Review: Better-Than-Chance Classification for Signal Detection

1 Paper Summary

This work addresses two-sample testing: the problem of detecting the difference between the means of two unknown distributions, given only an i.i.d. sample from each distribution. This is a classic problem well studied in the statistics literature with a wide range of applications. A well-known solution is the t-test. More recently, the accuracy of a binary classifier trained to distinguish the two distributions has also been used as the test statistic for two-sample testing.

It appears that the goal of this work is to provide simulation-based evidence that using a classifier's accuracy as the statistic (referred to as an "accuracy test") yields lower probability of detecting the difference than using a genuine two-sample test which directly checks for the mean difference (referred to as a "location test"). Accuracy tests considered are those derived from linear discriminant analysis (LDA), and support vector machine (SVM). For the location tests, many variants (e.g., for high dimensions, and low sample size) of the t-test are studied. The simulations specifically target the setting of $n = 40$ (sample size), and $p = 23$ dimensions (Section 3.1). Overall, the main conclusion is that the accuracy tests have less test power compared to the location tests. A brief explanation is given in Section 5.1, essentially stating that this phenomenon is due to the discrete nature of accuracy tests.

2 Review

Writing and description Overall, the paper is easy to understand. There are several minor typos. The description for the simulations is sufficiently detailed for others to replicate. The presentation of some parts can be improved easily. For instance, the names of the methods in Table 1 could have incorporated the parameter values. Concretely, use, say, **svm.resub.c10** instead of **svm.noCV.1** where 1 at the end is just an uninformative running index.

Empirical results The contributed simulation results in Section 3 are valuable for only the very specific settings considered (i.e. $n = 40$, $p = 23$, mean shift case). However,

collectively these settings are not diverse enough to provide definite evidence that "accuracy tests do not have any more flexibility than location tests." as claimed in Section 5.6.

Also the main statement the paper seeks to make is not clear. While the problem setup in Section 2.1 concerns a generic two-sample testing (i.e., test whether the two underlying distributions differ at all in any way, not just in the means), the provided simulation results in Section 3 only address the case where there is a difference in the means. All conclusions reached based on these simulations are thus valid only for the mean shift case. This is not clearly stated in the paper.

Simulations It is rather unclear why only LDA and SVM are studied as the accuracy tests, considering that there are many more commonly used binary classifiers. For the SVM, two parameter values for the cost parameter C are considered: $C = 10$ and $C = 0.1$. No justification is given for why these values are chosen. It is unclear what the neuroimaging result in Figure 10 implies about the main thesis of this paper: accuracy tests vs location tests. The paper only reports in Figure 10 that the location test detects 1232 brain regions and the accuracy test detects 441 regions. No further discussion is given. Other simulations provided are fairly thorough, for the settings considered.

Theoretical results The paper does not provide any theoretical result. As stated in the paper that small sample size is the target of study, it is understandable that a theoretical study is challenging since we cannot rely on asymptotics. Perhaps one way to proceed is by starting from the mean shift case under Gaussian assumptions for the two distributions. The exact finite-sample distributions under both null and alternative of the T-test statistic are well known, and can be found in standard text books (see Anderson [1958, Section 5.4], for instance). The behaviour of the LDA statistic in this case is also understood (see Anderson [1958, Section 6.4], for instance). An interesting part is the theoretical comparison of the two, which could be in the form of the exact test power, or Bahadur efficiency (or other relative efficiency measures as the authors see fit).

Related work Description on related work given in Section 5.8 only cover existing works related to neuroimaging. It is worth briefly describing recent development of nonparametric two-sample tests. To name a few, the works of Gretton et al. [2012] (the maximum mean discrepancy test), Székely and Rizzo [2004] (the energy distance), and Moulines et al. [2008] (kernel LDA) should be discussed. There were many more extensions of these works. If only the mean shift case is the main goal in this paper, then the paper should clearly state this in Section 3.1.

Claims in the paper Several claims or description are not accompanied by citations. To be concrete, the description in Sections 2.4 on estimates of a classifier's accuracy does not have appropriate sources. Others include

- The description on tie-breaking in Section 3.4.

- The first paragraph of Section 3.5: "*The Neyman-Pearson Lemma (NPL) type reasoning that favors the location test over accuracy tests may fail when the data is not multivariate Gaussian.*"
- Section 5.6: "*Accuracy tests do not have any more flexibility than location test.*" This is a strong statement. Either a citation or a proof (or both) is needed.

Many of the claims centered around the statement that location tests are better than accuracy tests. While the offered empirical results do support this statement, the design of the simulation settings is questionable. Firstly, the settings are not diverse enough, as previously mentioned. Secondly, it appears that the settings are designed so as to support this statement, rather than to find out the truth. All the synthetic problems considered are mean shift (i.e., shift alternative) problems, meaning that location tests are appropriate. Further, as stated in Section 5.1, "*We focused on shift alternative so that location tests are expectedly superior via an NPL type argument.*"

The explanation given in the paper that accuracy tests are essentially discrete, is reasonable. However, this observation alone and the provided empirical results are not sufficient for the conclusion. In my opinion, the statistic of an accuracy test can be decomposed into two parts: (1). the decision function (i.e., real-valued) (2). the thresholding function to turn the decision function into an actual classifier. This paper only offers an explanation which addresses the second part i.e., that thresholding turns the statistic into a discrete variable, and incurs a power loss.

In fact, one can use a rich, appropriate function class for the decision function to make the overall power of an accuracy test higher as well. This was studied in Lopez-Paz and Oquab [2017]. In particular, see table 1 on page 6 of Lopez-Paz and Oquab [2017], where an accuracy test constructed from a deep neural network (i.e., C2ST) has higher power than, for instance, MMD-quad (i.e., the maximum mean discrepancy test of Gretton et al. [2012]) which is a "bona fide" two-sample test. The reason is due to the discrepancy in (1) (the decision function) between the two types of tests. The state-of-the-art MMD test in this case relies on a Gaussian kernel on raw image pixels, while C2ST learns a neural network end-to-end to build a classifier. Naturally, one cannot expect a Gaussian kernel on raw pixels to perform well.

The point is that, to compare the two types of tests, it is important to isolate (1) and (2). Otherwise, we can always use (1) or (2) to make one type of test better than the other. Now, with (1) controlled to be the same in both types of tests, I tend to agree with this paper that location tests should be better in many cases. In the context of the MMD test, this was discussed in a review comment of Lopez-Paz and Oquab [2017] (on OpenReview) `https://openreview.net/forum?id=SJkXfE5xx&noteId=ry-le414x`. It is this kind of "fair" comparison (of course, under different, more diverse simulation settings) that the current paper does not emphasize.

3 Detailed Comments

- It might be useful to check Lopez-Paz and Oquab [2017], Yu et al. [2007] (classifier-based test) and other works cited therein.
- Last paragraph on page 2: At that point, it is unclear to the reader that one brain region corresponds to one dimension of the mean vector.
- Page 3, last paragraph: Is it possible to cite a peer-reviewed source rather than a comment on an online forum?
- Section 2.3: "The most prevalent interpretation of ....." This should be "The most prevalent assumption of ..."
- The "unbalanced V-fold CV" in Figure 1a is never defined precisely. What does it mean exactly?
- **A thought**: The paper claims that the accuracy tests are less powerful because they rely on discrete test statistics. In other words, they are not affected by minor perturbation of the input data (hence the insensitivity). However, optimistically, this would also imply that they are robust to outliers. The robustness of accuracy tests could be an interesting topic of future work.

References

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958. 2

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. 2, 3

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *ICLR*, 2017. 3, 4

Eric Moulines, Francis R Bach, and Zaïd Harchaoui. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2008. 2

Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 2004. 2

K. Yu, R. Martin, N. Rothman, T. Zheng, and Q. Lan. Two-sample comparison based on prediction error, with applications to candidate gene association studies. *Annals of Human Genetics*, 71(1):107–118, 2007. ISSN 1469-1809. doi: 10.1111/j.1469-1809.2006.00306.x. URL http://dx.doi.org/10.1111/j.1469-1809.2006.00306.x. 4