RN 006-68

NON-METRIC DISCRIMINANT ANALYSIS

Jan de Leeuw

April 1968

PSYCHOLOGICAL INSTITUTE/UNIVERSITY OF LEIDEN - NETHERLANDS

- 1 -

In the situations considered in this paper we have a set of n stimuli varying on m "physical" dimensions. Define an nxm matrix $A = (a_{ij})$ is the value of stimulus i on stimulus dimension j. The task of the subject in these experiments is to partition the set of stimuli into two mutually exclusive and exhaustive subsets A and B. These classifications define an n-element vector $\mathbf{k}$, with $k_i = 1$ if S classifies stimulus i as belonging to class A and $k_i = -1$ otherwise.

- 2 -

We shall say that a subject follows a linear cut-off strategy if there is an m-element vector of weights u and a real number p_c, such that

$$k_i = 1 \Leftrightarrow \sum_{j=1}^{m} u_j a_{ij} \geqslant p_c$$

$$k_i = -1 \Leftrightarrow \sum_{j=1}^{m} u_j a_{ij} \leqslant p_c$$

This can also be written as

$$k_i = 1 \Leftrightarrow \sum_{j=1}^{m} u_j a_{ij} - p_c \geqslant 0$$

$$k_i = -1 \Leftrightarrow \sum_{j=1}^{m} u_j (-a_{ij}) + p_c \geqslant 0$$

Define (for a particular subject) an nx(m+1) matrix X and an m+1-element vector w with

$$\begin{cases} x_{ij} = k_i a_{ij} & i=1, \ldots, n;\ j=1, \ldots, m \\ x_{ij} = -k_i & i=1, \ldots, n;\ j=m+1 \end{cases}$$

and

$$\begin{cases} w_j = u_j & j=1, \ldots, m \\ w_j = p_c & j=m+1 \end{cases}$$

Summarizing, our definition may be reformulated as follows:
We shall say that a subject follows a linear cut-off strategy if the system of homogeneous inequalities

$$Xw \gtrneqq 0$$

has a non-trivial solution.

Both the geometrical and the algebraic aspects of systems of linear inequalities are treated extensively in Kuhn and Tucker (1966). Some elegant algorithms for finding a solution vector w_o if the system is consistent, are discussed in Saaty (1959, p. 114-125).

- 3 -

Suppose that in a particular experiment there are four stimuli, characterized by the values 1/4, 1/2, 1 and 2 on one single physical dimension. The subject classifies the first three stimuli as examples of class B. We may say that he follows a linear cut-off strategy if the following four inequalities have at least one nontrivial solution.

$$1/4\ w_1 + w_2 \gtrneqq 0$$
$$1/2\ w_1 + w_2 \gtrneqq 0$$
$$w_1 + w_2 \gtrneqq 0$$
$$-2\ w_1 + w_2 \gtrneqq 0$$

Each of these inequalities cuts the space of all possible solutions in two halves. One of these is permissible, i.e., all points in this halfspace satisfy the defining inequality, all points outside it violate the inequality. For the four inequalities in our example the shaded areas in figures 1a-1d are the permissible halfspaces.

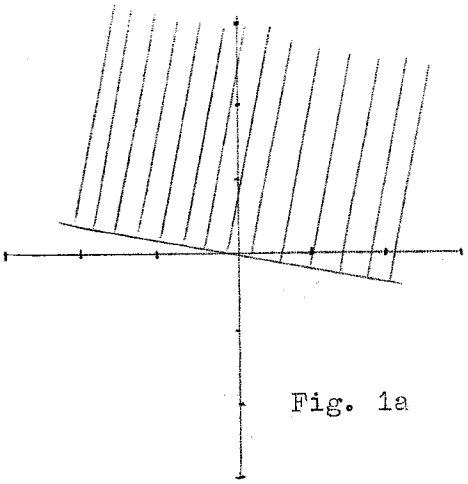

Fig. 1a

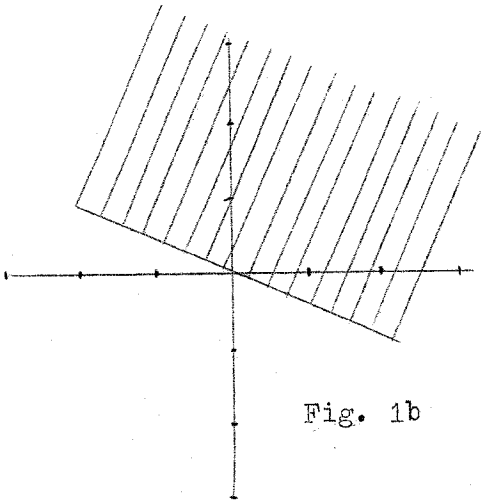

Fig. 1b

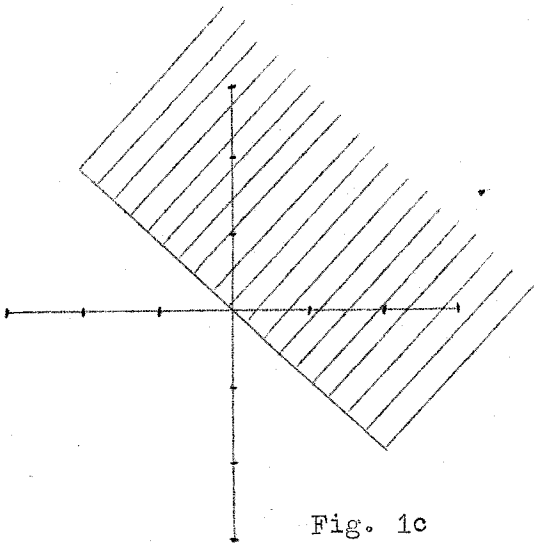

Fig. 1c

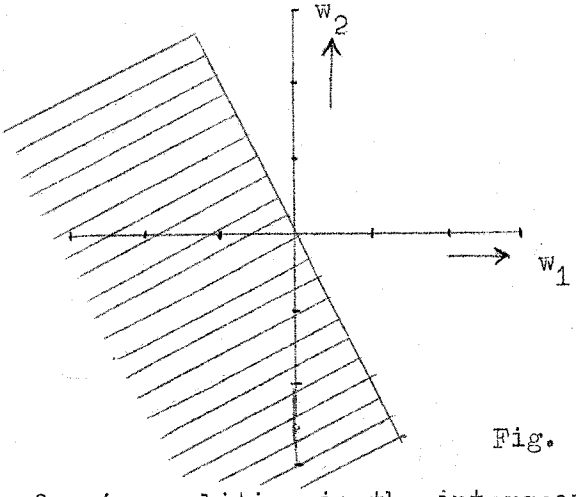

Fig. 1d

The solution set of a system of n inequalities is the intersection of the n permissible halfspaces. In the case of homogeneous inequalities this intersection always contains the trivial solution w=0. If the system is consistent the solution set is a polyhedral convex cone with apex at the origin (a polyhedral convex cone is defined geometrically as the intersection of a finite number of halfspaces, whose boundary hyperplanes pass through the origin). The solution set for our example is the shaded area in figure 2.

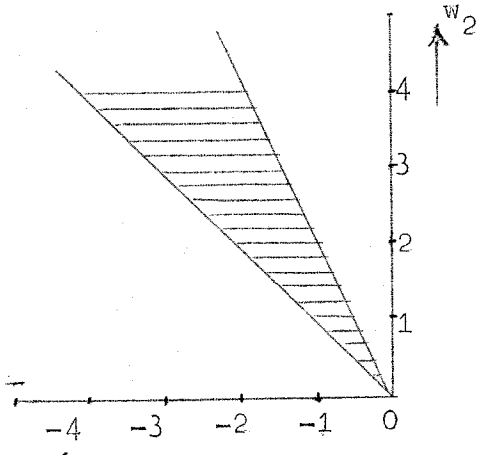

fig. 2.

- 4 -

In this section an algorithm will be constructed for solving systems of homogeneous linear inequalities, based on the theory of least squares. As a matter of fact it is a member of the class of alternating least squares algorithms discussed in De Leeuw (forthcoming). Introduce an n-element vector of slack variables s (with $s_i \geq 0$ for each i=1,...,n). The function to be minimized can be written as

$$F = (Xw-s)'(Xw-s)$$

Suppose that during the r^{th} iteration we have an estimate of s, denoted by $s^{(r)}$. Then the least squares estimate of $w^{(r)}$ is given by

$$w^{(r)} = (X'X)^{-1}X's^{(r)}$$

If the value of F for this w and this s equals zero, we have found a solution of the system. If not, another least squares process is used to find a new estimate $s^{(r+1)}$ which minimizes F for the current value of w, under the conditions that $s_i^{(r+1)} \geq 0$ for each i=1,...,n. Define $t=Xw^{(r)}$. Because F is a strictly convex function of s, we may write

$$\begin{cases} s_i^{(r+1)} = t_i & \text{if } t_i \geq 0 \\ s_i^{(r+1)} = 0 & \text{if } t_i < 0 \end{cases}$$

This $s^{(r+1)}$ is used to find a new value of w, and so on, until F=0. Evidently we may also proceed the other way around, i.e., start each iteration with an estimate of w and compute the corresponding value of s. Moreover, because the solution set is a convex polyhedral cone, it follows that, if w_o is a solution, then λw_o with $\lambda > 0$ is also a solution. This means that we may scale w (or s) while iterating (for example by requiring w'w = 1).

- 5 -

This approach generalizes very easily to non-linear decision boundaries. Consider for example the n x $\frac{1}{2}m(m+3)$ matrix B, with

$$b_{ij} = a_{ij} \qquad j=1,\ldots,m$$

$$b_{ij} = a_{ij}^2 \qquad j=m+1,\ldots,2m$$

$$b_{i,2m+1} = a_{i1}a_{i2}$$

$$\vdots$$

$$b_{i,\frac{1}{2}m(m+3)} = a_{i,m-1}a_{im}$$

By augmenting the matrix with an extra column of constants and by changing the sign in the appropriate rows, an $n \times (\frac{1}{2}m(m+3)+1) = n \times \frac{1}{2}(m+1)(m+2)$ matrix X is constructed upon which the algorithm may be applied. In this way we can test whether or not S uses a quadratic decision boundary.

- 6 -

Numerical experience with the algorithm outlined in section 4 indicates that it converges asymptotically to a point on the boundary of the permissible region (provided the system is consistent, and the initial point is chosen outside the permissible region and outside its polar cone). If the system is inconsistent, the algorithm converges asymptotically to the trivial solution $w_o = 0$, as it should do. In some applications however, it may be interesting to know to what extent the system of inequalities is consistent. A rough measure of the degree of inconsistency is the number of negative elements (say n_-) in the vector t. Define

$$R = 1 - \frac{n_-}{n}$$

R is a measure analogous to the coefficient of reproducibility in scalogram analysis. In some other cases it may be useful to obtain a non-trivial solution, which is "best" in some well defined sense, even if the system is inconsistent.

- 7 -

If the system is inconsistent we can always construct a consistent subsystem. If the number of deleted inequalities (i.e. rows of X) is equal to n_d, we want to find a subsystem, such that

$$S = 1 - \frac{n_d}{n}$$

is maximal, and such that the resulting system of $n-n_d$ equalities is consistent. In other words: we are looking for the maximal solvable subset. For methods to obtain such a solution, see De Leeuw (forthcoming).

- 8 -

Another alternative is to define

$$\phi = \frac{\sum_{i=1}^{n}(t_i - |t_i|)^2}{4\sum_{i=1}^{n} t_i^2}$$

We shall prove some simple properties of ϕ :

i) $0 \leqslant \phi \leqslant 1$

ii) $\phi = \frac{F}{\sum_{i=1}^{n} t_i^2}$

iii) For each w_o : $\phi(w_o) + \phi(-w_o) = 1$

iv) $\phi = 0$ iff w_o lies in the permissible region (excluding $w_o=0$).

v) $\phi = 1$ iff w_o lies in the polar cone of the permissible region excluding $w_o=0$).

vi) $\phi = \frac{1}{2}$ iff $\sum_{i=1}^{n} t_i |t_i| = 0$ (excluding $w_o=0$).

Proof:

i) $\phi \geqslant 0$ is trivial. Furthermore

$$\sum_{i=1}^{n}(t_i - |t_i|)^2 - 4\sum_{i=1}^{n} t_i^2 = -2\sum_{i=1}^{n}(t_i^2 + t_i|t_i|) = -2\sum_{i=1}^{n}|t_i|(|t_i| + t_i) \leqslant 0.$$

ii) $$F = \sum_{i=1}^{n} (t_i - s_i)^2 = \sum_{i=1}^{n} (t_i - \max(0, t_i))^2 = \sum_{i=1}^{n} (t_i - \tfrac{1}{2}(t_i + |t_i|))^2 =$$

$$= \tfrac{1}{4} \sum_{i=1}^{n} (t_i - |t_i|)^2$$

iii) $$\phi(w_o) + \phi(-w_o) =$$

$$\frac{\sum_{i=1}^{n} (t_i - |t_i|)^2 + \sum_{i=1}^{n} (-t_i - |t_i|)^2}{4 \sum_{i=1}^{n} t_i^2}$$

iv) $\phi = 0$ iff $\sum_{i=1}^{n} (t_i - |t_i|)^2 = 0$ and $\sum_{i=1}^{n} t_i^2 \neq 0$, iff $t_i = |t_i|$ for each i and $t_i \neq 0$ for at least one i. By definition the corresponding value of w_o lies in the permissible region (and $w_o \neq 0$).

v) From the proof of (i) if follows that $\phi = 1$ iff $|t_i| = -t_i$ for each i and $t_i \neq 0$ for at least one i. By definition the corresponding value of w_o lies in the polar cone of the permissible region (and $w_o \neq 0$).

vi) $\phi = \tfrac{1}{2}$ iff $\sum_{i=1}^{n} (t_i - |t_i|)^2 - 2 \sum_{i=1}^{n} t_i^2 = 0$ and $t \neq 0$; iff $\sum_{i=1}^{n} t_i |t_i| = 0$ and $t_i \neq 0$ for at least one i.

Two possible approaches towards minimizing ϕ are discussed in De Leeuw (forthcoming). The first approach is to use the algorithm discussed in section 4, and to scale t after each iteration so that $t't = 1$. An obvious disadvantage of this procedure is the unpredictable asymptotic behaviour. The second approach is a steepest descend process, which use the partial derivatives

$$\frac{\partial \phi}{\partial w_j} = \frac{4}{\sum_{i=1}^{n} t_i^2} \sum_{i=1}^{n} \left[\left\{ (1-2\phi) t_i - |t_i| \right\} x_{ij} \right]$$

Because $\phi = 1$ implies and is implied by $t_i = -|t_i|$ and $\phi = 0$ implies and is implied by $t_i = |t_i|$, for those cases all partial derivatives are equal to zero.

- 9 -

In this final section we shall give some numerical examples of the algorithm outlined in section 4. Data were taken from one of De Klerk's probabilistic concept learning experiments (De Klerk, 1968) in which 50 stimuli, varying on two independent dimensions were used. We investigated the quadratic case. A PL/I program was written for the IBM 360/50 and 20 iterations were performed for each of the 38 different blocks of classifications. The output for each iteration was:

i) number of negative elements in t

ii) value of F

iii) value of ϕ

Neither w nor t were scaled while iterating. In table II the results for a typical good subject are given, in table I the results for a bad one. Some tentative conclusions from these results (and from the other 36 cases) are the following:

i) minimization is fast for higher values of F, slow for lower ones.

ii) although the program was constructed to minimize F, the ϕ-values also seem to converge to a minimum.

iii) the number of errors may increase, but the size of these errors decreases consistently.

LITERATURE

De Klerk, L.F.W., (1968) Probabilistic Concept Learning. Doctor Dissertation. University of Leiden.

De Leeuw, J. (forthcoming) Algorithms for the numerical analysis of ordinal data structures. University of Leiden.

Kuhn, H & Tucker, A.W. (eds) (1956) Ann of Math. Studies, Princeton

Saaty, T.L. (1959) Mathematical methods of Operations Research, New York.

	n_-	F	ϕ
1	15	1.645880	0.178660
2	14	1.052479	0.168305
3	15	0.687707	0.158026
4	15	0.459402	0.148042
5	15	0.313695	0.138506
6	17	0.219028	0.129639
7	19	0.158736	0.123532
8	20	0.119453	0.120807
9	20	0.091462	0.119618
10	20	0.070596	0.119074
11	19	0.054661	0.118812
12	19	0.042390	0.118677
13	19	0.032900	0.118604
14	19	0.025546	0.118563
15	18	0.019840	0.118541
16	18	0.015410	0.118527
17	18	0.011970	0.118517
18	18	0.009299	0.118511
19	18	0.007224	0.118508
20	18	0.005612	0.118505
1	2	0.010405	0.000271
2	2	0.009086	0.000237
3	2	0.007948	0.000207
4	2	0.006960	0.000181
5	2	0.006105	0.000159
6	1	0.005359	0.000140
7	1	0.004704	0.000123
8	1	0.004129	0.000108
9	1	0.003622	0.000095
10	1	0.003177	0.000083
11	1	0.002786	0.000073
12	1	0.002443	0.000064
13	1	0.002141	0.000056
14	1	0.001876	0.000049
15	1	0.001644	0.000043
16	1	0.001439	0.000038
17	1	0.001260	0.000033
18	1	0.001102	0.000029
19	1	0.000964	0.000025
20	1	0.000843	0.000022

Notes for RN 006-68 : Non-metric Discriminant Analysis

p 4: line 17 and 18: read

$$s_i^{(r+1)} = t_i \qquad \text{if } t_i \geqslant 0$$
$$s_i^{(r+1)} = 0 \qquad \text{if } t_i < 0$$

p 6: The numerator of ϕ is a function that was already used to obtain solutions of systems of inequalities (for example to find an initial feasible point in mathematical programming routines, see P. Wolfe: Methods of Non-linear Programming. In J. Abadie (ed): Non-linear programming. Amsterdam, North Holland, 1967). It is similar to the exponential method discussed in Saaty (l.c.), but scaling is easier. Scaling is necessary because in almost all cases our systems are inconsistent.

p 6: Although there is no program for NDA in the G-L-series as yet, Professor Louis Guttman informs me that he would favor the absolute value principle. From the (up to now) rather sketchy accounts of this principle I gather that it is quite similar to my positive orthant method (RN 010-68).

p 7: minimzing ϕ is equivalent to maximizing $\sum |t_i| t_i$ under the condition that $\sum t_i^2$ is some constant value. I.e.

$$f(w) = w'X' |Xw| - \mu(w'X'Xw - 1)$$

and the conditions for an extreme value are

$$X' |Xw| = \mu X'Xw$$
$$w'X'Xw = 1$$

It follows that

$$w'X' |Xw| = \mu$$

p 8: at the moment of writing this note Kees Truyens is testing a new version of NDA which uses the Newton-Raphson method to minimize ϕ. Further progress will be reported in a note similar to this one.

p 8: line 18: read 'ϕ-values'

p 8: Further numerical experience proves that after a (comparatively) fast decrease in the beginning, ϕ begins to increase very slowly after a larger number of iterations. F, however, continues to decrease. In fact, writing C for $X(X'X)^{-1}X'$, t for $t^{(n)}$, and u for $t^{(n+1)}$, we obtain (because C is idempotent)

$$u'u = \tfrac{1}{2}\,(t'Ct + 2|t|'Ct + |t|'C|t|)$$

Because C is idempotent

$$t'Ct \leqslant t't$$

Moreover

$$|t|'C|t| \leqslant \sqrt{|t|'|t| \cdot |t|'C|t|} = \sqrt{t't \cdot |t|'C|t|}$$

so

$$\sqrt{|t|'C|t|} \leqslant \sqrt{t't}$$

and

$$|t|'C|t| \leqslant t't$$

Finally

$$2\,|t|'Ct \leqslant 2\sqrt{|t|'|t| \cdot t'Ct} \leqslant 2t't$$

Adding these inequalties gives

$$0 \leqslant u'u \leqslant t't$$

Equality iff

$$C|t| = Ct = t = |t|$$

in other words iff C has at least one eigenvector with all elements $\geqslant 0$. Otherwise $t't \longrightarrow 0$, in which case $t \longrightarrow 0$. Of course than we also have equality, i.e. reached convergence.

Jan de Leeuw

13-12-'68