Biostatistics (2019), **0**, 0, *pp.* 1–22
doi:10.1093/biostatistics/permutaing˙accuracy

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 suboptimal regularization. When the purpose of the analysis is the evaluation of a particular classifier, not signal detection, 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 [Friedman, 2003], and separately in the machine-learning literature [Eric et al., 2008, Ojala and Garriga, 2010, Lopez-Paz and Oquab, 2016]. Examples in the genetics literature include Golub et al. [1999], Slonim et al. [2000], Radmacher et al. [2002], Mukherjee et al. [2003], Juan and Iba [2004], Jiang et al. [2008], Yu et al. [2007]. Examples in the neuroscientific literature, which is our motivating use-case, include Golland and Fischl [2003], Pereira et al. [2009], Schreiber and Krekelberg [2013], Olivetti et al. [2013], Varoquaux et al. [2016], and especially the recently popularized *multivariate pattern analysis* (MVPA) framework in Kriegeskorte et al. [2006].

To fix ideas, we will adhere to a concrete example. In Gilron et al. [2016], the authors seek to detect brain regions that encode differences between vocal and non-vocal stimuli. Following the MVPA workflow, the localization problem is cast as a supervised learning problem: if the type of 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. [2009]:

> ... the problem of deciding whether the classifier learned to discriminate the classes can be subsumed into the more general question as to whether there is evidence that the underlying distributions of each class are equal or not.

A practitioner may thus approach the signal detection problem with a two-group hypothesis test. 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, Bai and Saranadasa [1996] noted that in high-dimension, multivariate tests tend to be low powered unless some regularization is involved. Since then, many high-dimensional tests have been proposed. These can be classified along the following lines: High-dim goodness of fit tests– Tests that seek for any difference between two multivariate distributions. GOF in short. High-dim location tests– Tests the seek for a shift in mean vectors. Shifts may be in many coordinates (dense), or only in a few (sparse). We collectively call GOF tests and location tests *two-group tests*.

At this point, it becomes unclear which test is preferable, in particular for genetics and neuroimaging: two-group tests or accuracy tests? In this manuscript, we do not provide a full answer to the matter. Instead, we seek to demonstrate that in the high dimensional regime **accuracy-tests never have more power than 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 Yu et al. [2007] 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. Olivetti et al. [2013] compare the kernel test of Gretton et al. [2012] to an accuracy-test based on logistic-regression. Their results are inconclusive with a slight advantage to the logistic regression. In Lopez-Paz and Oquab [2016], 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. [2016] 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 Ramdas et al. [2016] 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 [van der Vaart, 1998]. ARE is the limiting ratio of the sample sizes required by two statistics to achieve similar power. Ramdas et al. [2016] derive the asymptotic power functions of the two test statistics, 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 van der Vaart [1998] relates asymptotic power functions to ARE. Using this theorem and the results of Ramdas et al. [2016] we deduce that the ARE is lower bounded by $2\pi \approx 6.3$. This means that Fisher's LDA requires at least 6.3 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 [Lehmann, 2009], so that an ARE of 6.3 is strong evidence in favor of the two-group test.

We study the power of many accuracy, and two-sample methods, in a large scale simulation study. This allows us to evaluate theoretical results such as Ramdas et al. [2016], in various small-sample configurations. Our configurations include various two-group effect models. A particular emphasis is given to multivariate shift effects, but also include other effect models such as logistic regression and mixtures. We focus on two-group problems, because studying multi-group problems can be derived from multiple binary decisions [Zheng et al., 2018].

The simulation scenarios were designed with neuroimaging and genetic applications in mind. In these applications the sample acquisition is expensive, and the samples high-dimensional, leading to the high-dim–small-sample setup. Binary outcomes correspond to healthy/sick individuals, or active/inactive brain regions. Highly correlated contentious predictors correspond to blood oxygenation levels in a brain region, or gene expressions. Average blood oxygenation levels are expected to vary when a brain region is active, thus justifying our interest in shift alternatives. The same holds in genetics, where average expression levels of disease encoding genes are expected to vary between healthy and sick individuals.

We start with formalizing the problem in Section 2. The main findings are reported in Sections 3, and 4, with extensions in the online Supplementary Material. 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 with $\mathcal{F}$ 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 consider 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. We adhere to permutation tests and not parametric inference because in high-dim–small-sample problems 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: $\frac{1}{R}\sum_\pi I\{\mathcal{T}_\pi \geqslant \mathcal{T}\}$.
(f) Declare $\mathcal{F} \neq \mathcal{G}$ if the permutation p-value is smaller than α, which we set to $\alpha = 0.05$.

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 counterexample where equal means do not imply equal distributions). Difference in means only 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, 1931]. For more background see, for example, Anderson [2003].

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 Bai and Saranadasa [1996]. In these cases, high-dimensional versions of the T^2 should be applied, which essentially regularize the estimator of Σ, thus reducing the dimensionality of the problem and improving SNR and power. Examples of high-dim tests for (dense) shifts include Dempster [1958], Bai and Saranadasa [1996], Schäfer and Strimmer [2005], Goeman et al. [2006], Srivastava and Du [2008], Chen et al. [2010], Lopes et al. [2011], Ahmad [2014], Thulin [2014], Feng and Sun [2015].

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 Cai et al. [2013] and Chang et al. [2014].

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 Simes [1986], Donoho and Jin [2004], Zhong et al. [2013], Shen and Lin [2015], Moscovich et al. [2016].

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 Bickel [1969], Friedman and Rafsky [1979], Hall and Tajvidi [2002], Székely and Rizzo [2004], Biau and Gyorfi [2005], Rosenbaum Paul R. [2005], Eric et al. [2008], Pérez-Cruz [2009], Vayatis et al. [2009], Gretton et al. [2012].

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, $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$, is defined as the probability of $\mathcal{A}_\mathcal{S}$ making a correct prediction for a new data point. It is also known as (the complement of) the *test error*. The accuracy of a learning algorithm, $\mathcal{E}_\mathcal{A}$, is defined as the expected accuracy over all possible data sets $\mathcal{S}$. It is also known as (the complement of) the *expected test error*. 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.

If y is independent of x, then $\mathcal{A}_\mathcal{S}$ cannot capture any signal and is no more accurate than a coin toss (for balanced classes). This is known as *chance level*. A statistically significant better-than-chance-level estimate of $\mathcal{E}_\mathcal{A}$, or $\mathcal{E}_{\mathcal{A}_\mathcal{S}}$, 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)**.** Denote by $\mathcal{S}^v$ the v'th partition, or *fold*, of the dataset, and by $\mathcal{S}^{(v)}$ its complement. 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? 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, since it does not inflate the error rates of the accuracy-tests. 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, a.k.a. stratified [Ojala and Garriga, 2010, for example]. This means that each fold has the same number of samples 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 [e.g. Golland et al., 2005] .

How many folds Different authors suggest different rules for the number of folds. We fix the number of folds to $V = 4$. A different number of folds does not change our conclusions. We do not 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.noCV.c001	SVM	Resubstitution	cost=0.01
✣svm.noCV.c100	SVM	Resubstitution	cost=100
✣svm.CV.cCV	SVM	V-fold	cost=CV
✣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	Cai et al. [2013]	Resubstitution	–
Simes	Simes [1986]	Resubstitution	–
dCOV	Székely and Rizzo [2004]	Resubstitution	–
Gretton	Gretton et al. [2012]	Resubstitution	–
Srivastava	Srivastava and Du [2008]	Resubstitution	–
Goeman	Goeman et al. [2006]	Resubstitution	–
Schafer	Schäfer and Strimmer [2005]	Resubstitution	–
Hotelling	Hotelling [1931]	Resubstitution	–
Oracle	T^2 with Known Σ	Resubstitution	–

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 [Meyer et al., 2015]. *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 [Chang and Lin, 2011]: 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– Fisher's LDA*

The basic simulation setup is essentially the sampling distribution underlying Fisher's Linear Discriminant Analysis. 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 measurement, η_i, is distributed as $\mathcal{N}_p(0, \Sigma)$. The sample size is set to $n = 40$, and the dimension of the data set to $p = 23$. The covariance $\Sigma = I$.

In this basic setup, reported in Figure 1, the shift is denoted by μ. We set $\mu := c\,\mathbf{e}$ where $\mathbf{e}$ is a p-vector of ones. This implies that shifts are dense and equal in all p coordinates. We use the Mahalanobis norm between means as a measure of signal-to-noise (SNR): $\frac{n}{2}\|\mu\|_\Sigma^2 = \frac{n}{2}\mu'\Sigma^{-1}\mu$.

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 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 replicates where the permutation p-value fell below 0.05. The number of replicates is $1,000$, so that the standard errors of our estimates are no larger than 0.6% under the null, and 1.5% in general.

3.1.1 *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.

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

3.1.3 *Sample Size* We focus on high-dim–small-sample configurations because of our motivation in neuroimaging and genetics. Our results, however, also hold in the high-dim–large-sample configurations. To prove this point, we increase the scale of the problem by one order of magnitude: we fix p/n at 23/40, and set $n = 400, p = 230$. The results are qualitatively similar to the high-dim–small-sample in Fig.1, and reported in the online Supplementary Material. We also note that further increasing n and p requires tremendous amounts of computation. This is because the number of test statistics computed is in the order of the number of replications, times permutations, times data folds, times statistics to compute. In our current setup, this means roughly 8×10^6 operations, each scaling polynomially with n and p.

3.2 *Departure From Gaussianity*

Hotelling's T^2 is a generalized likelihood ratio test in the Gaussian shift class. This 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. To verify our conclusions in the non-Gaussian case, 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

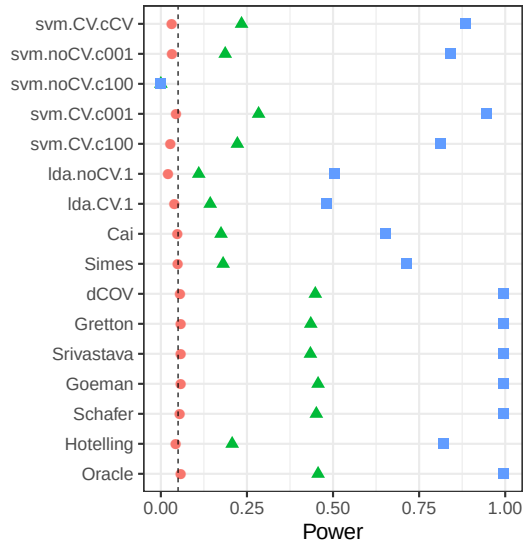


Fig. 1: [TODO: verify anti-conservatism.] The power of the permutation test with various test statistics. The power on the x-axis. Effects are color and shape coded. Effects vary over $\frac{n}{2}\|\mu\|_{\Sigma}^2 = 0$ (red circle), 25 (green triangle), and 100 (blue square). The various statistics on the y-axis. Their details are given in Table 1. Simulation details in Section 3.1. R code in `http://www.john-ros.com/permuting_accuracy/`.

Gaussian case (Figure 2).

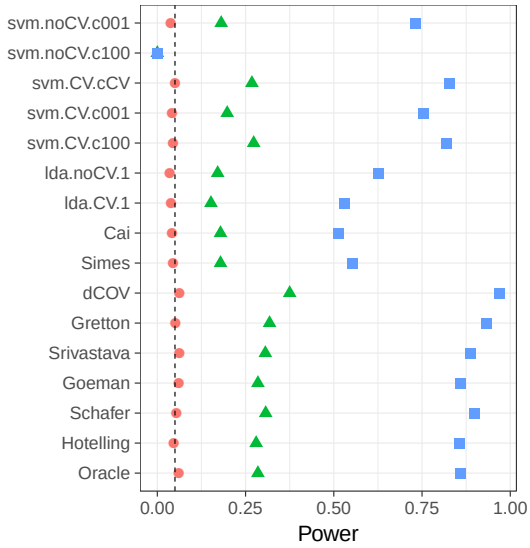


Fig. 2: **Heavytailed.** η_i is p-variate t distributed, with $df = 3$.

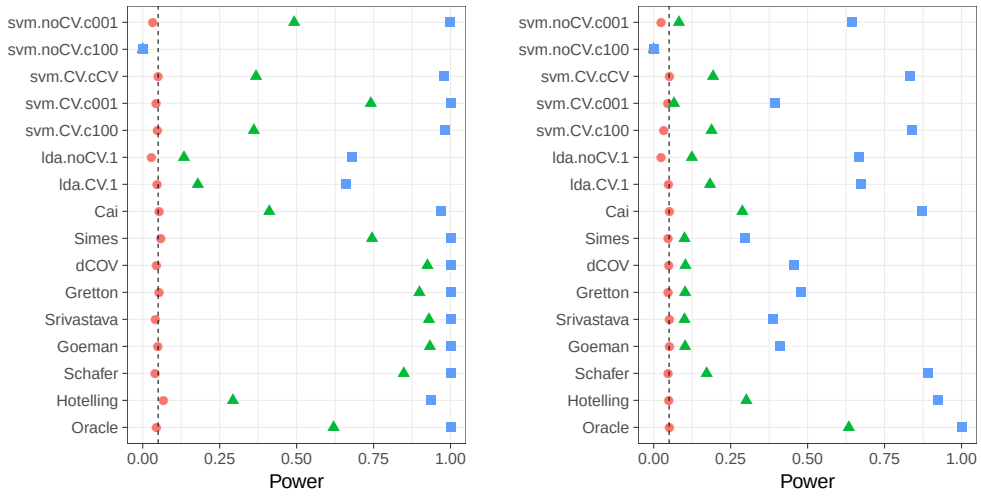
3.3 *Departure from Sphericity*

We now test the robustness of our results to 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 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 $\frac{n}{2}\|\mu\|_{\Sigma}^2$ fixed. This matter is discussed in Section 5.2.

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 accl acy-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 the Supplementary Material, and an arbitrary correlation, also in the Supplementary Material.

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 the Supplementary Material, and an arbitrary correlation, also in the Supplementary Material. We attribute this phenomenon to the bias introduced by the regularization, which masks the signal. This matter is further discussed in Section 5.3.



(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. $\Sigma_{k,l} = \rho^{|k-l|}; \rho = 0.6$.

3.4 *Departure From Shift Alternatives*

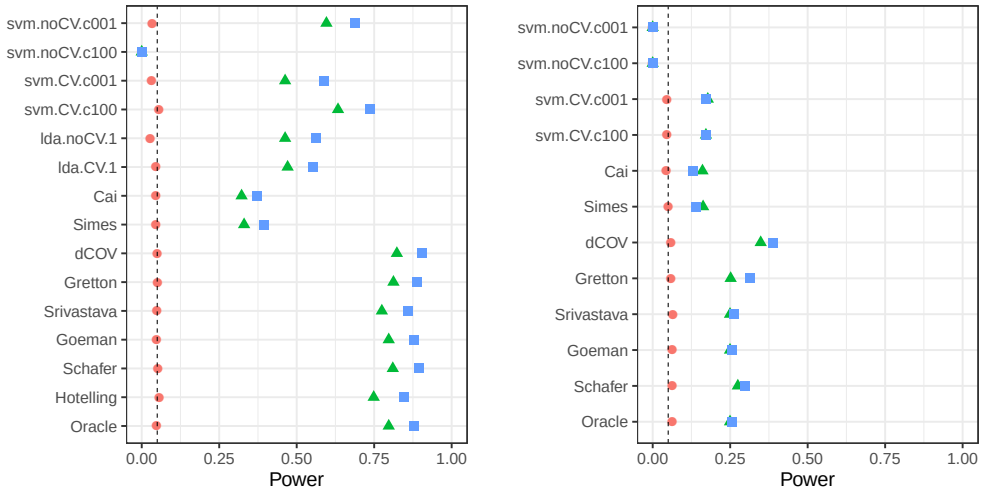
Shift alternatives are a popular signal model in the statistical literature. This is due to mathematical convenience, but also for empirical reasons: (a) Many effects are "pure shifts" after a

scale transformation. For instance, a multiplicative effect in log scale. (b) Many effects are not pure shifts, but have a shift component. In fact, it would be quite controversial to assume an effect is manifested in higher moments alone.

For completeness, we now report power for logistic regression. Logistic regression is not a shift class. This is because when fixing $P(y|x)$, there is no marginal distribution of x for which x_1 is a shifted version of x_0. In Figure 4 we report the usual power of our tests for a logistic model with main effects and second order interactions. We analyzed it both in the original space, x, and in an augmented space, $\tilde{x}$ with second order interactions:

$$\tilde{x} := \Phi(x) = (x_1, \ldots, x_j, \ldots, x_p, \ldots, x_1x_1, \ldots x_jx_{j'}, \ldots, x_px_p).$$

The figure demonstrates that two-group tests still dominate in power, even when the problem departs from the shift class. They also confirm that augmenting the feature space takes a toll in power, because many more covariance parameters need to be estimated. Sometimes, this toll is worthwhile, because the signal resides in the augmented space. Sometimes, this toll is needless, because the signal resides in the original space. Figure 4b is an example of the latter. In the Supplementary Material we provide an example of the former by simulating a logistic regression with main effects only.



(a) Data analyzed in the original space (x).

(b) Data analyzed in augmented interactions space ($\tilde{x}$). Some tests that are possible in x but not in $\tilde{x}$ are omited.

Fig. 4: **Logistic regression with second order interactions.** Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1 + \exp(\eta)]; \eta = x'\beta + x'Bx$ where β is a scaled vector of ones, and B a scaled identity matrix. Finally, $x \sim \mathcal{N}(0, I_{p\times p})$.

3.5 *Beyond 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*

replacement. An algorithm that samples test sets with replacement is the *leave-one-out bootstrap estimator*, and its derivatives, such as the *0.632 bootstrap*, and *0.632+ bootstrap* [Friedman et al., 2001, Sec 7.11].

Definition 3 (bLOO)**.** The *leave-one-out bootstrap* estimate, bLOO, is the average accuracy on the holdout observations, over all bootstrap samples. Denote by $\mathcal{S}^b$, a bootstrap sample b of size n, sampled with replacement from $\mathcal{S}$. Also denote by $C^{(i)}$ the index set of bootstrap samples not containing observation i. The leave-one-out bootstrap estimate, $\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO}$, is defined as: $\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{(i)}|}\sum_{b\in C^{(i)}}\mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i)=y_i\}$. 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 5 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 ✣.

3.6 *High-Dim Regularized Accuracy Tests*

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 thresholding the entries of $\hat{\Sigma}$, or inflating its diagonal, thus shrinking $\hat{\Sigma}^{-1}$. 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? The answer is affirmative and quite a lot of research effort has been devoted to the matter of covariance-regularized classifiers. See, for instance Bickel et al. [2004] or Dobriban et al. [2018]. We thus augment our simulations with some accuracy-tests that have explicit covariance regularization in them. These include shrinkage based LDA [Pang et al., 2009, Ramey et al., 2016], where Tikhonov regularization of $\hat{\Sigma}$ is used; just like the *Schafer* two-group test. We also try a diagonalized LDA [Dudoit et al., 2002], also known as *Gaussian Naïve Bayes*, which regularizes by canceling non-diagonal entries, similarly to the *Srivastava* and *Goeman* statistics.

Simulation results are reported in Figure 6 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.

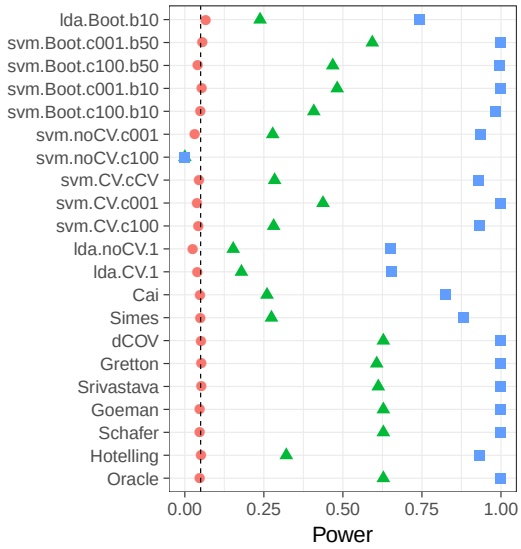


Fig. 5: **Bootstrap.** The power of a permutation test with various test statistics. The power on the x-axis. Effects are color and shape coded. The various statistics on the y-axis. Their details are given in tables 1 and 2. Effects vary over $\frac{n}{2}\|\mu\|_{\Sigma}^2 = 0$ (red circle), 25 (green triangle), and 100 (blue square). Simulation details in Section 3.1.

Optimizing the regularization parameter for classification does not result in a good test. The *svm.CV.cCV* statistic has a regularization parameter optimized with an inner CV. The *svm.CV.c001* statistic has a fixed, large, regularization. The better power of *svm.CV.c001* leads us to argue that the optimal regularization for prediction is larger than the optimal for testing.

Name	Algorithm	Resampling	Parameters
✣lda.highdim.Dudoit.CV	Dudoit et al. [2002]	V-fold	–
✣lda.highdim.Ramey.CV	Ramey et al. [2016]	V-fold	–
✣lda.highdim.Pang.CV	Pang et al. [2009]	V-fold	–
✣lda.highdim.Pang.b50	Pang et al. [2009]	bLOO	B=50

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

4. Neuroimaging Example

Figure 7 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. [2015]. The authors of Pernet et al. [2015] collected fMRI data while subjects were exposed to the sounds of human speech (vocal), and other non-vocal sounds. Each subject was exposed to 20 sounds of each type, totaling in $n = 40$ trials. The study was rather large and consisted of about 200 subjects. The data was kindly made available by the authors at the OpenNeuro website (`http://reproducibility.stanford.edu/`).

We perform group inference using within-subject permutations along the analysis pipeline of Stelzer et al. [2013]. Our test statistics account for dependence in space, but require independence in time. For this purpose, parameters were estimated with an orthogonal design, and an AR(6)

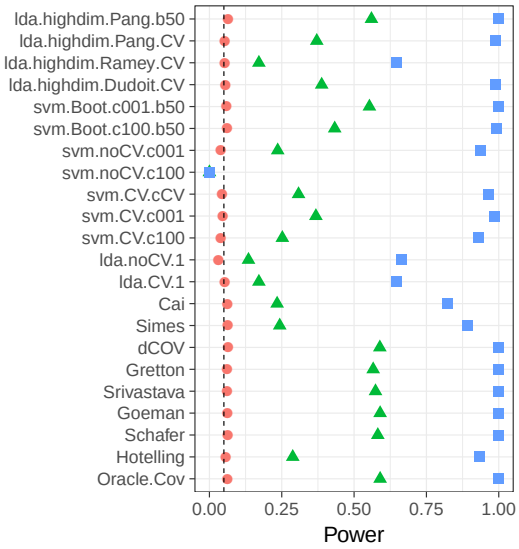


Fig. 6: [TODO: re-run to update Oracle caption] **HighDim Classifier.** The power of a permutation test with various test statistics. The power on the x-axis. Effects are color and shape coded. The various statistics on the y-axis. Their details are given in tables 1 and 3. Effects vary over $\frac{n}{2}\|\mu\|^2_\Sigma = 0$ (red circle), 25 (green triangle), and 100 (blue square). Simulation details in Section 3.1.

temporal model. The verification of temporal independence, and further details of the analysis, are reported in Gilron et al. [2016].

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

5. Discussion

We have set out to understand which of the tests is more powerful: accuracy-tests or two-group tests. Our current observation is that we have never found accuracy tests to be preferable in high-dim regimes; there was always a two-group test that dominated in power. We conjecture that accuracy are never preferred because of the needless discretization built in the test statistic. We also conjecture the advantage of two-group tests will increase when scaling from two-class to multi-class classification. Two-group tests are 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 multivariate 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.

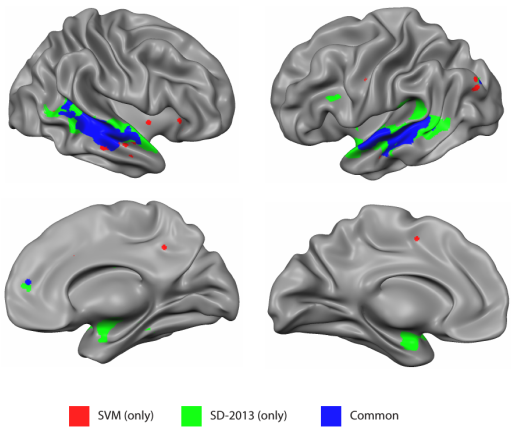


Fig. 7: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy-test u and a two-group test (*Srivastava*). The linear SVM was computed using 5-fold cross-validation, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of Stelzer et al. [2013], followed by region-wise $FDR \leqslant 0.05$ control using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. 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 Gilron et al. [2016].

5.1.1 *Data Splitting* Cross-validation splits 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 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 [Rosenblatt and Nadler, 2016].

The superiority of the Bootstrap over V-fold was independently observed in Yu et al. [2007]. 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 regularization for SVM (the inverse of the cost of *svm.CV.cCV*) was smaller then that of the most powerful SVM (*svm.CV.c001*). We thus conclude that testing requires *more* regularization than predicting. Why would this happen? Regularization introduces bias and reduces variance. For prediction we care about the bias in all coordinates of μ. For testing, we only care about the bias in the largest 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 Cheng et al. [2017], 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 [Hemerik and Goeman, 2018]. 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, overfitting may cause the resubstitution accuracy to be as high as 1 for the observed data [McLachlan, 1976, Theorem 1], but also for any permutation. The concentration of resubstitution accuracy near 1, and its discretization, render this test completely useless, with power tending to 0 for any (fixed) effect size, as the dimension of the model grows. 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 Ramdas et al. [2016], or Golland et al. [2005], will not capture power loss due to discretization. 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.

Using our simulations we may quantify the power loss due to discretization. This is because the resubstitution accuracy of Figher's LDA is equivalent to Hotelling's T^2 after binarizing predictions. From Figure 1 we see that for the intermediate signals strength, *Hotelling* has roughly twice the power of LDA (*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 *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 would have like to fix the power of the oracle test, but for some simulation setups that is possible only via a lengthy numerical search. We thus decided to fix $\frac{n}{2}\|\mu\|_\Sigma^2$, as a proxy. In the shifted Gaussian case, $\frac{n}{2}\|\mu\|_\Sigma^2$ is equivalent to fixing the Kullback–Leibler Divergence between sample means, $KL[\bar{x}_1, \bar{x}_0]$, which seems a good index of a problem's difficulty: it governs power, and is easy to compute. Exceptions to this rule are those where μ does not index signal, such as logistic regression, and the mixture class in the Supplementary Material. Another exception is the large-sample setup where fixing $\frac{n}{2}\|\mu\|_\Sigma^2$ was not enough to make power comparable. We attribute this to the fact that power depends on n and p, not only via $\frac{n}{2}\|\mu\|_\Sigma^2$.

Fixing $\frac{n}{2}\|\mu\|_\Sigma^2$ implies that the Euclidean norm of $\mu = \mathbb{E}(x_1) - \mathbb{E}(x_0)$ varies with Σ, with the sample size n, the dimension p, and with the direction of the signal. An initial intuition may suggest that detecting signal in the low variance PCs is easier than in the high variance PCs. This is true when fixing $\|\mu\|_2$, but not when fixing $\|\mu\|_\Sigma$. For verification, we report the power when fixing the Euclidean norm between population means in the Supplementary Material.

5.3 *Effect of Covariance Regularization*

Figure 3 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 already know from ridge regression that a Tikhonov regularization of the covariance shrinks estimates. We also know that shrinking is more aggressive in the low PCs of the design [Friedman et al., 2001]. If group means differ in the direction where shrinking is most aggressive, then regularization may mask the signal. This explains the fact that unregularized tests have more power than the regularized, as seen in Figures 3b and Supplementary Material.

5.4 *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. In the Supplementary Material we report the power when μ carries signal in a single coordinate, making it very sparse. As usual, two-group tests dominate accuracy-tests. This time, however, the winners are not the T^2 type statistics, but rather, tests for sparse shifts (*Cai*, *Simes*).

5.5 *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 [Gretton et al., 2012], is an example where a two-group test is performed after an implicit augmentation of x to some RKHS. We are also free, up to computational considerations, to augment the design matrix as we please. The logistic regression in Section 3.4 is such an example, where we analyzed the data both in the original space, and in an augmented space. We thus disagree with the argument that accuracy-tests have more flexibility than two-group tests. If one can perform an accuracy test after mapping the original features to some augmented space, then one can also perform a two-group test.

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, let alone in an augmented space. This is perhaps the reason why Harchaoui et al. [2009], who proposed using the covariance of the feature maps in RKHS, demonstrated their solution using an oracle covariance, and did not try to estimate it from data.

5.6 *A good accuracy-test*

Brain-computer interfaces and clinical diagnostics [Olivetti et al., 2012, Wager et al., 2013] 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. 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. There is a rich literature on optimal covariance regularization for prediction, but less so for testing. We find that the Shrinkage-based Diagonal Linear Discriminant Analysis of Pang et al. [2009] is a particularly good performer, but more research is required on optimal regularization for testing.

Resample with Replacement Smooth accuracy estimate by cross-validating with replacement. The bLOO estimator, in particular, is preferable over V-fold. This was also observed by Yu et al. [2007], albeit attributed to the stability of the accuracy estimate, and not to its smoothness. We believe bootstraping enjoys from both smoothing and stabilizing (compared to V-folding), but we currently cannot quantify the contributions of each.

5.7 *Epilogue*

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

> The recent popularity of machine learning has resulted in the extensive teaching and use of prediction in theoretical and applied communities and the relative lack of awareness or popularity of the topic of Neyman-Pearson style hypothesis testing in the computer science and related "data science" communities.

6. Supplementary Material

The Supplementary Material is available at [TODO: add url once online]. It includes further simulation results: large sample, extension of the non-spherical case, extension of the logistic regression case, a mixture class, sparse signal, heteroskedasticity, and tie breaking. It also includes a discussion of our choice of SNR measure.

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 was funded by the Israeli Science Foundation grants 900/16 and 924/16.

References

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*, 66(1):33–61, 2014.

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

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

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

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *IEEE Trans. Inf. Theor.*, 51(11):3965–3973, Nov. 2005. ISSN 0018-9448. doi: 10.1109/TIT.2005.856979. URL `http://dx.doi.org/10.1109/TIT.2005.856979`.

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

P. J. Bickel, E. Levina, et al. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10 (6):989–1010, 2004.

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*, 108 (501):265–277, Mar. 2013. doi: 10.1080/01621459.2012.758041.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

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.

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*, 38(2):808–835, 2010.

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

A. P. Dempster. A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, pages 995–1010, 1958.

E. Dobriban, S. Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 962–994, 2004.

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

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

L. Feng and F. Sun. A note on high-dimensional two-sample test. *Statistics & Probability Letters*, 105:29–36, 2015.

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

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

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

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.

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

P. Golland and B. Fischl. Permutation tests for classification: towards statistical significance in image-based studies. In *IPMI*, volume 3, pages 330–341. Springer, 2003.

P. Golland, F. Liang, S. Mukherjee, and D. Panchenko. Permutation Tests for Classification. In P. Auer and R. Meir, editors, *Learning Theory*, number 3559 in Lecture Notes in Computer Science, pages 501–515. Springer Berlin Heidelberg, June 2005. ISBN 978-3-540-26556-6 978-3-540-31892-7. doi: 10.1007/11503415_34.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, Oct. 1999. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.286.5439.531.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-sample Test. *J. Mach. Learn. Res.*, 13:723–773, Mar. 2012. ISSN 1532-4435.

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

Z. Harchaoui, E. Moulines, and F. R. Bach. Kernel change-point analysis. In *Advances in neural information processing systems*, pages 609–616, 2009.

J. Hemerik and J. Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018.

H. Hotelling. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2 (3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177732979.

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

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

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

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

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

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

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

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

A. Moscovich, B. Nadler, C. Spiegelman, et al. On the exact berk-jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

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

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

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

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

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

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

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

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

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

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

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

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)*, 67(4):515–530, Aug. 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00513.x.

J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

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

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

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

R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73 (3):751–754, 1986.

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

M. S. Srivastava and M. Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402, Mar. 2008. ISSN 0047-259X. doi: 10.1016/j.jmva.2006.11.002.

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

G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5 (16.10), 2004.

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

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

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

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

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

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

C. Zheng, R. Achanta, and Y. Benjamini. Extrapolating expected accuracies for large multi-class problems. *The Journal of Machine Learning Research*, 19(1):2609–2638, 2018.

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*, 41(6):2820–2851, 2013.

[*xxx*]