Better-Than-Chance Classification for Signal Detection

Jonathan D. Rosenblatt, Yuval Benjamini, Roee Gilron, Roy Mukamel, and Jelle J. Goeman,

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 the 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.

Index Terms

signal-detection, multivariate-testing, supervised-learning, hypothesis-testing, high-dimension.

I. Introduction

Many neuroscientists and geneticists detect signal by fitting a classifier and testing whether its prediction accuracy is better than chance. The workflow consists of fitting a classifier, estimating its predictive accuracy using cross validation, and testing the hypothesis that this accuracy can be attributed to chance alone. This general idea has been promoted in the statistical literature [1], and separately by in the machine-learning literature [2], [3], [4]. Examples in the genetics literature include [5], [6], [7], [8], [9], [10], [11]. Other examples include speaker verification [12], text classification [13], [4], distinguishing between facial expressions [4], data integration [12], record linkage in databases systems [12], [14], [15], optical character recognition [16], multimedia [17], and functional data analysis [14].

JDR was supported by the ISF 900/60 and 924/16 research grants.

JDR is with with the Dept. of IE&M at Ben-Gurion University of the Negev (johnros@bgu.ac.il). YB is with with the Dept. Statistics at the Hebrew University of Jerusalem. RG is with Starr Lab at the UCSF. RM is with the Psychology Dept. at Tel Aviv University. JJG is with the Leiden University Medical Center,

Manuscript received April 19, 2005; revised August 26, 2015.

Examples in the neuroscientific literature, which is our motivating use-case, include [18], [19], [20], [21], [22], and especially the recently popularized *multivariate pattern analysis* (MVPA) framework in [23].

To fix ideas, we will adhere to a concrete example. In [24], Gilron et al. 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 region'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 prediction accuracy as a test statistic.

This same signal detection task can also be approached as a multivariate *two-group* 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 by Pereira et al. [19]:

> ... 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. Multivariate two-groups hypothesis-tests may be divided into tests for equality of location (i.e. means), and two-sample goodness of fit tests (equality of the two whole distribution, GOF in short). The former generalizing the t-test, and the latter (roughly) generalizing Kolmogorov-Smirnov's test.

Crucially for our applications, we will assume that the number of samples is in the order of the dimension of each sample, if not smaller. 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 motivating example it means that the size of the brain's region of interest is large compared to the number of subjects in the experiment. It is thus a *high-dim–small-sample* problem.

In a seminal contribution, Bai and Saranadasa [25] 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. These can be classified along the following lines:

- **High-dim GOF**: Tests that seek for any difference between two distributions, such as [14], [26], [12].
- **High-dim location test for sparse shift**: Tests the seek for a sparse shift in mean vectors such as [27], [28].
- **High-dim location test for dense shift**: Tests the seek for a dense shift in mean vectors such as [29], [25], [30], [31], [32], [33], [34], [35], [36], [37], [38].
- **High-dim location test for shift with unknown sparsity**: Tests the seek for a shift in mean vectors, but adapt to the unknown sparsity, such as [39], [40], [41].

At this point, it becomes unclear which test is preferable, in particular for genetics and neuroimaging? In this manuscript, we do not provide a full answer to the matter. Instead, we merely seek to demonstrate that **accuracy-tests are never optimal, compared to high-dim two-group tests**. Out recommendations to the practitioner in these high-dim problems: (i) Prefer a two-group test over an accuracy-test. (ii) Appropriate regularization is crucial.

Various authors have compared accuracy-tests to two-group tests, often with contradicting conclusions. In [11] for instance, authors find that an accuracy-test based on a tree predictor is preferable over a two-group test. Their

simulated shift is sparse, so that it is no surprise that a tree-type predictor outperforms linear predictors and tests. Authors of [21] compare the kernel test of Gretton et al. [12] to an accuracy-test based on logistic-regression. Their results are inconclusive with a slight advantage to the logistic regression. In [4], authors compare several accuracy-tests to several two-group tests and conclude that an accuracy-test based on a neural-net 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 (even if large-sample), and such feature learning is typically impossible in high-dimension.

Ramdas et al. [42] currently offer the only analytic comparison; 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, [42] conclude LDA and T^2 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 [43]. ARE is the limiting ratio of the sample sizes required by two statistics to achieve similar power. The authors of [42] derive the asymptotic power functions of the two test statistics, with which we are able to compute the ARE between Hotelling's T^2 (two-group) test and Fisher's LDA (accuracy) test. Theorem 14.7 of [43] relates asymptotic power functions to ARE. Using this theorem and the results of [42] we deduce that the ARE is lower bounded by $2\pi \approx 6.3$.[TODO: verify] This means that Fisher's LDA requires at least 6.3 times more samples to achieve the same (asymptotic) power as the T^2 test. In this light, the accuracy-test is remarkably inefficient. For comparison, the t-test is only 1.04 more (asymptotically) efficient than Wilcoxon's rank-sum test [44], so that an ARE of 6.3 is strong evidence in favor of the two-group test.

The analysis in [42] is asymptotic. Since typical sample sizes in neuroscience and genetics are not large, we seek to study which test is to be preferred in finite samples, and not only asymptotically. Lacking a unifying mathematical framework for the finite sample power analysis, we opt for a simulation study.

We start with formalizing the problem in Section II. The main findings are reported in Sections III, and IV. We conclude with a discussion.

II. Problem setup

A. 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 $\mathcal{X} = \mathbb{R}^{27}$.

We denote a sample from x given y with x_y. We denote the distribution of x_1 by $\mathcal{F}$ and x_0 with $\mathcal{G}$. Denoting a dataset by $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n$, a two-group test amounts to testing whether $\mathcal{F} = \mathcal{G}$. For example, we can test whether multivariate voxel activation patterns are similarly distributed when given a vocal stimulus (x_1) or a non-vocal one (x_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 $\mathcal{F} \neq \mathcal{G}$.

B. From a Test Statistic to a Permutation Test

The multivariate tests we will be considering rely on fixing some test statistic, $\mathcal{T}$, 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 in our problems of interest central limit approximations are typically poor.

Because we focus on two-group testing under an independent sampling assumption, we know that a label-switching permutation test is valid. 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: $\mathbb{P}\{\mathcal{T}_\pi \geq \mathcal{T}\} := \frac{1}{R+1}\sum_\pi I\{\mathcal{T}_\pi \geq \mathcal{T}\}$.

(f) Declare $\mathcal{F} \neq \mathcal{G}$ 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.

C. Two-Group Tests

The most prevalent interpretation of $\mathcal{F} \neq \mathcal{G}$ is to assume they differ in means[1]. Difference in means leads to the *shift class* of alternatives, which is by far the most studied class in the statistical literature. In his seminal work in 1931, Harold Hotelling proposed the T^2 test as a straightforward generalization of the t-test, for testing the equality in means of two multivariate distributions [45]. Hotelling's statistic was later shown to be the generalized-likelihood-ratio statistic in the Gaussian shift class. It can also be thought of as the empirical Mahalanobis norm of the mean difference, or the empirical Kullback–Leibler divergence between the distribution of averages from two shifted Gaussian distributions. For more background see, for example, [46].

The major difficulty with the T^2 statistic is that it requires estimating a covariance matrix, thus introducing $p(p+1)/2 = \mathcal{O}(p^2)$ unknown parameters. If n is not much larger than p, or in low signal-to-noise (SNR), the test is very low powered, as shown in [25]. 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. Examples of high-dim tests for (dense) shifts include [29], [25], [30], [31], [47], [48], [33], [49], [35], [50].

If $\mathbb{E}(x_1)$ differs from $\mathbb{E}(x_0)$ in a small number of coordinates we say the *signal is sparse*. Examples of high-dim tests statistics for sparse shifts include [51] and [28].

It is possible that the practitioner is unaware of the amount of sparsity in the signal. Some high-dim test statistics that *adapt* to the level of (unknown) sparsity include [52], [39], [40], [36], [41].

[1]This is not a logical equivalence, but rather a prevalent convention. The Behrnes-Fisher problem is a counter example where equal means do not imply equal distributions.

It is possible that the signal is present not (only) in means. We would thus opt for a two-group GOF test, instead of a location test. Examples of multivariate GOF tests include [53], [54], [14], [55], [56], [57], [2], [16], [58], [12].

As previously mentioned, a classifier's accuracy may also be used as a test statistic. We now explain how an accuracy-test is constructed.

D. Prediction Accuracy as a Test Statistic

An accuracy-test amounts to using a predictor's accuracy as a test statistic. A predictor[2], $\mathcal{A}_\mathcal{S} : \mathcal{X} \to \mathcal{Y}$, is the output of a learning algorithm $\mathcal{A}$ when applied to the dataset $\mathcal{S}$. The accuracy of predictor[3], $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$, is defined as the probability of $\mathcal{A}_\mathcal{S}$ making a correct prediction. The accuracy of a learning algorithm[4], $\mathcal{E}_\mathcal{A}$, 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 $\mathcal{E}_{\mathcal{A}_\mathcal{S}} := \int_{(x,y)} \mathcal{I}\{\mathcal{A}_\mathcal{S}(x) = y\}\ d\mathcal{P}$, and $\mathcal{E}_\mathcal{A} := \int_\mathcal{S} \mathcal{E}_{\mathcal{A}_\mathcal{S}}\ d\mathcal{P}_\mathcal{S}$, where $\mathcal{I}\{A\}$ is the indicator function[5] 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 accuracy*[6], and the V-fold Cross Validation (CV) estimate.

Definition 1 (Resubstitution accuracy)**.** The resubstitution accuracy estimator of a learning algorithm $\mathcal{A}$, denoted $\hat{\mathcal{E}}^{Resub}_\mathcal{A}$, is defined as $\hat{\mathcal{E}}^{Resub}_\mathcal{A} := \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\{\mathcal{A}_\mathcal{S}(x_i) = y_i\}$.

Definition 2 (V-fold CV accuracy)**.** 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 $\hat{\mathcal{E}}^{Vfold}_\mathcal{A} := \frac{1}{V}\sum_{v=1}^{V} \frac{1}{|\mathcal{S}^v|}\sum_{i\in\mathcal{S}^v} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^{(v)}}(x_i) = y_i\}$, where $|A|$ denotes the cardinality of a set A.

E. 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.

a) 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 unbiased cross validated accuracies, and biased resubstitution accuracies.

[2]Also known as a *hypothesis* in the machine learning literature.

[3]Also known as (the complement of) the *test error*.

[4]Also known as (the complement of) the *expected test error*.

[5]Mutatis mutandis for continuous y.

[6]Also known as (the complement of) the *train-error*. We use *resubstitution* in accordance with Arlot et Celisse [59], and to emphasize the data has not been split to train-set and test-set.

b) Balanced folding: The standard practice in V-fold CV is to constrain the data folds to be balanced, i.e. stratified [3, for e.g.]. This means that each fold has the same number of examples from each class. We will report results only with balanced folding, mostly because we will conclude that V-fold CV should not be used for our detection problem.

c) 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[7] instead of labels, as done by [60].

d) 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 I collects an initial battery of tests we will be comparing. We selected the accuracy tests based on their popularity in the literature. We selected two-group tests based on their popularity, and so that various types of test statistics are represented: tests for dense and sparse shifts, and GOF tests.

Name	Algorithm	Resampling	Remark
✣svm.CV.cCV	SVM	V-fold	cost=CV
✣svm.noCV.c001	SVM	Resubstitution	cost=0.01
✣svm.noCV.c100	SVM	Resubstitution	cost=100
✣svm.CV.c001	SVM	V-fold	cost=0.01
✣svm.CV.c100	SVM	V-fold	cost=100
✣lda.noCV.1	LDA	Resubstitution	–
✣lda.CV.1	LDA	V-fold	–
Cai	[51]	Resubstitution	–
Simes	[52]	Resubstitution	–
dCOV	[55]	Resubstitution	–
Gretton	[12]	Resubstitution	–
Srivastava	[47]	Resubstitution	–
Goeman	[31]	Resubstitution	–
Schafer	[30]	Resubstitution	–
Hotelling	T^2	Resubstitution	–
Oracle	T^2	Resubstitution	Known Σ

TABLE I: This table collects the various test statistics we will be studying. Two-group tests for dense shifts include: *Oracle*, *Hotelling*, *Schafer*, *Goeman*, and *Srivastava*. Two-group tests for sparse shifts include *Cai*. Two-group adaptive tests for shifts include *Simes*. The rest are accuracy-tests, marked with a ✣, and details given in the table. For example, *svm.CV.c100* is a linear SVM, with V-fold cross validated accuracy, and cost parameter set at 100 [61]. *svm.CV.cCV* is a linear SVM, with V-fold CV accuracy, and cost parameter optimized with (an inner) CV. *lda.noCV.1* is Fisher's LDA, with a resubstituted accuracy estimate. Also recall that in LIBSVM, the *cost* is inversely proportional to the regularization [62]: larger cost implies less regularization.

[7]The difference between permuting labels, $\pi(y)$, or features, $\pi(X)$, 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.

III. 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 III-A. 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/.

A. Basic Simulation Setup and Notation

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

$$\mathbf{x}_i = \mu \mathbf{y}_i + \eta_i, \tag{1}$$

where $\mathbf{y}_i \in \mathcal{Y} = \{0, 1\}$ encodes the class of observation 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 dense and equal in all p coordinates of μ. We set $\mu := c\, e$ where e is a p-vector of ones. We will use c to index the signal's strength, and vary it over $c \in \{0, 1/4, 1/2\}$. With $\Sigma = I$ then the (squared) Euclidean and Mahalanobis norms of the signal are $\|\mu\|_2^2 = \|\mu\|_\Theta^2 = \mu' \Sigma^{-1} \mu = c^2 p \approx \{0, 1.4, 5.7\}$, where $\Theta := \Sigma^{-1}$. These can be thought as the effect's size.

Having generated the data, we compute each of the test statistics in Table I. 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 $\mathcal{F} = \mathcal{G}$ 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.

B. 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 $\mathcal{F} = \mathcal{G}$, 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.

C. Power

From Figure (1) we learn that in our simulation setup, two-group tests are more powerful than accuracy-tests. This is most notable for the intermediate signal strength (green triangles).

D. Large Sample

We focus on high-dim–small-sample configurations because of our motivation in neuroimaging and genetics. Our results, however, hold also in high-dim–large-sample configurations. To prove this point, we fix p/n at $23/40$, and set $n = 4,000, p = 2,300$. The results, reported in Figure 2, are qualitatively similar to the high-dim–small-sample of Figure 1.

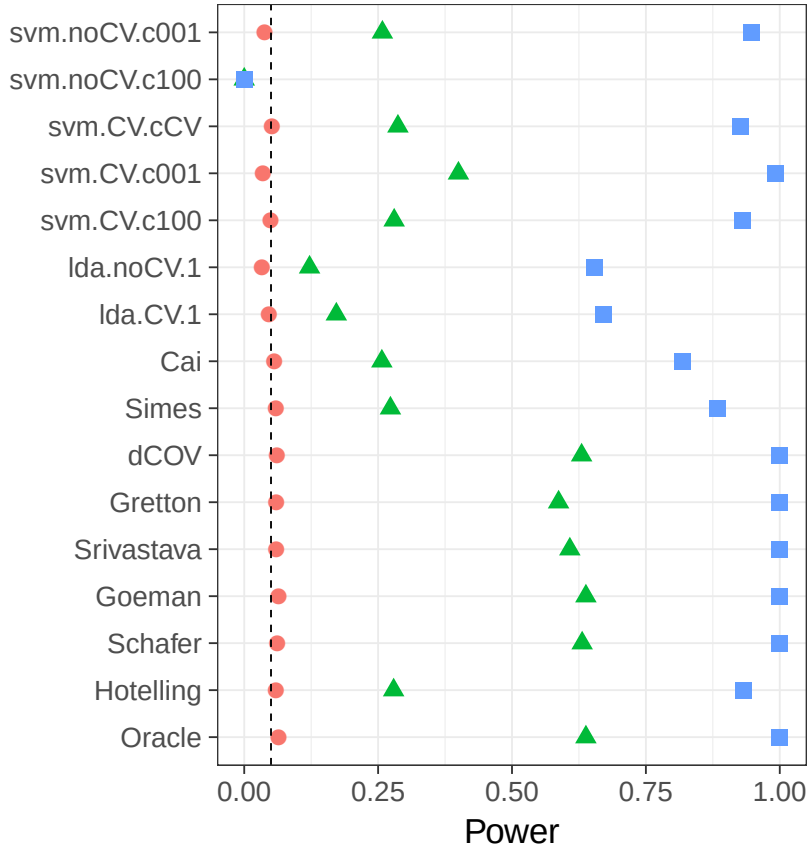


Fig. 1: The power of the permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. Effects vary over $c = 0$ (red circle), $c = 1/4$ (green triangle), and $c = 1/2$ (blue square). The various statistics on the y axis. Their details are given in Table I. Simulation details in Section III-A.

E. Departure From Gaussianity

The Neyman-Pearson Lemma (NPL) type reasoning that favors two-group location-tests over accuracy-tests in our simulations may fail when the data is not multivariate Gaussian. This is because Hotelling's T^2 statistic is no longer a generalized-likelihood-ratio test. To check this, we replaced the multivariate Gaussian distribution of η in Eq.(1) with a heavy-tailed multivariate-t distribution with 3 degrees of freedom. In this heavytailed setup, the dominance of the two-group tests was preserved, even if less evident than in the Gaussian case (Figure 3).

F. Departure from Sphericity

We now test the robustness of our results to the correlations in x. In terms of Eq.(1), Σ will no longer be the identity matrix. Some tests try to account for Σ by estimating it. Estimating Σ reduces possible bias at the cost of some variance. We thus do not know if the conclusions from the uncorrelated case (Fig. 1) repeat themselves in the presence of correlation.

We simulate using 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. To keep the comparisons fair

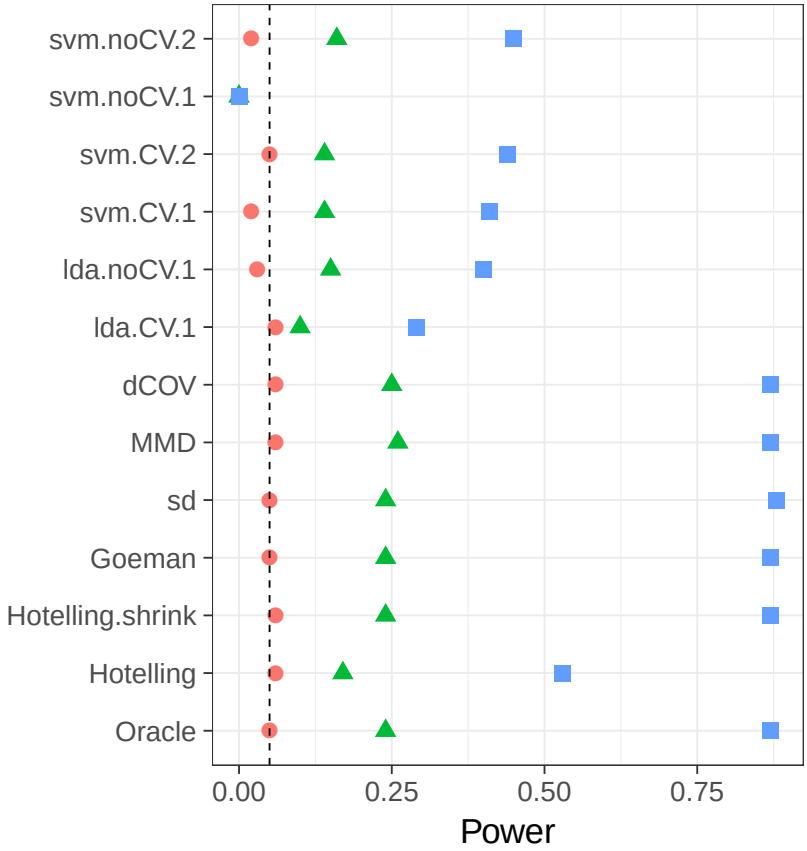


Fig. 2: [TODO: update figure] The same as Figure 1 with $n = 4,000; p = 2,300$.

as the correlations vary, we kept $\|\mu\|_\Theta := \sqrt{\mu'\Theta\mu}$ fixed. This matter is discussed in Section V-C.

The simulation results reveal some non trivial phenomena. First, when the signal is in the direction of the high variance PC, the high-dim two-group 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 two-group 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 V-C.

G. Departure from Homoskedasticity and Scalar Invariance

Our previous simulations assume variables have unit variance. Practitioners are already accustomed to z-score features before learning a regularized predictor (e.g. ridge regression) so this is not an unrealistic setup. Implicit z-scoring is sometime an integral part of a test statistic. This is known as *scalar invariance*. The *Srivastava* statistic,

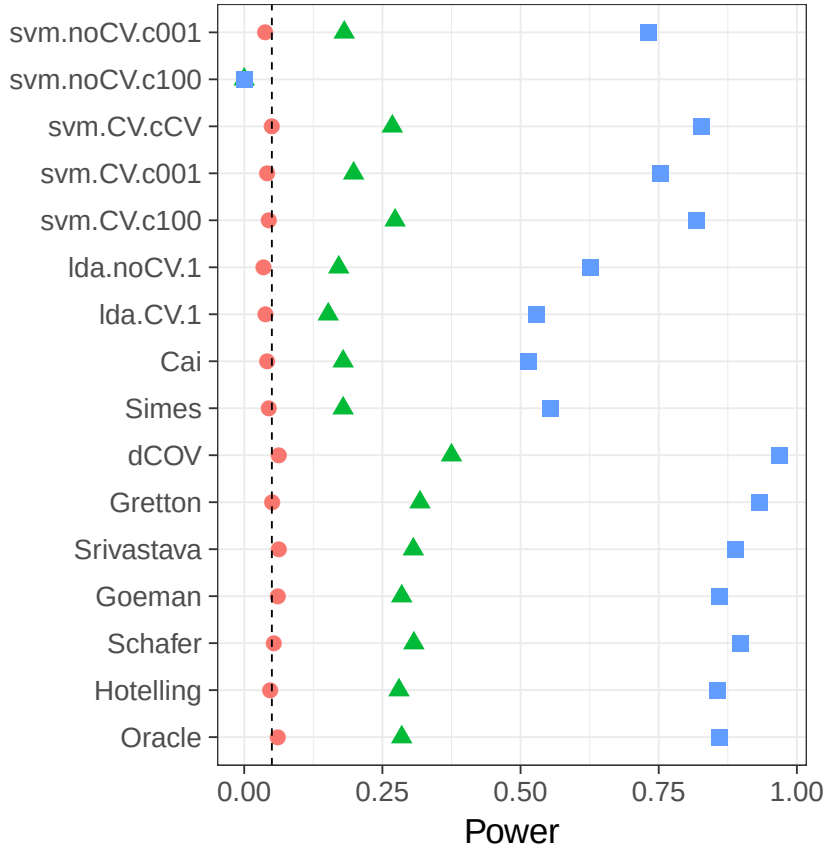


Fig. 3: **Heavytailed.** η_i is p-variate t, with $df = 3$.

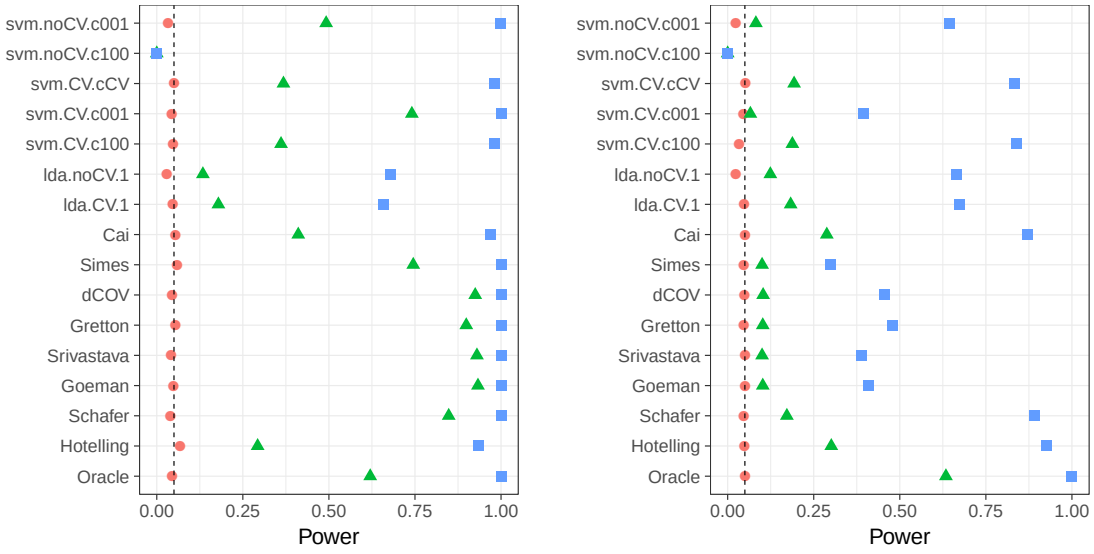
for instance, is scalar invariant. It can be (roughly) thought of as the l_2 norm of the p-vector of coordinate-wise t-statistics. The *Goeman* statistic, for instance, is not scalar invariant. It can be (roughly) thought of as the l_2 norm of the p-vector of variable-wise mean differences. Under heteroskedasticity, the *Goeman* statistic will give less importance to signal in the high-variance directions than signal in the low-variance directions. *Srivastava* will give all coordinates the same importance.

In Figure 7a we can see the difference between the scalar-invariant *Srivastava* and *Goeman* statistics. We also see that two-group tests dominate accuracy-tests also in the heteroskedastic case.

H. Departure from V-fold CV

The more a test statistic is discretized, the more it's power diminishes. This can be shown by observing that a classification is a quantization of a continuous prediction. The information processing inequality thus applies [63]. Because testing power is monotone in the mutual information, a non-quantized test statistic will have no less power than a quantized one. [TODO: verify]

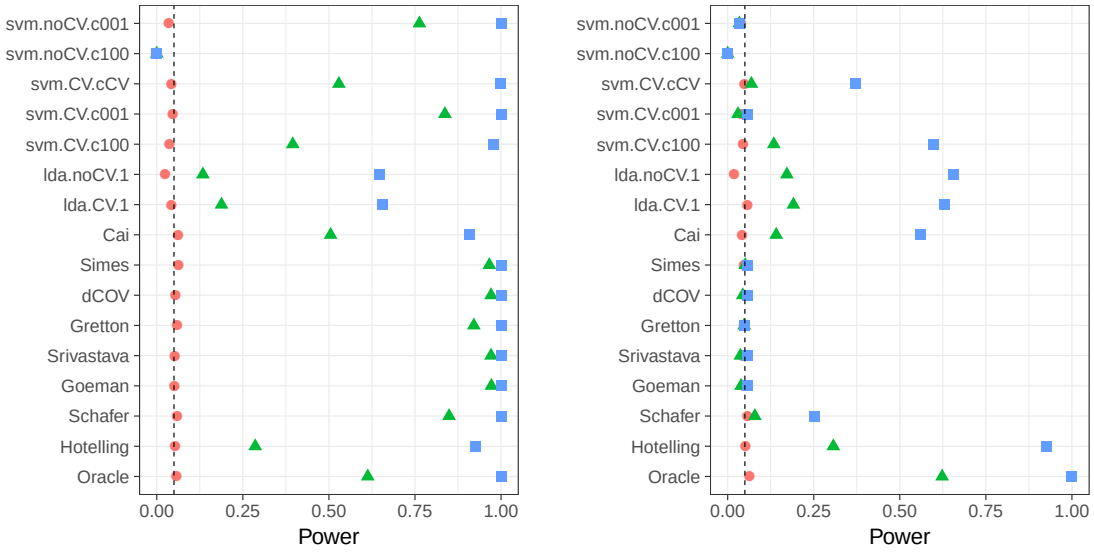
In V-fold CV, the discretization of the accuracy statistic is governed by the number of samples. This is the case whenever resampling without replacement. Intuition suggests we may alleviate the discretization of the accuracy statistic by replacing the V-fold CV, and resampling *with replacement*. An algorithm that samples test sets with



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

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

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



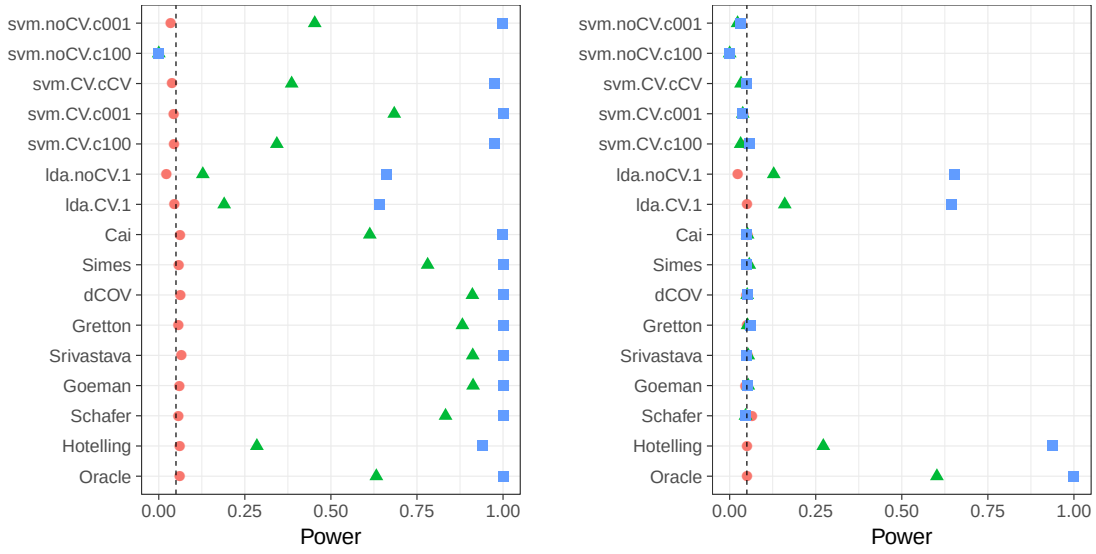
(a) Signal in direction of highest variance PC of Σ.

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

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

replacement is the *leave-one-out bootstrap estimator*, and its derivatives, such as the *0.632 bootstrap*, and *0.632+ bootstrap* [64, Sec 7.11].

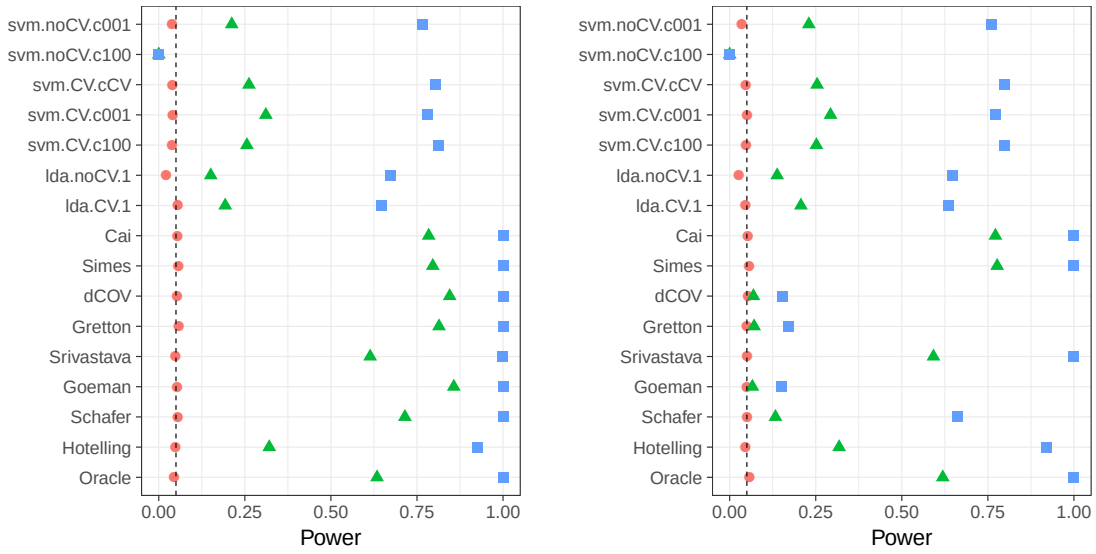
Definition 3 (bLOO)**.** The *leave-one-out bootstrap* estimate, bLOO, is the average accuracy of the holdout observations, over all bootstrap samples. Denote by $\mathcal{S}^b$, a bootstrap sample b of size n, sampled with replacement from



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

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

Fig. 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) μ in the high variance PC of Σ.

(b) μ in the low variance PC of Σ.

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

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

Simulation results are reported in Figure 8 with naming conventions in Table II. As expected, sampling test sets with replacement does increase the power of accuracy-tests, when compared to V-fold cross validation, but still falls short from the power of two-group 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	Remark
✣lda.Boot.b10	LDA	bLOO	10	–
✣svm.Boot.c001.b50	SVM	bLOO	10	cost=0.01
✣svm.Boot.c100.b50	SVM	bLOO	10	cost=100
✣svm.Boot.c001.b10	SVM	bLOO	50	cost=0.01
✣svm.Boot.c100.b10	SVM	bLOO	50	cost=100

TABLE II: The same as Table I for bootstrapped accuracy estimates. bLOO is defined in 3. B denotes the number of Bootstrap samples. Accuracy-tests marked with a ✣.

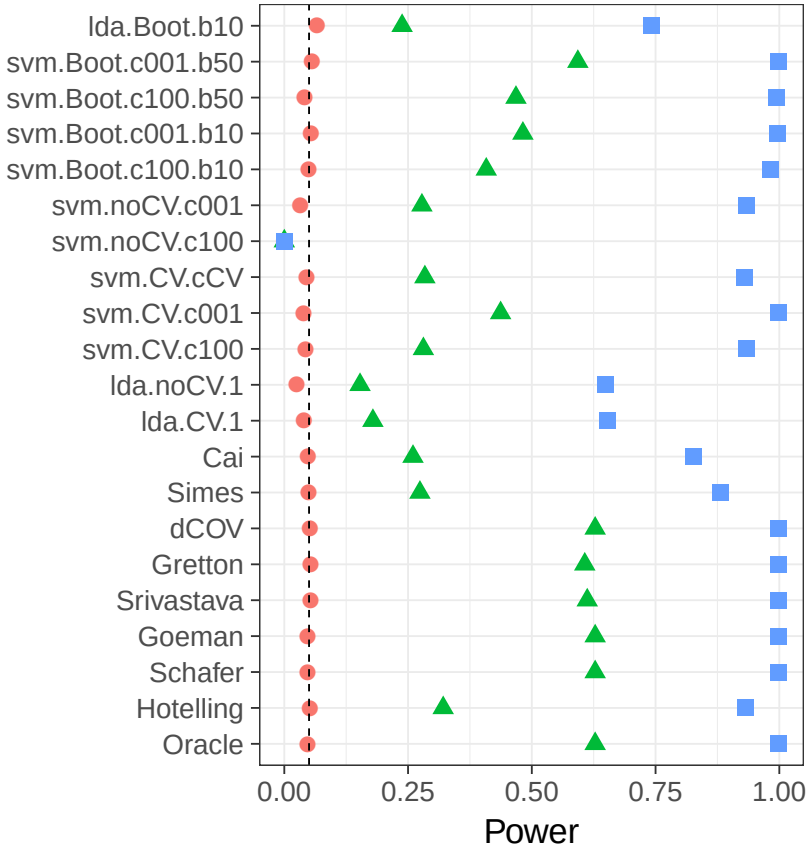


Fig. 8: **Bootstrap.** The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables I and II. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square).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

I. The Effect of high-dimension

Our best performing tests alleviate the high dimensionality of the problem by regularizing the estimation of Σ. By comparing the non-regularized T^2 to its regularized versions we see that in our high-dim setup, regularization adds power. Regularization is achieved by shrinking, or thresholding, the entries of $\hat{\Sigma}$. Shrinking is used in the *Schafer* statistic. Thresholding is used in the *Goeman* and *Srivastava* statistics.

Can we explicitly regularize the covariance estimate of a classifier? To answer this question we augment the simulation with some accuracy-tests that have explicit covariance regularization in them. These include shrinkage based LDA [65], [66], where Tikhonov regularization of $\hat{\Sigma}$ is used; just like the *Schafer* statistic. We also try a diagonalized LDA [67], also known as *Gaussian Naïve Bayes*, which regularizes by thresholding, similarly to the *Srivastava* and *Goeman* statistics.

Simulation results are reported in Figure 9 with naming conventions in Table III. The proper regularization of the covariance of a classifier, just like a two-group test, can improve power. See, for instance, *svm.CV.c001* 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.Pang.b50*. Even so, the out-of-the-box two-group tests outperform the accuracy-tests.

Optimizing the regularization parameter for classification does not result in a good test, as can be seem from the performance. In SVMs, the cost parameter governs the magnitude of the margins, and thus the regularization. The *svm.CV.cCV* statistic has a cost parameter optimized with an inner CV. The *svm.CV.c001* statistic has a large fixed regularization. The better power of *svm.CV.c001* leads us to argue that the optimal regularization for prediction is not the same as the optimal for testing.

Name	Algorithm	Resampling	Parameters
✣lda.highdim.Dudoit.CV	[67]	V-fold	–
✣lda.highdim.Ramey.CV	[66]	V-fold	–
✣lda.highdim.Pang.CV	[65]	V-fold	–
✣lda.highdim.Pang.b50	[65]	bLOO	B=50

TABLE III: The same as Table I for regularized (high-dimensional) predictors. Accuracy tests marked with a ✣.

J. Mixture Classes

When discussing the power of the resubstitution accuracy, the authors of [60] simulate power 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)

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

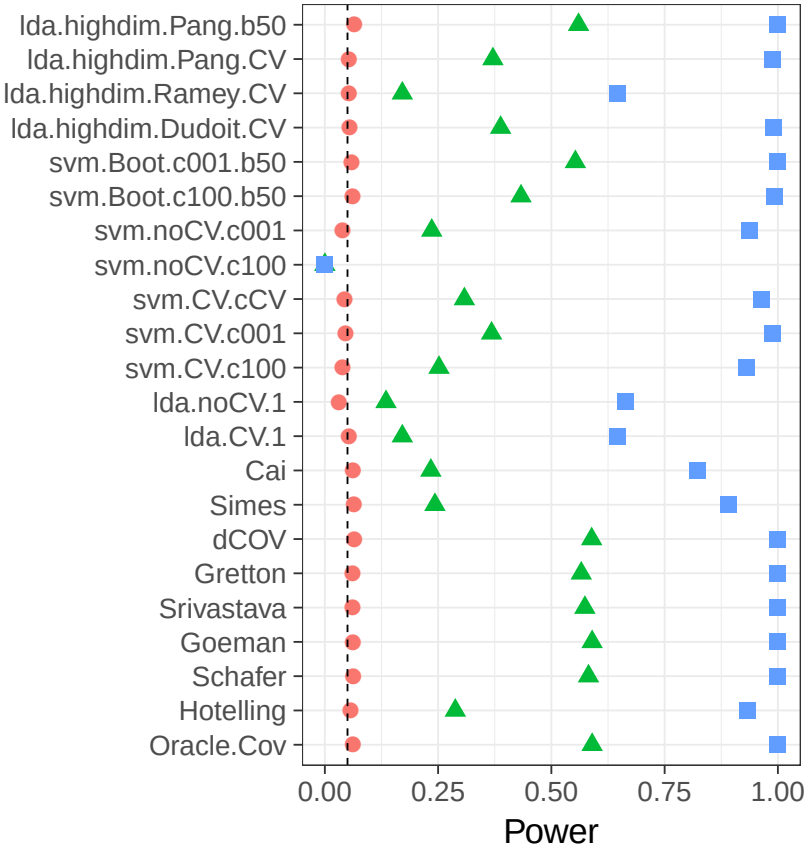


Fig. 9: **HighDim Classifier.** The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables I and III. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5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

Varying π interpolates between the null distribution $(\pi = 0.5)$ and a location shift model $(\pi = 0)$. We now perform the same simulation as [60], and in the same dimensionality as our previous simulations. We re-parameterize so that $\pi = 0$ corresponds to the null model:

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

From Figure 10, we see that also for the mixture class of [60] locations tests are to be preferred over accuracy-tests.

IV. Neuroimaging Example

Figure 11 is an application of (a) the Srivastava two-group test, and (b) a linear SVM accuracy-test, to the neuroimaging data of Pernet et al. [68]. The authors of [68] 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

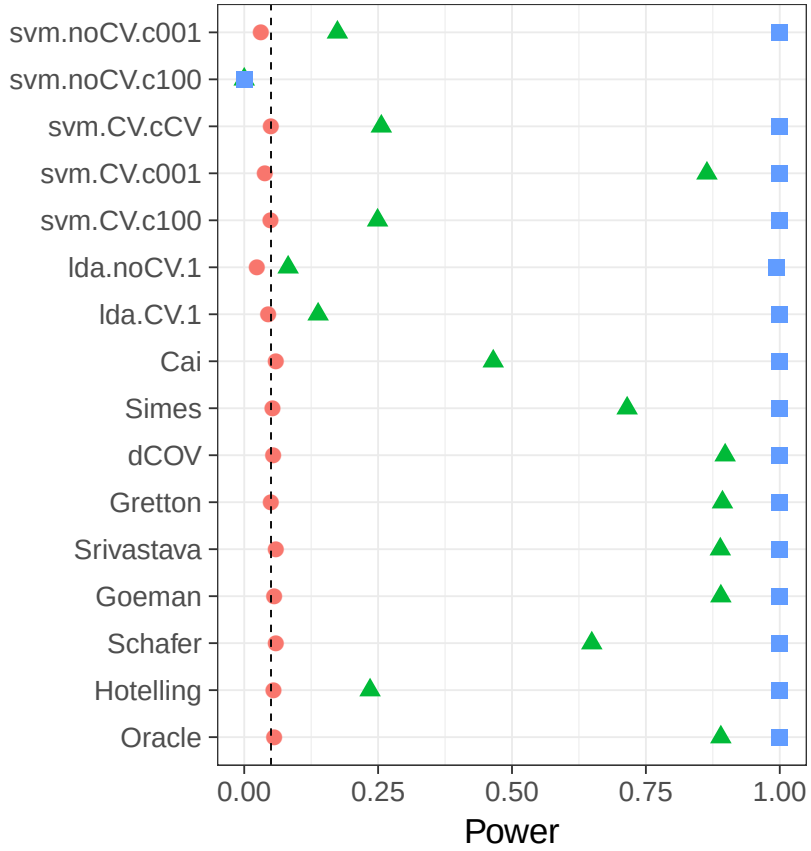


Fig. 10: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(2). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), $1/4$ (green triangle) and $1/2$ (blue square).

made available by the authors at the OpenNeuro website[8].

We perform group inference using within-subject permutations along the analysis pipeline of [69], which was also reported in [24].

In agreement with our simulation results, the two-group test (*Srivastava*) discovers more brain regions of interest when compared to an accuracy-test. The former discovers $1,232$ regions, while the latter only 441, as depicted in Figure 11. 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.

V. Discussion

We have set out to understand which of the tests is more powerful: accuracy-tests or two-group tests. Our current observation is that we have never found accuracy tests to be optimal in high-dim regimes; there was always a two-group test that dominated in power. We conjecture that they are never optimal, simply because of the needless discretization of the test statistic. Two-group tests are also typically easier to implement, and faster to run, since no

[8]http://reproducibility.stanford.edu/

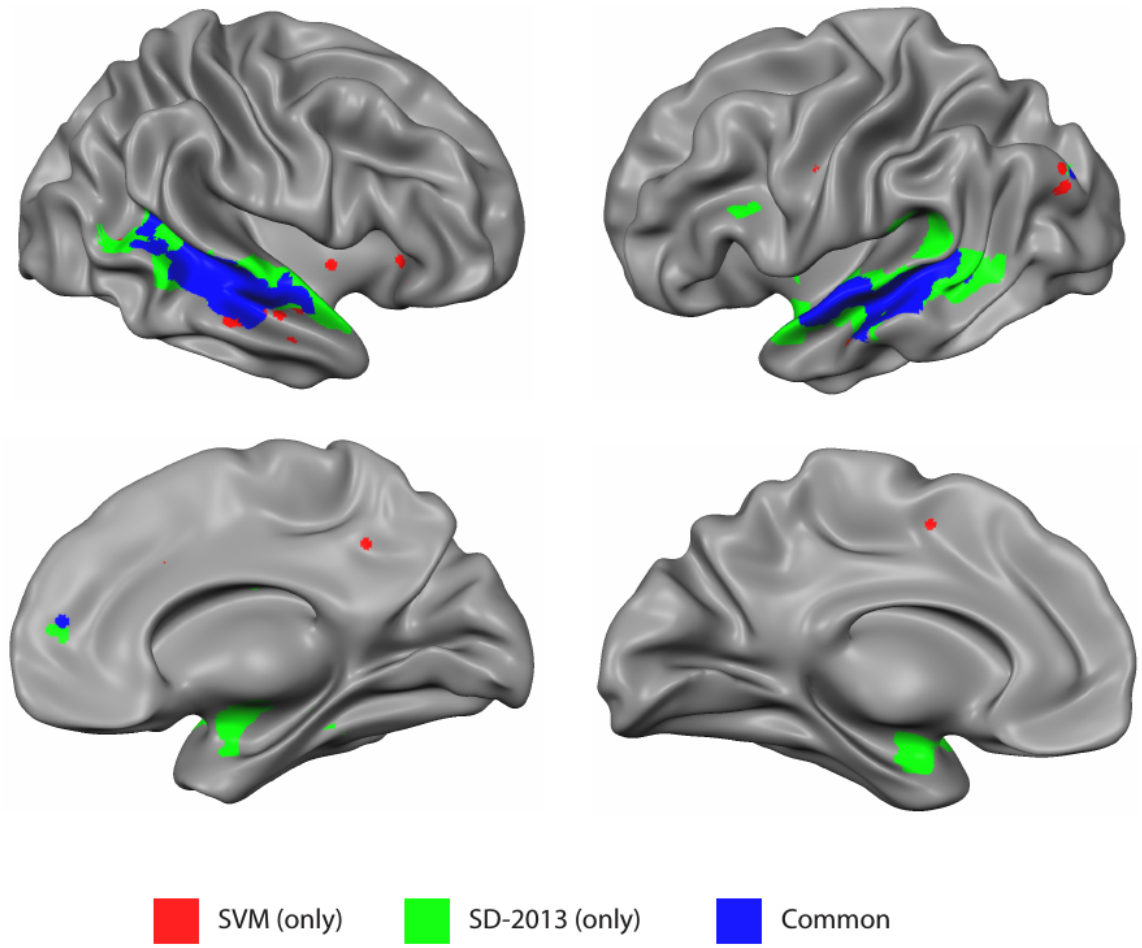


Fig. 11: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy-test u and a two-group test (*Srivastava*). The linear SVM was computed using 5-fold cross validation, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of [69], followed by region-wise $FDR \leq 0.05$ control using the Benjamini-Hochberg procedure [70]. Number of permutations equals 400. The two-group test detect $1,232$ regions, and the accuracy-test 441, 399 of which are common to both. For the details of the analysis see [24].

resampling is required. Statistics such as *Schafer*, *Goeman*, *Srivastava*, *dCOV*, and *Gretton*, are particularly well suited for detecting dense signal in high-dim.

A. Where do accuracy-tests Lose Power?

The low power of the accuracy-tests compared to two-group tests can be attributed to some of the following causes.

1) Data Splitting: Cross-validated statistics split the data. The train set serves to learn a statistic, and the test set to compute it. In a train-test validation scheme, the effective sample size is that of the test set. This is clearly inefficient. In V-fold validation scheme, the statistic is the average over all test sets, so the effective sample size is

less obvious. We argue that this is an inefficient use of the data, as seen in the distributed learning literature, where splitting the sample and averaging is less accurate then learning with the whole data [71].

The superiority of the Bootstrap over V-fold was independently observed in [11]. According to these authors, this superiority is due to the larger test-samples when Bootstrapping, compared to V-folding.

2) Inappropriate Regularization: From the fact that *svm.CV.cCV* is less accurate then *svm.CV.c001* we learn than testing requires different regularization than predicting. Does testing require more or less regularization? In our simulations, the optimal cross validated cost parameter for SVM (the cost of *svm.CV.cCV*) was larger then that of the most powerful SVM (*svm.CV.c001*). We thus conclude that that testing requires *more* regularization than predicting. Why would this happen? Regularization introduces bias and reduces bias. For testing, we only care about the bias in the largest coordinates of μ. For predictions we care about the bias in all coordinates of μ. This means that when testing, the bias introduced by regularization is not limited by the smaller coordinates of μ, permitting to remove move variance. This phenomenon was also observed in [72], which observe that recovering the support of a function requires different regularization (i.e. smoothing) than the *matched filter theorem*, optimal for recovering the whole function.

3) 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 [73]. Secondly, because of the presence of ties which does not allow to exhaust the permissible false positive rate, unless randomization is introduced. 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*, i.e. the train-accuracy, as a test statistic. In a very high-dimensional regime, the resubstitution accuracy may be as high as 1 for the observed data [74, 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. This explains the terrible power of *svm.noCV.c100*, which has barely any regularization[9].

The degree of discretization is governed by the sample size. For this reason, an asymptotic analysis such as [42], or [60], will not capture power loss due to discretization. This actually holds for all power analyses relying on a *contiguity* argument [43, Ch.6]. An asymptotic analysis, which eschews the discretization effect, may suggest resubstitution accuracy estimates are good test statistics, while they suffer from very low finite-sample power. One of the effects of discretization is ties. The canonical remedy for ties— random tie breaking — showed only a minor improvement (not reported herein).

Using our simulations we may quantify the power loss due to discretization, this is because Figher's LDA is equivalent to Hotelling's T^2 followed by a discretization stage. From Figure 1 we see that for the intermediate signal's strength, *Hotelling* has roughly twice the power of *lda.noCV.1*. We thus conclude that the effect of discretization may be considerable.

[9]Recall that the cost parameter in LIBSVM is inversely proportional to the regularization.

The matter of discretization was addressed by a 2011 post by Prof. Frank Harrell in CrossValidated[10]:

> ... 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.

B. Interpretation

Two-group tests, and location tests in particular, are easier to interpret. To do so we typically use a Neyman-Pearson Lemma type argument, and think: What type of signal is a test sensitive to? What is the direction of the effect? Accuracy-tests are seen as "black boxes", even though they can be analyzed in the same way. Gilron et at.[75] demonstrate that the type of signal captured by accuracy-tests is less interpretable to neuroimaging practitioners than two-group 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. Examples can be found, for instance, in [76] and references therein.

C. Fixed SNR

For a fair comparison between simulations, in particular between those with different Σ, we needed to fix the difficulty of the problem. For a fair comparison we fix the Kullback–Leibler Divergence between distributions of sample means. Abusing notation, we fix $KL[\bar{x}_1, \bar{x}_0] = c^2 p$, which is the same as fixing $\|\mu\|^2_\Theta$, with the exception of the large sample (III-D) and the heavytailed analysis (III-E).

Our choice implies that the Euclidean norm of $\delta := \mathbb{E}(x_1) - \mathbb{E}(x_0)$ varies with Σ, with the sample size, 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 $\|\mu\|_\Theta$.

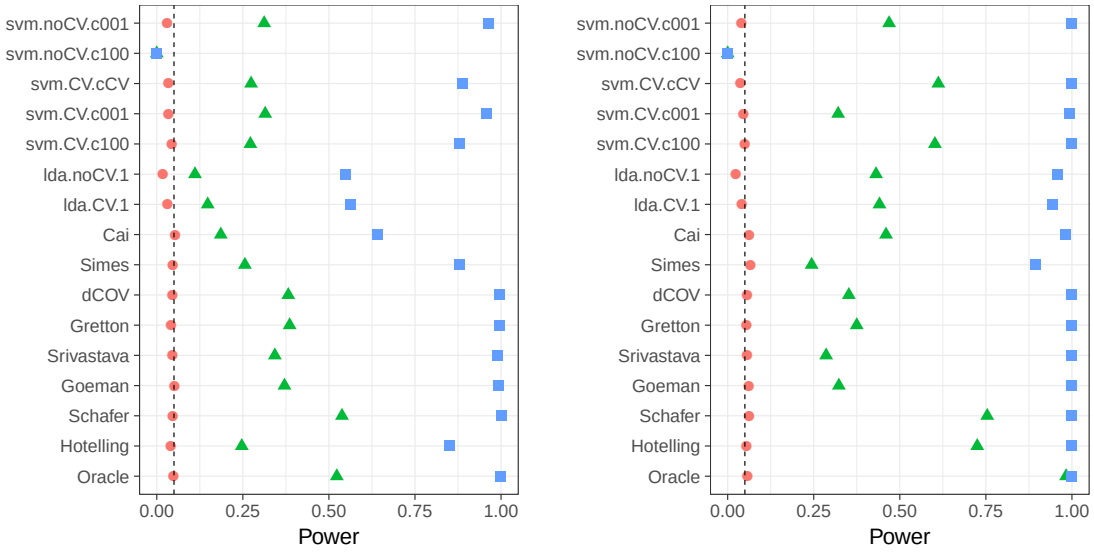
For completeness, Figure 12 reports the power analysis under $AR(1)$ correlations, but with $\|\mu\|_2$ fixed. We compare the power of a shift in the direction of some high variance PC (Figure 12a), versus a shift in the direction of a low variance PC (Figure 12b). The intuition that it is easier to detect signal in the low variance directions is confirmed.

Other authors have also observed the need for fixing the SNR for a fair comparison between tests. In [77], authors prefer to use sparse alternatives. With sparse alternatives, the difficulty of the problem is governed by the sparsity of the signal and not only the dimension of the data. In [78], authors fix $\|\mu\|_2^2/\|\Sigma\|^2_{Frob}$ where $\|\Sigma\|^2_{Frob} = \mathrm{Tr}(\Sigma'\Sigma)$ is the Frobenius matrix norm. Clearly, $\|\mu\|_2^2/\|\Sigma\|^2_{Frob}$ is invariant to the direction of the signal with respect to the noise. For this reason, we prefer fixing $\|\mu\|_\Theta$.

D. Effect of Covariance Regularization

Figures 4, 5, 6 and 12, 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?

[10] A Q&A website for statistical questions: http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier. And also "Problems with Classification of Predictions" in his book: http://www.fharrell.com/doc/bbr.pdf



(a) μ in PC7 of Σ.

(b) μ in PC15 of Σ.

Fig. 12: Short memory, AR(1) correlation. $\|\mu\|_2$ fixed.

We attribute this phenomenon to regularization. Whereas the signal, μ, varies in direction, the regularization of $\hat{\Sigma}$ does not. We borrow intuition from ridge regression, for which closed form solutions are available, and is equivalent to LDA with Tikhonov regularization. We first recall that in ridge regression

$$\hat{y} = X'(\hat{\Sigma} + \lambda I)^{-1} X' y,$$

so that penalizing $\|\beta\|_2^2$, ends up with a Tikhonov regularization of the covariance estimator. Using the SVD decomposition of X, then

$$\hat{y} = \sum_{j=1}^{p} \left(u_j \frac{d_j^2}{d_j^2 + \lambda} u_j' \right) y, \tag{3}$$

where $X = (u_1, \ldots, u_p) diag(d_1, \ldots, d_p)(v_1', \ldots, v_p')$. From Eq.(3) we see that the bias is larger in the directions of the smaller d_j^2, i.e., the smaller PCs of $\hat{\Sigma}$. This intuition explains the fact that unregularized tests have more power than the regularized, as seen in figures 4b, 5b, and 6b.

E. Sparse Alternatives

In our set of simulations we discussed "dense" alternatives, in the sense that all coordinates carry signal. Dense alternatives are motivated by neuroimaging where most brain locations in a regions carry signal. In a genetic application, a "sparse" alternative may be more plausible. Figure 13 reports the usual results, where μ is sparse. As usual, two-group tests dominate accuracy-tests, only this time, the winners are not the T^2 type statistics, but rather, the tests for sparse shifts (*Cai*, *Simes*).

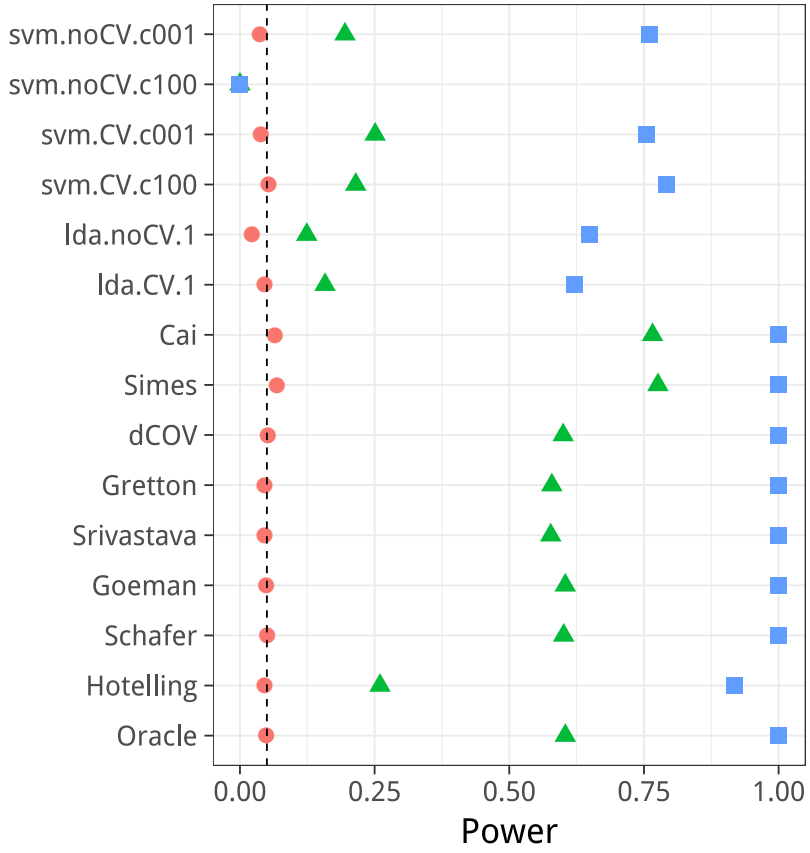


Fig. 13: Sparse μ.

F. 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_0 has a different distribution than x_1. This view is the one adopted in [31].

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

G. Feature Mapping

It may be argued that only accuracy-tests permit the separation between classes in augmented feature spaces, such as in *reproducing kernel Hilbert spaces* (RKHS). The *Gretton* statistic [12], is an example where the a two-group test is performed after augmenting the features to RKHS. We thus disagree with this argument: accuracy-tests do not have any more flexibility than two-group tests. One can always perform a two-group test after mapping the original features to some augmented space.

A different argument is that the feature mapping may not be known, but rather learned from the data. This is true but requires large amounts of data: in high-dim problems data is barely sufficient to learn covariances in the

original space. We are thus very skeptical of the possibility of learning covariances in augmented spaces. This is perhaps the reason why [79], who proposed using the covariance of the feature maps in RKHS, demonstrated their solution using a known covariance, and did not try to estimate it from data.

H. A Good accuracy-test

Brain-computer interfaces and clinical diagnostics [80], [81] are examples in which we want to know not only if information is encoded in a region, but rather, that a particular predictor can extract this information. In these cases an accuracy-test cannot be replaced by a two-group test. For the cases an accuracy-test cannot be replaced with other tests, we collect the following observations.

a) Sample size: The conservativeness of accuracy-tests, due to discretization, decrease with the size of the test set.

b) 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 [65] is a particularly good performer, but more research is required on optimal regularization for testing.

c) 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 [11], albeit attributed to the stability of the accuracy estimate, and not to its smoothness. We believe bootstraping enjoys from both smoothing and stabilizing (compared to V-folding), but we currently cannot quantify the contributions of each.

I. 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 two-group tests would naturally arise as the preferred method. As put by [42]:

> 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.

ACKNOWLEDGMENT

JDR wishes to thank, Jesse B.A. Hemerik, Yakir Brechenko, Omer Shamir, Joshua Vogelstein, Gilles Blanchard, and Jason Stein for their valuable inputs.

REFERENCES

[1] J. H. Friedman, "On multivariate goodness of fit and two sample testing," *eConf*, vol. 30908, no. SLAC-PUB-10325, pp. 311–313, 2003.

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

[3] M. Ojala and G. C. Garriga, "Permutation Tests for Studying Classifier Performance," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1833–1863, 2010.

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

[5] 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*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[6] 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*, ser. RECOMB '00. New York, NY, USA: ACM, 2000, pp. 263–272.

[7] M. D. Radmacher, L. M. McShane, and R. Simon, "A Paradigm for Class Prediction Using Gene Expression Profiles," *Journal of Computational Biology*, vol. 9, no. 3, pp. 505–511, Jun. 2002.

[8] 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*, vol. 10, no. 2, pp. 119–142, 2003.

[9] L. Juan and H. Iba, "Prediction of tumor outcome based on gene expression data," *Wuhan University Journal of Natural Sciences*, vol. 9, no. 2, pp. 177–182, Mar. 2004.

[10] 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*, vol. 7, no. 1, 2008.

[11] 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*, vol. 71, no. 1, pp. 107–118, 2007.

[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel Two-sample Test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[13] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1265–1287, 2003.

[14] P. Hall and N. Tajvidi, "Permutation tests for equality of distributions in high-dimensional settings," *Biometrika*, vol. 89, no. 2, pp. 359–374, 2002.

[15] H. H. Zhou, V. Singh, S. C. Johnson, G. Wahba, A. D. N. Initiative *et al.*, "Statistical tests and identifiability conditions for pooling and analyzing multisite datasets," *Proceedings of the National Academy of Sciences*, p. 201719747, 2018.

[16] F. Pérez-Cruz, "Estimation of information theoretic measures for continuous random variables," in *Advances in neural information processing systems*, 2009, pp. 1257–1264.

[17] 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*, 2004, pp. 1385–1392.

[18] P. Golland and B. Fischl, "Permutation tests for classification: towards statistical significance in image-based studies," in *IPMI*, vol. 3. Springer, 2003, pp. 330–341.

[19] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, vol. 45, pp. S199–S209, Mar. 2009.

[20] K. Schreiber and B. Krekelberg, "The statistical analysis of multi-voxel patterns in functional imaging," *PLoS One*, vol. 8, no. 7, p. e69328, 2013.

[21] 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*. IEEE, 2013, pp. 128–131.

[22] 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," Jun. 2016, working paper or preprint.

[23] 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*, vol. 103, no. 10, pp. 3863–3868, Jul. 2006.

[24] 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.

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

[26] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *The Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, Dec. 2009.

[27] T. Tony Cai, W. Liu, and Y. Xia, "Two-sample test of high dimensional means under dependence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 349–372, 2014.

[28] J. Chang, C. Zheng, W.-X. Zhou, and W. Zhou, "Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity," *arXiv preprint arXiv:1406.1939*, 2014.

[29] A. P. Dempster, "A high dimensional two sample significance test," *The Annals of Mathematical Statistics*, pp. 995–1010, 1958.

[30] 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*, vol. 4, no. 1, Jan. 2005.

[31] 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)*, vol. 68, no. 3, pp. 477–493, 2006.

[32] M. S. Srivastava, "Multivariate Theory for Analyzing High Dimensional Data," *Journal of the Japan Statistical Society*, vol. 37, no. 1, pp. 53–86, 2007.

[33] 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*, 2011, pp. 1206–1214.

[34] 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*, vol. 143, no. 11, pp. 1898–1911, 2013.

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

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

[37] G. Xu, L. Lin, P. Wei, and W. Pan, "An adaptive two-sample test for high-dimensional means," *Biometrika*, vol. 103, no. 3, pp. 609–624, 2016.

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

[39] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Annals of Statistics*, pp. 962–994, 2004.

[40] P.-S. Zhong, S. X. Chen, M. Xu *et al.*, "Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence," *The Annals of Statistics*, vol. 41, no. 6, pp. 2820–2851, 2013.

[41] A. Moscovich, B. Nadler, C. Spiegelman *et al.*, "On the exact berk-jones statistics and their p-value calculation," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 2329–2354, 2016.

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

[43] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, UK ; New York, NY, USA: Cambridge University Press, Oct. 1998.

[44] E. L. Lehmann, "Parametric versus nonparametrics: two alternative methodologies," *Journal of Nonparametric Statistics*, vol. 21, no. 4, pp. 397–405, 2009.

[45] H. Hotelling, "The Generalization of Student's Ratio," *The Annals of Mathematical Statistics*, vol. 2, no. 3, pp. 360–378, Aug. 1931.

[46] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken, NJ: Wiley-Interscience, Jul. 2003.

[47] M. S. Srivastava and M. Du, "A test for the mean vector with fewer observations than the dimension," *Journal of Multivariate Analysis*, vol. 99, no. 3, pp. 386–402, Mar. 2008.

[48] S. X. Chen, Y.-L. Qin, and others, "A two-sample test for high-dimensional data with applications to gene-set testing," *The Annals of Statistics*, vol. 38, no. 2, pp. 808–835, 2010.

[49] M. R. Ahmad, "A u -statistic approach for a high-dimensional two-sample mean testing problem under non-normality and behrens–fisher setting," *Annals of the Institute of Statistical Mathematics*, vol. 66, no. 1, pp. 33–61, 2014.

[50] L. Feng and F. Sun, "A note on high-dimensional two-sample test," *Statistics & Probability Letters*, vol. 105, pp. 29–36, 2015.

[51] T. Cai, W. Liu, and Y. Xia, "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 265–277, Mar. 2013.

[52] R. J. Simes, "An improved bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 73, no. 3, pp. 751–754, 1986.

[53] P. J. Bickel, "A distribution free version of the smirnov two sample test in the p-variate case," *The Annals of Mathematical Statistics*, vol. 40, no. 1, pp. 1–23, 1969.

[54] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *The Annals of Statistics*, pp. 697–717, 1979.

[55] G. J. Székely and M. L. Rizzo, "Testing for equal distributions in high dimension," *InterStat*, vol. 5, no. 16.10, 2004.

[56] G. Biau and L. Gyorfi, "On the asymptotic properties of a nonparametric l1-test statistic of homogeneity," *IEEE Trans. Inf. Theor.*, vol. 51, no. 11, pp. 3965–3973, Nov. 2005. [Online]. Available: http://dx.doi.org/10.1109/TIT.2005.856979

[57] Rosenbaum Paul R., "An exact distribution-free test comparing two multivariate distributions based on adjacency," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 4, pp. 515–530, Aug. 2005.

[58] N. Vayatis, M. Depecker, and S. J. Clémençcon, "AUC optimization and the two-sample problem," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 360–368.

[59] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[60] P. Golland, F. Liang, S. Mukherjee, and D. Panchenko, "Permutation Tests for Classification," in *Learning Theory*, ser. Lecture Notes in Computer Science, P. Auer and R. Meir, Eds. Springer Berlin Heidelberg, Jun. 2005, no. 3559, pp. 501–515.

[61] 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.

[62] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[63] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

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

[65] H. Pang, T. Tong, and H. Zhao, "Shrinkage-based Diagonal Discriminant Analysis and Its Applications in High-Dimensional Data," *Biometrics*, vol. 65, no. 4, pp. 1021–1029, Dec. 2009.

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

[67] 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*, vol. 97, no. 457, pp. 77–87, Mar. 2002.

[68] 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*, vol. 119, pp. 164–174, Oct. 2015.

[69] 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*, vol. 65, pp. 69–82, Jan. 2013.

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

[71] J. D. Rosenblatt and B. Nadler, "On the optimality of averaging in distributed statistical learning," *Information and Inference: A Journal of the IMA*, vol. 5, no. 4, pp. 379–404, 2016.

[72] D. Cheng, A. Schwartzman *et al.*, "Multiple testing of local maxima for detection of peaks in random fields," *The Annals of Statistics*, vol. 45, no. 2, pp. 529–556, 2017.

[73] J. Hemerik and J. Goeman, "Exact testing with random permutations," *TEST*, Nov 2017. [Online]. Available: https://doi.org/10.1007/s11749-017-0571-1

[74] G. J. McLachlan, "The bias of the apparent error rate in discriminant analysis," *Biometrika*, vol. 63, no. 2, pp. 239–244, Jan. 1976.

[75] 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*, vol. 146, pp. 113–120, 2017.

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

[77] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. A. Wasserman, "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions." in *AAAI*, 2015, pp. 3571–3577.

[78] S. X. Chen, Y.-L. Qin *et al.*, "A two-sample test for high-dimensional data with applications to gene-set testing," *The Annals of Statistics*, vol. 38, no. 2, pp. 808–835, 2010.

[79] Z. Harchaoui, E. Moulines, and F. R. Bach, "Kernel change-point analysis," in *Advances in neural information processing systems*, 2009, pp. 609–616.

[80] E. Olivetti, S. Greiner, and P. Avesani, "Induction in Neuroscience with Classification: Issues and Solutions," in *Machine Learning and Interpretation in Neuroimaging*, ser. Lecture Notes in Computer Science, G. Langs, I. Rish, M. Grosse-Wentrup, and B. Murphy, Eds. Springer Berlin Heidelberg, 2012, no. 7263, pp. 42–50.

[81] 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*, vol. 368, no. 15, pp. 1388–1397, Apr. 2013.

Jonathan D. Rosenblatt is a lecturer at the Industrial Engineering Department in Ben-Gurion University of the Negev, Israel. His research interests include computational statistics, distributed machine-learning algorithms, spatio-temporal data analysis (fMRI), and high-dimensional signal detection. JDR received the Ph.D. degree in Statistics from Tel-Aviv University, and did his post-doc at the Weizmann-Institute of Science. He is a member of Ben-Gurion University's Zlotowski Center for Neuroscience.

Yuval Benjamini is a senior lecturer at the Department of Statistics in the Hebrew University of Jerusalem, Israel. His research interests include high-dimensional prediction models and spatial statistics, specializing in the analysis of neuroscience and genomic data. YB received his Ph.D. degree in Statistics from UC Berkeley, and was a Stein Fellow at the Department of Statistics at Stanford University.

Roee Gilron is currently a postdoc in the Starr Lab at- UCSF. His research probes various aspects of motor control in health and disease using functional imaging and deep brain stimulation (DBS) in Parkinson's disease. He has developed methods for multidimensional analysis of neuroimaging data and is currently developing platforms for use in adaptive DBS. He received his B.S degree in Neuroscience from Brandeis University and completed his PhD in Cognitive Neuroscience from Tel Aviv University (TAU).

Roy Mukamel performed his undergraduate studies in computer science and biology and received his Ph.D. from the department of neuroscience at the Weizmann Institute of Science. Following post-doctoral training at the University of California at Los-Angeles (UCLA), he joined Tel-Aviv University. Today he is an associate Professor at the School of Psychological Sciences and Sagol School of Neuroscience at Tel-Aviv University, Israel.

Jelle J. Goeman is professor of biostatistics at Leiden University Medical Center in The Netherlands.His research focuses on high-dimensional data as they occur in e.g. transcriptomics and neuroimaging. He has worked on methods for penalized estimation and multiple hypothesis testing. He received his PhD from Leiden University and has worked at Imperial College in London and at Radboud University in Nijmegen.