Better than chance classification for signal detection

Jonathan Rosenblatt Roee Gilron Roy Mukamel

July 27, 2016

Abstract

[TODO]

1 Introduction

A common workflow in genetics or 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. Genetic examples include Jiang et al. [2008], Radmacher et al. [2002] [TODO: elaborate]. Neuroscience examples include *multivariate pattern analysis* (MVPA) Kriegeskorte et al. [2006], Varoquaux et al. [2016], Golland and Fischl [2003]. In both fields this method has become quite popular. Kriegeskorte et al. [2006] has already gained 956 citations[1], and Radmacher et al. [2002] has been cited 274 times.

To fix ideas, we will adhere to a Neuroscientific example: In Gilron et al. [2016], the authors seek to detect auditory brain regions which distinguish between vocal and non-vocal stimuli. According to the MVPA analysis 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*.

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. A practitioner may then call

[1] Based on GoogleScholar. Accesses on 26.7.2016.

upon a two-group location test such as Hotelling's T^2 Fujikoshi et al. [2011]. Alternatively, if the size of the 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 Srivastava [2013] or Schäfer et al. [2005]. In contrast to *accuracy tests*, we call these *location tests*.

At this point, it becomes unclear which is the preferred test. This 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 may differ in constants.

Those constants, governing the power of the tests, are crucial when dealing with typical sample sizes in neuroscience and genetics, and thus the focus of this study. In particular, which test is to be preferred in finite samples? Our conclusion is 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 [Hemerik and Goeman, 2014], since they cannot exhaust the permissible false positive rate. In accuracy tests, the degree of discretization of the accuracy statistic 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). Because the testing problem is high dimensional, the observed train accuracy will be close to 1. The same will occur in every permutation, for the same reason. The permutation p-value will thus be 1 for almost all data sets, the null will never be rejected, and the test will have no power.

Given these considerations, it is quite surprising that signal detection using accuracy tests is so popular in neuroscience and genetics. In the following, we quantify the power loss to be expected in typical studies, and identify the problems' characteristics that govern its severity. We start by establishing a best practice for permutation testing using the accuracy test statistic, We also discuss the problem characteristics that govern the mag-

nitude of the conservativeness, and try to offer an intuition to the scope of the observation that a multivariate test should always be preferred over a classification approach.

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, using effect coding, 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 governed by the number of voxels in a regions, which is the number of voxels in each brain region tested. We thus have $\mathcal{X} = \mathbb{R}^{27}$.

Given n pairs of (x_i, y_i), typically assumed i.i.d., the *testing* approach to localization amounts to testing whether $x|y = 1$ has the the same distribution as $x|y = -1$. I.e., the multivariate voxel activation pattern has the same distribution when given a vocal stimulus, as when given a non-vocal stimulus. The *classification* approach to the localization problem 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 $T^{acc}_{\hat{f}} := P\left(\hat{f}(x) = y\right)$ when given a new, randomly drawn data point, (x, y).

2.1 Candidate Tests

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

What test statistic?

Cross validated or not? Is the statistic cross validated or not?

Refolding? For a K-fold cross validated test statistic: is the data refolded in each permutation?

Permute labels of features? Should the y be permuted or should the x?

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

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.

What test statistic? Given a predictor $\hat{f}$, a natural test statistic is some estimate of its accuracy $T^{acc}_{\hat{f}}$. Then again, very low accuracies, even 0, is evidence that the classes are separated, and we only need to invert the predictions. We can thus consider some estimate of $|T^{acc}_{\hat{f}} - 0.5|$ 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 $|T^{acc}_{\hat{f}} - \hat{p}_{max}|$. Since we will latter be aggregating these statistic over random data foldings, where the dominant class may have varying frequencies, it seems appropriate to standardize the scale of this statistic. We thus also consider a z-scored accuracy: $|T^{acc}_{\hat{f}} - \hat{p}_{max}|/\sqrt{\hat{p}_{max}(1-\hat{p}_{max})}$.

Cross validated or not? Were we interested in an unbiased estimator of the prediction error, there is no question that some validation is in order. Since we are merely interested in detecting a difference between groups, a biased error estimate is not an issue provided that it 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.

Refolding? The standard practice in neuroimaging is to refold the data after each permutation. This is imperative if permuting labels while aiming at balanced data folds. This is not, however, imperative in general. In this work, 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 folds will probably not be balanced, and will thus

have to be refolded. If the features are permuted, then the labels conserve their original fold assignments, and the data need not be refolded. 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? 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. Then again, it may complicate matters, as can be seen from the above discussion on label versus feature permutation. In general , it is not imperative in general, and we will indeed be comparing the effect of balanced foldings versus unbalanced. We will thus report results with both balanced and unbalanced data foldings.

How many folds? Different authors suggest different rules for the number of folds. We will be varying the number of folds, since it will affect the concentration of the estimated accuracy, which will have a crucial effect on the conservativeness of the permutation 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.

By now, the reader will have observed that there are indeed many ways to perform 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

In the first of our battery of simulations we verify that various test statistics and permutation schemes control the type I error. Figure **??** demonstrates that this is indeed the case. All our candidate tests control the type I error, with varying degrees of conservativeness. In particular: (a) if the folds are balanced or not, (b) if the labels are permuted or the features, (c) if the test statistic is varied, (d) if the regularization level of the support vector machine classifier (SVM) is varied, (e) if the number of folds is varied.

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.

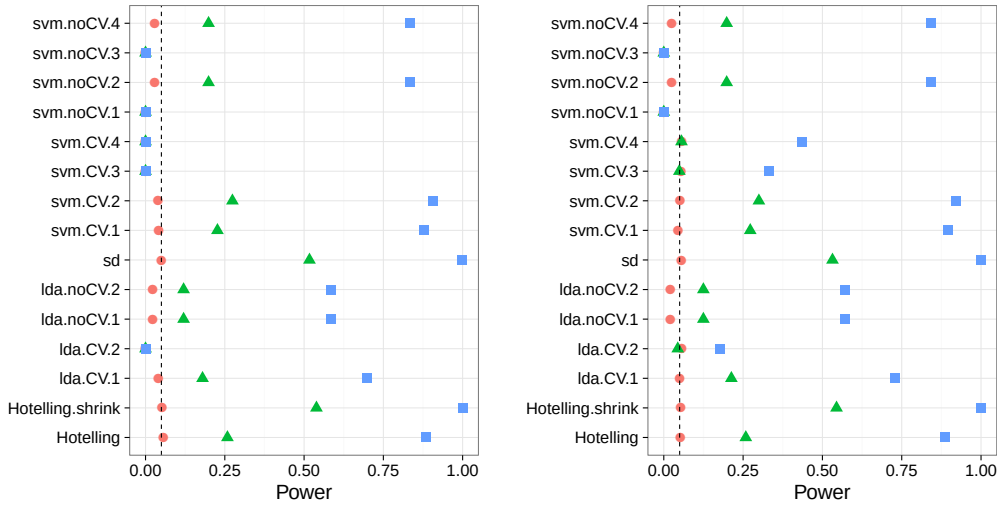
4 Power

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– at least when comparing the power of the various classifiers and multivariate tests. The results of Ramdas et al. [2016] suggest that power should be of the same order. 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]

We see by now that the use of accuracy tests for signal detection is underpowered compared to location tests. The above simulations can hardly support such a universal statement. We will thus verify on a neuroimaging dataset, and discuss the causes for this phenomenon thus the scope of the

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

statement.

5 Neuroimaging Example

Figure 2 is an application of our battery of tests 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[2].

To verify the observation that location tests have more power than accuracy tests, 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 the *sd* location test with the *svm.cv.1* accuracy test (see Table 1 for the definition of these statistics).

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

In agreement with our simulation results, the location test (*sd*) discovers more brain regions that encode information discriminating between vocal and non-vocal stimuli 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.

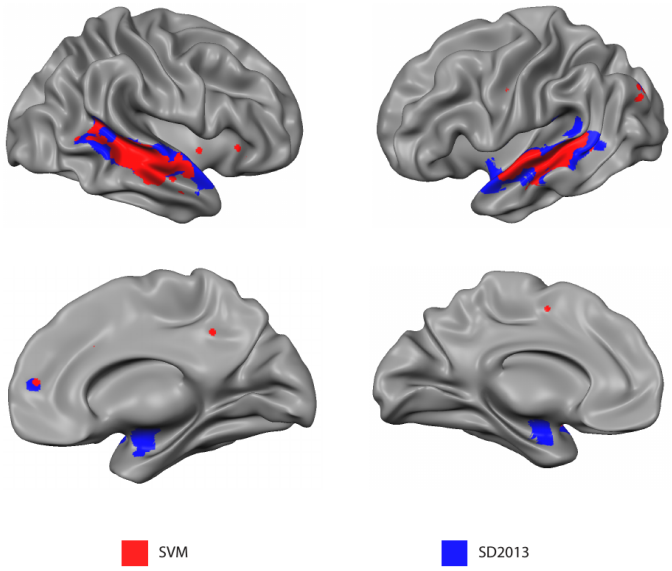


Figure 2: Brain regions encoding information discriminating between vocal and non-vocal stimuli. Map reports the centres of 27-voxel sized spherical regions, as discovered by an accuracy test (*svm.cv.1*), and a location test (*sd*). *svm.cv.1* was computed using 5-fold cross validation, and a cost parameter 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. The overlap is such that 90% of the accuracy test regions, are also detected by the location test. 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 qualitative difference.

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, for instance, with brain-machine interfaces, where the detection of a signal is not enough. In such cases, the performance of a particular classifier is the object of study, rendering the accuracy test the appropriate choice.

Secondly, there may be cases where the accuracy test does have more power then the location test. Our simulations were unable to point out such a scenario, but the fact that in our neuroimaging example (Section 5) some brain regions were detected with the accuracy test, and not the location test, suggest that the accuracy test does have more power for particular types of signal. [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, replacing a weak inequality with a strong inequality 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.

References

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.

Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu. *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. John Wiley & Sons, Aug. 2011. ISBN 978-0-470-53986-6.

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.

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

J. Hemerik and J. Goeman. Exact testing with random permutations. *arXiv:1411.7565 [math, stat]*, Nov. 2014.

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.

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.

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.

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.

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 $i = 1, \ldots, I$ the subject index, $v = 1, \ldots, V$ the voxel index, and $s = 1, \ldots, S$ the permutation index. Since regions[3] 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 $v \in 1, \ldots, V$ **do**
  for $i \in 1, \ldots, I$ **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.

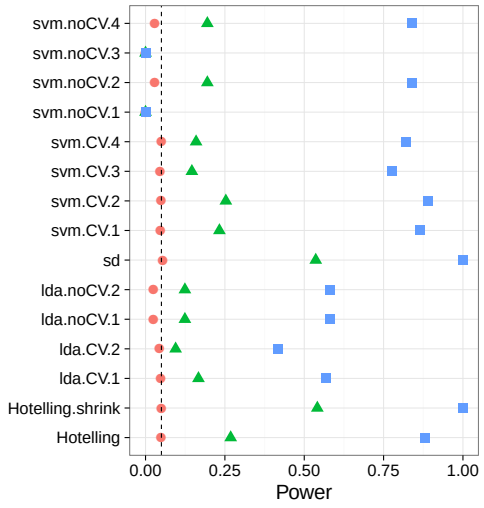
Data: fMRI scans of 20 subjects, experimental design.
Result: Brain map of permutation p-values: $\{p_v\}_{v=1}^{V}$

for $s \in 1, \ldots S$ **do**
  permute labels;
  $\bar{T}_v^s \leftarrow$ parametric map

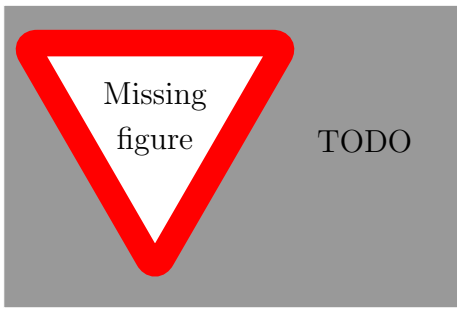
[3] *searchlight* or *sphere* in the MVPA parlance

B More Simulations

Figure 3: [TODO].



(a) 2 Folds.



(b) 20 Folds.