Zero-Mean Difference Discrimination and the Absolute Linear Discriminant Function

By

Peter A. Lachenbruch

University of California, Berkeley and
University of North Carolina, Chapel Hill

Institute of Statistics Mimeo Series No. 971
January 1975

Zero-Mean Difference Discrimination and the Absolute Linear Discriminant Function

Summary

A simple model for discriminating equal mean data is to perform a linear discriminant analysis on the absolute value of deviations from the mean. This avoids the necessity of writing a program to calculate a quadratic discriminant function, and it also seems to have some desirable robustness properties when long-tailed contamination is present. The method is applied to the well-known Stocks twins data.

Key Words: Discriminant analysis, equal means, robustness to contamination, selection of variables.

I. Introduction

The problem of assigning an individual to one of two populations which have the same mean is approached by an investigation of the covariance matrices in the two populations. Such a problem arises in the attempt to assign twin pairs to monozygotic or dizygotic groups on the basis of a set of measurements of physical characteristics (Stocks 1933).

The solution to the classification problem when the covariance matrices are different is to assign to the first population (Π_1) if (assuming normality)

$$\log \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = \frac{1}{2} \log |\underset{\sim}{\Sigma}_2|/|\underset{\sim}{\Sigma}_1| + \frac{1}{2} (\underset{\sim}{x}-\underset{\sim}{\mu}_2)' \underset{\sim}{\Sigma}_2^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2) - \frac{1}{2} (\underset{\sim}{x}-\underset{\sim}{\mu}_1)' \underset{\sim}{\Sigma}_1^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)$$

$$> \log \frac{1-p}{p} \qquad (1)$$

where p is the a priori probability of an unknown observation coming from Π_1. If $p = \frac{1}{2}$ and $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \underset{\sim}{0}$, (1) simplifies to

$$\frac{1}{2} \log |\underset{\sim}{\Sigma}_2|/|\underset{\sim}{\Sigma}_1| + \frac{1}{2} \underset{\sim}{x}' \left(\underset{\sim}{\Sigma}_2^{-1} - \underset{\sim}{\Sigma}_1^{-1}\right)\underset{\sim}{x} > 0$$

This was the formulation used by Bartlett and Please (1963). If the parameters are unknown, maximum likelihood estimates are used for the elements of $\underset{\sim}{\Sigma}_1$ and $\underset{\sim}{\Sigma}_2$. When sample sizes are small structural assumptions can be made on the covariance matrix to simplify estimation.

Two problems that the quadratic discriminant function suffers from are (1) no programs are generally available to perform the computations, and (2) it seems to be sensitive to long tailed contaminators. A possible solution for this is to consider performing a linear discriminant analysis on the absolute values of the observations. This offers the advantage of being able to use

available programs such as those in the BMD series, and as will be seen in section 3, is almost as good as the quadratic discriminant function when the populations are close together, and is much more resistant to contamination effects. I have termed this the Absolute Linear Discriminant Function.

2. Selection of Variables

A method for selecting variables might be based on estimated distances between populations. Cleveland and Lachenbruch (1974) suggested a measure based on the amount of overlap between Π_1 and Π_2. Let $f_{ij}(x)$ be the density of x_j in Π_i. The region in which $f_{1j} > f_{2j}$ is an interval $a_j < x < b_j$ (a_j or b_j may be infinite if $\sigma_{1j} = \sigma_{2j}$). In this case the probability of misclassification for Π_2 is

$$P_{2j} = \int_{a_j}^{b_j} f_{2j}(x)dx$$

Similarly, the probability of misclassification for Π_1 is

$$P_{1j} = \int_{-\infty}^{a_j} f_{1j}(x)dx + \int_{b_j}^{\infty} f_{2j}(x)dx$$

The total error is $P_{1j} + P_{2j}$. The variables with minimum error should be the ones which classify best. That is, one should select variable ℓ if $P_{1\ell} + P_{2\ell} = \min_j(P_{1j} + P_{2j})$. If desired, weighted forms of this measure can be used to take account of differing a priori probabilities. One could also use a conditional form given the previously entered variables. No studies of the use of this method for variable selection have been made. The measure is equivalent to Mahalanobis D^2 if the densities are normal and the covariance matrices are equal.

If the means are equal, this is equivalent to selecting variables on the basis of $F = \max_i(s_i^2)/\min_i(s_i^2)$.

3. Properties of the Absolute Linear Discriminant Function

The results given in this section are based on univariate studies only, and suggest what may happen in the multivariate case. Studies are in progress regarding multivariate properties of the absolute linear discriminant function. The work reported here is of two types: (1) the behavior of the absolute linear discriminant function when no contamination exists; (2) the relative behavior of the quadratic discriminant function and absolute linear discriminant function when contamination is present.

If x is $N(0,\sigma^2)$ then $y = |x|$ has the density

$$f(y) = \frac{2}{(2\pi)^{\frac{1}{2}}\sigma} e^{-y^2/2\sigma^2} \qquad \text{when } y \geq 0$$

which has the mean $\mu_y = \sigma(2/\pi)^{\frac{1}{2}}$ and variance $\sigma_y^2 = \sigma^2(1 - 2/\pi)$. The absolute linear discriminant function assigns y to Π_1 if

$$y - \frac{1}{2}(2/\pi)^{\frac{1}{2}}(\sigma_1 + \sigma_2) > 0$$

when $\sigma_1 > \sigma_2$, which is equivalent to

$$y > (\sigma_1 + \sigma_2)(2\pi)^{-\frac{1}{2}}$$

If $\sigma_1 < \sigma_2$ the inequality is reversed. The probabilities of misclassification may be calculated as follows:

$$P_1 = \text{pr}(Y < (\sigma_1 + \sigma_2)\ (2\pi)^{-\frac{1}{2}} | \Pi_1)$$

$$= \text{pr}\left[-(2\pi)^{-\frac{1}{2}}(\sigma_1 + \sigma_2) < X < (2\pi)^{-\frac{1}{2}}(\sigma_1 + \sigma_2)\right]$$

$$= \Phi\left[(2\pi)^{-\frac{1}{2}}\ \frac{(\sigma_1 + \sigma_2)}{\sigma_1}\right] - \Phi\left[-\ (2\pi)^{-\frac{1}{2}}\ \left[\frac{\sigma_1 + \sigma_2}{\sigma_1}\right]\right]$$

and

$$P_2 = \text{pr}(Y > (2\pi)^{-\frac{1}{2}}(\sigma_1 + \sigma_2) | \Pi_2)$$

$$= \text{pr}\left[X < -\ (2\pi)^{-\frac{1}{2}}(\sigma_1 + \sigma_2)\right] + \text{pr}\left[X > (2\pi)^{-\frac{1}{2}}(\sigma_1 + \sigma_2)\right]$$

$$= 2\Phi\left[-\ (2\pi)^{-\frac{1}{2}}(\sigma_1 + \sigma_2)/\sigma_2\right]$$

As $\sigma_1 \to \infty$ the absolute linear discriminant function has the following properties. The cutoff tends to $\sigma_1/(2\pi)^{-\frac{1}{2}}$, and so $P_2 \to 0$. P_1 tends to

$$\Phi((2\pi)^{-\frac{1}{2}}) - \Phi(-\ (2\pi)^{-\frac{1}{2}}) \approx .31$$

The quadratic discriminant function discriminant point is $(-2 \log \sigma_2/\sigma_1)/(1/\sigma_2^2 - 1/\sigma_1^2) = C$ and P_1 and P_2 both tend to zero when $\sigma_1 \to \infty$ (i.e., x is assigned to Π_1 if $x^2 > C$ when $\sigma_1 > \sigma_2$). Table 1 compares the error rates for the quadratic discriminant function and absolute linear discriminant function for various values of σ_1, and it is assumed that $\sigma_2 = 1$ and $\sigma_1 > \sigma_2$. From the table we may conclude that if σ_1/σ_2 is not too large, the absolute linear discriminant function is about as good as the quadratic discriminant function. In particular if $\sigma_1/\sigma_2 \leq 5$, $\bar{P}$ is never more than .015 worse for the absolute linear discriminant function. The cutoff point for the absolute linear discriminant function approaches ∞ too rapidly for large values of σ_1, and this is why the error rate does not go to zero.

[Table 1 goes here]

We next consider the behavior of the quadratic discriminant function and absolute linear discriminant function when contamination is present. The device we use is called the stylized sensitivity curve by Andrews et al. (1972) and is computed as follows. The observations from Π_1 are assumed to be equal to the expected value of the order statistics. Thus if a sample of size three were taken from Π_1, we would have $x_1 = -.846$, $x_2 = .000$, $x_3 = .846$. The observations from Π_2 have the expected values of the order statistics for a sample of size n_2-1. The remaining observation is then allowed to vary, and its effect on the error rates of the quadratic discriminant function and absolute linear discriminant function may be calculated by obtaining estimates of σ_1^2 and σ_2^2 for the quadratic discriminant function and the means of the absolute value distributions for the absolute linear discriminant function. Figure 1 gives such a picture for $n_1 = n_2 = 20$ and four combinations of σ_1 and σ_2. We see a clear picture in these curves. If $\sigma_1 < \sigma_2$ the quadratic discriminant function's performance shows a steady decline as the contaminating observation becomes larger. The quadratic discriminant function is slightly superior to the absolute linear discriminant function in this case; the average error rate for the absolute linear discriminant function is never more than .03 greater than that of the quadratic discriminant function. This is because the contaminating observation falls in the population with the larger variance. When $\sigma_1 > \sigma_2$ an entirely different picture emerges. For the quadratic discriminant function, the P_1 curve rises until the contaminating observation is large enough to cause s_1^2 to be less than s_2^2 at which point it drops sharply and levels off. The P_2 values are fairly small until the critical point when they become quite large (.995 or greater) and remain that way. The absolute linear discriminant function behaves in a similar way except that the critical point is much larger than that for the quadratic discriminant function.

The differences in $\bar{P}$ were quite large: when $\sigma_1=3$ $\sigma_2=1$ the largest difference was .41 in favor of the absolute linear discriminant function, and when $\sigma_1=5$ $\sigma_2=1$ a difference of .46 was observed. Similar curves are abailable for $n_1=n_2=10$. The critical point is much smaller, as one might expect, since in this case 10% of Π_2 is contaminated whereas in the data reported here only 5% is contaminated. In general, if the contaminating observation is such that it brings the estimated variances closer than they actually are, or reverses the magnitude (i.e., $s_1^2 > s_2^2$ when $\sigma_1^2 < \sigma_2^2$) the effects can be serious.

4. Examples

Stocks (1933) recorded a number of measurements on school children in London in 1925 to 1927. Bartlett and Please selected 10 of these variables to study their ability to distinguish monozygotic (Π_1) from dizygotic twins (Π_2). It should be noted here that modern methods based on blood groupings are more accurate for this purpose as Bartlett and Please point out. Desu and Geisser (1973) studied this problem from a Bayesian viewpoint. The variables used were

1. Height
2. Weight
3. Head Length
4. Head Breadth
5. Head Circumference
6. Interpupillary Distance
7. Systolic Blood Pressure
8. Pulse Interval
9. Strength of Grip, Right Hand
10. Strength of Grip, Left Hand

The data consisted of the difference between the measurement for the twins. There were 46 monozygotic twins (n_1) and 48 dizygotic twins (n_2). We assume that the a priori probability of a monozygotic twin pair is 1/2. Table 2 gives the values of the means and variances in each population and the F's for selection. Variables 7-10 add little to the analysis, although in conjunction with

other variables they might be useful. Variables 1, 5 and 6, Height, Head Circumference and Interpupillary Distance are the best variables.

[*Table 2 goes here*]

Three discriminant analyses were performed: one using the full set of ten variables, one using only variables 1-6, and one using variables 1, 5 and 6. To evaluate the discriminant functions thus produced, estimates of the error rates were obtained. The apparent error rate is found by substituting the initial observations into the discriminant function. The bias of this estimate is worse for a quadratic function than a linear function because of the greater number of parameters that must be estimated. The leaving-one-out method estimates the error rates by sequentially leaving each observation out of the calculations for the coefficients and classifying it, thus obtaining a less biased estimate of the error rates (Lachenbruch, 1974). Table 3 gives the results for each of the three analyses. The substantial bias in the apparent error rate primarily affects the error rate in Π_1. A 6 variable function seems preferable to the 10 variable function. Using 3 variables is somewhat less effective than using 6 variables but is slightly preferable to using 10 variables. Table 4 gives the apparent error rate and the leaving-one-out error rate for the absolute linear discriminant function. In this case the best function is the three variable rule, which performs as well as the 6 variate quadratic discriminant function. The absolute linear discriminant function performs slightly poorer than the quadratic function for 6 variables, and slightly better for 3 variables. The reason for this seems to be that the variables are not too non-normal. For some variables, the tails are slightly shorter than normal. As the absolute linear discriminant function is designed to protect against long-tailed contamination, this would operate in favor of the quadratic function. Plots of

the variables on normal probability paper did not reveal any major non-normality.

[*Tables 3 and 4 go here*]

To study the behavior of the two rules when contamination is present, I multiplied the components of the first five observations in Π_1 by 3 and performed the analyses for three, six and ten variables for each rule. The results of these analyses are given in Table 5. Again we see the substantial bias of the apparent error rate for the quadratic discriminant function. Compared to the uncontaminated cases in Table 3 there is an increase in the mean error rate (using the leaving-one-out method) of .103 for 3 variables and .181 for 6 variables. For the absolute linear discriminant function the increase is .105 for 3 variables and .074 for 6 variables. The error rate for the absolute linear discriminant function is lower than that for the quadratic discriminant function for 3 variables and for 6 variables but is slightly higher for 10 variables with this altered data. Multiplying the first five observations by 5 caused great disruption to the quadratic discriminant function and moderately great disruption to the absolute linear discriminant function. In no case would one want to use the quadratic discriminant function, as the error rate was almost .6 when 3 variables were used and almost .5 when 6 variables were used. The reason for this was that the variances in Π_1 were increased so that they were larger than those in Π_2 which causes the decline in performance.

[*Table 5 goes here*]

Further work is needed to evaluate the behavior of the absolute linear discriminant function and quadratic discriminant function. A study currently in progress will evaluate these procedures in a variety of contaminated situations.

References

1. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). Robust Estimates of Location. Princeton, New Jersey: Princeton University Press.
2. Bartlett, M.S. and Please, N.W. (1963). "Discrimination in the case of zero mean differences," Biometrika, 50, 17-21.
3. Cleveland, W.S. and Lachenbruch, P.A. (1974). "A measure of divergence among several populations," Comm. in Stat. 3, 201-211.
4. Desu, M.M. and Geisser, S. (1973). "Methods and applications of equal-mean discrimination," in Discriminant Analysis and Applications, T. Cacoullos, ed., New York: Academic Prese, 139-159.
5. Lachenbruch, P. (1974). Discriminant Analysis. New York: Hafner Press.
6. Stocks, P. (1933). "A biometric investigation of twins, Part II," Ann. Eugen., 5, 1-55.

Table 1

Performance of QDF and ALDF*

Error Rates

	QDF				ALDF			
σ_1	P_1	P_2	$\overline{P}$	Cutoff	P_1	P_2	$\overline{P}$	Cutoff
$\lim_{\sigma_1 \to 1}$	.683	.317	.500	1.0	.576	.424	.500	0.798
2	.503	.174	.339	1.848	.451	.230	.341	1.197
3	.397	.116	.257	2.472	.407	.110	.259	1.596
4	.333	.085	.209	2.956	.383	.046	.215	1.995
5	.289	.067	.178	3.352	.369	.017	.193	2.394
10	.174	.031	.103	4.652	.340	.000	.170	4.388
100	.024	.002	.013	9.210	.311	.000	.156	40.29
$\lim_{\sigma_1 \to \infty}$	.000	.000	.000	∞	.310	.000	.155	∞

*Notes: (1) QDF rule: assign X to Π_1 if $X^2 > -2 \ln \sigma_2/\sigma_1/(1/\sigma_2^2 - 1/\sigma_1^2)$

(2) ALDF rule: assign X to Π_1 if $|X| > (\sigma_1+\sigma_2)/(2\pi)^{\frac{1}{2}}$

(3) $p = 1/2$

(4) $\overline{P} = (P_1+P_2)/2$

(5) $\sigma_2 = 1$

Table 2

Means and Variances of Variables in Twins Data

Variable	$\bar{X}_1$	$\bar{X}_2$	s_1^2	s_2^2	s^2_{max}/s^2_{min}
Height	- .02	.21	5.04	19.45	3.86
Weight	-1.83	.27	403.08	793.95	1.97
Head Length	- .48	1.00	9.50	28.38	2.99
Head Breadth	.30	- .13	8.39	22.15	2.64
Head Circumference	-1.37	1.46	53.08	204.51	3.85
Interpupillary Distance	.41	- .06	2.25	10.74	4.77
Systolic Blood Pressure	-2.67	1.33	106.67	104.65	1.02
Pulse Interval	1.04	- .54	105.24	105.91	1.01
Strength of Grip - Right	- .43	- .44	7.10	9.32	1.31
Strength of Grip - Left	- .54	- .15	5.32	7.45	1.40

Table 3

Error Rates in Twins Data

Quadratic Discriminant Function

Number of Variables	Apparent Rates			Leaving-One-Out Rates		
	P_1	P_2	$\bar{P}$	P_1	P_2	$\bar{P}$
3*	.196	.333	.265	.239	.354	.297
6+	.152	.188	.170	.283	.250	.266
10	.109	.167	.138	.348	.250	.299

*Variables: Height, Head Circumference, Interpupiallary Distance

+Variables: Height, Weight, Head Length, Head Bredth, Head Circumference, Interpupillary Distance

Table 4

Error Rates in Twins Data

Absolute Linear Discriminant Function

Number of Variables	Apparent Rates			Leaving-One-Out Rates		
	P_1	P_2	$\bar{P}$	P_1	P_2	$\bar{P}$
3*	.196	.313	.254	.196	.333	.265
6+	.196	.313	.254	.239	.396	.317
10	.152	.313	.233	.304	.396	.356

*Variables: Same as above

+Variables: Same as above

Table 5

Error Rates in Altered Twins Data

Case	Apparent Rates			Leaving-One-Out Rates		
	P_1	P_2	$\bar{P}$	P_1	P_2	$\bar{P}$
3 variables						
QDF	.152	.500	.326	.174	.625	.400
ALDF	.217	.458	.338	.217	.500	.359
6 variables						
QDF	.261	.313	.287	.457	.438	.447
ALDF	.261	.458	.359	.261	.521	.391
10 variables						
QDF	.391	.104	.248	.348	.271	.309
ALDF	.239	.271	.255	.304	.354	.329

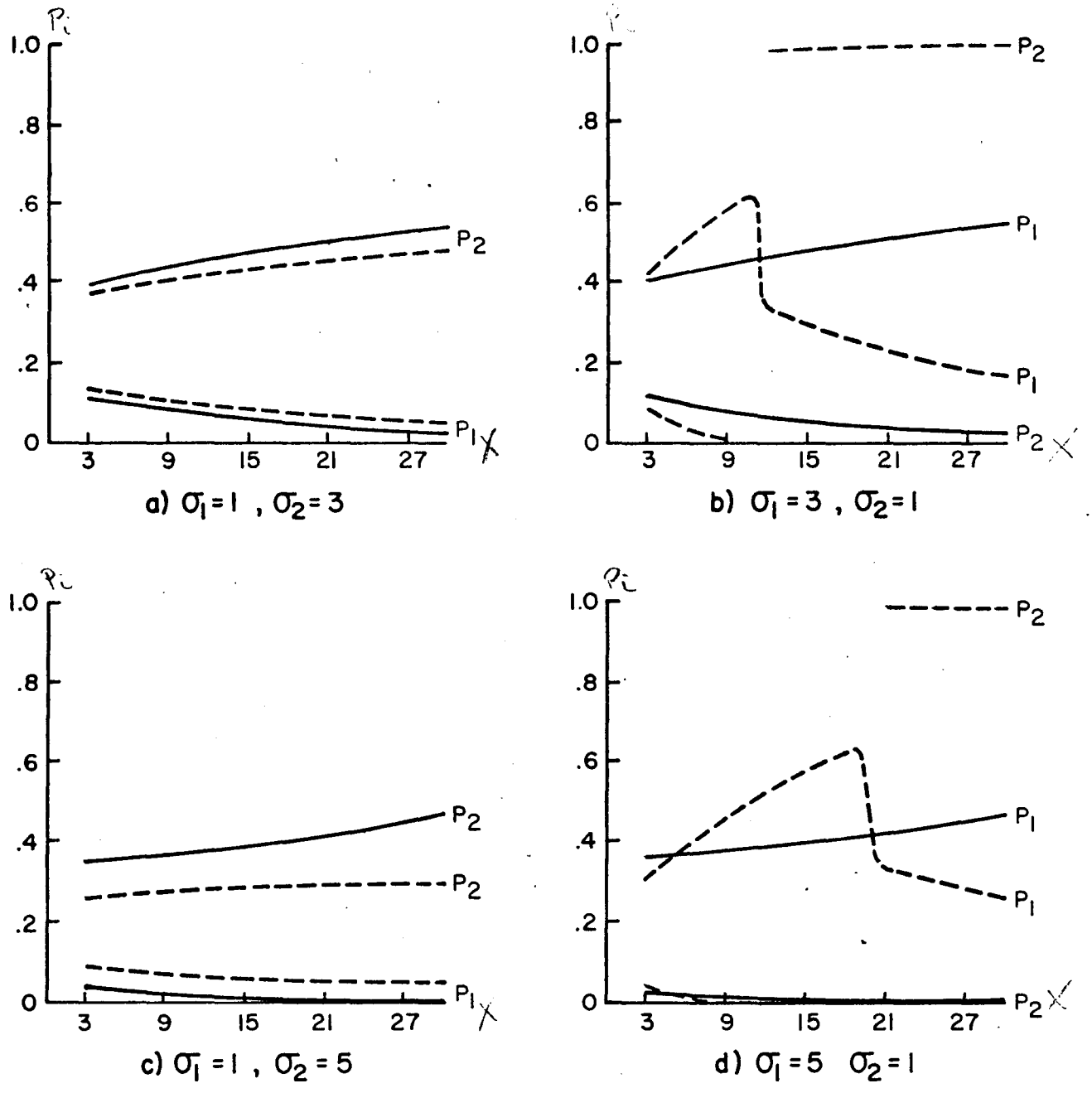

Figure 1

Stylized Sensitivity Curves