Pan. An adaptive two-sample test for high-dimensional means. *Biometrika*, 103(3):609–624, 2016.

K. Yu, R. Martin, N. Rothman, T. Zheng, and Q. Lan. Two-sample comparison based on prediction error, with applications to candidate gene association studies. *Annals of human genetics*, 71(1):107–118, 2007.

J. Zhang and M. Pan. A high-dimension two-sample test for the mean using cluster subspaces. *Computational Statistics & Data Analysis*, 97:87–97, 2016.