Behaviormetrika
1987, No. 21, 97-110

OPTIMAL LINEAR AND QUADRATIC CLASSIFIERS FOR TWO-GROUP DISCRIMINANT ANALYSIS[1)]

Yoshio Takane*

A simple algorithm was developed for estimating optimal linear and quadratic classifiers (OLC & OQC) for non-normal multivariate predictor variables in two-group discriminant analysis. The algorithm is based on the alternating least squares (ALS) principle. The optimal classifiers compared favorably with the linear and quadratic discriminant function (LDF & QDF) methods in true error rate. Possible generalizations of the optimal classifier approach (ridge regression, robust regression based on the weighted least squares, etc.) were discussed.

1. Introduction

Fisher's (1936) linear discriminant function (LDF) method is well established for equal covariance multivariate normal predictors (Anderson, 1958). Its optimality deteriolates, however, as the assumption of multivariate normality gets unrealistic (Krzanowski, 1975). In this paper we present an "optimal" linear classifier (OLC) method (along with a simple algorithm to estimate the classifier), which is applicable to non-normal cases. In particular the method allows a mixture of continuous and discrete predictor variables and their interactions.

In Section 2 we will present the basic method and the algorithm. They are generalized in various directions in Section 3. In Section 4 we compare the performance of OLC (and its quadratic analogue) with that of the conventional LDF method (and its quadratic counterpart). Discussion follows in Section 5.

2. Optimal linear classifier (OLC)

2.1 Optimization criterion and algorithm

It is well known that the sample analogue of Fisher's linear discriminant function (LDF) can be obtained by regression analysis by appropriately defining the dependent variable (Cramer, 1967; Tatsuoka, 1971). Let n_1 and n_2 represent the sample size of criterion groups 1 and 2, respectively, and define the vector of the dependent variable,

$$\boldsymbol{y}=\begin{pmatrix} \frac{1}{n_1}\boldsymbol{1}_{n_1} \\ -\frac{1}{n_2}\boldsymbol{1}_{n_2} \end{pmatrix} \tag{1}$$

where $\boldsymbol{1}_{n_1}$ and $\boldsymbol{1}_{n_2}$ are n_1- and n_2-component vector of ones. Let X be the matrix of predictor variables, and define a supermatrix X^* with a constant term appended to X;

Key Words and Phrases; linear discriminant function, alternating least squares (ALS), acceleration techniques, ridge regression, biased discrimination, quadratic discrimination, weighted least squares, robust regression, leaving-one-out method, bootstrap method

* Department of Psychology, McGill University, 1205 Avenue Docteur Penfield, Montreal, Quebec, H3A 1B1, Canada

1) The research reported in this paper was supported by Grant A6394 by Natural Sciences and Engineering Research Council of Canada.

i.e.,
$$X^* = [\mathbf{1}_n, X] \tag{2}$$
where $n = n_1 + n_2$. We apply regression analysis of $\boldsymbol{y}$ onto X^*, and obtain

$$\hat{\boldsymbol{b}} = (X^{*\prime} X^*)^{-1} X^{*\prime} \boldsymbol{y}, \tag{3}$$

the least squares (LS) estimate of regression coefficients, $\boldsymbol{b}$, which minimizes

$$f(\boldsymbol{b}) = (\boldsymbol{y} - X^* \boldsymbol{b})' (\boldsymbol{y} - X^* \boldsymbol{b}). \tag{4}$$

The estimate of regression coefficients given in (3) is known to be proportional to sample discriminant coefficients in LDF. (A more precise statement is given in the appendix). An observation $\boldsymbol{x}$ is assigned to group 1 if $\delta(\boldsymbol{x}) > 0$, and to group 2 if $\delta(\boldsymbol{x}) < 0$, where

$$\delta(\boldsymbol{x}) = \hat{\boldsymbol{b}}' \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} = \hat{b}_0 + \hat{\boldsymbol{b}}_1' \boldsymbol{x} \tag{5}$$

with $\hat{\boldsymbol{b}}' = (\hat{b}_0, \hat{\boldsymbol{b}}_1')$.

The dependent variable, $\boldsymbol{y}$, has to be defined in the specific way as in (1) in order for $\hat{\boldsymbol{b}}$ to have the direct relationship to the sample discriminant coefficients in LDF. However, in more general situations in which the assumption of normality is doubtful, there is no reason why we should stick with this definition of $\boldsymbol{y}$. We may even attempt to optimally scale $\boldsymbol{y}$. That is, we seek to find both $\boldsymbol{y}$ and $\boldsymbol{b}$ which simultaneously minimize a LS criterion analogous to (4). More specifically, define a normalized LS loss function,

$$g(\boldsymbol{y}, \boldsymbol{b}) = (\boldsymbol{y} - X^* \boldsymbol{b})' (\boldsymbol{y} - X^* \boldsymbol{b}) / \boldsymbol{y}' \boldsymbol{y} \tag{6}$$

where $\boldsymbol{y}$ (as well as $\boldsymbol{b}$) is considered a variable. The normalization factor (the denominator of (6)) is necessary, since the numerator of (6) can be made identically equal to zero by setting $\boldsymbol{y}$ to be a zero vector.

The criterion (6) is to be minimized with respect to $\boldsymbol{y}$ and $\boldsymbol{b}$ under some plausible restriction on $\boldsymbol{y}$. The restriction on $\boldsymbol{y}$ is necessary, since without it (6) can always attain its minimum at $\boldsymbol{y} = X^* \boldsymbol{b}$ for arbitary $\boldsymbol{b}$. Let $\boldsymbol{x}_i^*$ be the column vector of the i-th row of X^*. Since

$$\hat{y}_i = \delta(\boldsymbol{x}_i^*) = \hat{\boldsymbol{b}}' \boldsymbol{x}_i^* \tag{7}$$

has to be (at least) non-negative for case i to be classified into group 1, and non-positive to be classified into group 2, it seems natural to require:

$$\begin{cases} y_i \geqslant 0 & \text{if } i \text{ comes from group 1} \\ y_i \leqslant 0 & \text{if } i \text{ comes from group 2} \end{cases} \tag{8}$$

for $i = 1, \cdots, n$.

Several algorithms have been suggested along this line (Duda & Hart, 1973; Ho & Kashyap, 1965, 1966), mostly based on an iterative steepest descent method, to minimize (6) with respect to $\boldsymbol{y}$ and $\boldsymbol{b}$ under restriction (8). We propose a simpler and more efficient minimization algorithm based on the alternating least squares (ALS) method (de Leeuw, Young & Takane, 1976). A possible failure of the steepest descent algorithm was indicated by Hand (1981), who, in one of his examples, had to stop the iteration prematurely due to an excessive amount of computer time anticipated. The proposed ALS algorithm, on the other hand, took only a fraction of a second (on AMDAHL 5850) to obtain the optimal solution for the same problem.

The ALS algorithm we propose proceeds as follows:

Step 1 Initialize $\boldsymbol{y}$ by (1).

Step 2 For given $\boldsymbol{y}$ obtain the LS estimate of $\boldsymbol{b}$ that minimizes the numerator of (6). It is given by (3).

Step 3 For a given estimate of $\boldsymbol{b}$, obtain the LS estimate of $\boldsymbol{y}$ that minimizes the numerator of (6) under restriction (8).
It is given by:
$y_i = \hat{y}_i$, if i comes from group 1 and $y_i \geqslant 0$, or if i comes from group 2 and $y_i \leqslant 0$.
$y_i = 0$, if i comes from group 1 and $y_i < 0$, or if i comes from group 2 and $y_i > 0$.

Step 4 Normalize $\boldsymbol{y}$ so that $\boldsymbol{y}'\boldsymbol{y} = 1$.

Step 5 Convergence check. If the maximum change in $\boldsymbol{b}$ from the previous iteration is less than 10^{-6}, stop. Otherwise, go to Step 2.

The general convergence property of the ALS algorithm is discussed in de Leeuw, Young and Takane (1976). A crucial element is that in Step 2 and Step 3 the estimates of $\boldsymbol{b}$ and $\boldsymbol{y}$ indeed minimize the unnormalized LS criterion (the numerator of (6)) in turn. That Step 2 satisfies this condition is rather obvious. That Step 3 minimizes the numerator of (6) with respect to $\boldsymbol{y}$ for fixed $\hat{\boldsymbol{b}}$ follows from the separability of $\boldsymbol{y}$. That is,

$$(\boldsymbol{y} - X^*\hat{\boldsymbol{b}})'(\boldsymbol{y} - X^*\hat{\boldsymbol{b}}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{b}})^2, \tag{9}$$

so that the whole criterion is minimized by minimizing each term in summation separately. Hence we may set $y_i = \boldsymbol{x}_i'\hat{\boldsymbol{b}}$ if the sign of $\boldsymbol{x}_i'\hat{\boldsymbol{b}}$ agrees with case i's group membership, and $y_i = 0$ if it does not. (When the unconstrained estimate $y_i = \boldsymbol{x}_i'\hat{\boldsymbol{b}}$ does not satisfy the constraint, y_i is placed at the boundary of the feasible interval closest to $\hat{y}_i$.) That the normlized loss fruction (6) is in effect minimized by Step 4 following the two unnormalized minimization in Steps 2 and 3 has been discussed by de Leeuw (1977) and illustrated in Takane (1980, pp. 240-243).

The number of iterations to convergence may be significantly cut down by incorporating an acceleration technique (Ramsay. 1975) in updating $\hat{\boldsymbol{b}}$. Let $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{b}}^{(\mathrm{OLD})}$ be the current estimate (by (3)) and the old estimate of $\boldsymbol{b}$ in the previous iteration, respectively. Then the accelerated estimate $\hat{\boldsymbol{b}}^*$ in the current iteration is given by

$$\hat{\boldsymbol{b}}^* = \hat{\boldsymbol{b}}^{(\mathrm{OLD})} \times c + \hat{\boldsymbol{b}} \times (1 - c) \tag{10}$$

where c is the acceleration parameter. When $c = 0$, there is no acceleration. When $c < 0$, there is an acceleration. (When $c > 0$, a deceleration occurs. However, due to the monotonic convergence nature of the ALS algorithm, we never have to decelerate, and consequently we may set $c \leqslant 0$). The value of c is updated in every three iterations based on the behavior of $\hat{\boldsymbol{b}}^*$ in the previous three iterations. The exact updating formula for c is given in Ramsay (1975).

When the iterations are accelerated ($c < 0$), the monotonic convergence property of ALS no longer holds. To avoid divergence in practice we may set a minimum value for c somewhere between -2 and -5. In the first example below we set min $c = -3$. The

Table 1

Users of the University of London Computer Center divided into non-medical (1) and medical (2) users. The measurements are the logarithm of the numbers of units of computing under two different operating systems. Samples having zero values on either type unit are excluded. (Data from Hand, 1981, Table 2.1).

No.	Criterion group	Ln (Type I unit) x_1	Ln (Type II unit) x_2	Optimal discriminant score
1	1	5.938	5.407	−0.070*
2	1	4.304	4.883	−0.032*
3	1	0.000	3.761	0.088
4	1	6.620	6.849	−0.000*
5	1	7.686	8.157	0.043
6	1	5.916	5.914	−0.034*
7	1	6.986	9.264	0.153
8	1	7.098	8.268	0.078
9	1	8.581	8.447	0.022
10	1	7.473	8.520	0.078
11	1	4.754	8.591	0.209
12	1	6.812	9.392	0.170
13	1	6.818	9.657	0.188
14	1	5.513	6.782	0.046
15	1	4.220	5.347	0.005
16	1	4.595	8.454	0.206
17	1	3.932	8.168	0.217
18	1	4.143	9.207	0.280
19	1	4.554	7.653	0.152
20	1	5.118	8.703	0.200
21	1	5.717	8.178	0.135
22	1	5.652	8.127	0.134
23	1	4.533	5.112	−0.026*
24	1	4.407	8.960	0.251
25	1	1.386	8.019	0.324
26	1	3.555	6.537	0.119
27	1	1.099	3.258	0.002
28	1	6.805	6.807	−0.012*
29	2	4.898	4.522	−0.085
30	2	5.429	7.079	0.071*
31	2	3.989	4.317	−0.057
32	2	5.624	6.234	0.002*
33	2	5.389	7.409	0.096*
34	2	3.526	3.989	−0.059
35	2	5.694	6.075	−0.012
36	2	6.001	6.144	−0.021
37	2	5.684	2.890	−0.236
38	2	6.323	4.443	−0.156
39	2	5.394	6.295	0.017*
40	2	3.401	2.079	−0.188
41	2	3.434	4.820	0.004*
42	2	1.946	3.178	−0.043
43	2	5.878	6.568	0.014*
44	2	4.382	1.099	−0.302
45	2	5.771	6.114	−0.013
46	2	1.099	2.197	−0.073
47	2	5.924	6.100	−0.021
48	2	6.669	4.956	−0.136
49	2	3.850	0.000	−0.355

*misclassified

number of iteractions (19) was cut down to less than a half of what it took (43) without acceleration. (It happened that in this particular case further reduction (15) in the number of iterations was possible if no lower bound is imposed on c, but this practice may lead to more frequent divergences, and cannot be recommended generally.)

2.2 An example

The first example comes from Hand (1981, pp. 22-23), who applied a variety of discriminant analysis methods to the same set of data. Cases are users of the University of London Computer Center divided into non-medical (Group 1) and medical users (Group 2). Predictor variables are the numbers of units of computing used under two different operating systems (Type 1 units and Type 2 units). The measurements on the predictor variables are logarithmically transformed before the analysis is performed. Those cases with zero units on either of the two predictor variables are excluded from the analysis. There are 49 cases (28 in Group 1 and 21 in Group 2) left to be analyzed. The log-transformed data as well as group membership for the 49 cases are given in Table 1.

Table 2 presents the history of iterations, fairly typical of the OLC procedure. The convergence is smooth, particularly after the first iteration, and it is quick. It converged in 19 iterations with the acceleration technique described earlier. (It tends to take more iterations as the number of predictor variables gets larger). In order to avoid possible divergence the minimum of the c value was set to -3.0. The last column of Table 1 indicates the estimated y. This has to be non-negative for Group 1 to be correctly classified; it has to be non-positive for Group 2. There are six misclassifications in each of the two groups. (Apparent error rate is 6/28 for Group 1 and 6/21 for Group 2).

Fig. 1 displays the plot of the 49 cases in terms of the two predictor variables. Circles

Table 2
A typical iteration history with the OLC procedure.
(Data in Table 1).

Iteration No.	Optimization Criterion ($\sqrt{f}$)	Acceleration Parameter (c)	b_0	b_1	b_2
1	0.184450	0.000	−.051	−.006	.013
2	0.173584	0.000	−.251	−.034	.069
3	0.166406	0.000	−.241	−.035	.069
4	0.161495	0.000	−.232	−.036	.069
5	0.158137	0.000	−.223	−.037	.069
6	0.155841	0.000	−.216	−.038	.069
7	0.151387	−3.000	−.192	−.042	.069
8	0.150753	−3.000	−.181	−.043	.068
9	0.150620	−3.000	−.181	−.045	.070
10	0.150596	−2.722	−.166	−.042	.065
11	0.150592	−2.722	−.203	−.052	.080
12	0.150590	−2.722	−.099	−.025	.039
13	0.150590	0.000	−.174	−.045	.069
14	0.150589	0.000	−.174	−.045	.069
15	0.150589	0.000	−.174	−.045	.069
16	0.150589	−3.000	−.174	−.045	.069
17	0.150588	−3.000	−.174	−.045	.069
18	0.150588	−3.000	−.173	−.045	.069
19	0.150588	−3.000	−.173	−.045	.069

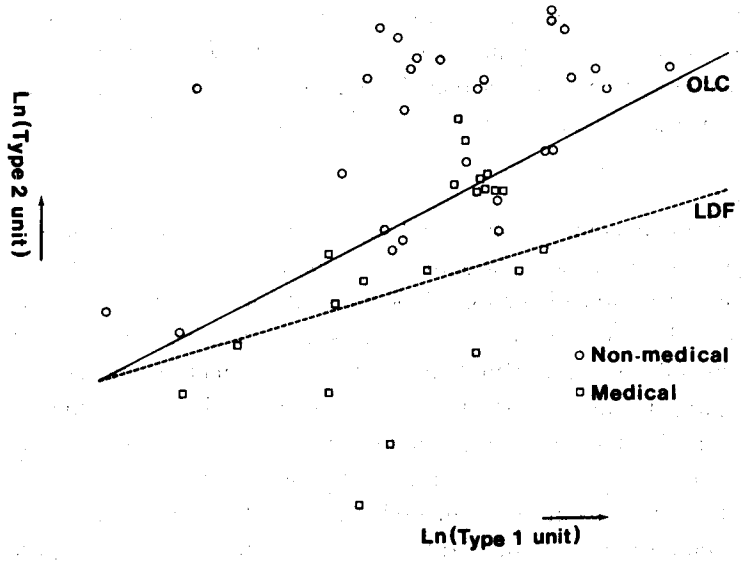

Fig. 1 Territorial maps for Hand's data obtained by the OLC and the LDF procedures.

represent cases in Group 1 and squares in Group 2. The boundary hyperplane obtained by the OLC procedure is shown by the solid line. For comparison the boundary hyperplane is also depicited for LDF and is shown by the dotted line. (This was obtained by Hand, 1981). This takes into account the observed sample sizes, n_1 and n_2. (That is, (A2) in the appendix is used rather than (A2′)). This actually worsens apparent error rate (the number of misclassifications in the current sample). The apparent error rate is zero in Group 1, while it is 13/21 in Group 2. However, the apparent error rate usually underestimates true error rate (simply because in the former the number of misclassifications is counted using the same sample used to estimate parameters in the discriminant function). A more rigorous comparison of the two procedures (OLC & LDF) will be made in Section 4 in terms of the true error rate.

3. Generalizations

The basic approach to OLC presented in the previous section may be generalized in various directions. In particular, since the discrimination problem has been formulated in terms of regression analysis, a variety of extensions which took place in regression analysis can be readily utilized.

3.1 Ridge regression and biased discrimination

The ridge regression estimate of $\boldsymbol{b}$ is obtained by

$$\tilde{\boldsymbol{b}}=(X^{*\prime}X^{*}+D^{2})^{-1}\,X^{*\prime}\boldsymbol{y} \tag{11}$$

where

$$D^{2}=\begin{bmatrix} s^{2} & \mathbf{0}' \\ \mathbf{0} & t^{2}I \end{bmatrix},$$

and s^2 and t^2 are some (non-negative) constant. (Comparability of measurement unit across predictor variables is assumed). Although this estimate of $\boldsymbol{b}$ is no longer unbiased (Whereas $\hat{\boldsymbol{b}}$ obtained by (3) is), it is often associated with a smaller mean square error (*e.g.*, Marquardt & Snee, 1975). In the Bayesian framework the smaller mean square error is understood as the Stein effect (Berger, 1982), which shrinks estimators toward zero. In the context of discriminant analysis ridge regression leads to so called biased discrimination (DiPillo, 1976).

The ridge estimator of $\boldsymbol{b}$ given by (11) may be used in connection with our OLC procedure. One potential problem is that it destroys the LS property of Step 2. In order for (11) to be consistent with the ALS procedure the optimization criterion, (6) has to be modified. The Bayesian framework is again helpful in this regard. Let the prior density of $\boldsymbol{b}$ multivariate normal with mean $\mathbf{0}$ and covariance D^2. Then minus twice the log of the posterior density of $\boldsymbol{b}$ is given by

$$g^*(\boldsymbol{y}, \boldsymbol{b})=(\boldsymbol{y}-X^*\boldsymbol{b})'(\boldsymbol{y}-X^*\boldsymbol{b})+\boldsymbol{b}'D^2\boldsymbol{b} \tag{12}$$

except for a constant term which does not involve $\boldsymbol{b}$ (Beck & Arnold, 1977). The MAP (Maximum *Aposteriori*) estimate of $\boldsymbol{b}$, which minimizes (12), is given by (11). Note that (12) is unnormalized; the normalization factor should be incorporated in the same way as in (6). The choice of s^2 and t^2 should be made taking this normalization factor into account. The optimal scaling of $\boldsymbol{y}$ is obtained in exactly the same way as before.

3.2 Optimal quadratic classifier

When the equal covariance assumption in multivariate normality is violated, quadratic discrimination (rather than the linear one) may be in order. This case can be easily accommodated into the regression framework by including quadratic terms such as x_1^2, x_2^2, and x_1x_2 in the set of predictor variables. Once the regression coefficients are estimated for the enlarged predictor set they are manipulated in a specific way to arrive at discrimination boundaries. Smith (1947) describes this process in detail. The derivation is particularly simple, when there are only two predictor variables *and* when they are mutually orthogonal. When they are orthogonal, we will not need cross-product terms such as x_1x_2. Consequently it may be wise to orthogonalize the predictor variables *apriori*, whenever the quadratic discrimination is required. Alternate applications of the quadratic LS regression and the optimal scaling of $\boldsymbol{y}$ lead to the optimal quadratic classifier (which is abbreviated as OQC).

Both OLC and OQC are applied to Laurie's data (given in Hand, 1981, p. 30). The data are concerned with estimated numbers of casualties caused by a nuclear strike on fifteen largest British cities (excluding London). The fifteen cities are divided into two groups, high casualty and low casualty cities, which are used as criterion groups. The high casualty cities (Group 1) are: 1. Birmingham, 2. Manchester, 3. Sheffield, 4. Liverpool, 5. Hull, 6. Nottingham, 7. New Castle (upon Tyne and Gateshead), and 8. Glasgow. The low casualty cities (Group 2) are: 9. Leeds, 10. Leicester, 11. Coventry, 12. Stoke-on-Trent and New Castle under Lyme, 13. Bristol, 14. Cardiff, and 15. Edinburgh. Predictor variables are population and area of the cities.

Fig. 2 depicts the territorial map obtained by OLC and OQC. While OLC misclassifies

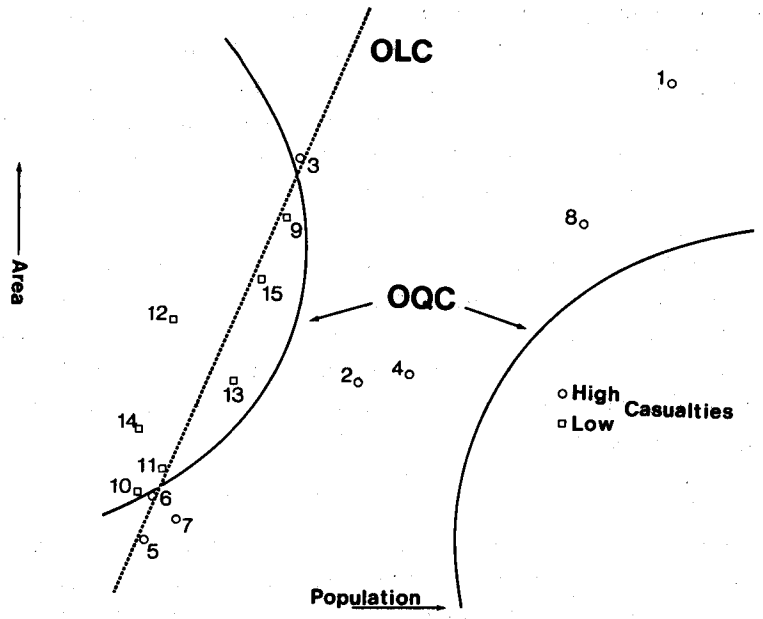

Fig. 2 Territorial maps obtained by OLC and OQC for Laurie's data (Hand, 1981, p. 30).

two cities (3 & 6) in Group 1 and three cities (9, 13 & 15) in Group 2, OQC perfectly discriminates the two groups. The superiority of OQC over OLC in this example is also confirmed by true error rate we will look at in Section 4. Note that the OQC boundaries happen to be a hyperbola in this example, although we cannot attach any meaning to the extreme right region that belongs to Group 2. There are no actual cities located in this region. Hand (1981, p. 29) applied kernal discriminant analysis to the same set of data, which gave a somewhat irregular boundary.

3.3 Weighted least squares and robust regression

The simple (unweighted) LS in (4) may be generalized to the weighted LS. This is particularly simple when the weight matrix is diagonal. The optimal scaling of $\boldsymbol{y}$ remains exactly the same as before, since in this case the (unnormalized) weighted LS criterion is separable; *i.e.*,

$$(\boldsymbol{y}-X^*\hat{\boldsymbol{b}})'W(\boldsymbol{y}-X^*\hat{\boldsymbol{b}})=\sum w_i(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{b}})^2, \tag{13}$$

where w_i is the i-th diagonal element of weight matrix W. The unweighted LS estimate of $\boldsymbol{b}$ in (3) must be replaced by the appropriate weighted LS estimate,

$$\hat{\boldsymbol{b}}=(X^{*\prime}WX^*)^{-1}X^{*\prime}W\boldsymbol{y}. \tag{14}$$

The normalization factor should also be modified; *i.e.*, we use $\boldsymbol{y}'W\boldsymbol{y}$ instead of $\boldsymbol{y}'\boldsymbol{y}$. When W is not diagonal, (14) is still valid. However, the optimal scaling of $\boldsymbol{y}$ requires a more sophisticated quadratic programming method (*e.g.*, Stoer, 1971).

For illustration we applied the weighted LS OLC to Neal's data reported in Smith (1947, p. 277). The data consist of two criterion groups, 25 normals (Group 1) and 25

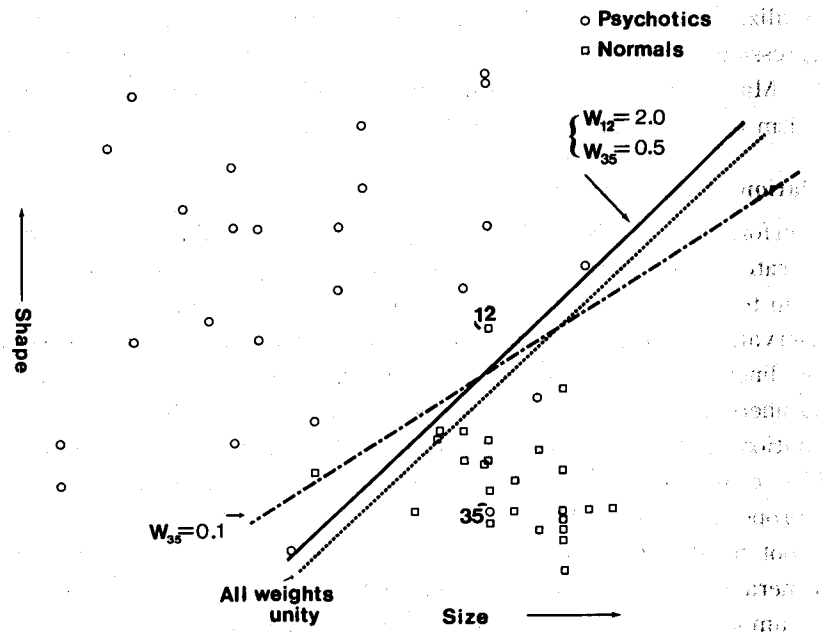

Fig. 3 Territorial maps obtained by OLC with three different weight matrices. The data are from Neal (reported in Smith, 1947, p. 277).

psychotics (Group 2). Predictor variables are some sort of size scores and shape scores. Fig. 3 displays the cases in terms of the two predictor variables used as coordinates. Normals (shown by squares) are clustered in the lower right corner. Smith (1947) previously applied the quadratic discriminant function (QDF) method to the data. Here we use the same set of data to demonstrate how the weights affect discrimination.

We first analyzed the data by the simple OLC (this corresponds with $W = I$). The boundary hyperplane for this case is indicated by the dotted line in the figure. Four cases in Group 1 are misclassified into Group 2, while two cases in Group 2 are misclassified into Group 1. Inspection of Fig. 1 suggests that the derived boundary hyperplane may be located too low. In fact by moving the boundary up slightly in parallel to the original, two of the misclassified cases in Group 1 can be saved. The boundary cannot be moved up too high, because two cases correctly classified in Group 2 would then be misclassified.

We have experimented on several sets of weights to come up with the set of weights that can lift the boundary to a desired position. The original boundary is lowered because of Case 35 which is deep in the wrong side of the boundary. The weight for this case was first lowered to 0.1 to reduce its influence while keeping all the other weights at unity. The derived boundary is shown by the broken line. The bottom part of the boundary has gone too far in this case. The weight for Case 35 was then increased to 0.5 and at the same time the weight for Case 12 is increased to 2.0 to lift the upper portion of the boundary. The resultant boundary is indicated by the solid line, which is more or less optimal. The total number of misclassifications is now reduced to four. As will be seen in Section 4, this weighting scheme also works better in terms of true error rate.

Although the search for an optimal set of weights sounds a bit arbitrary in the present

case (capitalizing too much on the data), an important implication is that we may use robust regression techniques (*e.g.*, Ramsay, 1977) in conjunction with our optimal scaling approach. Many techniques for robust regression adjust weights according to some built-in mechanism to control influence of influential cases.

4. Estimation of true error rate

The performance of discriminant analysis methods should be evaluated in terms of true error rate. More flexible data-oriented methods may do very well for the data at hand (*i.e.*, in terms of apparent error rate) but their performance may not carry over to future observations. In view of this an estimate of the true error rate was obtained for LDF/QDF (linear/quadratic discriminant functions) and OLC/OQC (optimal linear/quadratic classifiers) using the bootstrap method (Efron, 1983). While there are other available estimation methods (*e.g.*, McLachlan, 1976; the leaving-one-out method by Lachenbruch, 1975; cross-validation by Stone, 1974), the bootstrap method seems to have more desirable properties (Gong, 1986).

The bootstrap method attempts to correct bias in the apparent error rate. This is done by generating many sets of so called bootstrap samples. In the present context each bootstrap sample is generated in such a way that it consists of n_1 cases drawn randomly (with probability $1/n_1$) from group 1 with replacement and n_2 cases similarly drawn from group 2. For each bootstrap sample the difference is taken between the error rate in classifying the original sample (using the estimates derived from the bootstrap sample) and the apparent error rate in the bootstrap sample. This difference is averaged across bootstrap samples to obtain an estimate of bias in the apparent error rate in the original sample. The bootstrap estimates of true error rate in Table 3 are based on 1000 bootstrap

Table 3
Estimates of true error rate

	LDF/QDF	OLC/OQC	
	Bootstrap method	Bootstrap method	Leaving-one-out method
Example 1 (Hand's data)	.303 .197 (.258)	.211 .274 (.234)	.214 .286 (.245)
Example 2 (Laurie's deta) linear	.306 .035 (.180)	.255 .433 (.338)	.250 .429 (.334)
quadratic	.210 .043 (.132)	.091 .077 (.068)	.000 .000 (.000)
Example 3 (Neal's data) Weights all unity	.008 .174 (.091)	.157 .077 (.117)	.160 .080 (.120)
$w_{12}=20$; $w_{35}=.05$		.078 .079 (.079)	.080 .080 (.080)

In each cell the top figure indicates the estimated misclassification rate in group 1, the middle the misclassification rate group 2, and the bottom a weighted average of the top two.

samples each.

For Hand's data (discussed in Section 2.2) LDF and OLC perform approximately equally with the latter having a slight edge over the former. Both linear and quadratic discrimination methods (LDF/QDF and OLC/OQC) were applied to Laurie's data (discussed in Section 3.2). For linear discrimination LDF did considerably better than OLC. However, for quadratic discrimination which is deemed more appropriate for this data set, the optimal scaling method (OQC) outperformed the usual quadratic discriminant function method (QDF). That the quadratic discrimination methods are better than the linear ones for this data set is indicated by the smaller true error rates in the former. For Neal's data (discussed in Section 3.3) LDF was found slightly better than OLC with the simple (unweighted) LS, but with appropriate weights OLC could be made to perform at least as well as LDF. The weighting scheme discussed earlier indeed improved the discrimination in terms of the true error rate.

Overall the bias in the apparent error rate is only minimal for OLC/OQC (*i.e.*, the apparent and true error rates are not very much different from each other). Also, for OLC/OQC there is a good agreement between the bootstrap method and the leaving-one-out method. This may indicate some robustness of the optimal classifier approach. In fact it can be shown that for this approach the apparent and true error rates estimated by the leaving-one-out method are always identical (there is no bias in the estimate of the apparent error rate). It may be due to the fact that in the optimal classifier approach only misclassified cases contribute to the loss function. On the other hand, it may imply that the optimal classifier approach is stable when the error rate is small, but it may not be so for more difficult discrimination problems. (This may be the reason why OLC was rather poor for Laurie's data). This point needs further investigations.

5. Discussion

In this paper we presented a relatively robust method for two-group discriminant analysis. The method enjoys wider applicability than the ordinary discriminant function method, since no distributional assumptions are made on the predictor variables. Hence, for example, a mixture of discrete and continuous predictors can be accommodated quite naturally. Extensions to biased discrimination, quadratic (and higher order polynomial) discrimination, and robust discrimination are also rather straightforward. An extension to multiple-group discriminant analysis seems also quite straightforward, although this case was not explicitly discussed. Obviously in order to claim superiority of the present method in more general contexts much work needs to be done, particularly in reference to performance of other discriminant analysis methods, such as kernel discriminant analysis (Hand, 1982), biased discrimination (DiPillo, 1976), classification trees (Breiman, *et al.*, 1983), logistic discrimination (Cox, 1966 ; Anderson, 1972), the location model (Krzanowski, 1975) and ideal point discriminant analysis (Takane & Shibayama, 1984).

A couple of optimal scaling procedures have been proposed previously for discriminant analysis based on the ALS framework. One is MORALS/CORALS (multiple and canonical optimal regression by ALS) by Young, de Leeuw and Takane (1976) and the other called CRIMINALS (Gifi, 1981, pp. 234-235). MORALS is, however, based on the monotonicity principle, and CRIMINALS on the homogeneity principle. These procedures are thus

distinct from our OLC, which is based on fixed boundary constraints (or sign constraints). Comparisons between OLC and the other ALS procedures (MORALS/CORALS and CRIMINALS) would undoubtedly be interesting.

Appendix

The $\hat{\boldsymbol{b}}$ given in (3) is proportional to sample discriminant coefficients in LDF. This is always true for the $\hat{\boldsymbol{b}}_1$ part (coefficients applied to X). However, it is strictly true for $\hat{b}_0$ (coefficient for the constant term) only when $n_1=n_2$.

The sample analogue of LDF is given by

$$\lambda(\boldsymbol{x})=\hat{\beta}_0+\hat{\boldsymbol{\beta}}_1'\boldsymbol{x}$$

with

$$\hat{\boldsymbol{\beta}}_1=S^{-1}\boldsymbol{d} \tag{A1}$$

$$\hat{\beta}_0=\ln(n_1/n_2)-\hat{\boldsymbol{\beta}}_1'(\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2)/2 \tag{A2}$$

where $\boldsymbol{d}=\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2$, S is the pooled within-group sample covariance matrix, and $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$ are the sample means of group 1 and group 2, respectively. The assignment is to group 1 if $\lambda(\boldsymbol{x})>0$, and to group 2 if $\lambda(\boldsymbol{x})<0$. The $\hat{\beta}_0$ in (A2) reduces to

$$\hat{\beta}_0=-\hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}} \tag{A2'}$$

when $n_1=n_2$, where $\bar{\boldsymbol{x}}$ is the vector of sample grand means, $\bar{\boldsymbol{x}}=\frac{1}{n}X'\boldsymbol{1}_n$.

From (1), (2) and (3) in the main text we have

$$\hat{\boldsymbol{b}}=\begin{pmatrix}\hat{b}_0\\ \hat{\boldsymbol{b}}_1\end{pmatrix}=\begin{bmatrix}n & \boldsymbol{1}_n'X\\ X'\boldsymbol{1}_n & X'X\end{bmatrix}^{-1}\begin{pmatrix}0\\ \boldsymbol{d}\end{pmatrix}$$

But because of a well known inversion formula for partitioned matrices (*e.g.*, Yanai & Takeuchi, 1983, p. 15), we obtain

$$\begin{pmatrix}\hat{b}_0\\ \hat{\boldsymbol{b}}_1\end{pmatrix}=\begin{bmatrix}\frac{1}{n}+\bar{\boldsymbol{x}}'S_T^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}'S_T^{-1}\\ -S_T^{-1}\bar{\boldsymbol{x}} & S_T^{-1}\end{bmatrix}\begin{pmatrix}0\\ \boldsymbol{d}\end{pmatrix}=\begin{pmatrix}-\bar{\boldsymbol{x}}'S_T^{-1}\boldsymbol{d}\\ S_T^{-1}\boldsymbol{d}\end{pmatrix}=\begin{pmatrix}-\hat{\boldsymbol{b}}_1'\bar{\boldsymbol{x}}\\ S_T^{-1}\boldsymbol{d}\end{pmatrix} \tag{A3}$$

where S_T is the total sum of squares matrix given by $S_T=X'(I-\frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n')X$. But since S and S_T are related by

$$(n-2)\,S=S_T-k\boldsymbol{d}\boldsymbol{d}'$$

where $k=n/n_1n_2$,

$$S^{-1}=(n-2)\,(S_T^{-1}-m\,S_T^{-1}\,\boldsymbol{d}\boldsymbol{d}'\,S_T^{-1})$$

where $m=(1/k+\boldsymbol{d}'S^{-1}\boldsymbol{d})^{-1}$ (See, for example, Yanai & Takeuchi, 1983, p. 18), we have

$$\hat{\boldsymbol{\beta}}_1=S^{-1}\boldsymbol{d}=qS_T^{-1}\boldsymbol{d}=q\hat{\boldsymbol{b}}_1 \tag{A4}$$

where $q=(n-2)\,(1-mc)$ with $c=\boldsymbol{d}'S_T^{-1}\boldsymbol{d}$. That is, $\hat{\boldsymbol{\beta}}_1$ is always proportional to $\hat{\boldsymbol{b}}_1$ with q as the constant of proportionality. From (A2') we have

$$\hat{\beta}_0=-\hat{\boldsymbol{\beta}}_1'\bar{\boldsymbol{x}}=-q\hat{\boldsymbol{b}}_1'\bar{\boldsymbol{x}}=q\hat{b}_0 \tag{A5}$$

That is, $\hat{\beta}_0$ is also proportional to $\hat{b}_0$ with the same constant of proportionality q. However, this is true for $\hat{\beta}_0$ only when $n_1 = n_2$ and hence

$$(\bar{x}_1 + \bar{x}_2)/2 = \bar{x} = \frac{1}{n}(n_1\bar{x}_1 + n_2\bar{x}_2).$$

REFERENCES

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika,* **59**, 19-35.

Anderson, T.W. (1958). *An introduction to multivariate statistical analysis.* New York : Wiley.

Beck, J.V., and Arnold, K.J. (1977). *Parameter estimation in engineering and science.* New York : Wiley.

Berger, J. (1982). Bayesian robustness and the Stein effect. *Journal of the American Statistical Association,* **77**, 358-368.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and regression trees.* Belmont, CA : Wadsworth.

Cox, D.R. (1966). Some procedure connected with the logistic qualitative response curve. In David, F.N. (Ed.), *Research papers in statistics : Festschrift for J. Neyman.* New York : Wiley, 55-71.

Cramer, M.E. (1967). Equivalence of two methods of computing discriminant function coefficients. *Biometrics,* **23**, 153.

de Leeuw, J. (1977). A normalized cone regression approach to alternating least squares algorithms. Unpublished manuscript, Department of Data Theory, University of Leiden.

de Leeuw, J., Young, F.W., and Takane, Y. (1976). Additive structure in qualitative data : An alternating least squares method with optimal scaling features. *Psychometrika,* **41**, 471-503.

DiPillo, P.J. (1976). The application of bias to discriminant analysis. *Commun. Statist Theor. Meth., A5(9),* 843-854.

Duda, R., and Hart, P. (1973). *Pattern classification and scene analysis.* New York : Wiley.

Efron, B. (1983). Estimating the error rate of a prediction rule : improvement on cross-validation. *Journal of the American Statistical Association,* **78**, 316-331.

Fisher, R.A. (1936). The use of multiple measurement in taxonomic problems. *Annals of Eugenics,* **7**, 179-188.

Gifi, A. (1981). *Non-linear multivariate analysis.* Dept. of Data Theory, Univ. of Leiden.

Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap : excess error estimation in forward logistic regression. *Journal of the American Statistical Association,* **81**, 108-113.

Hand, D.J. (1981). *Discrimination and classification.* Chichester : Wiley.

Hand, D.J. (1982). *Kernel discriminant analysis.* Chichester : Research Studies Press.

Ho, Y.C., and Kashyap, R.L. (1965). An algorithm for linear inequalities and its applications. *IEEE Transactions on Electronic Computers,* **14**, 683-688.

Ho, Y.C., and Kashyap, R.L. (1966). A class of iterative procedures for linear inequalities. *Journal of SIAM Control,* **4**, 112-115.

Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association,* **70**, 782-790.

Lachenbruch, P.A. (1975). *Discriminant analysis.* New York : Hafner.

Marquardt, D.W., and Snee, R.D. (1975). Ridge regression in practice. *The American Statistician,* **29**, 3-20.

McLachlan, G.J. (1976). A criterion for selecting variables for the linear discriminant function. *Biometrics,* **32**, 529-534.

Ramsay, J.O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika,* **40**, 337-360.

Ramsay, J.O. (1977). A comparative study of several robust estimates of slope, intercept, and scale in linear regression. *Journal of the American Statistical Association,* **72**, 608-615.

Smith, C.A.B. (1947). Some examples of discrimination. *Annals of Eugenics,* **13**, 272-282.

Stoer, J. (1971). On the numerical solution of constrained least squares problems. *SIAM-Journal of Numerical Analysis,* 8, 382-411.

Stone, M. (1974). Cross validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society,* **36**, 111-147.

Takane, Y. (1980). *Multidimensional scaling.* Tokyo: University of Tokyo Press (in Japanese).

Takane, Y., and Shibayama, T. Multiple discriminant analysis for predictor variables measured at various scale levels. Proceedings of the 11th annual meeting of the Behaviormetric Society of Japan.

Tatsuoka, M.M. (1971). *Multivariate analysis.* New York: Wiley.

Yanai, H., and Takeuchi, K. (1983). *Projection matrices, generalized inverse and singular value decomposition.* Tokyo: University of Tokyo Press, (in Japanese).

Young, F.W., de Leeuw, J., and Takane, Y. (1976). Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. *Psychometrika,* **41**, 505-529.

(Received July, 1986)