Biostatistics (2019), **0**, 0, *pp.* **??**–22
doi:10.1093/biostatistics/samplebibtex

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.

Summary

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 purpose 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.

Key words:

1. 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

*johnros@bgu.ac.il

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 [**?**], and separately in the machine-learning literature [**???**]. Examples in the genetics literature include **???????**. Other examples include speaker verification [**?**], text classification [**??**], distinguishing between facial expressions [**?**], data integration [**?**], record linkage in databases systems [**???**], optical character recognition [**?**], multimedia [**?**], and functional data analysis [**?**].

Examples in the neuroscientific literature, which is our motivating use-case, include **?????**, and especially the recently popularized *multivariate pattern analysis* (MVPA) framework in **?**.

To fix ideas, we will adhere to a concrete example. In **?**, 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 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 **?**:

> ... 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-group 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 replications of a treatment/stimulus. It is thus a *high-dim–small-sample* problem.

In a seminal contribution, **?** 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 **???**.
- **High-dim location test for sparse shifts**: Tests the seek for a sparse shift in mean vectors such as **??**.
- **High-dim location test for dense shifts**: Tests the seek for a dense shift in mean vectors such as **???????????**.
- **High-dim location test for shifts with unknown sparsity**: Tests the seek for a shift in mean vectors, but adapt to the unknown sparsity, such as **???**.

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**. Our 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 **?** 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, which may be favorable for tree type predictors, over linear ones. Authors of **?** compare the kernel test of Gretton et al. **?** to an accuracy-test based on logistic-regression. Their results are inconclusive with a slight advantage to the logistic regression. In **?**, 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 may be impossible in high-dimension.

Ramdas et al. **?** currently offer the only analytic analysis; comparing Hotelling's T^2 location test to *Fisher's linear discriminant analysis* (LDA) accuracy-test. By comparing the consistency rates **?** 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 [**?**]. ARE is the limiting ratio of the sample sizes required by two statistics to achieve similar power. **?** 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 **?** relates asymptotic power functions to ARE. Using this theorem and the results of **?** 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 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 [**?**], so that an ARE of 6.3 is strong evidence in favor of the two-group test.

The analysis in **?** 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 2. The main findings are reported in Sections 3, and 4. 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 $\mathcal{X} = \mathbb{R}^{27}$.

We denote with x_y a sample of x from group y . We denote the distribution of x_1 by $\mathcal{F}$ and x_0 with $\mathcal{G}$. 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 $\mathcal{F} \neq \mathcal{G}$.

2.2 *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 its observed value to it's permutation distribution. Tests differ in the statistic they employ. Comparison metric is 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 (b)-(c) R times.
(e) The permutation p-value is the proportion of $\mathcal{T}_\pi$ larger than the observed $\mathcal{T}$. Formally: $\mathbb{P}\{\mathcal{T}_\pi \geqslant \mathcal{T}\} := \frac{1}{R+1}\sum_\pi I\{\mathcal{T}_\pi \geqslant \mathcal{T}\}$.
(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.

2.3 *Two-Group Tests*

The most prevalent interpretation of $\mathcal{F} \neq \mathcal{G}$ is to assume they differ in means (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). 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 **?**. 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, **?**.

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 by **?**. 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 SNR and power. Examples of high-dim tests for (dense) shifts include **??????????**.

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 test statistics for sparse shifts include **?** and **?**.

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 **?????**.

If the signal is present not (only) in means we opt for a two-group GOF test, instead of a location test. Examples of multivariate GOF tests include **??????????**.

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

2.4 *Prediction Accuracy as a Test Statistic*

An accuracy-test amounts to using a predictor's accuracy as a test statistic. Denoting a dataset by $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n$, a predictor, $\mathcal{A}_\mathcal{S} : \mathcal{X} \to \mathcal{Y}$, is the output of a learning algorithm $\mathcal{A}$ when applied to the dataset $\mathcal{S}$. The accuracy of a predictor, also known as (the complement of) the *test error.*, $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$, is defined as the probability of $\mathcal{A}_\mathcal{S}$ making a correct prediction. The accuracy of a learning algorithm, also known as (the complement of) the *expected test error.*, $\mathcal{E}_\mathcal{A}$, 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 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*, also known as (the complement of) the *train-error*, and the V-fold Cross Validation (CV) estimate.

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

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

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.

Balanced folding The standard practice in V-fold CV is to constrain the data folds to be balanced, i.e. stratified [**?**, 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.

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 instead of labels, as done by **?**.

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. 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	**?**	Resubstitution	–
Simes	**?**	Resubstitution	–
dCOV	**?**	Resubstitution	–
Gretton	**?**	Resubstitution	–
Srivastava	**?**	Resubstitution	–
Goeman	**?**	Resubstitution	–
Schafer	**?**	Resubstitution	–
Hotelling	**?**	Resubstitution	–
Oracle	T^2	Resubstitution	Known Σ

Table 1: 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 [**?**]. *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 [**?**]: larger cost implies less regularization.

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 and Notation*

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

$$\mathbf{x}_i = \mu \mathbf{y}_i + \eta_i, \tag{3.1}$$

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 SNR.

Having generated the data, we compute each of the test statistics in Table 1. 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.

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

3.3 *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).

3.4 *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 $n = 4,000, p = 2,300$. This simulation required roughly 11 years of computing time. The results are qualitatively similar to the high-dim–small-sample of Figure 1 and thus not reported herein.

3.5 *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 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.(3.1) 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 light-tailed Gaussian case (Figure 2).

3.6 *Departure from Sphericity*

We now test the robustness of our results to the correlations in x. In terms of Eq.(3.1), 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

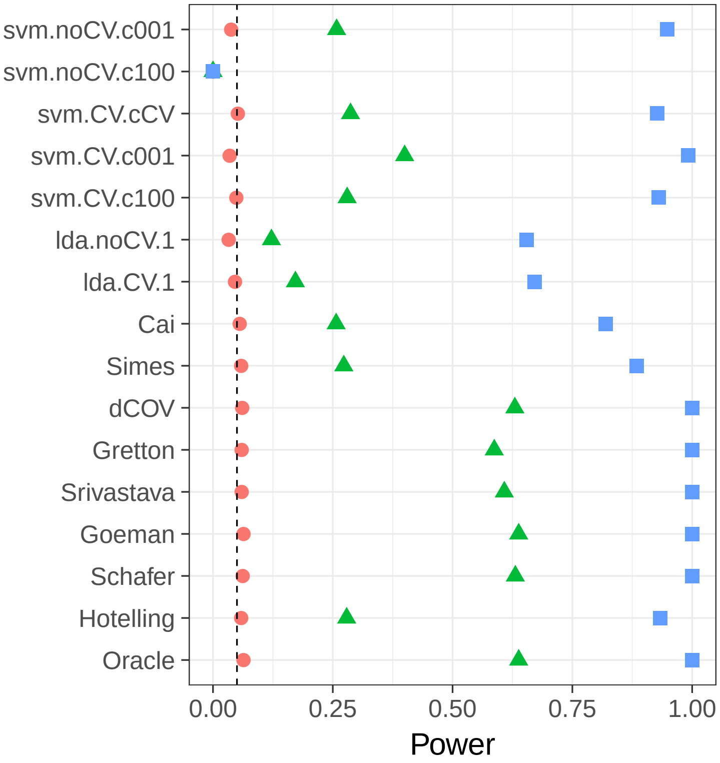


Fig. 1: The power of the permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. Effects vary over $c = 0$ (red circle), $c = 1/4$ (green triangle), and $c = 1/2$ (blue square). The various statistics on the y axis. Their details are given in Table 1. Simulation details in Section 3.1.

the uncorrelated case (Fig. 1) repeat themselves in the presence of correlation.

We simulate 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. To keep the comparisons fair as the correlations vary, we kept $\|\mu\|_\Theta := \sqrt{\mu'\Theta\mu}$ fixed. This matter is discussed in Section 5.3.

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 3a, the long memory correlations of a Brownian motion in Figure 4a, and the arbitrary correlation in Figure 5a.

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 3b, the long memory correlations of a Brownian motion in Figure 4b, and the arbitrary correlation in Figure 5b. We attribute this phenomenon to the bias introduced by the regularization, which masks the signal. This matter is further discussed in Section 5.3.

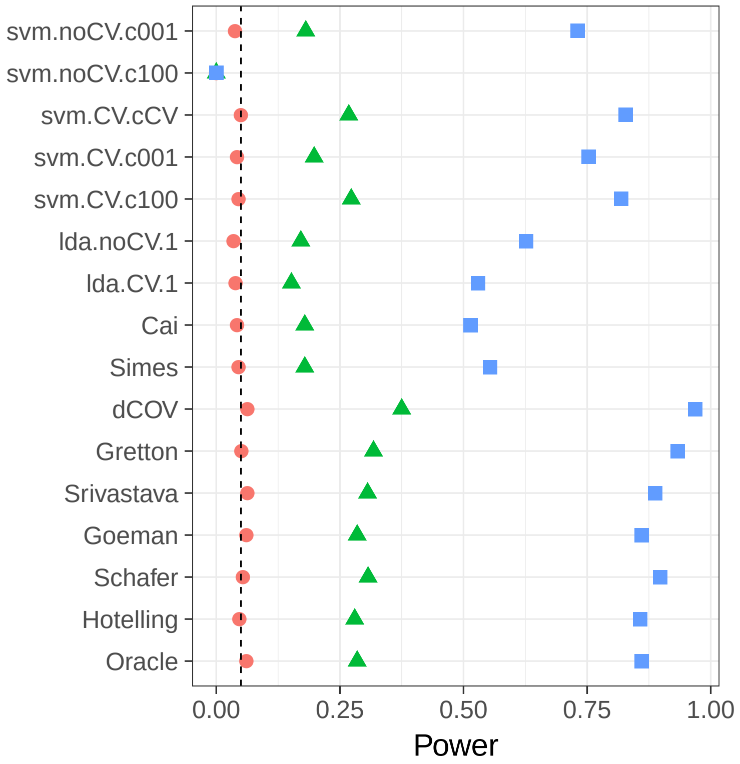


Fig. 2: **Heavytailed.** η_i is p-variate t distributed, with $df = 3$.

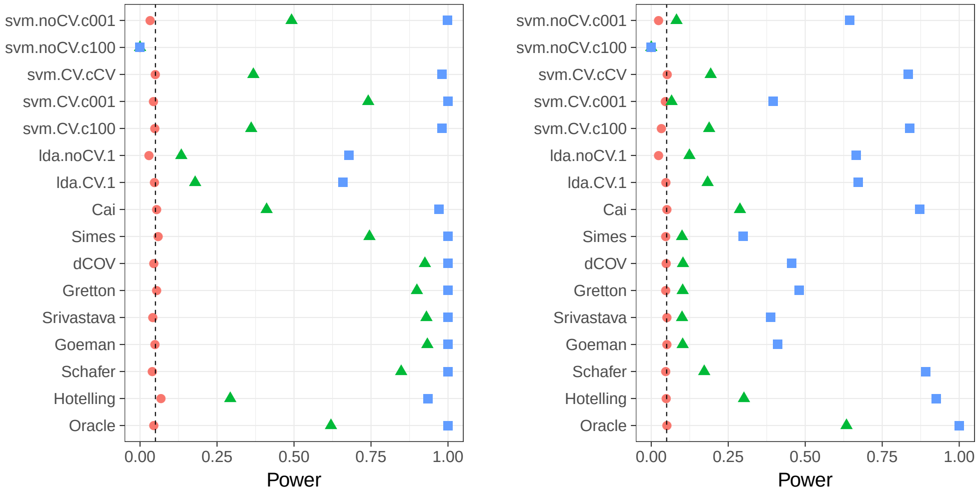
3.7 *Departure from Homoskedasticity and Scalar Invariance*

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, 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 6a 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.

3.8 *Departure from V-fold CV*

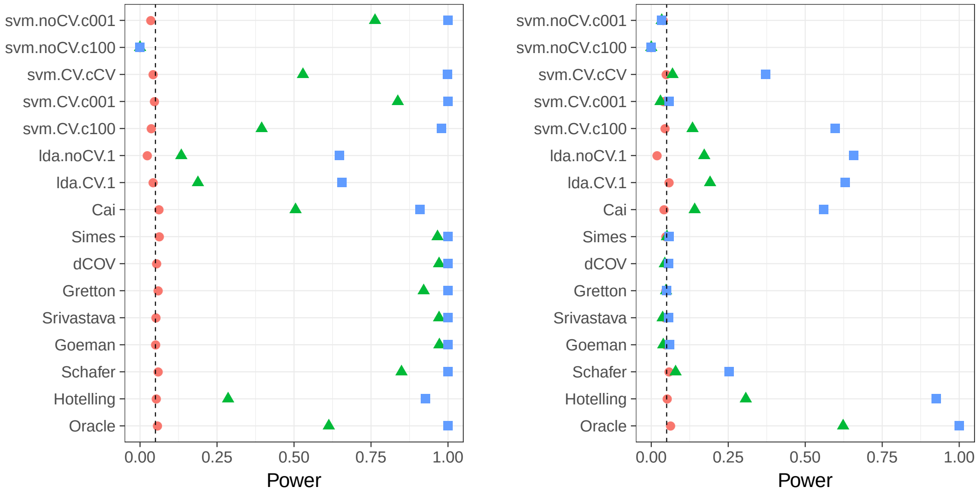
In V-fold CV, the discretization of the accuracy statistic is governed by the number of samples. This is the case whenever resampling without replacement. Intuition suggests we may alleviate the discretization of the accuracy statistic by replacing the V-fold CV, and resampling *with*



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

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

Fig. 3: Short memory, AR(1) correlation. $\Sigma_{k,l} = \rho^{|k-l|}; \rho = 0.6$.



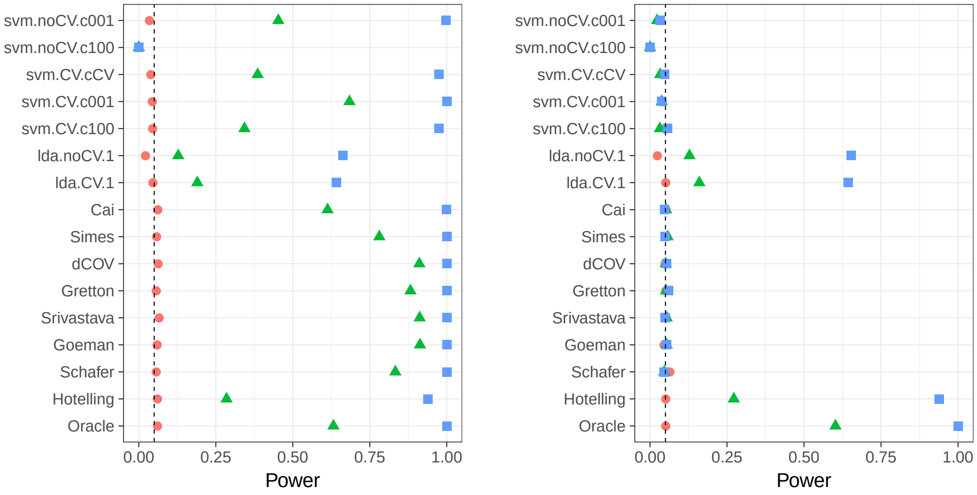
(a) Signal in direction of highest variance PC of Σ.

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

Fig. 4: 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. 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* [**?**, Sec 7.11].

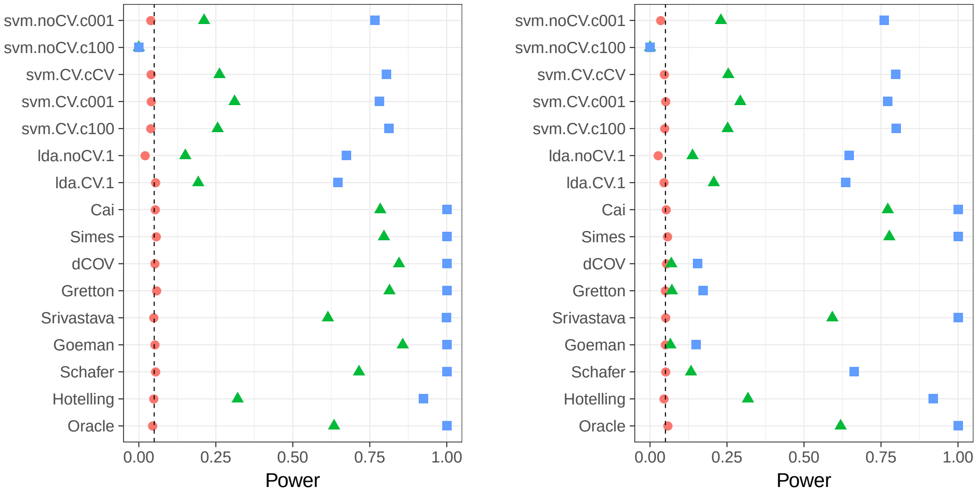
Definition 3 (bLOO)**.** The *leave-one-out bootstrap* estimate, bLOO, is the average accuracy of the holdout observations, over all bootstrap samples. Denote by $\mathcal{S}^b$, a bootstrap sample b of size n, sampled with replacement from $\mathcal{S}$. Also denote by $C^{(i)}$ the index set of bootstrap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



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

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

Fig. 5: 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. 6: Heteroskedasticity: Σ is diagonal with $\Sigma_{jj} = j$.

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 7 with naming conventions in Table 2. 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 2: The same as Table 1 for bootstrapped accuracy estimates. bLOO is defined in 3. B denotes the number of Bootstrap samples. Accuracy-tests marked with a ✠.

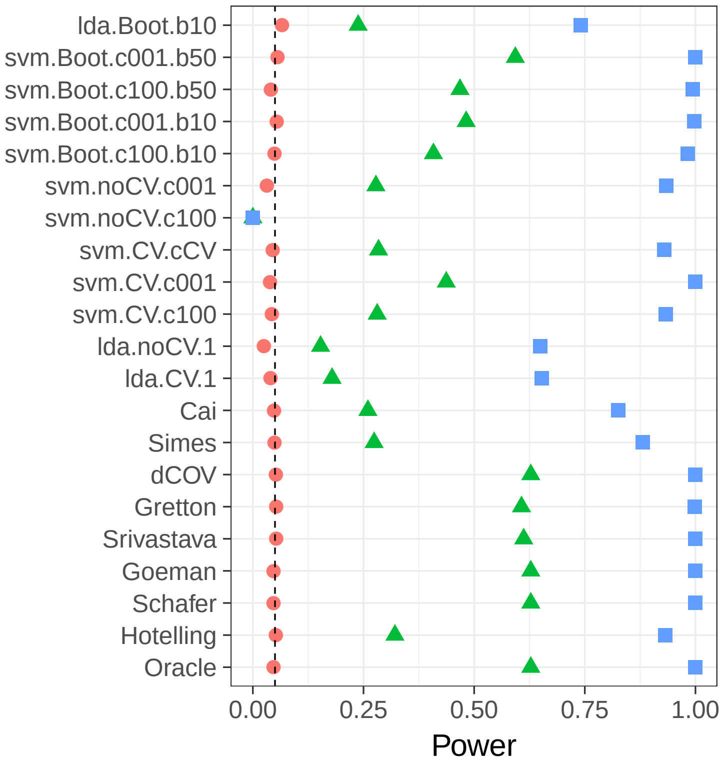


Fig. 7: **Bootstrap.** The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 2. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Appendix 3.1.

3.9 *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 [**??**], where Tikhonov regularization of $\hat{\Sigma}$ is used; just like the *Schafer* statistic. We also try a diagonalized LDA [**?**], also known as *Gaussian Naïve Bayes*, which regularizes by thresholding, similarly to the *Srivastava* and *Goeman* statistics.

Simulation results are reported in Figure 8 with naming conventions in Table 3. 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	**?**	V-fold	–
✠lda.highdim.Ramey.CV	**?**	V-fold	–
✠lda.highdim.Pang.CV	**?**	V-fold	–
✠lda.highdim.Pang.b50	**?**	bLOO	B=50

Table 3: The same as Table 1 for regularized (high-dimensional) predictors. Accuracy tests marked with a ✠.

3.10 *Mixture Classes*

When discussing the power of the resubstitution accuracy, **?** 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)

$$\begin{aligned} x_1 &\sim \pi\mathcal{N}(\mu_1, I) + (1-\pi)\mathcal{N}(\mu_2, I)\,, \\ x_0 &\sim (1-\pi)\mathcal{N}(\mu_1, I) + \pi\mathcal{N}(\mu_2, I)\,. \end{aligned}$$

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

$$\begin{aligned} x_1 &\sim (1/2-\pi)\mathcal{N}(\mu_1, I) + (1/2+\pi)\mathcal{N}(\mu_2, I)\,, \\ x_0 &\sim (1/2+\pi)\mathcal{N}(\mu_1, I) + (1/2-\pi)\mathcal{N}(\mu_2, I)\,. \end{aligned} \tag{3.2}$$

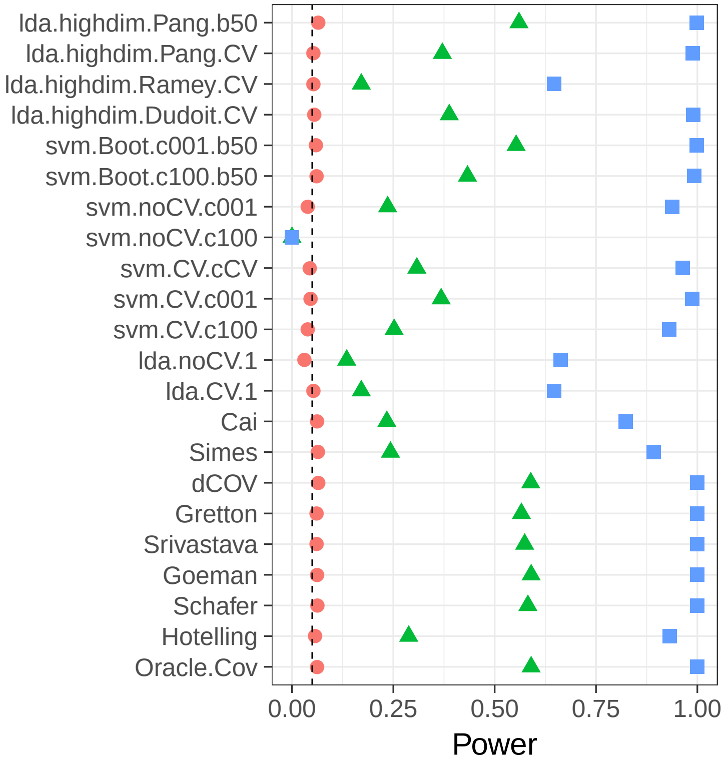


Fig. 8: **HighDim Classifier.** The power of a permutation test with various test statistics. The power on the x axis. Effects are color and shape coded. The various statistics on the y axis. Their details are given in tables 1 and 3. Effects vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). Simulation details in Section 3.1.

From Figure 9, we see that for the mixture class of **?** locations tests are still preferred over accuracy-tests.

4. Neuroimaging Example

Figure 10 is an application of (a) the Srivastava two-group test, and (b) a linear SVM accuracy-test, to the neuroimaging data of **?**. The authors of **?** 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 OpenNeuro website (`http://reproducibility.stanford.edu/`).

We perform group inference using within-subject permutations along the analysis pipeline of **?**, which was also reported in **?**.

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 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.

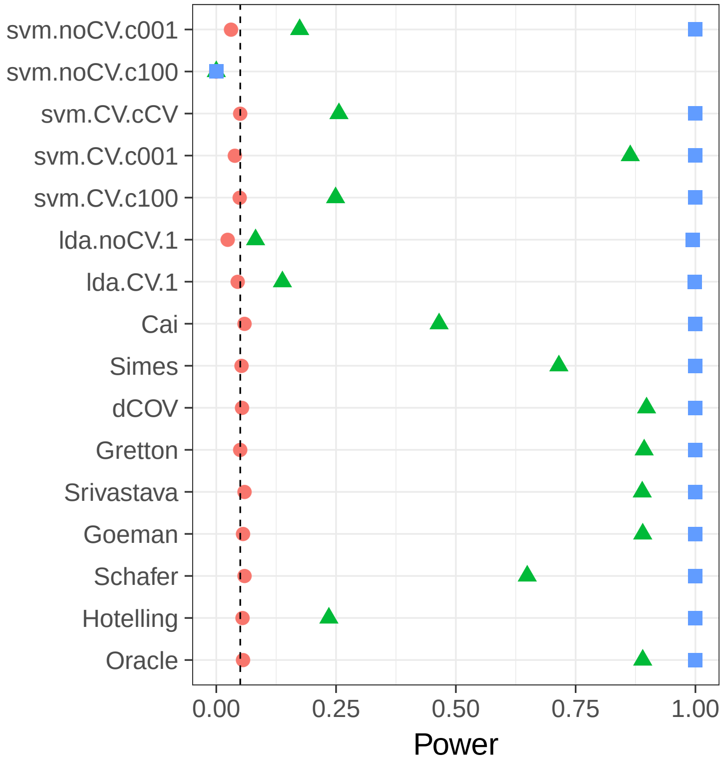


Fig. 9: **Mixture Alternatives.** $\mathbf{x}_i$ is distributed as in Eq.(3.2). μ is a p-vector with $3/\sqrt{p}$ in all coordinates. The effect, π, is color and shape coded and varies over 0 (red circle), 1/4 (green triangle) and 1/2 (blue square).

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

5.1 *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.

5.1.1 *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

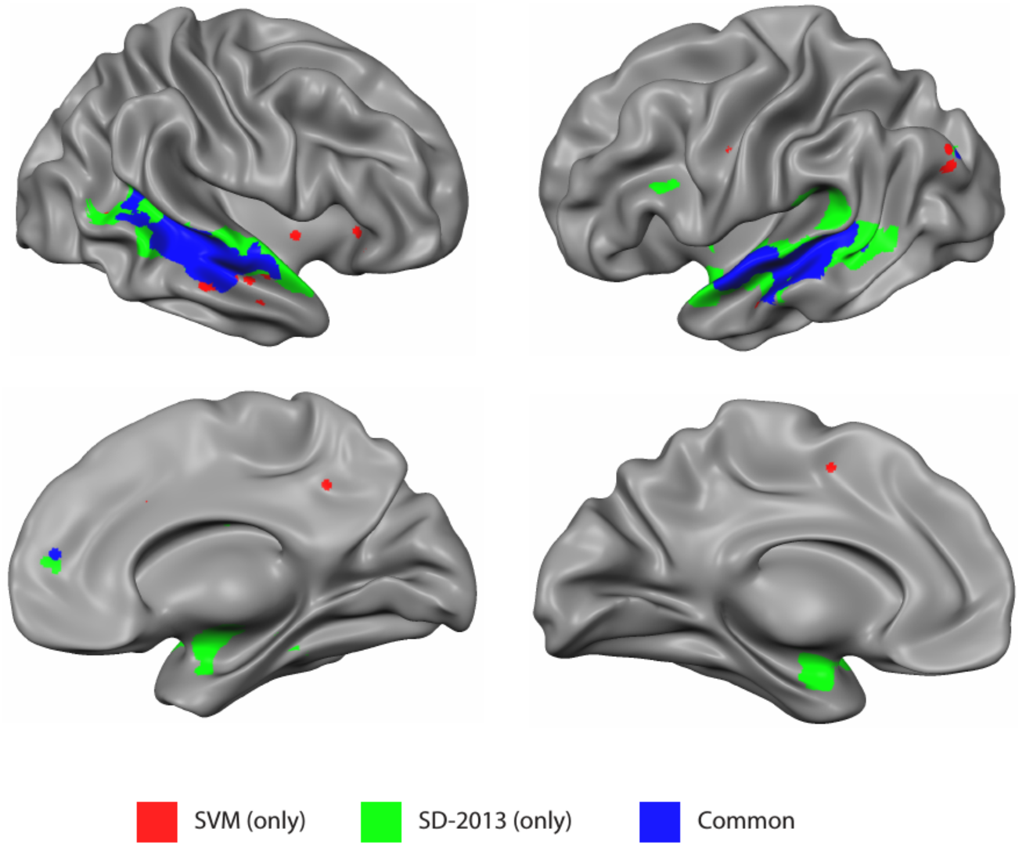


Fig. 10: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy-test u and a two-group test (*Srivastava*). The linear SVM was computed using 5-fold cross validation, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of **?**, followed by region-wise $FDR \leqslant 0.05$ control using the Benjamini-Hochberg procedure **?**. 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 **?**.

still 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 [**?**].

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

5.1.2 *Inappropriate Regularization* From the fact that *svm.CV.cCV* is less powerful than *svm.CV.c001* we learn that 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 more variance. This phenomenon was also observed in **?**, 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.

5.1.3 *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 [**?**]. 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 [**?**, 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 (recall that the cost parameter in LIBSVM is inversely proportional to the regularization).

The degree of discretization is governed by the sample size. For this reason, an asymptotic analysis such as **?**, or **?**, will not capture power loss due to discretization. This actually holds for all power analyses relying on a *contiguity* argument [**?**, 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 signals strength, *Hotelling* has roughly twice the power of *lda.noCV.1*. We thus conclude that the effect of discretization may be considerable.

The matter of discretization was addressed in a 2011 post by Prof. Frank Harrell in CrossValidated; a Q&A website for statistical questions `http://stats.stackexchange.com/questions/17408/how-to-assess-statistical-significance-of-the-accuracy-of-a-classifier`. :

> ... 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*

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. **?** demonstrate that the type of signal captured by accuracy-tests is less interpretable to neuroimaging practitioners than two-group tests.

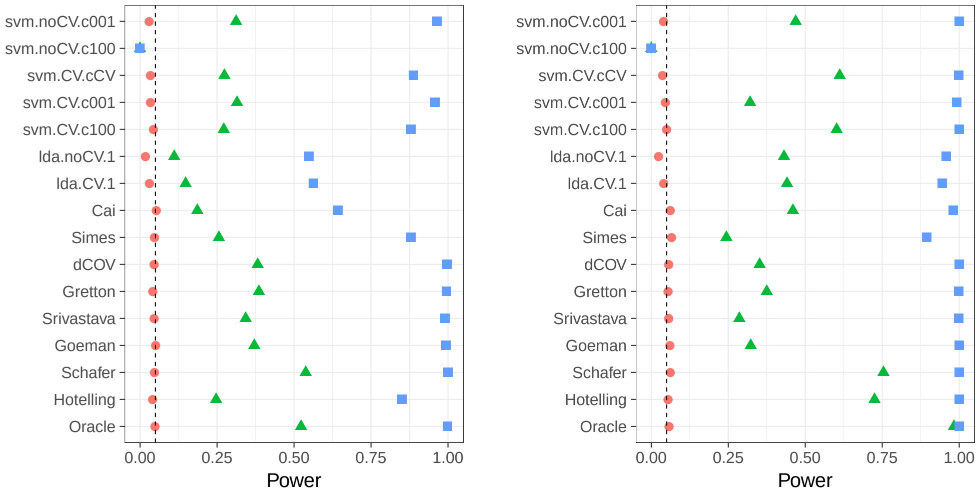
Some authors prefer accuracy-tests because they can be seen as effect-size estimates, invariant to the sample size. This is true, but the multivariate-statistics literature provides many multivariate effect-size estimators. Examples can be found, for instance, in **?** and references therein.

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 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 (3.4) and the heavytailed analysis (3.5).

Our choice implies that the Euclidean norm of $\mu := \mathbb{E}(x_1) - \mathbb{E}(x_0)$ 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 11 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 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.



(a) μ in PC7 of Σ.

(b) μ in PC15 of Σ.

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

Other authors have also observed the need for fixing the SNR for a fair comparison between tests. In **?**, 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 **?**, 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 $\|\mu\|_\Theta$.

5.4 *Effect of Covariance Regularization*

Figures 3, 4, 5 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. 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

$$\hat{y} = \sum_{j=1}^{p} \left(u_j \frac{d_j^2}{d_j^2 + \lambda} u_j' \right) y, \tag{5.3}$$

where $X = (u_1, \dots, u_p) diag(d_1, \dots, d_p)(v_1', \dots, v_p')$. From Eq.(5.3) 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 3b, 4b, and 5b.

5.5 *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 12 reports power when 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*).

5.6 *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 **?**.

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.

5.7 *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 **?**, 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 **?**, 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.

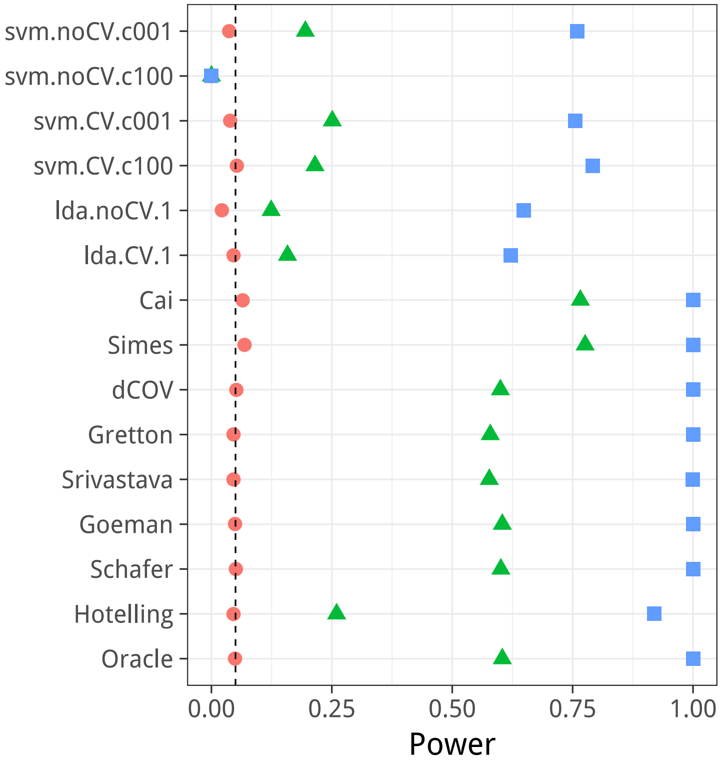


Fig. 12: Sparse μ.

5.8 *A Good accuracy-test*

Brain-computer interfaces and clinical diagnostics [**??**] 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.

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

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 **?** is a particularly good performer, but more research is required on optimal regularization for testing.

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 **?**, 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.

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 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 **?**:

> 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. JDR also thanks ISF grant 924/16 for the computing infrastructure used in this research.

$[xxx]$