Review of "Better-Then-Chance Classification for Signal Detection"

Summary: This paper attempts to shed some light on the common practice of using statistics based on prediction accuracy to test the presence of signal (and quantify the uncertainty of the detection). From the existing literature - scattered across many domains, it seems there are many contradicting results. In this paper the authors focus on the high-dimensional small sample setting (also known as the large p - small n setting). The authors make a strong claim that accuracy-tests are never optimal relative to high-dimensional two-group tests. This strong statement is, in the referees' opinion, not wise, as their conclusions are based on a simulation study in a rather specific setting (therefore they cannot really support such a statement, and it is possible to come up with different (but relevant) settings leading to the reverse conclusion). Nevertheless, the premise of the paper still leads to a worthwhile discussion.

The paper is generally well written, although the English usage can be improved. The authors attempt to cite almost any result that is remotely related to the work here, but often fail to put it in context or give enough insight to the reader to appreciate their relevance in the discussion. I feel the authors are overselling their claims quite a bit while supporting them on a specific simulation study that is chosen because it supports their claim. Note that, the main goal of the paper is to compare two statistical approaches to address the same problem. One approach is indirect (accuracy-based testing), while the other is direct (two-group testing). Each has advantages and disadvantages: the first approach is sensible to use even without making strong assumptions (the assumptions are indirect - e.g., the choice of prediction method, that might or might not lead to a powerful enough test). The second approach takes a more direct stance, and is only sensible (i.e., powerful) if the setting matches the statistical model assumptions made by the test statistic. Naturally, a direct approach will always be preferable, but it will require a better/more accurate knowledge of the statistical model. In some settings this is fine, but in other settings one does not really "know" a good model. One can easily argue that, in certain settings, the first approach might be better than the second, if the two group test is based on wrong assumptions for the problem (e.g., testing for a dense shift in the mean when the shift is actually sparse). On the other hand, some accuracy based tests that take the sparse nature of the problem into consideration will fare significantly better. In a nutshell - the conclusion heavily depends on the context, and therefore it can be scientifically misleading to make such blanket statements. Nevertheless, these comparisons are still useful to get insights on the strengths and weaknesses of each approach.

As a side-remark (not as important as the above): I cannot understand the choice of venue by the authors. Who is your target audience? To me it seems it should be either the practitioners in neuroscience and genetics, practitioners in machine learning, or statisticians (both applied and theoretical). The choice of TSP does not seem quite fit. Furthermore, none of your references are in signal processing venues. That is, by itself, not a problem, but seems odd given this should appeal to that community.

Below I outline some more detailed comments, some of them supporting the arguments above.

Detailed comments:

1. Page 1, first line of the introduction: “detect a signal” or “detect signals”.

2. Page 2, line 18-19: the sentence does not make sense as the word sample appears twice with two difference meanings. Do you mean “the number of samples is on the order of the number of parameters of the model”? It is important to be a bit careful here.

3. Page 2, bulleted list: “tests that seek” instead of “tests the seek” (this happens in three places)

4. Page 2, line -6: “neuroimaging?” This is not a question, but a statement.

5. Page 2, line -4: the statement in bold is rather strong, and in my opinion the rest of the manuscript does not support it.

6. Page 2, line -3: “Prefer” and “Appropriate” should not be capitalized. Also, before (ii) you should use a semi-colon.

7. Page 2, line -1: “is preferable over a ***specific*** two-group test”. Actually this discussion already highlights some of the problems I have with the strong statement in the paper. Here the choice of accuracy-based statistic is well-suited for the problem, while the choice to two-group test is not. So, it is no surprise (as you state) that the first approach is better, even if the accuracy-based test is indirect.

8. Page 3, lines 3-6: again, the conclusions are context specific. Furthermore, I don’t understand the last comment. Neural networks can (and are) often used for dimensionality reduction, so these are indeed “learning” features. This is possible even when the number of parameters is large, compared to the number of samples. The statement might only be somewhat sensible if by “high-dimensional” you also mean small-sample.

9. Page 3, lines 11-24: this discussion and argument is quite relevant for the story you are trying to tell. Since this is not published anywhere it should be part of your paper, with all the necessary details. You are actually

making a valuable formal statement here, but right now it is not obvious to the reader this is a valid argument. I urge the authors to write the argument in full detail.

10. Page 3, line -5: although in the preceding paragraph you use a general notation, here you are already particularizing to the case $\mathcal{Y} = \{0, 1\}$. What this the intention? Also, you notation has a problem: you use both x_1 to denote a sample from x given $y = 1$, and to denote a datapoint in $\mathcal{S}$. Please be more careful! Furthermore, here there are already some modeling assumptions. Namely you assume the dataset can be views as a collection of two independent i.i.d. samples from two distribution (that should be the same distribution under the null). It appears here there are no further assumptions on the class labels. It would be useful to clarify if the labeling is balanced or unbalanced (or if you consider both cases).

11. Page 4, line 11: not clear why you are using $R + 1$. I guess you are including the un-permuted labels in the summation. Also, the description is a bit poor. The sum over π is over which permutations? It would be useful for the reader to have a full and clear picture here.

12. Page 4, line -9: "If the sample size n is not much larger than the dimension p"

13. Page 4, line -6: this statement is confusing. The SNR is a property of the distributions (e.g., how separated are the null and alternative classes). If the test statistic is not appropriately chosen it will indeed result in a test with low power. But that doesn't mean all tests will have low power. Furthermore, the high-dimensional tests you mention all make assumptions about the covariance structure (e.g., sparsity). So their power will depend on the specific setting.

14. Page 4, line -4: explain the notation $\mathbb{E}(x_0)$ and $\mathbb{E}(x_1)$.

15. Page 4, lines -1, -3 and -4/5: the lists of references do not match the ones in the introduction - why?

16. Page 5, line 2: same as previous comment. Why? It seems the authors are just dumping a lot of references, without putting them into context.

17. Page 5, line 6: you should use $\mapsto$, namely $\mathcal{X} \mapsto \mathcal{Y}$.

18. Page 5, line 7: maybe make it clear you are talking about binary classification with the 0/1 loss.

19. Page 5, line 9: here you are making further assumptions, in particular, you are assuming there is a joint distribution in $\mathcal{X} \times \mathcal{Y}$. So you are actually introducing a model for the marginal distribution of the class labels.

20. Page 5, footnote 5: what is the analogue for continuous y? I know what you mean, but this is very poorly explained.

21. Page 5, line 10: $\varepsilon_{\mathcal{A}_S} = \mathcal{P}(\mathcal{A}_S(x) = y)$ (maybe easier to read than the expression with the indicator).

22. Page 5, Definition 1: This is also known as the empirical error if you ask someone in machine learning/learning theory. Maybe emphasize $\mathcal{S}$ is used both to train a prediction rule and to access the error (typically leading to negatively biased estimation of the risk).

23. Page 5, Definition 2: reword the statement - suppose one has a partition of $\mathcal{S}$ into V so that... Also, use fold in math mode, so you have V-fold and not $V fold$.

24. Page 6, line 2: (e.g, see [3]). Also, commonly the folds are obtained by partitioning the data uniformly in V equal sized sets. This will lead to a stochastically balanced partition. Is this considered to be balanced folds? In light of point (c) I guess this is not the case. However, it does seem more sensible than enforcing the folds to be exactly balanced (and will avoid the issue mentioned in (c)).

25. Page 6, line 10: dot should be not.

26. Page 6, Table I: dCOV and Gretton are not mentioned in the caption.

27. Page 6, footnote 7: Explain better the permutation of labels or features. For the binary classification case you are simply sampling n balls without replacement from a urn with $\sum_{i=1}^{n} y_i$ balls with a 1 and $\sum_{i=1}^{n}(1-y_i)$ balls with a zero, right? If one doesn't refold the set of features in each fold will be the same (regardless of the label permutation). On the other hand, permuting the features instead will make the set of features in each fold to change.

28. Section III: maybe structure this section better. Begin by introducing the general model (in (1)). Then consider the specific variations you consider in each subsection. Begin by dealing with the dense case. Finally, how do you generate y_i? Is this an i.i.d. sample from a Bernoulli distribution, or you generate them deterministically (e.g., $n/2$ samples from each class). This is not explained.

29. Page 7, line 18: replication should be replications.

30. Section III-C: you should at least explain why the Hotelling approach is performing so poorly here (at least remind the reader of the discussion earlier on). I guess you are not using the knowledge of Σ, right?

31. Page 7, line -2: this is a completely nonsensical statement !!! You claim that the results hold for large p large n, but then state the simulation took 11 years of computing time (meaning you started it in 2007). Or does it mean you have not done the simulation but are claiming the conclusions still hold (on what basis then)?

32. Page 8, line -1: this needs to be better explained, so one can interpret the results of Figures 3 and 4.

33. Page 9, line -9: I would say most accuracy tests are performing slightly better.

34. Page 10: General remark: the Hotelling's test has more or less the same power, not matter what the situation is so far. This is not strange, but should be remarked at some point.

35. Page 12, Definition 3: would be good to give a better insight about the idea here. Does this provide a way to reduce the high variance of leave-one-out cross validation without increasing the bias?

36. Table II: what is meant by cost? It also shows up in the results in Figure 10, but I could not find an explanation about it (maybe I missed it)?

37. Page 13, line -5: seem should be seen.

38. Page 13, line -1: this is not strange - there is a significant difference between testing and estimation in sparse settings. See for instance the work of Yuri Ingster, or your reference [39].

39. Page 14, line -4: as ***in*** our

40. Page 14, line -2: you are using π to denote both a permutation and the mixing probabilities. Be careful!!!

41. Section IV: what does (a) and (b) refer too? You don't use that terminology anywhere, in particular in the figures.

42. Figure 10: what is accuracy test u? Also, there are 4 panels. What does each panel represent?

43. Page 17, point 2): this comment is strange. In many of your figures svm.CV.cCV has more power than svm.CV.c001. So, I don't see any support for such an argument.

44. Page 18, footnote 9: this is the only reference to cost I encountered, but not particularly informative.

45. Page 19, lines 1-4: in addition, in the low PC setting accuracy-based tests seem to perform slightly better than many of the two-group testing competitors.

46. Figure 12: in the introduction (page 3) you stated it is no surprise accuracy-based tests work better in the sparse regime. These results contradict that statement, but this is clearly because the accuracy-based tests you are using are not able to capitalize on the sparse nature of the problem. Had you used a tree-predictor the story would be different. So, I don't think these comparisons are actually fair.

47. Page 20, line -3: Bernoulli, not Brenoulli.