Biostatistics (2019) **0**, 0, *pp.* 1–16
doi:10.1093/biostatistics/kxz035

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, P.O. 653, Beer Sheva, 84105 Israel

johnros@bgu.ac.il

YUVAL BENJAMINI

Department of Statistics, Hebrew University, Mount Scopus, Jerusalem 9190501, Israel

ROEE GILRON

Movement Disorders and Neuromodulation Center, University of California, 1635 Divisadero St, San Francisco, CA 94115, USA

ROY MUKAMEL

School of Psychological Sciences, and Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv 69978, Israel

JELLE J. GOEMAN

Department of Biomedical Data Sciences, Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, 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.

Keywords: High dimension; Multivariate testing; Neuroimaging; Statistical genetics; Supervised learning.

*To whom correspondence should be addressed.

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 its predictive accuracy using cross-validation (CV), 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 (e.g. Eric *and others*, 2008; Lopez-Paz and Oquab, 2016). Examples in the genetics literature include Golub *and others* (1999), Yu *and others* (2007), Jiang *and others* (2008), and many more. Examples in the neuroscientific literature, which is our motivating use-case, include Golland and Fischl (2003), Pereira *and others* (2009), the very popular *multivariate pattern analysis* (MVPA) framework in Kriegeskorte *and others* (2006), and many more.

To fix ideas, we will adhere to a concrete example. In Gilron *and others* (2017), 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. 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 distributions, 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 *and others* (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 *and others* (2013) compare the kernel test of Gretton *and others* (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 *and others* (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 *and others* (2016) conclude LDA and T^2 are rate-equivalent. Rates, however, are only a first stage when comparing test statistics. Two statistics may be rate-equivalent, yet one much more efficient than the other.

We study the power of many accuracy, and two-sample tests, in a large scale simulation study. This allows us to evaluate theoretical results such as Ramdas *and others* (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 the study of multi-group problems can be derived from multiple binary decisions (Zheng *and others*, 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. The problem is formalized in Section 2. The main findings are reported in Sections 3 and 4, with extensions in the Supplementary material available at *Biostatistics* online. We conclude with a discussion.

2. Problem setup

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 tests is *power*, the probability to infer that $\mathcal{F} \neq \mathcal{G}$.

From a test statistic to a permutation test: The tests we consider rely on fixing some test statistic, $\mathcal{T}$, and comparing its observed value to its 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.

The sketch of our permutation test is the following: (i) Fix a test statistic $\mathcal{T}$ with a right tailed rejection region. (ii) Sample a random permutation of the class labels, $\pi(y)$. (iii) Permute labels and recompute the statistic $\mathcal{T}_\pi$. (iv) Repeat (ii)–(iii) R times. (v) The permutation p-value is the proportion of $\mathcal{T}_\pi$ larger than the observed: $\frac{1}{R}\sum_\pi I\{\mathcal{T}_\pi \geq \mathcal{T}\}$. (vi) Declare $\mathcal{F} \neq \mathcal{G}$ if the permutation p-value is smaller than α, which we set to $\alpha = 0.05$.

Two-group tests: The most prevalent interpretation of $\mathcal{F} \neq \mathcal{G}$ is to assume they differ in means, i.e., a *shift class* of alternatives. 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). 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).

In high dimension, when n is not much larger than p, the T^2 test is very low powered (Bai and Saranadasa, 1996). Many high-dimensional versions of the T^2 test exist, which consist of regularizing the estimator of Σ. Examples of high-dim tests for (dense) shifts include Dempster (1958), Bai and Saranadasa (1996), Schäfer and Strimmer (2005), Goeman *and others* (2006), Srivastava and Du (2008), and many more. If $\mathbb{E}(x_1)$ differs from $\mathbb{E}(x_0)$ in few coordinates, we say the *signal is sparse*. Examples of high-dim test statistics for sparse shifts include Cai *and others* (2013) and Chang *and others* (2017). 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), and many more.

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), Gretton *and others* (2012), and many more.

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

Prediction accuracy as a test statistic: An accuracy-test amounts to using a predictor's accuracy as a test statistic. Denoting a data set 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 data set $\mathcal{S}$. 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, let $\mathcal{P}$ be 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 $\mathcal{E}_{\mathcal{A}_\mathcal{S}} := \int_{(x,y)} \mathcal{I}\{\mathcal{A}_\mathcal{S}(x) = y\}\, \mathrm{d}\mathcal{P}$, and $\mathcal{E}_\mathcal{A} := \int_\mathcal{S} \mathcal{E}_{\mathcal{A}_\mathcal{S}}\, \mathrm{d}\mathcal{P}_\mathcal{S}$, 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 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 data set, 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.

How to estimate accuracies? Estimating $\hat{\mathcal{E}}_\mathcal{A}$ requires design choices regarding the estimation of accuracies. In particular, are the accuracies estimated via CV? 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. We will thus be considering both unbiased cross-validated accuracies, and biased resubstitution accuracies. For V-fold CV, we will use V = 4 and will constrain the data folds to be balanced, a.k.a. stratified. More on resampling estimators of accuracy, in the Supplementary material available at *Biostatistics* online.

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.

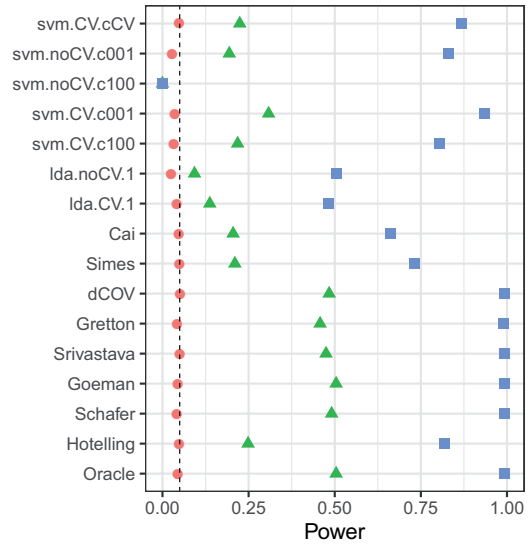

Fig. 1. The power of the permutation test with various test statistics. The power on the x-axis. Effects are color and shape coded. Effects vary over $\frac{n}{2}\|\mu\|_\Sigma^2 = 0$ (red circle), 25 (green triangle), and 100 (blue square). The various statistics on the y-axis. Their details are given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

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 https://github.com/johnros/better_than_chance_code (commit 13ceaf).

3.1. *Basic simulation setup—Fisher's LDA*

The basic simulation setup is essentially the sampling distribution underlying Fisher's LDA. 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 permutations (with one exception, explained in Section 3.1.1). 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. We use $R = 1000$ replicates, so that the standard errors of our estimates are $\leq$0.6% under the null and $\leq$1.5% in general.

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 and others, 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*

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 *and others* (2013)	Resubstitution	—
Simes	Simes (1986)	Resubstitution	—
dCOV	Székely and Rizzo (2004)	Resubstitution	—
Gretton	Gretton *and others* (2012)	Resubstitution	—
Srivastava	Srivastava and Du (2008)	Resubstitution	—
Goeman	Goeman *and others* (2006)	Resubstitution	—
Schafer	Schäfer and Strimmer (2005)	Resubstitution	—
Hotelling	Hotelling (1931)	Resubstitution	—
Oracle	T^2 with Known Σ	Resubstitution	—

3.1.1. *False positive rate* We start with a sanity check. Theory suggests that a (random) permutation test with the identity permutation is slightly conservative, and without the identity, it is slightly liberal. Theory also suggests that this bias vanishes with the number of permutations (Hemerik and Goeman, 2018). We thus ran the initial simulation setup with 1000 permutations, and confirmed that all permutation tests control their false positive rates. This can be seen in Figure 1, where the power under the null (red circles) is no larger than the nominal error rate of $\alpha = 0.05$. We may thus proceed and compare the power of each test statistic.

3.1.2. *Power* In our first simulation setup, two-group tests are more powerful than accuracy-tests (Figure 1). 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 Figure 1, and reported in the Supplementary material available at *Biostatistics* online.

3.2. *Departure from Gaussianity*

Hotelling's T^2 is a generalized likelihood ratio test in the Gaussian shift class. This Neyman–Pearson Lemma type reasoning that favors two-group location-tests over accuracy-tests in our simulations may

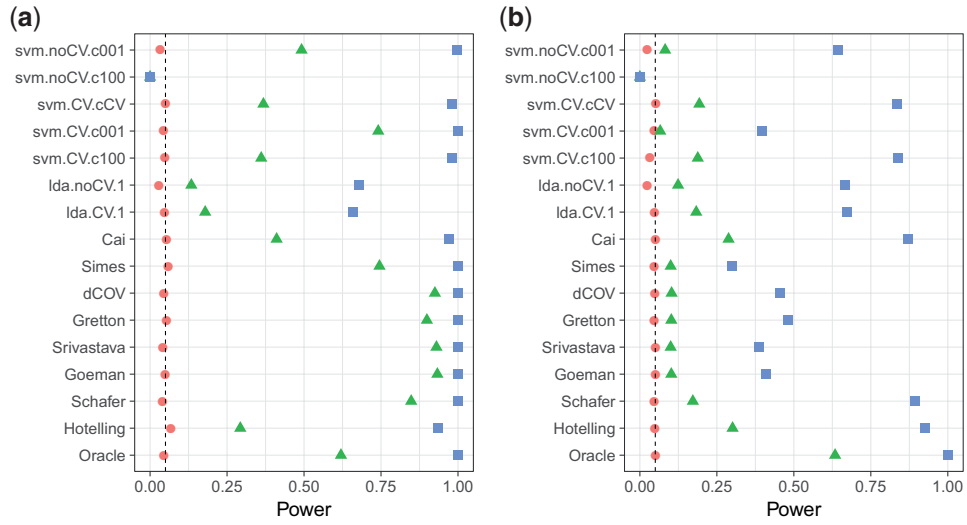

Fig. 2. Short memory, AR(1) correlation. $\Sigma_{k,l} = \rho^{|k-l|}$; $\rho = 0.6$. (a) Signal in direction of highest variance PC of Σ. (b) Signal in direction of lowest variance PC of Σ.

fail when the data are not Gaussian. To verify our conclusions in the non-Gaussian case, we replaced the multivariate Gaussian distribution of η in (3.1) with a heavy-tailed multivariate-t distribution with 3 degrees of freedom. In this heavy-tailed setup, the dominance of the two-group tests was preserved, even if less evident than in the light-tailed Gaussian case. Results are in the Supplementary material available at *Biostatistics* online.

3.3. *Departure from sphericity*

We now test the robustness of our results to correlations in x, In terms of (3.1), we use various correlation structures in Σ. 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, we kept $\frac{n}{2}\|\mu\|_{\Sigma}^2$ fixed. Note that this induces differences in the Euclidean norm between population means $\|\mu\|_2$ between the two settings. In the Supplementary material available at *Biostatistics* online, we report the power when fixing $\|\mu\|_2$ instead.

The simulation results reveal some non-trivial phenomena. When the signal is in the direction of the high variance PC, the high-dim two-group tests are far superior than accuracy-tests (Figure 2(a)). When the signal is in the direction of the low variance PC, there is no clear preference between two-group or accuracy-tests (Figure 2(b)). Instead, the non-regularized tests are the clear victors. We attribute this phenomenon to the bias introduced by the regularization, which masks the signal (see Section 5.3.1).

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: (i) Many effects are "pure shifts" after a scale transformation. For instance, a multiplicative effect in log scale. (ii) 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.

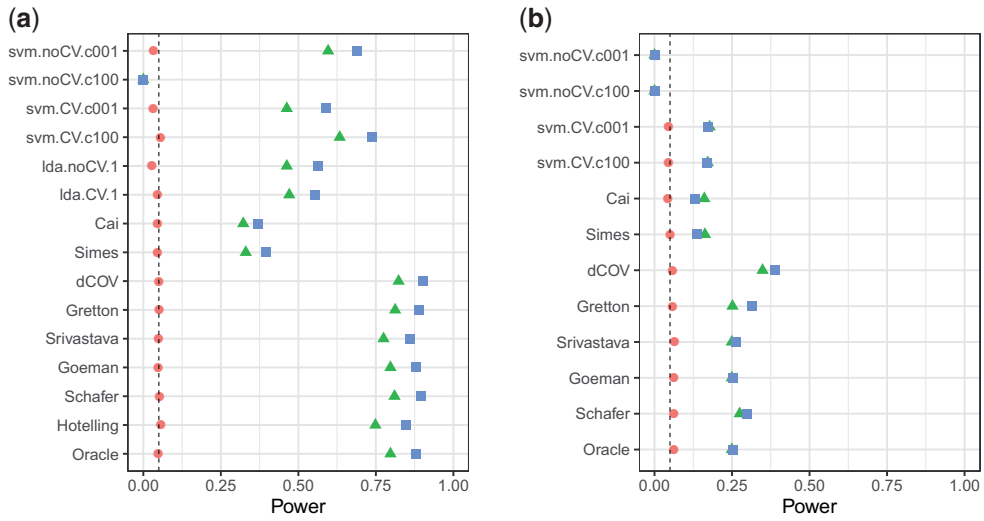

Fig. 3. Logistic regression with second order interactions. Data generated via $y|x \sim Binom(1, p(x)); p(x) = \exp(\eta)/[1+\exp(\eta)]; \eta = x'\beta + x'Bx$ where β is a scaled vector of ones, and B a scaled identity matrix. Finally, $x \sim \mathcal{N}(0, I_{p\times p})$. (a) Data analyzed in the 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 omitted.

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 3, 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 3b is an example of the latter. In the Supplementary material available at *Biostatistics* online, we provide an example of the former by simulating a logistic regression with main effects only.

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* (Friedman *and others*, 2001, Section 7.11).

DEFINITION 3 (bLOO) 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: $\hat{\mathcal{E}}_{\mathcal{A}}^{bLOO} := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|C^{(i)}|} \sum_{b\in C^{(i)}} \mathcal{I}\{\mathcal{A}_{\mathcal{S}^b}(x_i) = y_i\}$.

Simulation results are reported in Figure 4 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 CV, but

Table 2. *The same as Table 1 for bootstrapped accuracy estimates. bLOO is defined in Definition 3. B denotes the number of Bootstrap samples. Accuracy-tests marked with a* ✣

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

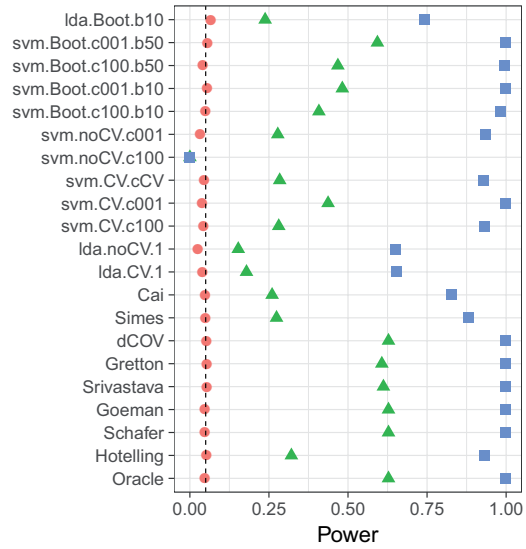

Fig. 4. Bootstrap. The power of a permutation test with various test statistics. The power on the x-axis. Effects are color and shape coded. The various statistics on the y-axis. Their details are given in Tables 1 and 2. Effects vary over $\frac{n}{2}\|\mu\|^2_\Sigma = 0$ (red circle), 25 (green triangle), and 100 (blue square). Simulation details in Section 3.1.

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.

3.6. *High-dim regularized accuracy-tests*

Our best performing tests regularize the estimation of Σ. In our high-dim setup regularization adds power, as seen by comparing the non-regularized T^2 to its regularized versions. Regularization is achieved by thresholding the entries of $\hat{\Sigma}$ (*Goeman*, *Srivastava* statistics), or inflating the diagonal of $\hat{\Sigma}$ (*Schaffer*).

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 and Levina (2004) or Dobriban and Wager (2018). We thus augment our simulations with some accuracy-tests that have explicit covariance regularization in them. These include shrinkage-based LDA (Pang *and others*, 2009; Ramey *and others*, 2016), where Tikhonov regularization of $\hat{\Sigma}$ is used;

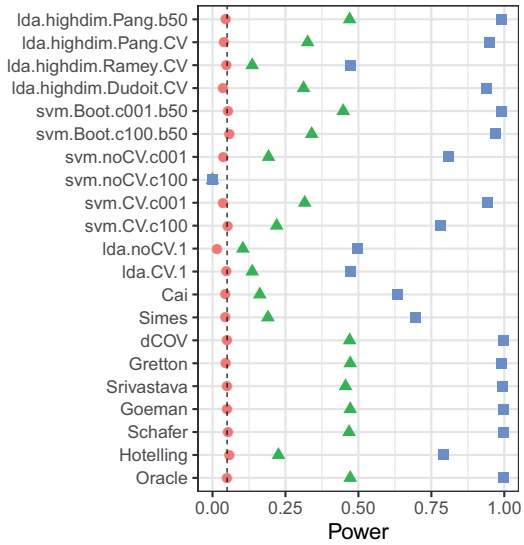

Fig. 5. HighDim Classifier. The power of a permutation test with various test statistics. The power on the x-axis. Effects are color and shape coded. The various statistics on the y-axis. Their details are given in Tables 1 and 3. Effects vary over $\frac{n}{2}\|\mu\|_{\Sigma}^2 = 0$ (red circle), 25 (green triangle), and 100 (blue square). Simulation details in Section 3.1.

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

Name	Algorithm	Resampling	Parameters
✣lda.highdim.Dudoit.CV	Dudoit *and others* (2002)	V-fold	—
✣lda.highdim.Ramey.CV	Ramey *and others* (2016)	V-fold	—
✣lda.highdim.Pang.CV	Pang *and others* (2009)	V-fold	—
✣lda.highdim.Pang.b50	Pang *and others* (2009)	bLOO	B=50

just like the *Schafer* two-group test. We also try a diagonalized LDA (Dudoit *and others*, 2002), a.k.a *Gaussian Naïve Bayes*, which regularizes by canceling non-diagonal entries.

Simulation results are reported in Figure 5 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. 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.

4. NEUROIMAGING EXAMPLE

Figure 6 is an application of (i) the Srivastava two-group test, and (ii) a linear SVM accuracy-test, to the neuroimaging data of Pernet *and others* (2015). The authors of Pernet *and others* (2015) collected fMRI

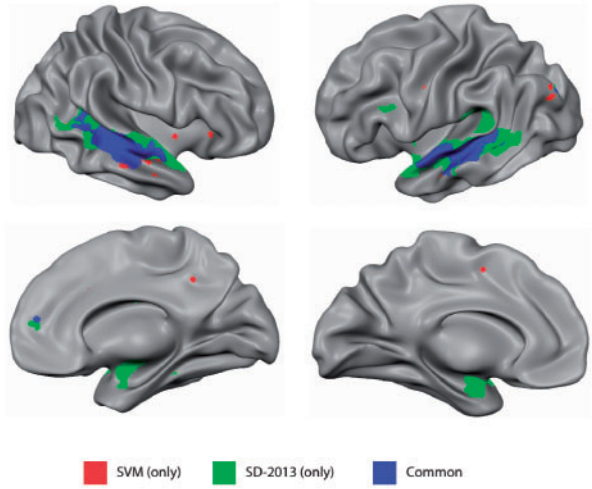

Fig. 6. Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centers of 27-voxel sized spherical regions, as discovered by an accuracy-test u and a two-group test (*Srivastava*). The linear SVM was computed using 5-fold CV, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of Stelzer *and others* (2013), followed by region-wise FDR ≤ 0.05 control using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). Number of permutations equals 400. The two-group test detect 1232 regions, and the accuracy-test 441, 399 of which are common to both. For the details of the analysis, see Gilron *and others* (2017).

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 were 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 *and others* (2013). Our test statistics account for dependence in space, but require independence in time. Parameters were estimated with an orthogonal design, and an AR(6) temporal model. Further details of the analysis, are reported in Gilron *and others* (2017).

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 1232 regions, while the latter only 441, as depicted in Figure 6. 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 tests 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.

5.1.1. *Data splitting* CV 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 a 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).

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 than 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 and Schwartzman (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 tests with discrete test statistics are known to be conservative. Firstly, a Monte-Carlo sample of permutations is conservative compared to a full enumeration of permutations (Hemerik and Goeman, 2018). Secondly, the presence of ties does not allow to exhaust the permissible false positive rate, unless randomization is introduced. Thirdly, a discrete test statistic is less sensitive to mild perturbations of the data. For intuition, consider using *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 both the observed data and most label-permuted data. The concentration of accuracy scores near 1, and its discretization, render this test completely useless: power tends to 0 for any (fixed) effect size as p grows. This explains the terrible power of *svm.noCV.c100* that is effectively unregularized.

We observe that the power loss due to discretization may be considerable. We compare Figher's LDA to Hotelling's T^2, which have comparable resubstitution accuracy after binarizing the predictions. For intermediate signals strength (Figure 1), *Hotelling* has roughly twice the power of LDA (*lda.noCV.1*). Note that this power loss due to discretization will not be captured by asymptotic analyses such as Ramdas *and others* (2016), because the discretization decreases with sample size.

5.2. *A good accuracy-test*

Often we want to know if a particular predictor can extract information in a region. Examples include brain–computer interfaces and clinical diagnostics (Olivetti *and others*, 2012; Wager *and others*, 2013). In those cases, we may prefer accuracy-tests. Here are some observations for increasing power in accuracy-tests:

Test-set size: Larger test-sets reduce the effect of discretization on the power of accuracy-tests.

Regularize: Regularization proves crucial to detection power in low SNR regimes ($n \approx p$) or under strong correlations. We find that shrinkage-based diagonal LDA (Pang *and others*, 2009) performs well overall. More research is required on optimal regularization for testing.

Resample with replacement: Smoothing the accuracy estimates by cross-validating with replacement (e.g. the bLOO method) improves power for accuracy-tests compared to V-Fold. We believe this is primarily due to the smoothing effect.

5.3. *Additional comments*

5.3.1. *Effect of covariance regularization* Figure 2 demonstrates that detecting signal in the direction of the high variance PCs is very different than detecting in the low variance PCs. We attribute this phenomenon to regularization. Whereas the signal, μ, varies in direction, the regularization of $\hat{\Sigma}$ does not. From ridge regression, we know that Tikhonov regularization of the covariance shrinks estimates more aggressively in the low PCs of the design. Signal is thus masked if the difference between group means is the directions of smaller variance. In those cases, unregularized tests dominate the regularized ones.

5.3.2. *Sparse alternatives* 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 available at *Biostatistics* online, 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, tests for sparse shifts (*Cai*, *Simes*) dominate the T^2 type statistics.

5.3.3. *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 *and others*, 2012), is an example where a two-group test is performed after an implicit augmentation of x to some RKHS. More generally, the design matrix may be augmented as we please, up to computational considerations. We thus disagree with the argument that accuracy-tests have more flexibility than two-group tests. For example, in Section 3.4, we analyze the data both in the original space and in an augmented space.

A different argument is that the feature mapping may not be known, but rather learned from the data. This is true in low-dimension, where data are abundant compared to the model's complexity. In high-dim problems data are barely sufficient to learn covariances in the original space, let alone to learn a space augmentation and covariances in the augmented space.

5.4. *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.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

J.D.R. wishes to thank, Jesse B.A. Hemerik, Yakir Brechenko, Omer Shamir, Joshua Vogelstein, Gilles Blanchard, and Jason Stein for their valuable inputs.
Conflict of Interest: None declared.

FUNDING

The Israeli Science Foundation (900/16 and 924/16 to J.D.R.); NIH (R01GM083084 to Y.B.).

REFERENCES

ANDERSON, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*, 3rd edition. Hoboken, NJ: Wiley-Interscience. ISBN 978-0-471-36091-9.

BAI, Z. AND SARANADASA, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B* **57**, 289–289.

BIAU, G. AND GYORFI, L. (2005). On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, **51**, 3965–3973.

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

BICKEL, P. J. AND LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

CAI, T., LIU, W. AND XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association* **108**, 265–277.

CHANG, C.-C. AND LIN, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**, 27.

CHANG, J., ZHENG, C., ZHOU, W.-X. AND ZHOU, W. (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics*, **73**, 1300–1310.

CHENG, D. AND SCHWARTZMAN, A. (2017). Multiple testing of local maxima for detection of peaks in random fields. *The Annals of Statistics*, **45**, 529–556.

DEMPSTER, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, **29**, 995–1010.

DOBRIBAN, E. AND WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, **46**, 247–279.

DONOHO, D. AND JIN, J. S. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, **32**, 962–994.

DUDOIT, S., FRIDLYAND, J. AND SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

ERIC, M., BACH, F. R. AND HARCHAOUI, Z. (2008). Testing for homogeneity with kernel fisher discriminant analysis. *Advances in Neural Information Processing Systems*, 609–616.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2001). *The Elements of Statistical Learning*, Volume 1. Springer Series in Statistics. New York: Springer.

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

FRIEDMAN, J. H. AND RAFSKY, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, **7**, 697–717.

GILRON, R., ROSENBLATT, J., KOYEJO, O., POLDRACK, R. A. AND MUKAMEL, R. (2017). What's in a Pattern? Examining the type of signal multivariate analysis uncovers at the group level. *NeuroImage* **146**, 113–120.

GOEMAN, J. J., VAN DE GEER, S. A. AND VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 477–493.

GOLLAND, P. AND FISCHL, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. In: *Information Processing in Medical Imaging*, Volume 3. Springer, pp. 330–341.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. AND LANDER, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. AND SMOLA, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, **13**, 723–773.

HALL, P. AND TAJVIDI, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, **89**, 359–374.

HEMERIK, J. AND GOEMAN, J. (2018). Exact testing with random permutations. *TEST*, **27**, 811–825.

HOTELLING, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, **2**, 360–378.

JIANG, W., VARMA, S. AND SIMON, R. (2008). Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, **7**, doi: 10.2202/1544-6115.1322.

KRIEGESKORTE, N., GOEBEL, R. AND BANDETTINI, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3863–3868.

LOPEZ-PAZ, D. AND OQUAB, M. (2016). Revisiting classifier two-sample tests. ICLR.

MEYER, D., DIMITRIADOU, E., HORNIK, K., WEINGESSEL, A. AND LEISCH, F. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7.

OLIVETTI, E., GREINER, S. AND AVESANI, P. (2012). Induction in neuroscience with classification: issues and solutions. In: LANGS, G., RISH, I., Grosse-WENTRUP, M. and MURPHY, B. (editors), *Machine Learning and Interpretation in Neuroimaging*, number 7263. Lecture Notes in Computer Science. Berlin Heidelberg: Springer, pp. 42–50.

OLIVETTI, E., BENOZZO, D., KIA, S. M., ELLERO, M. AND HARTMANN, T. (2013). The kernel two-sample test vs. brain decoding. In: *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 128–131.

PANG, H., TONG, T. AND ZHAO, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, **65**, 1021–1029.

PEREIRA, F., MITCHELL, T. AND BOTVINICK, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, **45**, S199–S209.

PERNET, C. R., MCALEER, P., LATINUS, M., GORGOLEWSKI, K. J., CHAREST, I., BESTELMEYER, P. E. G., WATSON, R. H., FLEMING, D., CRABBE, F., VALDES-SOSA, M. AND BELIN, P. (2015). The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, **119**, 164–174.

RAMDAS, A., SINGH, A. AND WASSERMAN, L. (2016). Classification accuracy as a proxy for two sample testing. *arXiv:1602.02210 [cs, math, stat]*.

RAMEY, J. A., STEIN, C. K., YOUNG, P. D. AND YOUNG, D. M. (2016). High-dimensional regularized discriminant analysis. *arXiv preprint arXiv:1602.01182*.

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

SCHÄFER, J. AND STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1–32.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

SRIVASTAVA, M. S. AND DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, **99**, 386–402.

STELZER, J., CHEN, Y. AND TURNER, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage*, **65**, 69–82.

SZÉKELY, G. J. AND RIZZO, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, **5**, 1249–1272.

WAGER, T. D., ATLAS, L. Y., LINDQUIST, M. A., ROY, M., WOO, C.-W. AND KROSS, E. (2013) An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, **368**, 1388–1397.

YU, K., MARTIN, R., ROTHMAN, N., ZHENG, T. AND LAN, Q. (2007). Two-sample comparison based on prediction error, with applications to candidate gene association studies. *Annals of Human Genetics*, **71**, 107–118.

ZHENG, C., ACHANTA, R. AND BENJAMINI, Y. (2018). Extrapolating expected accuracies for large multi-class problems. *The Journal of Machine Learning Research* **19**, 2609–2638.

[*Received November 9, 2018; revised August 9, 2019; accepted for publication August 14, 2019*]