Better-Than-Chance Classification for Signal Detection

Jonathan Rosenblatt Roee Gilron Roy Mukamel

August 4, 2016

Abstract

[TODO]

1 Introduction

A common workflow in neuroimaging consists of fitting a classifier, and estimating its predictive accuracy using cross validation. Given that the cross validated accuracy is a random quantity, it is then common to test if the cross validated accuracy is significantly better than chance using a permutation test. Examples in the neuroscientific literature include Golland and Fischl [2003], Pereira et al. [2009], Varoquaux et al. [2016], and especially the recently populirized *multivariate pattern analysis* (MVPA) framework of Kriegeskorte et al. [2006]. This practice is also observed in the genetics literature, but to a lesser extent [Radmacher et al., 2002, Jiang et al., 2008].

To fix ideas, we will adhere to a concrete example. In Gilron et al. [2016], the authors seek to detect brain regions which encode differences between vocal and non-vocal stimuli. Following the MVPA workflow, the localization problem is cast as a supervised learning problem: if the type of the stimulus can be predicted from the spatial activation pattern significantly better than chance, then a region is declared to encode vocal/non-vocal information. We call this an *accuracy test*, a.k.a. *class prediction* in Simon et al. [2003], or *pattern discrimination* in Pereira et al. [2009].

This same signal detection task can be also approached as a two-group multivariate test. Inferring that a region encodes vocal/non-vocal information, is essentially inferring that the spatial distribution of brain activations is different given a vocal/non-vocal stimulus. As put in Pereira et al. [2009]:

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

A practitioner may then call upon a two-group location test such as Hotelling's T^2 [Anderson, 2003]. Alternatively, if the size of a brain region is too large compared to the number of observations, so that the spatial covariance cannot be fully estimated, then a high dimensional version of Hotelling's test can be called upon, such as in Schäfer et al. [2005] or Srivastava [2013]. For breivity, and in contrast to *accuracy tests*, we will call these two-sample multivaraite tests simply *location tests*, also termed *class comparisons* in Simon et al. [2003].

At this point, it becomes unclear which is the preferred test. The comparison between location and accuracy tests was precisely the topic of Ramdas et al. [2016], who compared the Hotelling location test to the accuracy of *Fisher's linear discriminant analysis* classifier (LDA) [Hastie et al., 2003]. Using an asymptotic analysis, Ramdas et al. [2016] concluded that accuracy and location tests are equivalent with respect to their order of convergence to a consistent test, while they differ in constants. Judging by rate of convergence alone, this result may suggest that not much is (asymptoticaly) lost by using an accuracy test. On the other hand, asymptotic relative efficency measures (ARE) such as *Pitman's*, *Bahadur's*, and *Hodges-Lehman's*, always assume equivalent convergence rates [van der Vaart, 1998].

In Ramdas et al. [2016] setup, the ARE between Hotelling's T^2 (location) test and Fisher's LDA (accuracy) test is lower bounded by $\sqrt{2\pi} \approx 2.5$. This means that Fisher's LDA requires at least 2.5 more samples to achieve the same (asymptotic) power than the T^2 test. Clearly, the accuracy test is remarakbly inneficient, even when the discretization effecet has been cancelled by asymptotics. For comparison, the t-test is only 1.04 more (asymptotically) efficienct than Wilcoxon's rank-sum test [Lehmann, 2009]. Admittidly, Ramdas et al. [2016]'s results hold for LDA with a half-sample holdout. This suggests that the ARE of leave-one-out validation, for instance, will be closer to 1. We revisit this matter in the discussion section.

The relative efficiency, governing the power of the tests, may prove crucial when dealing with the finite sample sizes in neuroscience and genetics, and thus the focus of this study. We thus seek to study which test is to be preferred in finite samples? Our conclusion will be quite simple: *location tests almost always have more power than accuracy tests.*

The main argument for our statement rests upon the observation that with typical sample sizes, the accuracy test statistic is highly discrete. Discrete test statistics are known to be conservative [**?**], since they cannot exhaust the permissible false positive rate. For accuracy tests, the degree of discretization is governed by the number of samples. In our running neuroscience example [Gilron et al., 2016], the classification is performed based on 40 trials, so that the test statistic may assume only 40 possible values. This number of examples is not unusual if considering this is the number of subject in a genetic study, or the number of trial-repeats in an fMRI brain scan.

The discretization effect is aggravated if the test statistic is highly concentrated. For an intuition consider the usage of the *train* accuracy test statistic (i.e., not cross validated). In Section 4 we then address our main question- which test has more power? Based on the finding that the location test is typically more powerful, we try to offer an intuition for this phenomenon in the Discussion section.

2 Problem setup

Adhering to our neuroscientific example, we now formalize terminology and notation. Let $y \in \mathcal{Y}$ be a class encoding. In our vocal/non-vocal example we have $\mathcal{Y} = \{-1, 1\}$. Let $x \in \mathcal{X}$ be a p dimensional feature vector. In our vocal/non-vocal example p is the number of voxels in a brain region. We thus have $\mathcal{X} = \mathbb{R}^{27}$.

Given n pairs of (x_i, y_i), typically assumed i.i.d., a location test amounts to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$ (or at least the same location). I.e., the multivariate voxel activation pattern has the same distribution when given a vocal stimulus, as when given a non-vocal stimulus. An accuracy test amounts to learning a predictive model $\hat{f}(x)$ from some assumed model class $\hat{f} \in \mathcal{F}$. The prediction accuracy, denoted $T^{acc}_{\hat{f}}$, is defined as the probability of a given classifier $\hat{f}$ of making a correct prediction $T^{acc}_{\hat{f}} := Prob(\hat{f}(x) = y)$ when given a new, randomly drawn data point, (x, y). A statistically significant "better than chance" estimate of $T^{acc}_{\hat{f}}$ is evidence that the classes are distinct.

2.1 Candidate Tests

The design of a permutation test using the prediction accuracy, requires the following design choices:

1. How to estimate accuracy?

2. Is the statistic cross validated or not?

3. For a K-fold cross validated test statistic: should the data be refolded in each permutation?

4. Permute labels of features?

5. For a K-fold cross validated test statistic: should the data folding balanced? (a.k.a. stratified).

6. How many folds?

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 the prediction error, but rather in the mere detection of a difference between two groups, leading to a better-than-chance accuracy.

How to estimate accuracy? Given a predictor $\hat{f}$, a natural test statistic is some estimate of its accuracy $T_{\hat{f}}^{acc}$. Complicating matters: very low accuracies, even 0, is evidence that the classes are separated, and we only need to invert the predictions. We can thus consider $|T_{\hat{f}}^{acc} - 0.5|$ as the test statistic. This, however, implies that if the classes are identical, random guessing has a 0.5 accuracy. This is not true if the classes are not balanced. The chance level in which case is the prevalence of the dominant class, we denote by $\hat{p}_{max}$. This suggests the following test statistic $|T_{\hat{f}}^{acc} - \hat{p}_{max}|$. Since we will be aggregating these statistic over random data sets where the dominant class may have varying frequencies, it seems appropriate to standardize the scale of this statistic. We thus also consider the z-scored accuracy: $|T_{\hat{f}}^{acc} - \hat{p}_{max}|/\sqrt{\hat{p}_{max}(1 - \hat{p}_{max})}$.

Cross validate or not? Were we interested in an unbiased estimator of the prediction error, there is no question that some independent validation is in order. Since we are merely interested in detecting a difference between classes, a biased error estimate is not an issue provided that bias is consistent over all permutations. The underlying intuition is that if the exact same computation is performed over all permutations, then a permutation test will be “fair”, i.e., will not inflate the false positive rate. We will thus be considering both cross validated accuracies, and *train* accuracies as our test statistics, a.k.a. *resubstitution classification* in Ramdas et al. [2016].

Refolding? The standard practice in neuroimaging is to refold the data after each permutation [Pereira et al., 2009]. This is imperative if permuting labels while aiming at balanced data folds. This is not, however, imperative in general. For simplicity, we will adhere to the standard practice of refolding the data within each permutation.

Permute labels of features? While seemingly identical, the compounding of permutations with data foldings renders these two approaches distinct. As an example, consider balanced (stratified) K-fold cross validation where the initial data folding is balanced. After a label permutation, the original folds will probably not be balanced. If the *features* are permuted, then the labels conserve their original fold assignments, and the original folds are balanced after each permutation. Since we only report results while refolding the data in each permutation, then the only difference between permuting labels and permuting features seems to be a computational one. We thus adhere to the more common, albeit less efficient practice, of permuting labels.

Balanced folding? As already implied, a standard practice when cross validating is to constrain the data folds to be balanced (i.e. stratified). This is well justified when aiming at unbiased accuracy estimation. This also simplifies matter when aiming at signal detection, as can be seen from the above discussion of the appropriate test statistic. On the other hand, it may complicate matters, as can be seen from the above discussion on label versus feature permutation. We will report results with both balanced and unbalanced data foldings, only to discover, it does not really matter.

How many folds? Different authors suggest different rules for the number of folds. We will be varying the number of folds. This will affect the concentration of permutation distribution of the estimated accuracy, which will have a crucial effect on the conservativeness of the accuracy test. Our intuition suggests that since more folds imply a less concentrated estimate, then leave-one-out should be the less conservative, and 2-fold should be the most conservative.

There are indeed many design choices when performing a permutation test using a cross validated statistic. The subset of tests we will be comparing is collected for convenience in Table 1.

3 Controlling the False Positive Rate

We start by verifying that the battery of tests in Table 1 control the false positive rate at the desired 0.05 level, with varying conservativeness levels. Figure 1 demonstrates that this is indeed the case. All our candidate tests

Name	Basis	CV	Accuracy	Parameters
Hotelling	Hotelling	–	–	shrink=FALSE
Hotelling.shrink	Hotelling	–	–	shrink=TRUE
lda.CV.1	LDA	TRUE	accuracy	–
lda.CV.2	LDA	TRUE	z-accuracy	–
lda.noCV.1	LDA	FALSE	accuracy	–
lda.noCV.2	LDA	FALSE	z-accuracy	–
sd	SD	–	–	–
svm.CV.1	SVM	TRUE	accuracy	cost=1e1
svm.CV.2	SVM	TRUE	accuracy	cost=1e-1
svm.CV.3	SVM	TRUE	z-accuracy	cost=1e1
svm.CV.4	SVM	TRUE	z-accuracy	cost=1e-1
svm.noCV.1	SVM	FALSE	accuracy	cost=1e1
svm.noCV.2	SVM	FALSE	accuracy	cost=1e-1
svm.noCV.3	SVM	FALSE	z-accuracy	cost=1e1
svm.noCV.4	SVM	FALSE	z-accuracy	cost=1e-1

Table 1: This table enumerates the various test statistics we will be studying. Three are location tests: Hotelling, Hotelling.shrink, and sd. *Hotelling* is the classical two-group T^2 statistic. *Hotelling.shrink* is a high dimensional version with the regularized covariance in Schäfer et al. [2005]. *sd* is another high dimensional version of the T^2, from Srivastava et al. [2013]. The rest of the tests are variations of the linear SVM, and Fisher's LDA, with varying accuracy measures, cross validated or not, and varying tuning parameters. For example, *svm.CV.4* is a linear SVM, with *libsvm*'s cost parameter set at 0.1, using the cross validated z-scored accuracy ($|T^{acc}_{\hat{f}} - \hat{p}_{max}|/\sqrt{\hat{p}_{max}(1-\hat{p}_{max})}$, see Section 2.1). Another example is *lda.noCV.1*, which is Fisher's LDA, returning the train accuracy, without cross validation, and without z-scoring.

control the type I error, with varying degrees of conservativeness. In particular: (a) if the folds are balanced or not, (b) if the tuning parameters of some test statistic are varied, (d) if the number of folds is varied.

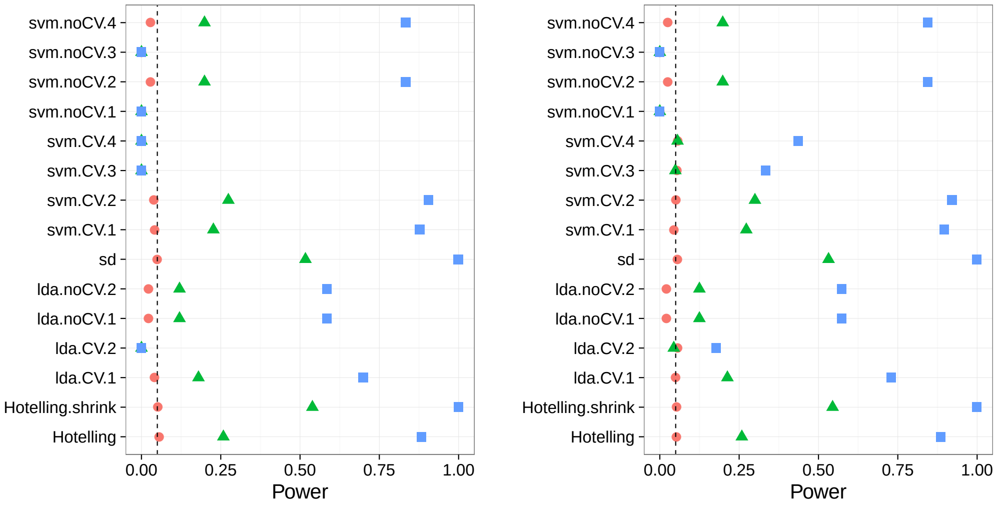
4 Power

Having established that all of the tests in our battery control the false positive rate, it remains to be seen if they have similar power; Especially when comparing the power of location tests to accuracy tests. On the other hand, the results of our previous sections suggest that the conservativeness of some of the considered tests can be considerable, rendering them underpowered.

[TODO: discuss power of various tests after finishing simulations]

We see by now that the use of accuracy tests for signal detection is un-

Figure 1: The power of a permutation test with various test statistics. The power on the x axis. Effect are color and shape coded. They are assumed to be equal in all the 23 dimensions, and vary over 0 (red circle), 0.25 (green triangle), and 0.5 (blue square). The various statistics on the y axis. Their details are given in Table 1. Simulation code available at [TODO].



(a) Unbalanced.

(b) Balanced.

derpowered compared to location tests. Simulations alone cannot, however, support such a universal statement. We will thus verify on a neuroimaging dataset, and discuss the causes for this phenomenon with implications on the scope of our statement.

5 Neuroimaging Example

Figure 2 is an application of both a location and an accuracy test to the 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, totalling in $n = 40$ trials in each scan. The study was rather large and consisted of about 200 subjects. The data was kindly made available by the authors at the OpenfMRI website[1].

We perform permutation inference using the pipeline of Stelzer et al. [2013], which was also used in Gilron et al. [2016]. For completeness, the pipeline is described in Appendix A. To demonstrate our point, we compare

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

the *sd* location test with the *svm.cv.1* accuracy test (see Table 1 for the definition of these statistics).

In agreement with our simulation results, the location test (*sd*) discovers more brain regions when compared to an accuracy test (*svm.cv.1*). The former discovers $1,232$ regions, while the latter only 441, as reported in Figure 2. 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.

Having established that accuracy tests are underpowered both in simulation and in application, we wish to identify the conditions under which this will occur, and discuss implications on the practice of accuracy tests.

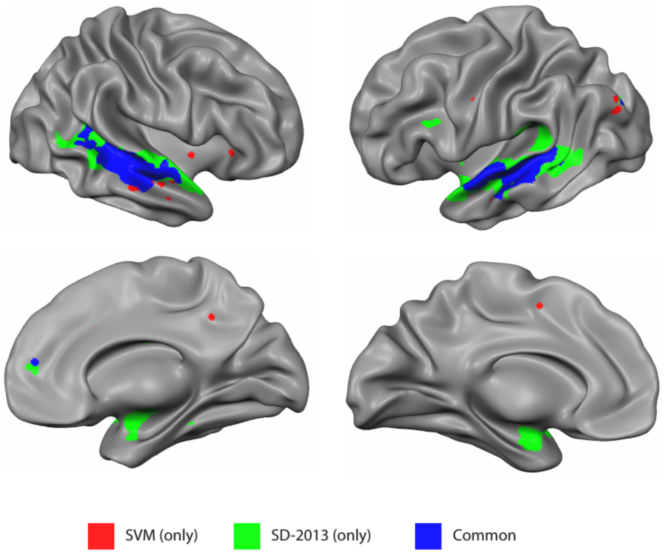


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centres of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a location test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost parameter of 1. Region-wise significance was determined using the permutation scheme of Stelzer et al. [2013], followed by region-wise $FDR \leq 0.05$ control using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. Number of permutations equals 400. The location test detect $1,232$ regions, and the accuracy test 441, 399 of which are common to both. For the details of the analysis see Appendix A and Gilron et al. [2016].

6 Discussion

We have set out to understand which of the tests is more powerful: the accuracy test or the location test. Using simulations, we have concluded that the location tests are preferable. We attribute this to the discretization introduced in finite samples by the accuracy test statistic. This also explains why an asymptotic analysis, such as Ramdas et al. [2016], did not find a rate difference. Their results however are limited in that: (a) they are asymptotic, thus eschew the discretization effect. (b) They assume a half-sample holdout, so that half of the data is available for estimation. (c) They assume a linear classifier.

The linear classifier assumption, (c), is immaterial since for every non-linear clasifier, one may design a non-linear location test. See Gretton et al. [2012] for an example of a location test in RKHS space. [TODO: relate to large sample simulation] [TODO: discuss ARE, and holdout versus leave one out effect]. [TODO: non-linear classification and testing].

Olivetti et al. [2012] and Olivetti et al. [2014] also looked into a similar problem as we do, namely, what is the preferred accuracy test? They propose a new test they call an *independence test*, and demonstrate by simulation that it has more power than other accuracy tests, and can deal with non-balanced data sets. We did not include this test in the battery we compared, but we note the following: (a) The independence test of Olivetti et al. [2012] relies on a discrete test statistic. This means that in the cases that the accuracy test is called upon for discriminating populations, it will probably be under-powered compared to location tests. (b) The problem of the accuracy test with unbalanced data-sets, which motivates Olivetti et al. [2012]'s independence test, can also be remedied by replacing the accuracy statistic with its z-score, as suggested in Section 2.1.

At this point some reservations to the generality of our findings are in order. Firstly, not all accuracy tests are concerned with signal detection. Indeed, it is possible that the purpose of the test is not to detect a difference between classes, but to actually test is a particular classifier is better than chance. This would be the case in decoding applications, like brain-machine interfaces, where the localization a signal is not enough. Clinical diagnosis is another application, where the presence of a medical condition is "predicted" from imaging data. [e.g. Olivetti et al., 2012, Wager et al., 2013]

Secondly, not all signals are manifested in a shift of the null distrubiton. Put differently, the preferred alternative to an accuracy test is not always a location test. Indeed, one may consider signal, i.e. effects, as a change in scale, such as the *spiked covariance* model. In this case, other-than-Hotelling type tests are appropriate [TODO: cite change in covariance alternative].

Tests have been proposed even when the nature of the difference between populations is left unspecified [e.g. Gretton et al., 2012]. The fact that in our neuroimaging example (Section 5) some brain regions were detected with the accuracy test, and not the location test, is consistent with this observation. On the other hand, the far greater power of the location test, certianly in our example, does serve as en empirical evidence that changes in location are a prevalent phenomenon. [TODO: signal in scale? heavy tails?]

A very important point is the ease of implementation. The need for cross validation of the accuracy test greatly increases its computational complexity. Moreover, anyone who has actually implemented tests with discrete statistics, will attest they are considerably harder to implement. This is because their unforgiveness to the type of inequality. Indeed, mistakenly replacing a weak inequality with a strong inequality in one's program may considerably change the results. This is not the case for continuous test statistics.

Given all the above, we find the popularity of accuracy tests 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 location tests would naturally arise as the preferred method. As put by Ramdas et al. [2016]:

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

References

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

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.

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.

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.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13 (Mar):723–773, 2012. ISSN ISSN 1533-7928.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, July 2003. ISBN 0-387-95284-5.

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.

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.

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, S. Greiner, and P. Avesani. Statistical independence for the evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19, Dec. 2014. ISSN 2198-4018, 2198-4026. doi: 10.1007/s40708-014-0007-6.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209, Mar. 2009. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.11.007.

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. Schäfer, K. Strimmer, and others. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005.

R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of the National Cancer Institute*, 95(1):14–18, Jan. 2003. ISSN 0027-8874, 1460-2105. doi: 10.1093/jnci/95.1.14.

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

M. S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358, Feb. 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.08.014.

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.

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

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

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

A Analysis pipeline

Here is the analysis pipeline of Stelzer et al. [2013] we for the auditory data in Gilron et al. [2016]. Denoting by $i = 1, \dots, I$ the subject index, $v = 1, \dots, V$ the voxel index, and $s = 1, \dots, S$ the permutation index. Since regions[2] are centred around a unique voxel, the voxel index v also serves as a unique region index. Algorithm 1 computes a region-wise test statistic, which is compared to its permutation null distribution computed by Algorithm 2.

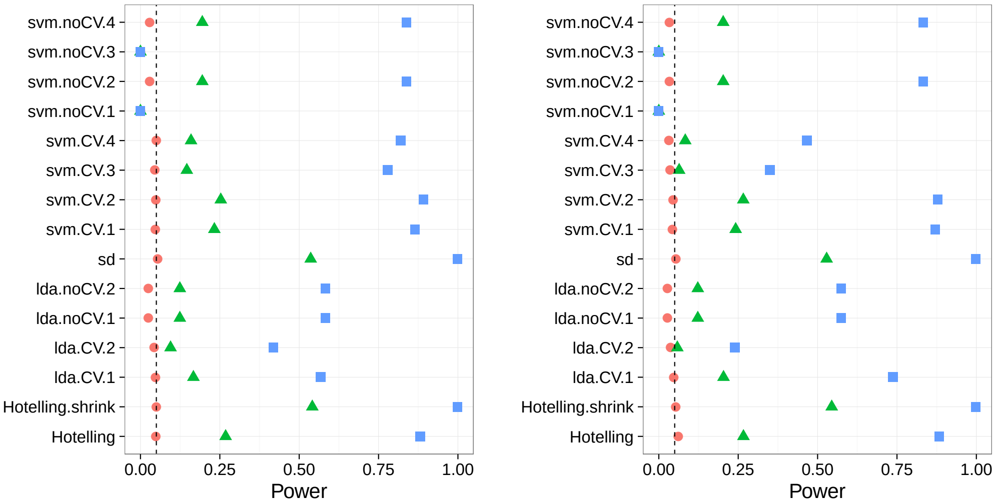
```
Algorithm 1: Compute a group parametric map.
  Data: fMRI scans, and experimental design.
  Result: Brain map of group statistics: $\{\bar{T}_v\}_{v=1}^V$
for $v \in 1, \dots, V$ do
    for $i \in 1, \dots, I$ do
        $T_{i,v} \leftarrow$ test statistic for subject $i$ in a region centered at $v$.
    $\bar{T}_v \leftarrow \frac{1}{I} \sum_{i=1}^{I} T_{i,v}$.
```

```
Algorithm 2: Compute a permutation p-value map.
  Data: fMRI scans of 20 subjects, experimental design.
  Result: Brain map of permutation p-values: $\{p_v\}_{v=1}^V$
for $s \in 1, \dots S$ do
    permute labels;
    $\bar{T}_v^s \leftarrow$ parametric map
```

[2] *searchlight* or *sphere* in the MVPA parlance

B More Simulations

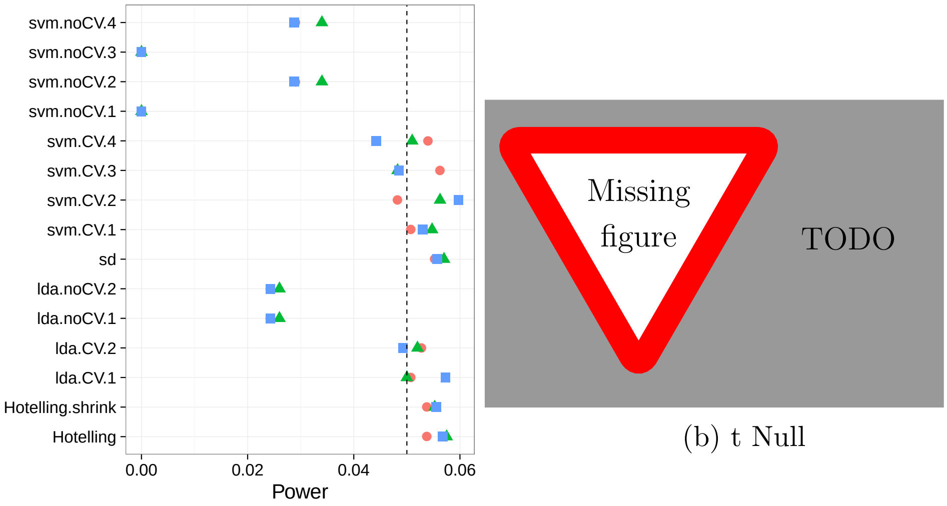
Figure 3: [TODO].



(a) 2 Folds.

(b) 20 Folds.

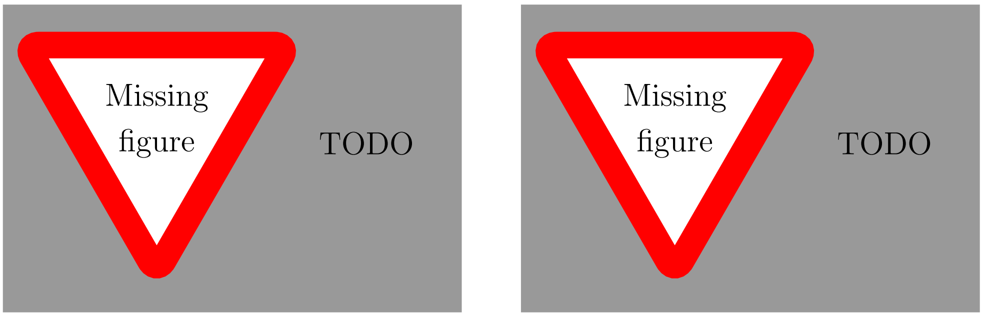
Figure 4: [TODO].



(a) Scale Change.

(b) t Null

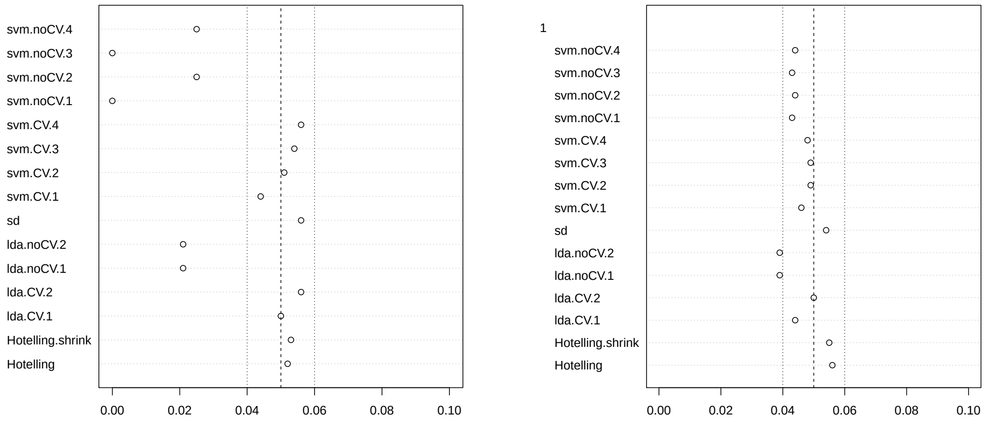
Figure 5: [TODO].



(a) Compound symmetry

(b) AR(1)

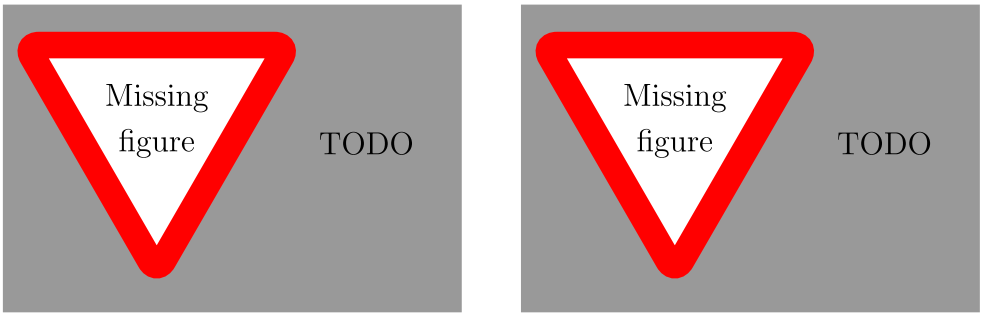
Figure 6: [TODO].



(a) n=40, zoom

(b) n=400, zoom

Figure 7: [TODO].



(a) n=400, smaller effects

(b) n=400, smaller effect, zoom