DEPARTMENT OF MATHEMATICS

UNIVERSITY OF HOUSTON HOUSTON, TEXAS

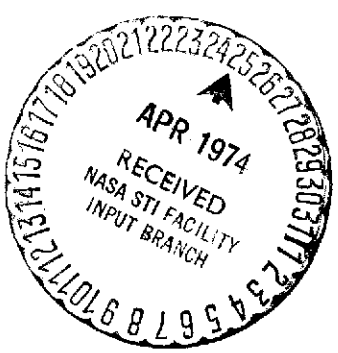
CR 134217

N74-20178

(NASA-CR-134217) ON DIFFERENTIATING THE PROBABILITY OF ERROR IN MULTIPOPULAR FEATURE SELECTION (Houston Univ.) 17 p CSCL 12A Unclas
HC $4.00 G3/19 35032

ON DIFFERENTIATING THE PROBABILITY OF ERROR IN MULTIPOPULAR FEATURE SELECTION

BY B. CHARLES PETERS FEB. 1974



PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-12777

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Report #30

On Differentiating the Probability of Error In The Multipopulation Feature Selection Problem

by

B. Charles Peters
Mathematics Department
Texas A & M University

February 1974
NAS-9-12777 MOD 1S

ABSTRACT

The use of techniques for feature selection allows one to treat classification problems in spaces of lower dimension. In this note we consider a method of linear feature selection for n dimensional observation vectors which belong to one of m populations. Where each population has a known apriori probability and is described by a known multivariate normal density function. Specifically we consider the problem of finding a $k \times n$ matrix B of rank k $(k < n)$ for which the transformed probability of misclassification is minimized.

Subject to the condition that the transformed a posterior probabilities are distinct we obtain theoretical results which, for the case $k = 1$, give rise to a numerically tractable formula for the derivative of the probability of misclassification. It is shown that for the two population problem this condition is also necessary. Finally, we investigate the dependence of the minimum probability of error on the a priori probabilities and show that the minimum probability of error satisfies a uniform Lipschitz condition with respect to the a priori probabilities.

On Differentiating the Probability of Error in the Multipopulation Feature Selection Problem

1. Introduction

Let $\pi_1,\ldots,\pi_m$ be populations in R^n with apriori probabilities $\alpha_1,\ldots,\alpha_m$ and conditional densities $p_i(x)$, $i = 1,\ldots,m$, defined for $x = (x_1,\ldots,x_n)^T \in R^n$ by

$$p_i(x) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}$$

If B is a $k \times n$ matrix of rank k, then the transformed conditional densities are defined for $y = (y_1,\ldots,y_k)^T \in R^k$ by

$$p_i(y,B) = \frac{1}{(2\pi)^{k/2}|B\Sigma_iB^T|^{1/2}} e^{-\frac{1}{2}(y-B\mu_i)(B\Sigma_iB^T)^{-1}(y-B\mu_i)}$$

Let $g(B)$ denote the probability of misclassification in R^k as a function of B, with a Bayes optimal (maximum likelihood) classification rule. If B_o minimizes $g(B)$ and the Gateaux differential, [3, p. 171],

$$\delta g(B_o;C) = \lim_{s\to o} \frac{g(B_o + sC) - g(B_o)}{s}$$

exists for a $k \times n$ matrix C, then $\delta g(B_o,C) = 0$. Thus it is desirable to obtain a formula for $\delta g(B;C)$. Such a formula has been obtained for the case $m = 2$, $\alpha_1 = \alpha_2 = 1/2$, by Guseman and Walker [1], [2]. In this

note we obtain a formula for the general case subject to the condition that the functions $\alpha_i p_i(y,B)$ are all distinct. Unless otherwise stated, this assumption will be made.

2. Differentiating the Probability of Error.

Using a maximum likelihood classification rule, the probability of error in R^k as a function of a feature selection matrix B of rank k may be expressed as

$$g(B) = \int_{R_1(B)} f_1(y,B)dy + \ldots + \int_{R_m(B)} f_m(y,B)dy$$

where

$$f_i(y,B) = \sum_{\substack{j=1 \\ j\neq i}}^{m} \alpha_j p_j(y,B)$$

and

$$R_i(B) = \{y \in R^k | \alpha_i p_i(y,B) > \alpha_j p_j(y,B) \quad \text{for all} \quad j \neq i\}.$$

$$= \{y \in R^k | f_i(y,B) < f_j(y,B) \quad \text{for} \quad j \neq i\}.$$

Since the functions $\alpha_i p_i(y,B)$ are distinct, the $R_i(B)$ are disjoint open sets which cover R^k except for a set of measure zero; i.e., their boundaries.

Let

$$r(y,B) = \min_i f_i(y,B).$$

Then $R_i(B)$ is the interior of the set

$$\{y \in R^k | f_i(y,B) = r(y,B)\}$$

and

$$g(B) = \int_{R^k} r(y,B)dy$$

Let C be a $k \times n$ matrix. If $y \in R_i(B)$ and $|s|$ is sufficiently small, then

$$r(y,B + sC) = f_i(y,B + sC)$$

Hence, for $y \in R_i(B)$,

$$\lim_{s\to 0} \frac{r(y,B + sC) - r(y,B)}{s} = \lim_{s\to 0} \frac{f_i(y,B + sC) - f_i(y,B)}{s}$$

$$= \delta f_i(y,B;C)$$

$$= \sum_{\substack{\ell=1 \\ \ell\neq i}}^{m} \alpha_\ell \delta p_\ell(y,B;C)$$

Thus, provided that

$$(1)\qquad \lim_{s\to o}\int_{R_i(B)}\frac{r(y,B+sC)-r(y,B)}{s}\,dy=\int_{R_i(B)}\lim_{s\to o}\frac{r(y,B+sC)-r(y,B)}{s}\,dy$$

we have

$$(2)\qquad \delta g(B;C)=\sum_{i=1}^{m}\sum_{\substack{\ell=1\\ \ell\neq i}}^{m}\alpha_\ell\int_{R_i(B)}\delta p_\ell(y,B;C)\,dy$$

It is shown in [2], that

$$(3)\qquad \delta p_\ell(y,B;C)=p_\ell(y,B)\{(y-B\mu_\ell)^T(B\Sigma_\ell B^T)^{-1}[C\mu_\ell+C\Sigma_\ell B^T(B\Sigma_\ell B^T)^{-1}(y-B\mu_\ell)]-\mathrm{tr}[C\Sigma_\ell B^T(B\Sigma_\ell B^T)^{-1}]\}.$$

Combining (2) and (3) gives the required formula for $\delta g(B;C)$. For $k>1$ this formula is numerically intractable because of the integrals which appear. For $k=1$, however, it is possible to obtain an integral free expression for $\delta g(B;C)$. Indeed, when $k=1$, (3) becomes

$$(4)\qquad \delta p_\ell(y,B;C)=p_\ell(y,B)\{\frac{C\Sigma_\ell B^T}{(B\Sigma_\ell B^T)^2}(y-B\mu_\ell)^2+\frac{C\mu_\ell}{B\Sigma_\ell B^T}(y-B\mu_\ell)-\frac{C\Sigma_\ell B^T}{B\Sigma_\ell B^T}\}.$$

Integrating (4) by parts yields

$$\int_{R_i(B)} \delta p_\ell(y,B;C)dy = - p_\ell(y,B)\left[\frac{C\Sigma_\ell B^T}{B\Sigma_\ell B^T}(y - B\mu_\ell) + C\mu_\ell\right]\Big|_{R_i(B)}$$

where $\Big|_{R_i(B)}$ means the sum of the values of the function at the right endpoints of the intervals comprising $R_i(B)$ minus the sum of its values at the left endpoints. Thus, for $k = 1$,

$$(5) \qquad -\delta g(B;C) = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} \alpha_j p_j(y,B)\left[\frac{C\Sigma_j B^T}{B\Sigma_j B^T}(y - B\mu_j) + C\mu_j\right]\Big|_{R_i(B)}$$

The remainder of this section is devoted to showing that (1) is true. To do this we require three lemmas. The first two of these are generalizations of well known facts from calculus and integration theory. If f is a real valued function defined in a neighborhood of a real number x, let $\overline{f}(x)$ and $\underline{f}(x)$ denote respectively its upper and lower derivates at x defined by, [4, p.96],

$$\overline{f}(x) = \limsup_{y\to x} \frac{f(y) - f(x)}{y - x}$$

$$\underline{f}(x) = \liminf_{y\to x} \frac{f(y) - f(x)}{y - x}$$

Lemma 1: If f is continuous on an interval $[a,b]$, then there exists $c \in (a,b)$ such that

$$\underline{f}(c) \le \frac{f(b) - f(a)}{b - a} \le \overline{f}(c).$$

Lemma 2: Let (X,μ) be a measure space. Suppose $h(y,s)$ is a real valued function on $X \times [-\delta,\delta]$ such that for each s, $h(y,s)$ is absolutely integrable on X and for each y, $h(y,s)$ is continuous in s. Suppose also that there exists an absolutely integrable function $\beta(y)$ such that

$$|\overline{h}_s(y,s)| \le \beta(y)$$

$$|\underline{h}_s(y,s)| \le \beta(y)$$

for all y and s and that for each y, the partial derivative $h_s(y,o)$ exists. Then

$$\frac{d}{ds}\int_X h(y,s)d\mu \Big|_{s=o} = \int_X h_s(y,o)d\mu.$$

Proof: Apply Lemma 1 and the Lebesgue dominated convergence theorem, [4, p.229].

Lemma 3: If $\delta > o$ is small enough that $B + sC$ is rank k for $|s| \le \delta$, then there exists a function $\beta(y)$, integrable on R^k, such that

$$|\delta f_j(y,B + sC;C)| \le \beta(y)$$

for all $y \in R^k$, $|s| \le \delta$, $j = 1, \ldots, m$.

Proof: By (3),

$$\delta f_j(y,B + sC;C) = \sum_{\substack{\ell=1 \\ \ell \neq j}} \alpha_\ell \delta p_\ell(y,B + sC;C)$$

$$= -\sum_{\ell \neq j} \alpha_\ell p_\ell(y,B + sC)\{[y - (B + sC)\mu_\ell]^T[(B + sC)\Sigma_\ell(B + sC)^T]^{-1}$$

$$[C\mu_\ell + C\Sigma_\ell(B + sC)^T(B + sC)\Sigma_\ell(B + sC)^T)^{-1}(y - (B + sC)\mu_\ell)]$$

$$-\mathrm{tr}[C\Sigma_\ell(B + sC)^T((B + sC)\Sigma_\ell(B + sC)^T)^{-1}]\}.$$

Since the means and covariances of the density functions $p_\ell(y,B + sC)$, as well as the coefficients of the terms in $\{\ \}$, are continuous functions of s, they form compact sets. From this fact, it is clear that the required function $\beta(y)$ exists. Since the actual construction of $\beta(y)$ is tedious it will be omitted.

Now let $h(y,s) = r(y,B + sC)$. We want to show that

$$\frac{d}{ds} \int_{R_1(B)} h(y,s)dy \Big|_{s=o} = \int_{R_1(B)} h_s(y,o)dy.$$

Let $\delta > o$ be small enough that for $|s| \le \delta$, $B + sC$ is rank k and the functions $\alpha_j p_j(y,B + sC)$ are all distinct. Let $\beta(y)$ be the function in Lemma 3. Clearly, $h(y,s)$ is integrable on $R_1(B)$ for each fixed s

and continuous on $[-\delta,\delta]$ for each fixed y. Thus the result follows from Lemma 2 once it is shown that

$$|\overline{h}_s(y,s)| \le \beta(y) \tag{6}$$

$$|\underline{h}_s(y,s)| \le \beta(y)$$

for all $y \in R_i(B)$, $|s| \le \delta$. For $y \in R_i(B)$ and $|s| \le \delta$, there are two possibilities:

Case 1, $y \in R_j(B + sC)$ for some j: Then $h_s(y,s) = \delta f_j(y,B + sC;C)$ and (6) follows from Lemma 3.

Case 2: y is not in any $R_j(B + sC)$: Then $h(y,s) = f_j(y,B + sC)$ for more than one index j. Let $J(y)$ be the set of indices j such that $h(y,s) = f_j(y,B + sC)$. Then for sufficiently small $|t| > o$

$$h(y,s + t) = r(y,B + sC + tC) = f_j(y,B + sC + tC)$$

for some j, depending on t, in $J(y)$. Thus,

$$\frac{h(y,s + t) - h(y,s)}{t} = \frac{f_j(y,B + sC + tC) - f_j(y,B + sC)}{t}.$$

Since $J(y)$ is a finite set, there are indices j and k in $J(y)$ such that

$$\overline{h}_s(y,s) = \delta f_j(y,B + sC;C)$$

$$\underline{h}_s(y,s) = \delta f_k(y,B + sC;C)$$

and (6) follows again from Lemma 3.

This concludes the proof.

3. The Case of Non-Distinct Transformed Densities

In this section we show that the requirement that the $\alpha_i p_i(y,B)$ be distinct cannot be eliminated. Specifically, consider a two population problem where $\alpha_1 = \alpha_2 = 1/2$, and $p_1(y,B) \equiv p_2(y,B)$; that is, $B\mu_1 = B\mu_2$ and $B\Sigma_1 B^T = B\Sigma_2 B^T$. Let C be a $k \times n$ matrix such that $C\mu_1 \neq C\mu_2$ or $C\Sigma_1 B^T \neq C\Sigma_2 B^T$. We will show that $\delta g(B:C)$ does not exist. Indeed, using the formula

$$\min\{f_1,f_2\} = \frac{1}{2}[f_1 + f_2 - |f_1 - f_2|]$$

we see that

$$g(B + sC) = \frac{1}{2}\int_{R^k} \min\{p_1(y,B+sC),\ p_2(y,B+sC)\}dy$$

$$= \frac{1}{2} - \frac{1}{4}\int_{R^k} |p_1(y,B+sC) - p_2(y,B+sC)|dy.$$

$g(B) = \frac{1}{2}$.

Hence, for $s > 0$,

$$\frac{g(B+sC) - g(B)}{s} = -\frac{1}{4}\int_{R^k} \frac{1}{s}\,|p_1(y,B+sC) - p_2(y,B+sC)|dy$$

$$= -\frac{1}{4}\int_{R^k} \left| \frac{p_1(y,B+sC) - p_1(y,B)}{s} - \frac{p_2(y,B+sC) - p_2(y,B)}{s} \right| dy$$

which tends to $-\frac{1}{4}\int_{R^k}|\delta p_1(y,B;C) - \delta p_2(y,B;C)|dy$ as $s \to 0$. On the other hand, for $s < 0$,

$$\frac{g(B+sC) - g(B)}{s} = \frac{1}{4}\int_{R^k}\left|\frac{p_1(y,B+sC) - p_1(y,B)}{s} - \frac{p_2(y,B+sC) - p_2(y,B)}{s}\right|dy$$

which tends to

$$\frac{1}{4}\int_{R^k}|\delta p_1(y,B;C) - \delta p_2(y,B;C)dy.$$

Hence $\delta g(B;C)$ exists if and only if

$$\int_{R^k}|\delta p_1(y,B;C) - \delta p_2(y,B;C)|dy = 0.$$

That is, if and only if $\delta p_1(y,B;C) = \delta p_2(y,B;C)$ almost everywhere. But

$$\delta p_1(y,B;C) = p_1(y,B)\{(y-B\mu_i)^T(B\Sigma_1B^T)^{-1}[C\mu_1 + C\Sigma_1B^T(B\Sigma_1B^T)^{-1}(y-B\mu_1)] - tr[C\Sigma_1B^T(B\Sigma_1B^T)^{-1}]\}.$$

$$\delta p_2(y,B;C) = p_1(y,B)\{(y-B\mu_1)^T(B\Sigma_1B^T)^{-1}[C\mu_2 + C\Sigma_2B^T(B\Sigma_1B^T)^{-1}(y-B\mu_1)] - tr[C\Sigma_2B^T(B\Sigma_1B^T)^{-1}]\}.$$

Since the polynomial parts of these two expressions have different coefficients, they cannot be equal almost everywhere. Hence, $\delta g(B;C)$ does not exist.

Notice that the problem of non differentiability does not arise if the apriori probabilities are distinct, since the functions $\alpha\, p_i(y,B)$ are

distinct in this case. This suggests that if some of the apriori probabilities are equal then one might attempt to find a B which nearly minimizes $g(B)$ by changing the apriori probabilities slightly and insuring that the new apriori probabilities are distinct. The following theorem shows that this approach is valid. Let $\alpha = (\alpha_1,\ldots\alpha_m)$ denote the vector of apriori probabilities and write $g(B,\alpha)$ to show the dependence of the probability of error on α as well as on the feature selection matrix B. Let $f_i(y,B)$ be defined as in Section 2, and let

$$f(y,B) = \sum_{i=1}^{m} \alpha_i p_i(y,B).$$

Then
$$\begin{aligned} g(B,\alpha) &= \int_{R^k} \min_i f_i(y,B)dy \\ &= \int_{R^k} \min_i (f(y,B) - \alpha_i p_i(y,B))dy \\ &= \int_{R^k} [f(y,B) - \max_i \alpha_i p_i(y,B)]dy \\ &= 1 - \int_{R^k} \max \alpha_i p_i(y,B)dy. \end{aligned}$$

Theorem: For all α and β,

$$|\min_B g(B,\alpha) - \min_B g(B,\beta)| \leq \|\alpha-\beta,\|$$

where $\|\alpha\ \beta\| = |\alpha_1-\beta_1| + \ldots + |\alpha_m-\beta_m|$.

Proof: In view of the formula for $g(B,\alpha)$ given above, it clearly suffices to show that if $q_1(y)$, ..., $q_m(y)$ are probability density functions on R^k

and α,β are m-tuples of real numbers, then

$$(7) \qquad \int_{R^k} \left| \max_{1\le i\le m} \alpha_i q_i(y) - \max_{1\le i\le m} \beta_i q_i(y) \right| dy \le \|\alpha-\beta\| .$$

This inequality is clear for $m = 1$. For $m > 1$ write

$$\max_{i\le m} \alpha_i q_i(y) = \frac{1}{2}\{\alpha_m q_m(y) + \max_{i\le m-1} \alpha_i q_i(y) + \left|\alpha_m q_m(y) - \max_{i\le m-1} \alpha_i q_i(y)\right|\}$$

On substituting this and the corresponding expansion for $\max_{i\le m} \beta_i q_i(y)$ into the left hand side of (7) it follows easily that

$$\int_{R^k} \left|\max_{i\le m} \alpha_i q_i(y) - \max_{i\le m} \beta_i q_i(y)\right| dy$$

$$\le |\alpha_m-\beta_m| \int_{R^k} q_m(y)dy + \int_{R^k} \left| \max_{i\le m-1} \alpha_i q_i(y) - \max_{i\le m-1} \beta_i q_i(y)\right| dy$$

$$= |\alpha_m-\beta_m| + \int_{R^k} \left| \max_{i\le m-1} \alpha_i q_i(y) - \max_{i\le m-1} \beta_i q_i(y)\right| dy.$$

Thus the result follows by induction.

4. Concluding Remarks

It will be shown in a subsequent report that the condition that the $\alpha_i p_i(y,B_o)$ be distinct is necessary as well as sufficient for the differentiability of $g(B)$ at B_o. Thus the following conjecture is of importance whether it

is intended to solve the variational equation directly for the minimizing B or to use a steepest discent method and use the expression for $\delta g(B;C)$ developed in Section 2 to compute the gradient at each step.

Conjecture: If $\alpha_i p_i(x) \neq \alpha_j p_j(x)$ for $i,j = 1,\ldots,m$ and B_o minimizes $g(B)$, then the functions $\alpha_i p_i(y,B_o)$ are distinct.

REFERENCES

1. L.F. Guseman, Jr. and H.F. Walker, On Minimizing the Probability of Misclassification for Linear Feature Selection, JSC Internal Technical Note. JSC-08412, August, 1973.

2. _____, The Differentiability of the Probability of Misclassification as a Function of a Linear Feature Selection Matrix, Report #28 NAS-9-12777, University of Houston, Department of Mathematics, August, 1973.

3. David G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, Inc., New York, 1969.

4. H.L. Royden, Real Analysis (2nd ed.), Collier-Macmillan Ltd., London, 1970.