CMA

Centre for Mathematics and its Applications
The Australian National University
Canberra, A.C.T. 0200, Australia

8th November 2004

Professor A.T.A. Wood
School of Mathematical Sciences
Joint Editor, *JRSSB*
The University of Nottingham
University Park
Nottingham NG7 2RD, UK

Dear Andy,

B5879: Hall, Marron and Neeman

Thank you for your letter of 31st August, regarding our paper. We have carefully revised the paper, and have addressed all the reviewers' points. Attached are detailed comments for the reviewers, indicating where we have made the alterations. In addition to those changes we have added a number of references to theoretical work where dimension is permitted to increase.

Would you please consider the revised paper for publication in *JRSSB*?

Sincerely

Peter Hall

Referee I

1. We have added the new Figure 6, with appropriate discussion, summarising some of the simulations we have done in the unequal sample size case.

2. We have added the following remark immediately after the paragraph condition condition (6):

> As will shortly be seen, the second part of (6) is especially relevant to accurate classification. If μ in (6) is too small, and in particular if it equals zero, then a classifier of any conventional type (support vector machine, distance weighted discrimination, nearest neighbour, etc) operates asymptotically in a degenerate fashion, without respecting the population, with probability converging to 1 as $d \to \infty$, from which a new datum comes. That is, the classifier assigns the new datum to the same population, regardless of the actual population from which it came. In such instances the classifier is overwhelmed by the stochastic noise that accrues from a very large number of dimensions. The case $\mu = 0$ can arise when there is only a finite number of truly discriminating components.

Referee II

1. We have added the new Section 6, entitled "Summary and Conclusions," immediately before the references.

2. Discussion of this point has been added to the new Section 6. See also the following material, added to the paragraph immediately below display (4):

> Assumption 3 is a simple way of permitting the amount of information available for discrimination to diverge to infinity as d increases. (In conventional asymptotics, information diverges through increasing sample size.) However, it is also of interest to explore more marginal cases where conditions such as assumption 3 fail; see Section 6.

3. We have re-stated these results as theorems, as requested.

4. The reason we did not give a proof of this result is we do not have a simple argument. We shall give a proof below. It has the advantage that it is constructive, and so is perhaps appealling to practitioners. But we feel that its length makes it unsuitable for a statistical journal. However, we would be happy to include the proof if the editor felt it appropriate, or to include a shorter one if the editor or the referee knows of such a proof.

Here is the proof: It suffices to treat the case $N = d (= m+n)$. In this case the assumption that "no k data points lie in a $k-2$ dimensional hyperplane" ensures that the set $\mathscr{Z}_1$, say, of d points is a linearly independent set. Hence, given any set, $\mathscr{Z}_2$ say, of another d linearly independent points in d-variate space, there is an nonsingular linear transformation, T say, that takes $\mathscr{Z}_1$ to $\mathscr{Z}_2$.

Partition $\mathscr{Z}_1$ into two disjoint subsets, $\mathscr{Z}_{11}$ and $\mathscr{Z}_{12}$ say, of respective sizes m and n. Let $\mathscr{Z}_{21}$ and $\mathscr{Z}_{22}$ denote the respective images of $\mathscr{Z}_{11}$ and $\mathscr{Z}_{12}$ under T. If we can construct a $(d-1)$-variate hyperplane that separates $\mathscr{Z}_{21}$ and $\mathscr{Z}_{22}$, then the inverse of T, applied to this hyperplane, will produce a hyperplane that separates $\mathscr{Z}_{11}$ and $\mathscr{Z}_{12}$. So, it is necessary only to show that, for a suitable choice of $\mathscr{Z}_2$, we can separate $\mathscr{Z}_{21}$ and $\mathscr{Z}_{22}$. This is clear if we take $\mathscr{Z}_2$ to be the following set of d points:

$$(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)\,.$$

Indeed, two subsets of these points, containing m and $n = d-m$ points, respectively, are contained in parallel $(d-1)$-dimensional hyperplanes.

5. We agree with the referee's suggestion, and have made the recommended changes.

6.

- Section 4.1 etc: Change made.
- $(9,-16)$: Correction made.
- $(9,-15)$: Correction made.
- $(12,-1)$: Sorry, we omitted to make the qualification, "for data from at least one of the populations" at this point. That error has been corrected.
- $(14,7\text{–}8)$: We have qualified the remark and added a parenthetical comment of

explanation. The material now reads as follows:

Also as suggested by the theory, the error rates for SVM, DWD and CRD come together for increasing d, although the convergence is perhaps faster than expected. (Recall, from Theorems 1 and 2 and the first paragraph of Section 4.2, that in the case $m = n$ which we are considering here, the classification probabilities for SVM, DWD and CRD all converge to 1 as d increases.)

- $(17, 8)$: Change made.