1 NP

K. Adinkey

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

C.R. 134248

ON DIFFERENTIATING
THE PROBABILITY OF ERROR IN
MULTIPOPULAR FEATURE SELECTIONII
BY B. CHARLES PETERS FEB. 1974
REPORT #31

(NASA-CR-134248) ON DIFFERENTIATING THE PROBABILITY OF ERROR IN THE MULTIPOPULATION FEATURE SELECTION PROBLEM, 2 (Houston Univ.) 12 p HC $4.00 CSCL 12A G3/19

N74-23153

Unclas
37898

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Report #31

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

by

B. Charles Peters

Mathematics Department

Texas A & M University

March, 1974

NAS-9-12777 MOD 1S

|

ABSTRACT

In this note we give a necessary and sufficient condition for the Gateaux differentiability of the probability of misclassification as a function of a feature selection matrix B, assuming a maximum likelihood classifier and normally distributed populations. It is also shown that if the probability of error has a local minimum at B then it is differentiable at B.

On Differentiating the Probability of Error in the Multipopulation Feature Selection Problem, II.

1. Introduction.

Let $\pi_1, \ldots, \pi_m$ be populations in R^n with a priori probabilities $\alpha_1, \ldots, \alpha_m$ and multivariate normal conditional density functions,

$$P_i(x) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp[-\tfrac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)].$$

$i = 1, \ldots, m$. If B is a $k\times n$ matrix of rank k then the transformed conditional densities are, for $y \in R^k$,

$$P_i(y,B) = \frac{1}{(2\pi)^{k/2}|B\Sigma_i B^T|^{1/2}} \exp[-\tfrac{1}{2}(y-B\mu_i)^T(B\Sigma_i B^T)^{-1}(y-B\mu_i)].$$

Let $g(B)$ denote the probability of misclassifying an observation $x \in R^n$ using the Bayes optimal classifier: classify x in π_i if $\alpha_i P_i(Bx, B) \geq \alpha_j P_j(Bx, B)$ for each $j = 1, \ldots, m$. Then $g(B) = 1 - h(B)$, where

$$h(B) = \int_{R^k} \max_{1\leq i\leq m} \alpha_i P_i(y,B)dy.$$

is the probability of correct classification.

If the transformed probability of error is to be used as a feature selection criterion we require a method for obtaining a $k\times n$ matrix B_o of rank k which minimizes $g(B)$. If B_o minimizes $g(B)$ then the Gateaux differential, [2,p.178],

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

vanishes for all $k\times n$ matrices C for which it exists. If $\delta g(B_o:C)$ exists for all $k\times n$ matrices C, then g is said to be Gateaux differentiable at B_o. Thus it is desireable to have necessary and sufficient conditions for Gateaux differentiability of g as well as a formula for $\delta g(B;C)$.

2. Main Results.

For a given $k\times n$ matrix B partition the set $\{\alpha_i P_i(x)\}_{i=1}^m$ into disjoint sets

$$S_1 = \{\alpha_{11}P_{11}(x),\ \alpha_{12}P_{12}(x),\ \ldots,\ \alpha_{1n_1}P_{1n_1}(x)\}$$

$$\vdots$$

$$S_r = \{\alpha_{r1}P_{r1}(x),\ \alpha_{r2}P_{r2}(x),\ \ldots,\ \alpha_{rn_r}P_{rn_r}(x)\}$$

where the S_q are defined by

$$\alpha_{qj}P_{qj}(y,B) \equiv \alpha_{qi}P_{qi}(y,B) \qquad 1 \le i,j \le n_q$$

$$\alpha_{qj}P_{qj}(y,B) \not\equiv \alpha_{\ell i}P_{\ell i}(y,B) \qquad q \ne \ell$$

For $\ell = 1, \ldots, r$ let

$$R_\ell = \{y \in R^k | \alpha_{\ell 1}P_{\ell 1}(y,B) > \alpha_{k1}P_{k1}(y,B) \quad , \quad k \ne \ell\}.$$

The R_ℓ are disjoint open sets which cover R^k except for a set M of measure zero.

For a given $k\times n$ matrix C write $P_{ij}(y,s)$ for $P_{ij}(y,B+sC)$ and $h(s)$ for $h(B+sC)$. That is, $h(s) = \int_{R^k} \max_{i,j} \alpha_{ij} P_{ij}(y,s)dy$.

Theorem 1: h is Gateaux differentiable at B if and only if for each ℓ such that $R_\ell \ne \emptyset$, $\mu_{\ell i} = \mu_{\ell j}$ and $\Sigma_{\ell i}B^T = \Sigma_{\ell j}B^T$ for each $i,j \le n_\ell$.

Proof: By repeating some of the members of the S_q's if necessary, we can assume $n_1 = n_2 = \ldots = n_r = n_o$. Thus

$$h(s) = \int_{R^k} \max_{1\le j\le n_o} \max_{1\le i\le r} \alpha_{ij}P_{ij}(y,s)dy$$

$$\int_{R^k} \max_{1\le j\le n_o} f_j(y,s)dy,$$

where $f_j(y,s) = \max\limits_{1\le i\le r} \alpha_{ij} P_{ij}(y,s)$

The $f_j(y,s)$ have the properties:

1) $f_1(y,0) \equiv f_2(y,0) \equiv \ldots \equiv f_{n_o}(y,0)$

and

2) $\frac{\partial f_j}{\partial s}(y,0)$ is defined for all $y \notin M$, $j = 1, \ldots n_o$. By an argument in [3], it can be shown that for sufficiently small $|s|$, the difference quotients

$$\frac{f_j(y,s) - f_j(y,o)}{s}$$

are bounded by an integrable function $\beta(y)$ for $y \notin M$. Hence, for $s>0$,

$$\frac{h(s) - h(o)}{s} = \int_{R^k} \frac{1}{s}[\max_{j\le n_o} f_j(y,s) - \max_{j\le n_o} f_j(y,o)]dy$$

$$= \int_{R^k} \frac{1}{s} \max_{j\le n_o} [f_j(y,s) - f_j(y,o)]dy$$

$$= \int_{R^k} \max_{j\le n_o} \frac{f_j(y,s) - f_j(y,o)}{s} dy$$

$$\rightarrow \int_{R^k} \max_{j\le n_o} \frac{\partial f_j}{\partial s}(y,o)dy$$

as $s \to 0+$. On the other hand, for $s < 0$,

$$\frac{h(s) - h(o)}{s} = \int_{R^k} \min_{j\le n_o} \frac{f_j(y,s) - f_j(y,o)}{s} dy$$

$$\rightarrow \int_{R^k} \min_{j\le n_o} \frac{\partial f_j}{\partial s}(y,o)dy.$$

as $s \to 0-$. Thus the Gateaux differential $h'(0)$ exists if and only if

$$\max_{j \le n_o} \frac{\partial f_j}{\partial s}(y,o) = \min_{j \le n_o} \frac{\partial f_j}{\partial s}(y,o) \qquad \text{a.e.}$$

That is, if and only if

$$\frac{\partial f_j}{\partial s}(y,o) = \frac{\partial f_i}{\partial s}(y,o) \qquad \text{a.e.}$$

for all $i,j \le n_o$. For $y \in R^{\ell}$ it is readily verified that

$$\frac{\partial f_i}{\partial s}(y,o) = \alpha_{\ell i}\frac{\partial P_{\ell i}}{\partial s}(y,o).$$

Hence, $h'(0)$ exists if and only if

$$\alpha_{\ell i}\frac{\partial P_{\ell i}}{\partial s}(y,o) = \alpha_{\ell j}\frac{\partial P_{\ell j}}{\partial s}(y,o)$$

for $i,j \le n_o$, almost all $y \in R^{\ell}$, $\ell = 1, \ldots, r$.

It is shown in [1], that

$$\alpha_{\ell j} \frac{\partial P_{\ell j}}{\partial s}(y,o) = \alpha_{\ell j} P_{\ell j}(y,o)\{(y-B\mu_{\ell j})^T(B\Sigma_{\ell j}B^T)^{-1}$$

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

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

Since $B\mu_{\ell j} = B\mu_{\ell i}$, $B\Sigma_{\ell j}B^T = B\Sigma_{\ell i}B^T$, $\alpha_{\ell j} = \alpha_{\ell i}$,

$$\alpha_{\ell j} \frac{\partial P_{\ell j}}{\partial s}(y,o) = \alpha_{\ell i}P_{\ell i}(y,o)\{(y - B\mu_{\ell i})^T(B\Sigma_{\ell i}B^T)^{-1}$$

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

$$- \operatorname{tr}[C\Sigma_{\ell j}B^T(B\Sigma_{\ell i}B^T)^{-1}]\}.$$

If $R_\ell \neq \emptyset$, then R_ℓ has positive measure. Thus it is easily seen that if $R_\ell \neq \emptyset$,

$$\alpha_{\ell i}\frac{\partial P_{\ell i}}{\partial s}(y,o) = \alpha_{\ell j}\frac{\partial P_{\ell i}}{\partial s}(y,o) \qquad \text{a.e. in } R_\ell$$

if and only if $C\mu_{\ell j} = C\mu_{\ell i}$, $C\Sigma_{\ell j}B^T = C\Sigma_{\ell i}B^T$ for all $i, j \leq n_o$. Thus h is Gateaux differentiable at B if and only if $\mu_{\ell i} = \mu_{\ell j}$, $\Sigma_{\ell i}B^T = \Sigma_{\ell j}B^T$ $\forall\, i,j \leq n_o$, $\forall\, \ell$ such that $R_\ell \neq \emptyset$. This concludes the proof.

It is clear that if h is Gateaux differentiable at B, then

$$\delta h(B:C) = \sum_{i=1}^{r}\alpha_{i1} \int_{R_i} \delta P_{i1}(y,B:C)dy$$

Thus the Gateaux differential of the probability of error is

$$\delta g(B:C) = -\sum_{i=1}^{r}\alpha_{i1} \int_{R_i} \delta P_{i1}(y,B:C)\,dy.$$

<u>Theorem</u> 2: If h has a local maximum at B, then h is Gateaux differentiable at B.

<u>Proof</u>: It is evident from the proof of Theorem 1 that for any $k\times n$ matrix C,

$$\limsup_{s\to o} \frac{h(B+sC) - h(B)}{s} = \lim_{s\to 0+} \frac{h(B+sC) - h(B)}{s}$$

$$= \int_{R^k} \max_{j\le n_o} \frac{\partial f_j}{\partial s}(y,o)\,dy$$

and

$$\liminf_{s\to o} \frac{h(B+sC) - h(B)}{s} = \lim_{s\to o-} \frac{h(B+sC) - h(B)}{s}$$

$$= \int_{R^k} \min_{j\le n_o} \frac{\partial f_j}{\partial s}(y,o)\,dy.$$

If h has a maximum at B, then since $\lim\limits_{s\to 0-} \frac{h(B+sC) - h(B)}{s}$ exists,

$$\limsup_{s\to o} \frac{h(B+sC) - h(B)}{s} = \lim_{s\to o-} \frac{h(B+sC) - h(B)}{s}$$

$$= \liminf_{s\to o} \frac{h(B+sC) - h(B)}{s}$$

Thus h is Gateaux differentiable at B. Q.E.D.

3. Concluding Remarks.

The meaning of the necessary and sufficient condition for differentiability of $g(B)$ becomes a little more obvious when it is applied to the two population problem. Let π_1 and π_2 be normally distributed populations in R^n with class statistics α_1, μ_1, Σ_1 and α_2, μ_2, Σ_2, respectively.

Case 1: $\alpha_1 \neq \alpha_2$. Then $g(B)$ is differentiable for all B.

Case 2: $\alpha_1 = \alpha_2$, $\mu_1 \neq \mu_2$. Then g is differentiable at B if and only if $B\mu_1 \neq B\mu_2$ or $B\Sigma_1 B^T \neq B\Sigma_2 B^T$.

Case 3: $\alpha_1 = \alpha_2$, $\mu_1 = \mu_2$, $\Sigma_1 - \Sigma_2$ is invertible. Then g is differentiable at B if and only if $B\Sigma_1 B^T \neq B\Sigma_2 B^T$.

Case 4: $\alpha_1 = \alpha_2$, $\mu_1 = \mu_2$, $\Sigma_1 - \Sigma_2$ is not invertible. Then g is differentiable at B if and only if $B\Sigma_1 B^T \neq B\Sigma_2 B^T$ or $\Sigma_1 B^T = \Sigma_2 B^T$.

As a special case of Case 4, we have the degenerate case in which the class statistics for π_1 and π_2 are the same. Then g is differentiable for all B and has derivative 0. Finally, we remark that it is mistakenly asserted in [3] that the condition $\alpha_i P_i(y,B) \not\equiv \alpha_j P_j(y,B)$ is necessary as well as sufficient for differentiability of $g(B)$. As the analysis above shows, this is not even true in the two population probelm.

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

3. B. C. Peters, Jr., On Differentiating the Probability of Error in the Multipopulation Feature Selection Problem, Report #30, NAS-9-12777, University of Houston, Department of Mathematics, February, 1974.