N76-30891

UNIVERSITY OF HOUSTON • Founded 1927 •

DEPARTMENT OF MATHEMATICS CR151699
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

(NASA-CR-151699) THE NUMERICAL EVALUATION OF MAXIMUM-LIKELIHOOD ESTIMATES OF THE PARAMETERS FOR A MIXTURE OF NORMAL DISTRIBUTIONS FROM PARTIALLY IDENTIFIED SAMPLES Final Report (Houston Univ.) 218 p G3/64

N78-21842

Unclas
14088

FINAL REPORT
NAS-9-15000
FEBRUARY 28, 1978

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15000

HOUSTON, TEXAS 77004

Final Report

NAS-9-15000

February 28, 1978

The Numerical Evaluation of Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions from Partially Identified Samples

by

Homer F. Walker
Department of Mathematics, University of Houston
Houston, Texas 77004

June, 1976
Report #54

The Numerical Evaluation of Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions from Partially Identified Samples

by

Homer F. Walker
Department of Mathematics, University of Houston
Houston, Texas 77004

1. Introduction.

Let $\pi_1,\ldots,\pi_m$ be populations whose multivariate observations in $\mathbb{R}^n$ are distributed with respective normal density functions

$$p_i(x) = \frac{1}{(2\pi)^{n/2}|\Sigma_i^o|^{1/2}} e^{-\frac{1}{2}(x-\mu_i^o)^T\Sigma_i^{o-1}(x-\mu_i^o)}, \quad i = 1,\ldots,m.$$

If π_o is a given mixture of members of these populations, then observations on π_o are distributed in $\mathbb{R}^n$ with density function

$$p(x) = \sum_{i=1}^{m} \alpha_i^o p_i(x)$$

for an appropriate set of proportions $\{\alpha_i^o\}_{i=1,-,m}$. These proportions necessarily satisfy $\sum_{i=1}^{m} \alpha_i^o = 1$ and $\alpha_i^o \geq 0$, $i = 1,-,m$. In this note, we also assume that each α_i^o is strictly positive.

We address here the problem of numerically approximating the maximum-likelihood estimates of the parameters $\{\alpha_i^o,\mu_i^o,\Sigma_i^o\}_{i=1,\ldots,m}$ determined by samples of two types. Samples of both types consist of sets $\{x_{ik}\}_{k=1,\ldots,N_i}$

of independent observations on π_i, $i = 0,\ldots,m$. (The sets $\{x_{ik}\}_{k=1,\ldots,N_i}$, $i = 1,\ldots,m$, comprise the *identified observations* of such samples, and such samples are said to be *partially identified*.) We distinguish samples of the two types according to whether the numbers N_i of identified observations contain information about the proportions α_i^0, $i = 1,\ldots,m$. If the numbers of identified observations contain no information about the proportions, then the sample is of the first type; otherwise, the sample is of the second type. The following are examples of how samples of the first and second types, respectively, might be obtained:

(1) For $i = 0,\ldots m$, numbers N_i are arbitrarily choosen and independent observations $\{x_{ik}\}_{k=1,-,N_i}$ are obtained from π_i.

(2) A number K_0 of observations are obtained from π_0. For some $N_0 < K_0$, N_0 of these observations are left unidentified, while the remaining $K_0 - N_0$ observations are identified. For $i = 1,\ldots,m$, a subset $\{x_{ik}\}_{k=1,\ldots,N_i}$ of the identified observations is determined whose member observations come from π_i.

In the following, we consider likelihood equations determined by the two types of samples which are necessary conditions for a maximum-likelihood estimate. These equations, which were derived by Coberly [1], suggest certain successive-approximations iterative procedures for obtaining maximum-likelihood estimates. These procedures, which are generalized steepest ascent (deflected gradient) procedures, contain those of Hosmer [2] as a special case. Using arguments that parallel those of [3], we show that, with probability 1 as

N_o approaches infinity (regardless of the relative sizes of N_o and N_i, $i = 1,\ldots,m$), these procedures converge locally to the strongly consistent maximum-likelihood estimates* whenever the step-size is between 0 and 2. Furthermore, the value of the step-size which yields optimal local convergence rates is bounded from below by a number which always lies between 1 and 2.

2. Samples of the first type.

We first assume that numbers $\{N_i\}_{i=0,\ldots,m}$ are given and that, for $i = 0,\ldots,m$, N_i independent observations $\{x_{ik}\}_{k=1,\ldots,N_i}$ are drawn on π_i. The log-likelihood function for a sample of this type is

$$L_1(\Theta) = \sum_{i=1}^{m} \sum_{k=1}^{N_i} \log p_i(x_{ik}) + \sum_{k=1}^{N_o} \log p(x_{ok}) \ .$$

In this expression, the parameter vector Θ (with components α_i, μ_i, Σ_i, $i = 1,\ldots,m$) belongs to the vector space $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ defined in [3], and the density functions on the right-hand side are evaluated with the true parameter vector Θ^o (with components α_i^o, μ_i^o, Σ_i^o, $i = 1,\ldots,m$) replaced by Θ.

*As in [3], one can show that, given any sufficiently small neighborhood of the true parameters, there is, with probability 1 as N_o approaches infinity (regardless of the relative sizes of N_o and N_i, $i = 1,\ldots,m$), a unique solution of the likelihood equations for either type of sample in that neighborhood, and this solution is a maximum-likelihood estimate.

Differentiating $L_1(\Theta)$ and setting its partial derivatives to zero gives the likelihood equations

$$\text{(1.a)} \quad \alpha_i = A_i(\Theta) \equiv \frac{\alpha_i}{N_o} \sum_{k=1}^{N_o} \frac{p_i(x_{ok})}{p(x_{ok})}$$

$$\text{(1.b)} \quad \mu_i = M_i(\Theta) \equiv \Big\{ \sum_{k=1}^{N_i} x_{ik} + \sum_{k=1}^{N_o} x_{ok} \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \Big\} \Big/ \Big\{ N_i + \sum_{k=1}^{N_o} \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \Big\}$$

$$\text{(1.c)} \quad \Sigma_i = S_i(\Theta) \equiv \Big\{ \sum_{k=1}^{N_i} (x_{ik}-\mu_i)(x_{ik}-\mu_i)^T + \sum_{k=1}^{N_o} (x_{ok}-\mu_i)(x_{ok}-\mu_i)^T \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \Big\} \Big/$$

$$\Big\{ N_i + \sum_{k=1}^{N_o} \frac{\alpha_i p_i(x_{ok})}{p(x_{ok})} \Big\}$$

for $i = 1,\ldots,m$.

We set

$$A(\Theta) = \begin{pmatrix} A_1(\Theta) \\ \cdot \\ \cdot \\ \cdot \\ A_m(\Theta) \end{pmatrix} , \quad M(\Theta) = \begin{pmatrix} M_1(\Theta) \\ \cdot \\ \cdot \\ \cdot \\ M_m(\Theta) \end{pmatrix} , \quad S(\Theta) = \begin{pmatrix} S_1(\Theta) \\ \cdot \\ \cdot \\ \cdot \\ S_m(\Theta) \end{pmatrix}$$

and define an operator Φ_ϵ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ by

$$\Phi_\epsilon(\Theta) = (1-\epsilon)\Theta + \epsilon \begin{pmatrix} A(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix} .$$

Clearly, for any non-zero ϵ, the likelihood equations are satisfied by a vector $\Theta \in \mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ if and only if $\Theta = \Phi_\epsilon(\Theta)$.

We consider the following iterative procedure: Beginning with some starting value $\Theta^{(1)}$, define successive iterates inductively by

$$\text{(2)} \qquad \Theta^{(j+1)} = \Phi_\epsilon(\Theta^{(j)})$$

for $j = 1, 2, 3, \ldots$. Our local convergence result for this iterative procedure, as stated in the introduction, follows immediately from the theorem below.

Theorem 1: With probability 1 as N_o approaches infinity, Φ_ϵ is a locally contractive operator (in some norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$) near the strongly consistent maximum-likelihood estimate whenever $0 < \epsilon < 2$.

In saying that Φ_ϵ is a locally contractive operator near a point $\Theta \in \mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$, we mean that there is a vector norm $\|\ \|$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ and a number λ, $0 \le \lambda < 1$, such that

$$\|\Phi_\epsilon(\Theta') - \Theta\| \le \lambda\|\Theta' - \Theta\|$$

whenever Θ' lies sufficiently near Θ.

Proof of Theorem 1: Let

$$\Theta = \begin{pmatrix} \overline{\alpha} \\ \\ \overline{\mu} \\ \\ \overline{\Sigma} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \\ \vdots \\ \mu_1 \\ \vdots \\ \mu_m \\ \\ \Sigma_1 \\ \vdots \\ \Sigma_m \end{pmatrix}$$

be the strongly consistent maximum-likelihood estimate. We assume that

$\alpha_i \neq 0$, $i = 1,\ldots,m$. (As N_0 approaches infinity, the probability is 1 that this is the case.) As in [3], it suffices to show that, with probability 1, $\nabla\Phi_\epsilon(\Theta)$ converges to an operator which has operator norm less than 1 with respect to a suitable vector norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$.

Now

$$\nabla\Phi_\epsilon(\Theta) = (1-\epsilon)I + \epsilon\,\nabla\begin{pmatrix} A(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix} ,$$

and we write

$$\nabla\begin{pmatrix} A \\ M \\ S \end{pmatrix} = \begin{pmatrix} \nabla_{\overline{\alpha}}A & \nabla_{\overline{\mu}}A & \nabla_{\overline{\Sigma}}A \\ \nabla_{\overline{\alpha}}M & \nabla_{\overline{\mu}}M & \nabla_{\overline{\Sigma}}M \\ \nabla_{\overline{\alpha}}S & \nabla_{\overline{\mu}}S & \nabla_{\overline{\Sigma}}S \end{pmatrix} .$$

Define inner products $<\ ,\ >_i'$ on $\mathcal{M}$, $<\ ,\ >_i''$ on $\mathcal{S}$, and $<\ ,\ >$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ as in [3]. Setting

$$\beta_i(x) = \frac{p_i(x)}{p(x)},\ \gamma_i(x) = (x-\mu_i),\ \delta_i(x) = [\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T - I], K_i = N_i + \alpha_i N_0$$

for $i = 1,\ldots,m$, one calculates

$$\nabla_{\overline{\alpha}}A(\Theta) = I - (\text{diag }\alpha_i)\ \frac{1}{N_0}\sum_1^{N_0}\begin{pmatrix}\beta_1\\ \vdots \\ \beta_m\end{pmatrix}\begin{pmatrix}\beta_1\\ \vdots \\ \beta_m\end{pmatrix}^T$$

$$\nabla_{\overline{\mu}}A(\Theta) = -\ (\text{diag }\alpha_i)\ \frac{1}{N_0}\sum_1^{N_0}\begin{pmatrix}\beta_1\\ \vdots \\ \beta_m\end{pmatrix}\begin{pmatrix}<\beta_1\gamma_1,\cdot>_1'\\ \vdots \\ <\beta_m\gamma_m,\cdot>_m'\end{pmatrix}^T$$

$$\nabla_{\overline{\Sigma}}A(\Theta) = -\ (\text{diag }\alpha_i)\ \frac{1}{N_0}\sum_1^{N_0}\begin{pmatrix}\beta_1\\ \vdots \\ \beta_m\end{pmatrix}\begin{pmatrix}<\beta_1\delta_1,\cdot\ ''_1\\ \vdots \\ <\beta_m\delta_m,\cdot>_m''\end{pmatrix}^T$$

$$\nabla_{\bar{\alpha}}M(\Theta) = (\operatorname{diag} \frac{1}{K_i} \sum_1^{N_o} \beta_i\gamma_i) - (\operatorname{diag} \frac{\alpha_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1\gamma_1 \\ \vdots \\ \beta_m\gamma_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\mu}}M(\Theta) = (\operatorname{diag} \frac{\alpha_i}{K_i} \sum_1^{N_o} \gamma_i\gamma_i^T\Sigma_i^{-1}\beta_i) - (\operatorname{diag} \frac{\alpha_i}{K_i}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1\gamma_1 \\ \vdots \\ \beta_m\gamma_m \end{pmatrix} \begin{pmatrix} <\beta_1\gamma_i,\cdot>'_1 \\ \vdots \\ <\beta_m\gamma_m,\cdot>'_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\Sigma}}M(\Theta) = (\operatorname{diag} \frac{1}{K_i} \sum_1^{N_o} \beta_i\gamma_i<\delta_i,\cdot>''_i) - (\operatorname{diag} \frac{\alpha_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1\gamma_1 \\ \vdots \\ \beta_m\gamma_m \end{pmatrix} \begin{pmatrix} <\beta_1\delta_1,\cdot''_1 \\ \vdots \\ <\beta_m\delta_m,\cdot''_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\alpha}}S(\Theta) = (\operatorname{diag} \frac{\Sigma_i}{K_i} \sum_1^{N_o} \beta_i\delta_i) - (\operatorname{diag} \frac{\alpha_i\Sigma_i}{K_i})\left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1\delta_1 \\ \vdots \\ \beta_m\delta_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\mu}}S(\Theta) = (\operatorname{diag} \frac{1}{K_i} \{-\sum_1^{N_i}[(\cdot)\gamma_i^T+\gamma_i(\cdot)^T] - \alpha_i \sum_1^{N_o}[(\cdot)\gamma_i^T+\gamma_i(\cdot)^T]\beta_i+\Sigma_i \sum_1^{N_o} \delta_i<\beta_i\gamma_i,\cdot>'_i\}) -$$

$$- (\operatorname{diag} \frac{\alpha_i\Sigma_i}{K_i}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1\delta_1 \\ \vdots \\ \beta_m\delta_m \end{pmatrix} \begin{pmatrix} <\beta_1\gamma_1,\cdot>'_1 \\ \vdots \\ <\beta_m\gamma_m,\cdot>'_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\Sigma}}S(\Theta) = (\operatorname{diag} \frac{\Sigma_i}{K_i} \sum_1^{N_o} \beta_i\delta_i<\delta_i,\cdot>''_i) - (\operatorname{diag} \frac{\alpha_i\Sigma_i}{K_i}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1\delta_1 \\ \vdots \\ \beta_m\delta_m \end{pmatrix} \begin{pmatrix} <\beta_1\delta_1,\cdot> \\ \vdots \\ <\beta_m\delta_m,\cdot> \end{pmatrix}^T \right\}$$

Here, the arguments of β_i,γ_i and δ_i can be determined from the indices of summation, e.g.,

$$\sum_1^{N_o} \beta_i\gamma_i = \sum_{k=1}^{N_o} \beta_i(x_{ok})\gamma_i(x_{ok}) .$$

Setting

$$V = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \beta_1\gamma_1 \\ \vdots \\ \beta_m\gamma_m \\ \beta_1\delta_1 \\ \vdots \\ \beta_m\delta_m \end{pmatrix}$$

one obtains at Θ

$$\nabla \begin{pmatrix} A \\ M \\ S \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{pmatrix} - \begin{pmatrix} (\mathrm{diag}\ \frac{\alpha_i}{N_o}) & 0 & 0 \\ 0 & (\mathrm{diag}\ \frac{\alpha_i}{K_i}) & 0 \\ 0 & 0 & (\mathrm{diag}\ \frac{\alpha_i \Sigma_i}{K_i}) \end{pmatrix} \{\sum_1^{N_o} V(x_{ok}) \cdot V(x_{ok}), \cdot\} ,$$

where

$$B_{21} = (\mathrm{diag}\ \frac{1}{K_i} \sum_1^{N_o} \beta_i\gamma_i)$$

$$B_{22} = (\mathrm{diag}\ \frac{\alpha_i}{K_i} \sum_1^{N_o} \gamma_i\gamma_i^T\Sigma_i^{-1}\beta_i)$$

$$B_{23} = (\mathrm{diag}\ \frac{1}{K_i} \sum_1^{N_o} \beta_i\gamma_i<\delta_i,\cdot>''_i)$$

$$B_{31} = (\mathrm{diag}\ \frac{\Sigma_i}{K_i} \sum_1^{N_o} \beta_i\delta_i)$$

$$B_{32} = (\mathrm{diag}\ \frac{1}{K_i}\{-\sum_1^{N_i}[(\cdot)\gamma_i^T+\gamma_i(\cdot)^T] - \alpha_i\sum_1^{N_o}[(\cdot)\gamma_i^T+\gamma_i(\cdot)^T]\beta_i + \Sigma_i\sum_1^{N_o}\delta_i<\beta_i\gamma_i,\cdot>''_i\})$$

$$B_{33} = (\mathrm{diag}\ \frac{\Sigma_i}{K_i} \sum_1^{N_o} \beta_i\delta_i<\delta_i,\cdot>''_i) .$$

We have assumed that Θ is the strongly consistent maximum-likelihood estimate. Then, regardless of the relative sizes of N_i and N_o, one can show as in [3] that, with probability 1, $\{\nabla\Phi_\epsilon(\Theta) - E(\nabla\Phi_\epsilon(\Theta^o))\}$ converges to zero as N_o approaches infinity. Now

$$E(\nabla \begin{pmatrix} A(\Theta^o) \\ M(\Theta^o) \\ S(\Theta^o) \end{pmatrix}) = \begin{pmatrix} I & 0 & 0 \\ 0 & (\mathrm{diag}\ \frac{\alpha_i^o N_o}{K_i}\ I) & 0 \\ 0 & 0 & (\mathrm{diag}\ \frac{\alpha_i N_o}{K_i}\ I) \end{pmatrix} -$$

$$- \begin{pmatrix} (\mathrm{diag}\ \alpha_i^o) & 0 & 0 \\ 0 & (\mathrm{diag}\ \frac{\alpha_i^o N_o}{K_i}\ I) & 0 \\ 0 & 0 & (\mathrm{diag}\ \frac{\alpha_i^o N_o}{K_i}\ \Sigma_i^o) \end{pmatrix} \{ \int_{\mathbb{R}^n} V(x) \langle V(x), \cdot \rangle p(x)dx \}$$

$$= B(I - QR),$$

where

$$B = \begin{matrix} I & 0 & 0 \\ 0 & (\mathrm{diag}\ \frac{\alpha_i^o N_o}{K_i}\ I) & 0 \\ 0 & 0 & (\mathrm{diag}\ \frac{\alpha_i^o N_o}{K_i}\ I) \end{matrix}$$

$$Q = \begin{matrix} (\mathrm{diag}\ \alpha_i^o) & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (\mathrm{diag}\ \Sigma_i^o) \end{matrix}$$

$$R = \int_{\mathbb{R}^n} V(x) \ \langle V(x), \cdot \rangle p(x)dx .$$

It was shown in [3] that QR is positive-definite and symmetric with operator norm less than 1 with respect to the inner product $<\cdot,Q^{-1}\cdot>$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$. It follows that I-QR is positive-definite and symmetric with norm less than 1 with respect to $<\cdot,Q^{-1}\cdot>$. Since B and Q commute, $<\cdot,Q^{-1}B^{-1}\cdot>$ is an inner product on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$, and one sees that $<W,Q^{-1}W> \leq <W,Q^{-1}B^{-1}W>$ for $W \in \mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$. Consequently, B(I-QR) is positive-definite and symmetric with norm less than 1 with respect to the inner product $<\cdot,Q^{-1}B^{-1}\cdot>$. One concludes that

$$E(\nabla\Phi_{\epsilon}(\Theta^{O})) = (1 - \epsilon)I + \epsilon\, E(\nabla \begin{pmatrix} A(\Theta^{O}) \\ M(\Theta^{O}) \\ S(\Theta^{O}) \end{pmatrix})$$

has norm less than 1 with respect to $<\cdot,Q^{-1}B^{-1}\cdot>$ whenever $0 < \epsilon < 2$. This completes the proof of the theorem.

We remark that, reasoning as in [3], one may determine a particular value of ϵ (the "optimal ϵ") which yields, with probability 1 as N_O approaches infinity, the fastest asymptotic uniform rates of local convergence of the iterative procedure (2) near Θ. This optimal ϵ is given by

$$\epsilon = \frac{2}{2 - (\tau+\rho)}$$

where ρ and τ are, respectively the largest and smallest eigenvalues of B(I-QR) regarded as an operator on $\mathcal{E}\oplus\mathcal{M}\oplus\mathcal{S}$ ($\mathcal{E}$ is the subspace of $\mathcal{A}$ whose components sum to zero.) Since ρ and τ lie between zero and 1, one sees that the optimal ϵ is always greater than 1. If the component populations are "widely separated," then ρ and τ are near zero and, hence, the optimal ϵ is near 1. If two or more of the component populations are nearly indistinguishable and if N_o is large relative to the N_i's, then τ is near zero, and the optimal ϵ cannot be much smaller than 2.

3. Samples of the second type.

We now assume that K_o observations are obtained from the mixture population π_o, and that, for some $N_o < K_o$, N_o of these observations are left unidentified, while the remaining $K_o - N_o$ observations are identified. For $i = 1,\ldots,m$, let $\{x_{ik}\}_{k=1,\ldots,N_i}$ denote the subset of the identified observations which come from π_i, and let $\{x_{ok}\}_{k=1,\ldots,N_o}$ be the set of unidentified observations from π_o. The log-likelihood function for this sample is

$$L_2(\Theta) = \log \left\{\frac{(\sum_{i=1}^{m} N_i)!}{N_1!\ldots N_m!} \alpha_1^{N_1}\ldots\alpha_m^{N_m}\right\} + \sum_{i=1}^{m}\sum_{k=1}^{N_i} \log p_i(x_{ik}) + \sum_{k=1}^{N_o} \log p(x_{ok})$$

$$= \log \left\{\frac{(\sum_{i=1}^{m} N_i)!}{N_1!\ldots N_m!}\right\} + \sum_{i=1}^{m}\sum_{k=1}^{N_i} \log[\alpha_i p_i(x_{ik})] + \sum_{k=1}^{N_o} \log p(x_{ok}) .$$

Differentiating L_2 and setting its partial derivatives to zero gives the likelihood equations

$$\alpha_i = \tilde{A}_i(\Theta) \equiv \frac{N_i}{K_o} + \frac{\alpha_i}{K_o}\sum_{k=1}^{N_o}\frac{p_i(x_{ok})}{p(x_{ok})} \tag{3.a}$$

$$\mu_i = M_i(\Theta) \tag{3.b}$$

$$\Sigma_i = S_i(\Theta) \tag{3.c}$$

for $i = 1,\ldots,m$.

We set

$$\tilde{A}(\Theta) = \begin{pmatrix} \tilde{A}_1(\Theta) \\ \vdots \\ \tilde{A}_m(\Theta) \end{pmatrix}$$

and define an operator $\tilde{\Phi}_\epsilon$ on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$ by

$$\tilde{\Phi}_\epsilon(\Theta) = (1-\epsilon)\Theta + \epsilon \begin{pmatrix} A(\Theta) \\ M(\Theta) \\ S(\Theta) \end{pmatrix} .$$

Our iterative procedure is the following: Beginning with some starting value $\Theta^{(1)}$, define successive iterates inductively by

$$\Theta^{(j+1)} = \tilde{\Phi}_\epsilon(\Theta^{(j)}) \tag{4}$$

for $j = 1,2,3,\ldots$. As before, the desired local convergence result for this iterative procedure follows from the theorem below.

Theorem 2: With probability 1 as N_0 approaches infinity, $\tilde{\Phi}_\epsilon$ is a locally contractive operator (in some norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$) near the strongly consistent maximum-likelihood estimate whenever $0 < \epsilon < 2$.

Proof of Theorem 2: If Θ is the strongly consistent maximum-likelihood estimate, then, as before, it suffices to show that, with probability 1, $\nabla\tilde{\Phi}_\epsilon(\Theta)$ converges as N_0 approaches infinity to an operator which has operator norm less than 1 with respect to some vector norm on $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{S}$. Proceeding as before, one sees that

$$\nabla_{\bar{\alpha}}\tilde{A}(\Theta) = (\mathrm{diag}\ (1 - \frac{N_i}{\alpha_i K_o})) - (\mathrm{diag}\ \frac{\alpha_i}{K_o}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\mu}}\tilde{A}(\Theta) = -(\mathrm{diag}\ \frac{\alpha_i}{K_o}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \begin{pmatrix} <\beta_1\gamma_1,\cdot>'_1 \\ \vdots \\ <\beta_m\gamma_m,\cdot>'_m \end{pmatrix}^T \right\}$$

$$\nabla_{\bar{\Sigma}}\tilde{A}(\Theta) = -\ (\mathrm{diag}\ \frac{\alpha_i}{K_o}) \left\{ \sum_1^{N_o} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \begin{pmatrix} <\beta_1\delta_1,\cdot>''_1 \\ \vdots \\ <\beta_m\delta_m,\cdot>''_m \end{pmatrix}^T \right\}$$

The remaining Fréchet derivatives, i.e., the derivatives at Θ of M and S with respect to $\bar{\alpha}$, $\bar{\mu}$, and $\bar{\Sigma}$, are unchanged, except that K_i must be replaced by $\alpha_i K_o$ wherever it appears.

One obtains at Θ

$$(4) \qquad \nabla \begin{pmatrix} \tilde{A} \\ M \\ S \end{pmatrix} = \begin{pmatrix} (\mathrm{diag}(1 - \frac{N_i}{\alpha_i K_o})) & 0 & 0 \\ \tilde{B}_{21} & \tilde{B}_{22} & \tilde{B}_{23} \\ \tilde{B}_{31} & \tilde{B}_{32} & \tilde{B}_{33} \end{pmatrix} -$$

$$\begin{pmatrix} (\mathrm{diag}\ \frac{\alpha_i}{K_o}) & 0 & 0 \\ 0 & \frac{1}{K_o} I & 0 \\ 0 & 0 & (\mathrm{diag}\ \frac{\Sigma_i}{K_o}) \end{pmatrix} \left\{ \sum_{k=1}^{N_o} V(x_{ok}) < V(x_{o,k}),\cdot \right\}$$

In this expression, each $\tilde{B}_{jk}$ is the same as the corresponding B_{jk} defined previously, except that each K_i in the latter is replaced by $\alpha_i K_0$ in the former. One verifies that, with probability 1 as N_0 approaches infinity, (4) has the same limit as $\tilde{B}(I-QR)$, where Q and R are as before and $\tilde{B} = \frac{N_0}{K_0} I$. Repeating our earlier reasoning, one verifies that $\tilde{B}(I-QR)$ is positive-definite and symmetric with norm less than 1 with respect to the inner product $<\cdot, Q^{-1}\tilde{B}^{-1}\cdot>$. Hence

$$\nabla\tilde{\Phi}_\epsilon(\Theta) = (1-\epsilon) + \epsilon\nabla\begin{pmatrix}\tilde{A}(\Theta)\\ M(\Theta)\\ S(\Theta)\end{pmatrix}$$

converges to an operator which has norm less than 1 with respect to $<\cdot, Q^{-1}\tilde{B}^{-1}\cdot>$ whenever $0 < \epsilon < 2$. This completes the proof of the theorem.

The remarks concerning the "optimal ϵ" at the conclusion of the preceding section are valid here verbatim.

BIBLIOGRAPHY

1. W. H. Coberly, private communication.

2. D. W. Hosmer, Jr., "A comparison of iterative maximum-likelihood estimates of the parameters of a mixture of two normal distributions under three different types of samples," Biometrics 29 (1973), pp. 761-770.

3. B. C. Peters, Jr., and H. F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions," Report #51, NASA contract NAS-9-12777, University of Houston, Department of Mathematics.

FEATURE COMBINATIONS AND THE BHATTACHARYYA CRITERION

by

Henry P. Decell, Jr. and Salma K. Marani
Department of Mathematics
University of Houston
Houston, Texas

Report #55
NASA Contract NAS-9-15000

FEATURE COMBINATIONS AND THE BHATTACHARYYA CRITERION

Henry P. Decell, Jr. and Salma K. Marani

Department of Mathematics
University of Houston

ABSTRACT

We develop a procedure for calculating a kxn rank k matrix B for data compression using the Bhattacharyya bound on the probability of error and an iterative construction using Householder transformations. Two sets of remotely sensed agricultural data are used to demonstrate the application of the procedure. The results of the applications give some indication of the extent to which the Bhattacharyya bound on the probability of error is affected by such transformations for multivariate normal populations.

1. INTRODUCTION

For n-dimensional normal classes $N(\mu_i \Sigma_i)$ $i = 1,\ldots,m$, the Bhattacharyya coefficient (Andrews, 1972) for class i and j is given by:

$$\rho(i,j) = (q_i q_j)^{\frac{1}{2}} \int_{R^n} \{p_i(x) p_j(x)\}^{\frac{1}{2}} dx$$

and the _Bayes probability of error_ (Anderson, 1958) (Andrews, 1972) by

$$P_e = 1 - \int_{R^n} \max_{1 \leq i \leq m} \{q_i p_i(x)\} dx$$

where $p_i(x)$ denotes the conditional density of the random variable X given that $X \sim N(\mu_i, \Sigma_i)$ and $q_1, \ldots, q_m$,respectively, denote the (known) _a priori_ probabilities of the classes $N(\mu_i \Sigma_i)$ $i = 1, \ldots, m$.

It has been shown (Andrews, 1972) (Kaileth, 1967) that

$$P_e \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \{q_i q_j\}^{\frac{1}{2}} \int_{R^n} \{p_i(x) p_j(x)\}^{\frac{1}{2}} dx$$

If one considers a kxn rank k linear transformation B of the random variable X (i.e., Y≡BX), then the Bhattacharyya coefficient for class i and j for the classes $N(B\mu_i, B\Sigma_i B^T)$, $i = 1, \ldots, m$ is:

$$\rho_B(i,j) \equiv \{q_i q_j\}^{\frac{1}{2}} \int_{R^k} \{p_i(y,B) p_j(y,B)\}^{\frac{1}{2}} dy$$

and the Bayes probability of error for the classes $N(B\mu_i, B\Sigma_i B^T)$, $i = 1, \ldots, m$ is:

$$P_e(B) = 1 - \int_{R^k} \max_{1 \leq i \leq m} \{p_i(y,B)\} dy$$

where $p_i(y,B)$, $i = 1, \ldots, m$ denotes the conditional density of the random variable $Y = BX$ given that $Y \sim N(B\mu_i, B\Sigma_i B^T)$. It follows, since $P_e \leq \rho \equiv \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho(i,j)$, that

$$P_e(B) \leq \rho(B) \equiv \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_B(i,j)$$

and moreover, (Decell and Quirein, 1973) (Kaileth, 1967), that

(1) $P_e \leq P_e(B) \leq \rho(B)$.

(2) $P_e = P_e(B)$ if and only if $\rho = \rho(B)$.

2. THEORETICAL PRELIMINARIES

Let k be an integer $(0 < k < n)$, and $N(\mu_i, \Sigma_i)$ $i = 1,\ldots,m$ be n-variate normal populations with a priori probabilities $q_1,\ldots,q_m$. We would like to construct a kxn rank k matrix B that will minimize $\rho(B)$. The theoretical extent to which this is possible and the basis for the construction (Decell and Smiley, to appear) is summarized in the following theorem. Let $C = \{ u \in R^n : ||u|| = 1\}$ and $T(H) = \{H = I - 2uu^T : u \in C\}$ denote the set of Householder transformations on R^n (Householder, 1958).

Theorem. For each positive i, let $H_i \in T(H)$ be chosen such that

$$\rho((I_k|Z)H_1) = \underset{H \in T(H)}{\text{g.l.b}}\ \rho((I_k|Z)H)$$

and

$$\rho((I_k|Z)H_{i+1}H_i\cdots H_1) = \underset{H \in T(H)}{\text{g.l.b.}}\ \rho((I_k|Z)HH_i\cdots H_1)$$

then,

(1) $\rho((I_k|Z)H_{i+1}H_i\cdots H_1) \leq \rho((I_k|Z)H_i\cdots H_1)$.

(2) $\rho((I_k|Z)H_{i+1}\cdots H_1) \leq \rho((I_k|Z)H_i\cdots H_1H,\ H \in T(H))$.

(3) $\rho((I_k|Z)H_{i+1}H_i\cdots H_1) \leq \rho((I_k|Z)HH_i\cdots H_1,\ H \in T(H))$.

(4) $\rho((I_k|Z)H\cdots H_{i-(p-1)}HH_{i-(p+1)}H_1) \leq \rho((I_k|Z)H_{i+1}H_i\cdots H_1)$, $H \in T(H)$ and $p = 0,\ldots,i-2$.

(5) The monotone sequence of real numbers $\{\rho(B_i)\}_{i=1}^{\infty}$ where

$B_i = (I_k|Z)H_i \cdots H_1$ is bounded below by P_e and hence

$$\lim_{i\to\infty} \rho(B_i) = \underset{i}{\text{g.l.b.}}\{\rho(B_i)\}$$

We know (Decell and Quirein, 1973) that there is some kxn rank k matrix, say $\hat{B}$, that minimizes $\rho(B)$. If $\rho(B) < \underset{i}{\text{g.l.b.}}\{\rho(B_i)\}$ we will call the sequence $\{B_i\}_{i=1}^{\infty}$ sub optimal (optimal in the case of equality). There are several results (Decell and Smiley, to appear) that lend credibility to the conjecture that the sequence is optimal and cofinallv constant beyond the index $i = \min\{k,n-k\}$. We will proceed with the development of an iterative procedure for constructing the subject sequence and, finally, tabulate results of applications to remotely sensed agricultural data with equal a priori class probabilities. The approach (and its merit) will depend upon the bound provided by the inequality $P_e \le \rho(B_i)$ $i = 1,2,\ldots$, the non-increasing nature of the sequence $\{\rho(B_i)\}_{i=1}^{\infty}$, and the ability to manipulate the expressions for $\rho(B_i)$, $i = 1,2,\ldots$ in the case of normal populations.

3. THE GRADIENT OF $\rho((I_k|Z)H)$

We will develop an expression (for the case of normal n-variate populations $N(\mu_i,\Sigma_i)$, $i = 1,\ldots,m$) for the gradient of $\rho((I_k|Z)H)$ where $H \varepsilon T(H)$ has the form $H = I-2\frac{xx^T}{x^Tx}$, $x \neq \theta$. This expression will be used in a steepest descent procedure to calculate each Householder transformation H_1, H_2, H_3,... described in the preceding theorem. For m populations $N(\mu_i\Sigma_i)$, $i = 1,\ldots,m$ it is easy to establish that in order to calculate H_{i+1}, one need only apply the steepest descent procedure to the Bhattacharyya coefficient determined by the populations $N(H_i\cdots H_1\mu_j, H_i\cdots H_1\Sigma_j H_1\cdots H_i)$ $j = 1,\ldots,m$.

The expression for $\rho_{(I_k|Z)H}(i,j)$ is given by (Andrews, 1972) (Kaileth, 1967) (for the case of equal a priori probabilities $q_i = 1/m$, $i = 1,\ldots,m$):

$$\rho_{(I_k|Z)H}(i,j) = \frac{1}{m}\exp - \frac{1}{4}\delta_{ij}^T(\Sigma_i+\Sigma_j)^{-1}\delta_{ij} \quad - \frac{1}{2}\ln\left(\frac{|\hat{\Sigma}_i+\hat{\Sigma}_j|}{2^k|\hat{\Sigma}_i|^{1/2}|\hat{\Sigma}_j|^{1/2}}\right)$$

where $\hat{\delta}_{ij} = (I_k|Z)H(\mu_i-\mu_j)$ and $\hat{\Sigma}_i = (I_k|Z)H\Sigma_iH(I_k|Z)^T$, in which case,

$$\rho((I_k|Z)H) = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\rho_{(I_k|Z)H}(i,j).$$

If we define

$$F_{ij} = -\frac{1}{4}\hat{\delta}_{ij}^T(\hat{\Sigma}_i+\hat{\Sigma}_j)^{-1}\hat{\delta}_{ij} \quad \text{and} \quad G_{ij} = -\frac{1}{2}\ln\left(\frac{|\hat{\Sigma}_i+\hat{\Sigma}_j|}{2^k|\hat{\Sigma}_i|^{1/2}|\hat{\Sigma}_j|^{1/2}}\right)$$

we have that the differential of $\rho_{(I_k|Z)H}(i,j)$ is

$$d(\rho_{(I_k|Z)H}(i,j)) = \frac{1}{m}\exp(F_{ij}+G_{ij})(d(F_{ij}) + d(G_{ij})).$$

from whence it follows that

$$d(\rho((I_k|Z)H)) = \frac{1}{m}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\exp(F_{ij}+G_{ij})(d(F_{ij}) + d(G_{ij})).$$

In order to simplify the notation, define $\Sigma_{ij} = \Sigma_i + \Sigma_j$ and $\Delta_{ij} = (\mu_i-\mu_j)(\mu_i-\mu_j)^T$.

Let tr(•) denote the trace of (•) and $|\cdot| = \det(\cdot)$. With a bit of matrix algebra it follows that

$$F_{ij} = -\frac{1}{4}\mathrm{tr}\{((I_k|Z)H\Sigma_{ij}H(I_k|Z)^T)^{-1}(I_k|Z)H\Delta_{ij}H(I_k|Z)^T\}$$

and

$$G_{ij} = -\frac{1}{2} \ln|(I_k|Z)H\Sigma_{ij}H(I_k|Z)^T| + \frac{1}{4} \ln|(I_k|Z)H\Sigma_i H(I_k|Z)^T|$$
$$+ \frac{1}{4} \ln|(I_k|Z)H\Sigma_j H(I_k|Z)^T| + \frac{k}{2} \ln 2.$$

We will now develop expressions for $d(F_{ij})$ and $d(G_{ij})$, $i,j = 1,\ldots,m$.

According to Decell and Quirein (1973)

$$d(F_{ij}) = -\frac{1}{2} \operatorname{tr}\{d((I_k|Z)H)Q_{ij}\}$$

where $B = (I_k|Z)H$ and

$$Q_{ij} = [\Delta_{ij}B^T - \Sigma_{ij}B^T(B\Sigma_{ij}B^T)^{-1}B\nabla_{ij}B^T](B\Sigma_{ij}B^T)^{-1}.$$

Since $H = I - 2\frac{xx^T}{x^Tx}$ it follows that

$$d((I_k|Z)H) = d((I_k|Z)(I - 2\frac{xx^T}{x^Tx})) = -2(I_k|Z)d\left(\frac{xx^T}{x^Tx}\right)$$

$$= -2(I_k|Z)\left\{\frac{x^Txd(xx^T) - xx^Td(x^Tx)}{(x^Tx)^2}\right\}$$

$$= \frac{-2(I_k|Z)}{(x^Tx)^2}\{x^Tx(d(x)x^T+xd(x)^T)-xx^T(d(x)^Tx+x^Td(x))\}$$

$$= \frac{-2(I_k|Z)}{(x^Tx)^2}\{(d(x)x^Txx^T+xx^Txd(x)^T-xx^Td(x)x^T-xd(x)^Txx^T\}$$

$$= \frac{-2(I_k|Z)}{(x^Tx)^2}\{(d(x)x^T-xd(x)^T)xx^T-xx^T(d(x)x^T-xd(x)^T)\}.$$

Substituting the latter in the expression

$$d(F_{ij}) = -\frac{1}{2}\,\mathrm{tr}\{d((I_k|Z)H)Q_{ij}\}$$

and using the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we have

$$d(F_{ij}) = -\frac{1}{2}\mathrm{tr}\left\{\frac{-2(I_k|Z)}{(x^Tx)^2}[(d(x)x^T-xd(x)^T)xx^T-xx^T(d(x)x^T-xd(x)^T)]\,Q_{ij}\right\}$$

$$= \frac{1}{(x^Tx)^2}\mathrm{tr}\{Q_{ij}(I_k|Z)[(d(x)x^T-xd(x)^T)xx^T-xx^T(d(x)x^T-xd(x)^T)]\}$$

$$= \frac{1}{(x^Tx)^2}\,\mathrm{tr}\{xx^TQ_{ij}(I_k|Z)(d(x)x^T-xd(x)^T)-Q_{ij}(I_k|Z)xx^T(d(x)x^T-xd(x)^T)\}.$$

With a little matrix algebra (and some patience) it follows that

$$d(F_{ij}) = \frac{1}{(x^Tx)^2}\,\mathrm{tr}\{[(xx^TQ_{ij}(I_k|Z) - Q_{ij}(I_k|Z)xx^T)^T - (xx^TQ_{ij}(I_k|Z) - Q_{ij}(I_k|Z)xx^T)]xd(x)^T\}$$

We now find an expression for $d(G_{ij})$. First, recall (Kullback, 1968) that

$$d(\ln|B\Sigma B^T|) = 2\mathrm{tr}\{d(B)\Sigma B^T(B\Sigma B^T)^{-1}\}$$

so that

$$d(G_{ij}) = -\mathrm{tr}\{d((I_k|Z)H)\Sigma_{ij}H(I_k|Z)^T((I_k|Z)H\,\Sigma_{ij}H(I_k|Z)^T)^{-1}\}$$
$$-\frac{1}{2}\mathrm{tr}\{d((I_k|Z)H)\Sigma_iH(I_k|Z)^T((I_k|Z)H\Sigma_iH(I_k|Z)^T)^{-1}$$
$$+\frac{1}{2}\,\mathrm{tr}\{d((I_k|Z)H)\Sigma_jH(I_k|Z)^T((I_k|Z)H\Sigma_jH(I_k|Z)^T)^{-1}\}.$$

Obviously, the summands in the expression for $d(G_{ij})$ differ from the expression

$$d(F_{ij}) = -\frac{1}{2}\,\mathrm{tr}\{d((I_k|Z)H)Q_{ij}\}$$

only by multiplicative constants and the matrix Q_{ij}. Hence, we may use the final expression for $d(F_{ij})$ to obtain the expression for $d(G_{ij})$ by simply adjusting the multiplicative constants and replacing Q_{ij} (in each summand in $d(G_{ij})$) with the expressions

$$J_{ij} = \Sigma_{ij}H(I_k|Z)^T[(I_k|Z)H\Sigma_{ij}H(I_k|Z)^T]^{-1}$$

$$K_{ij} = \Sigma_i H(I_k|Z)^T[(I_k|Z)H\Sigma_i H(I_k|Z)^T]^{-1}$$

$$L_{ij} = \Sigma_j H(I_k|Z)^T[(I_k|Z)H\Sigma_j H(I_k|Z)^T]^{-1}$$

At this point we will simplify the notation. Let

$$\hat{Q}_{ij} = (xx^TQ_{ij}(I_k|Z)-Q_{ij}(I_k|Z)xx^T)^T-(xx^TQ_{ij}(I_k|Z)-Q_{ij}(I_k|Z)xx^T)$$

and let $\hat{J}_{ij}$, $\hat{K}_{ij}$, and $\hat{L}_{ij}$ be similarly defined by substituting, respectively, J_{ij}, K_{ij}, and L_{ij} for Q_{ij} in the expression for $\hat{Q}_{ij}$, $i,j = 1,\ldots,m$. It follows that

$$d(F_{ij}) = \frac{1}{(x^Tx)^2}\,\mathrm{tr}(\hat{Q}_{ij}xd(x)^T)$$

$$d(Gij) = \frac{2}{(x^Tx)^2}\,\mathrm{tr}(\hat{J}_{ij}xd(x)^T) - \frac{1}{(x^Tx)^2}\,\mathrm{tr}(\hat{K}_{ij}xd(x)^T)$$

$$- \frac{1}{(x^Tx)^2}\,\mathrm{tr}(\hat{L}_{ij}xd(x)^T).$$

In order that x be extremal, it is sufficient that x satisfy

$$G(x) \equiv \frac{1}{m} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{\exp(F_{ij}+G_{ij})}{(x^T x)^2} (\hat{Q}_{ij} + 2\hat{J}_{ij} - \hat{K}_{ij} - \hat{L}_{ij})x = \theta.$$

Of course, the function $G(x)$ is the gradient of

$\rho((I_k|Z)(I - 2\frac{xx^T}{x^T x}))$ with respect to x.

With $G(x)$, we use a steepest descent technique to construct H_1. The process is repeated for the construction of H_2 since, given H_1, the problem of constructing H_2 is identical to that of constructing H_1 provided the populations are taken to be $N(H_1\mu_i, H_1\Sigma_i H_1)$ $i = 1,\ldots,m$.

Test results are presented in the following tables for nine twelve channel, C-1 flight line agricultural classes: soybeans, corn, oats, red-clover, alfalfa, rye, bare soil, and two types of wheat. The Hill County data is sixteen channel data for five agricultural classes: winter wheat, fallow crop, barley, grass, and stubble.

C-1 FLIGHT LINE DATA

$n = 12$, $m = 9$, $k = 6$, $\rho = .024$

Iteration	H_{B_1}	H_{B_2}	H_{B_3}
0	.327	.109	.134
1	.223	.060	.034
2	.171	.062	.033
3	.135	.068	.032
4	.116	.058	.031
5	.1157	.055	.0309
6	.1150	.054	.0303

HILL COUNTY DATA

$n = 16, m = 5, k = 6. \rho = .107$

Iteration	H_{B_1}	H_{B_2}	H_{B_3}
0	.872	.336	.299
1	.785	.310	.287
2	.525	.286	.232
3	.439	.273	.227
4	.576	.267	.226
5	.386	.265	.224
6	.363	.264	.223

BIBLIOGRAPHY

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, Inc.

Andrews, H.C. (1972). Introduction to Mathematical Techniques in Pattern Recognition. New York: Wiley-Interscience.

Decell, H.P. Jr. and Quirein, J.A. (March, 1973). "An Iterative Approach to the Feature Selection Problem". IEEE Cat. #CH0834-2, pp. 3B-1--3B-12.

Decell, H.P. and Smiley, W. "Householder Transformations and Optimal Linear Combinations". Report # 38 NAS-9-12777, Dept. of Mathematics, Univ. of Houston, Texas.

Householder, Alston S. (1958). "Unitary Triangularization of a Non-Symmetric Matrix". J. Assoc. Comput. Mech., 339-342.

Kaileth, T. (Feb. 1967). "The Divergence and Bhattacharyya Distance Measures in Signal Selection". IEEE Transaction on Communications Theory, Vol. 15, NO. 1, pp 52-60.

Kullback, Solomon (1968). Information Theory and Statistics. New York: Dover Publications.

FEATURE COMBINATIONS AND THE DIVERGENCE CRITERION

by

Henry P. Decell, Jr. and Shailesh M. Mayekar
Department of Mathematics
University of Houston
Houston, Texas

Report #56
NASA Contract NAS-9-15000

FEATURE COMBINATIONS AND THE DIVERGENCE CRITERION

Henry P. Decell, Jr. and Shailesh M. Mayekar

Department of Mathematics
University of Houston

ABSTRACT

Classifying large quantities of multidimensional data (e.g., remotely sensed agricultural data)(Remote, 1968) requires efficient and effective classification techniques and the construction of certain transformations of a dimension-reducing, information-preserving nature. This paper will deal with the construction of transformations that minimally degrade information (i.e., class separability). We will only consider the construction of linear dimension-reducing transformations for multivariate normal populations and information content will be measured by divergence (Kullback, 1968).

1. INTRODUCTION

For n-dimensional normal classes $N(m_i, V_i)$ $i = 1, \ldots, m,$ the divergence between class i and j (Kullback, 1968) is given by

$$D_{ij} = \frac{1}{2}\,\mathrm{tr}[(V_i - V_j)(V_j^{-1} - V_i^{-1})] + \frac{1}{2}\mathrm{tr}[(V_i^{-1} + V_j^{-1})(m_i - m_j)(m_i - m_j)^T]$$

Let $\delta_{ij} = m_i - m_j$. Then

$$\begin{aligned} D_{ij} &= \frac{1}{2}\mathrm{tr}[(V_i - V_j)(V_j^{-1} - V_i^{-1})] + \frac{1}{2}\mathrm{tr}[(V_i^{-1} + V_j^{-1})(\delta_{ij})(\delta_{ij})^T] \\ &= \frac{1}{2}\mathrm{tr}[V_i^{-1}(V_j + \delta_{ij}\delta_{ij}^T)] + \frac{1}{2}\mathrm{tr}[V_j^{-1}(V_i + \delta_{ij}\delta_{ij}^T)] - n. \end{aligned}$$

The <u>interclass divergence</u> (Decell and Quirein, Oct. 1973) for m populations is given by

$$D = \sum_{i=1}^{m-1} \sum_{\substack{j=1 \\ i \neq j}}^{m} D_{ij}$$

and it follows that

$$\begin{aligned} D &= \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} V_i^{-1}(\sum_{\substack{j=1 \\ i \neq j}}^{m} (V_j + \delta_{ij}\delta_{ij}^T))] - \frac{m(m-1)}{2}\, n \\ &= \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} V_i^{-1}\; S_i] - \frac{m(m-1)}{2}\, n, \end{aligned}$$

where

$$S_i = \sum_{\substack{j=1 \\ i \neq j}}^{m} (V_j + \delta_{ij}\delta_{ij}^T).$$

If B is a $k \times n$ rank k matrix, the <u>B-interclass divergence</u> (Decell and Quirein, Oct. 1973) is given by

$$D_B = \sum_{i=1}^{m-1} \sum_{\substack{j=1 \\ i \neq j}}^{m} D_B(i,j)$$

$$D_B = \frac{1}{2}\,\mathrm{tr}\Big[\sum_{i=1}^{m} (BV_iB^T)^{-1}(BS_iB^T)\Big] - \frac{m(m-1)}{2}k.$$

As in the case of average interclass divergence, the B-interclass divergence is a measure of the "separation" in the classes $N(Bm_i, BV_iB^T)$ $i = 1,\ldots,m$, and is a useful tool for constructing rank k linear transformations that preserve "class separability". It has been shown (Decell and Quirein, Oct. 1973) that whenever $D = D_B$, the probability of misclassification (Anderson, 1958) for the classes $N(Bm_i, BV_iB^T)$, $i = 1,\ldots,m$ is the same as the probability of misclassification for the classes $N(m_i, V_i)$, $i = 1,\ldots,m$.

2. THEORETICAL PRELIMINARIES

We will assume that k is an integer $(k < n)$ and develop a procedure for selecting a $k \times n$ rank k matrix B such that D_B is maximum. The procedure will be based upon the following theorem (Decell and Smiley, to appear). We will let $C = \{u \in R^n\colon ||u||=1\}$ and $T(H) = \left\{H = I-2uu^T\colon u \in C\right\}$ denote the set of Householder transformations defined on R^n (Householder, 1968).

Theorem. For each positive integer i let $H_i \in T(H)$ be inductively chosen such that

$$D_{(I_k|Z)H_iH_{i-1}\cdots H_1} = \underset{H\in T(H)}{\text{l.u.b.}}[D_{(I_k|Z)HH_{i-1}\cdots H_1}]$$

where

$$D_{(I_k|Z)H_1} = \underset{H\in T(H)}{\text{l.u.b.}}\; D_{(I_k|Z)H}.$$

The following hold:

(1) $D_{(I_k|Z)H_iH_{i-1}\cdots H_1} \leq D_{(I_k|Z)H_{i+1}H_i\cdots H_1}$.

(2) $D_{(I_k|Z)H_iH_{i-1}\cdots H_1H} \leq D_{(I_k|Z)H_{i+1}H_i\cdots H_1}$, for every $H \in T(H)$.

(3) $D_{(I_k|Z)HH_iH_{i-1}\cdots H_1} \leqq D_{(I_k|Z)H_{i+1}H_i\cdots H_1}$, for every $H \varepsilon T(H)$.

(4) $D_{(I_k|Z)H_iH_{i-1}\cdots H_{i-(p-1)}HH_{i-(p+1)}\cdots H_1} \leqq D_{(I_k|Z)H_{i+1}\cdots H_1}$,

for every $H \varepsilon T(H)$, $p = 0,1,\ldots,i-2$.

(5) The monotone sequence

$$\{D_{B_i}\}_{i=1}^{\infty} \equiv \{D_{(I_k|Z)H_i\cdots H_1}\}_{i=1}^{\infty} \text{ is bounded above,}$$

and hence

$$\lim_{i\to\infty} D_{(I_k|Z)H_i\cdots H_1} = \text{l.u.b.}_i \{D_{(I_k|Z)H_i\cdots H_1}\}.$$

We would, of course, be pleased if it were the case that $\text{l.u.b.}_i \{D_{(I_k|Z)H_i\cdots H_1}\} = D$. This, unfortunately, is not always the case for some choice of $k < n$ and is not possible, in general, for any $k < n$. We do know that there is some $k \times n$ rank k matrix B for which D_B is maximum and, in general, that $D_B \leqq D$ (Decell and Quirein, Oct. 1973). It follows, moreover, that since the matrices of the form $(I_k|Z)H_i\cdots H_1$ have rank k,

$$D_{(I_k|Z)H_i\cdots H_1} \leqq D_B \leqq D \text{ for every integer } i.$$

We will call the sequence $\{D_{(I_k|Z)H_i\cdots H_1}\}_{i=1}^{\infty}$ <u>suboptimal</u> whenever

$$\text{l.u.b.}_i \{D_{(I_k|Z)H_i\cdots H_1}\} < D_B$$

(and <u>optimal</u> in the case of equality).

There are several open theoretical questions that deal with the conjecture that the sequence is, in general, optimal and cofinally constant beyond the index $i = \min\{k,n-k\}$ (Decell and Smiley, to appear). In what follows we will develop a procedure for constructing the subject sequence and demonstrate its application to agricultural data.

3. THE GRADIENT OF D_B

It has been shown (Quirein, Nov. 1972) that the differential dD_B of D_B (regarded as a function of the $k \times n$ matrix B) can be expressed in the form $dD_B = F + G$, where, when the indicated inverses exist,

$$F = \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BV_iB^T)^{-1}(dB\ S_iB^T + BS_i dB^T)\ ]$$

$$= \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (dB\ S_iB^T)(BV_iB^T)^{-1}]$$

$$+ \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BS_i\ dB^T)(BV_iB^T)^{-1}]$$

$$= \mathrm{tr}[\sum_{i=1}^{m} (dB\ S_iB^T)(BV_iB^T)^{-1}]$$

and

$$G = -\frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BV_iB^T)^{-1}(dB\ V_iB^T + BV_i dB^T)(BV_iB^T)^{-1}(BS_iB^T)\ ]$$

$$= -\frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (dB\ V_iB^T)(BV_iB^T)^{-1}(BS_iB^T)(BV_iB^T)^{-1}]$$

$$- \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BV_iB^T)^{-1}(BS_iB^T)(BV_iB^T)^{-1}(BV_i dB^T)]$$

$$= -\ \mathrm{tr}[\sum_{i=1}^{m} (dB\ V_iB^T)(BV_iB^T)^{-1}(BS_iB^T)(BV_iB^T)^{-1}\ ].$$

Thus,

$$-dD_B = \mathrm{tr}[\sum_{i=1}^{m} dB\{S_iB^T - V_iB^T(BV_iB^T)^{-1}(BS_iB^T)\}(BV_iB^T)^{-1}]$$

$$= \mathrm{tr} \sum_{i=1}^{m} dB\ Q_i$$

where

$$Q_i = [\{S_iB^T - V_iB^T(BV_iB^T)^{-1}(BS_iB^T)\}(BV_iB^T)^{-1}].$$

We are, of course, interested in extremizing D_B over the particular subclass of $k \times n$ rank k matrices of the form $(I_k|Z)H$ where $H \varepsilon T(H)$ (e.g., for $i = 1$ we find H_1 that maximizes $D_{(I_k|Z)H}$). Actually, one need only consider what is required to compute H_1. The computation of H_2 is accomplished by the same procedure as that for H_1. It is simply a matter of, after selecting H_1, redefining the m classes to be $N(H_1m_i, H_1V_iH_1)$, $i = 1,\ldots,m$ and proceeding as in the selection of H_1.

With these facts in mind we will simply calculate the gradient of D_B where B is restricted to having the form $B = (I_k|Z)H$, $H \varepsilon T(H)$. The restrictions $H \varepsilon T(H)$ can be accomplished by considering those $k \times n$ rank k matrices of the form

$$B = (I_k|Z)(I - 2\frac{ww^T}{w^Tw}), \quad w \varepsilon R^n (w \neq \theta)$$

It follows that

$$dB = d[(I_k|Z)\ (I - 2\frac{ww^T}{w^Tw})] = -2(I_k|Z)\ d(ww^T/w^Tw)$$

$$= -2(I_k|Z)\ [\frac{w^Twd(ww^T) - ww^Td(w^Tw)}{(w^Tw)^2}]$$

$$= -\frac{2(I_k|Z)}{(w^Tw)^2}[w^Tw(dw\,w^T + wdw^T) - ww^T(w^Tdw + dw^T\,w)]$$

$$= -\frac{2(I_k|Z)}{(w^Tw)^2}[dw\,w^Tw\,w^T + w\,w^Tw\,dw^T - w\,w^Tdw\,w^T - w\,dw^Tw\,w^T]$$

$$= -\frac{2(I_k|Z)}{(w^Tw)^2}[(dw\,w^T - wdw^T)ww^T - ww^T(dw\,w^T - wdw^T)]$$

Substituting the latter in the expression for dD_B,

$$dD_B = \text{tr}\sum_{i=1}^{m}[-\frac{2(I_k|Z)}{(w^Tw)^2}\{(dw\,w^T - wdw^T)ww^T - ww^T(dw\,w^T - wdw^T)\}Q_i]$$

$$= \text{tr}\sum_{i=1}^{m}[-\frac{2Q_i(I_k|Z)}{(w^Tw)^2}\{(dw\,w^T - wdw^T)ww^T - ww^T(dw\,w^T - wdw^T)\}]$$

$$= \text{tr}\sum_{i=1}^{m}\frac{-2}{(w^Tw)^2}[ww^T\,Q_i\,(I_k|Z)(dw\,w^T - wdw^T)$$

$$- Q_i(I_k|Z)ww^T(dw\,w^T - wdw^T)]$$

$$= \frac{-2}{(w^Tw)^2}\,\text{tr}\sum_{i=1}^{m}[M_idw\,w^T - M_iwdw^T - N_idw\,w^T + N_iwdw^T]$$

Where $M_i = ww^TQ_i(I_k|Z)$ and $N_i = Q_i(I_k|Z)ww^T$.

$$dD_B = \frac{-2}{(w^Tw)^2}\,\text{tr}[\sum_{i=1}^{m}\{w^T\,M_i\,dw - w^T\,N_i\,dw + N_i\,w\,dw^T - M_i\,w\,dw^T\}]$$

$$= \frac{-2}{(w^TW)^2}\,\text{tr}[\sum_{i=1}^{m}\{dw^T\,M_i^T\,w - dw^T\,N_i^T\,w + N_i\,w\,dw^T - M_i\,w\,dw^T\}]$$

$$dD_B = \frac{-2}{(w^T w)^2}\, \mathrm{tr}[\sum_{i=1}^{m} \{M_i^T\, w\, dw^T - N_i\, w\, dw^T + N_i\, w\, dw^T - M_i\, w\, dw^T\}]$$

$$= \frac{-2}{(w^T w)^2}\, \mathrm{tr}[\sum_{i=1}^{m} \{(M_i - N_i)^T - (M_i - N_i)\}w\, dw^T].$$

The necessary condition that w be extremal is then,

$$G(w) = \frac{-2}{(w^T w)^2} \sum_{i=1}^{m} \{(M_i - N_i)^T - (M_i - N_i)\}w = \theta \text{ (the zero vector).}$$

We note that $G(w)$ is the gradient of $D_{(I_k|Z)(I - 2\frac{ww^T}{w^T w})}$ and use a steepest descent procedure for finding the extremal w. The process is repeated for each sequential index until corresponding values of divergence "stabilize." Test results are presented in the following tables. The C-1 flight line data is twelve channel data for nine agricultural classes: soybeans, corn, oats, red-clover, alfalfa, rye, bare soil, and two types of wheat. The Hill County data is sixteen-channel data for five agricultural classes: winter wheat, fallow crop, barley, grass, and stubble.

The starting value w_o for the steepest descent procedure for selecting each successive Householder transformation $H_1, H_2, H_3 \cdots$ was arbitrarily chosen to be $w_o = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^T$. Choosing starting values in this arbitrary fashion is certainly not the most clever thing to do in the presence of the monotone behavior of the sequence $D_{(I_k|Z)H_i \cdots H_1}$. One would expect, for example, that the starting values for the selection of H_{i+1} should depend upon the unit vectors previously selected as generators of $H_i, H_{i-1}, \ldots, H_1$ in such a way as to guarantee that the starting value w_o, for the descent procedure for selecting H_{i+1}, satisfies

$$D(I_k|Z)H_i\cdots H_1 \leq D(I_k|Z)(I - 2\frac{w_o w_o^T}{w_o^T w_o})H_i\cdots H_1.$$

This rather arbitrary selection of the starting vector does, as the examples demonstrate, violate the latter inequality. The question about how to choose starting vectors, according to the latter inequality, is still an open one and its answer would certainly decrease computation time.

C-1 Flight Line Date

n=12, k=6, m=9, D=10,660

Iteration for H_1

No*	Divergence D_B
1	1982
2	3536
3	4533
4	5781
5	6910
6	7522
7	7710
8	7790
9	7838
10	7865
11	7881
12	7892

Hill County Data

n=16, k=8, m=5, D=636

Iteration for H_1

No*	Divergence D_B
1	114.58
2	136.66
3	152.27
4	179.69
5	223.81
6	247.42
7	252.78
8	257.12
9	260.74
10	263.95

*Iteration counter

C-1 Flight Line Data (cont.)

Iteration for H_2

No*	Divergence D_B
1	7815
2	8797
3	9542
4	9785
5	9901
6	9966
7	10,005
8	10,031
9	10,048

Iteration for H_3

No*	Divergence D_B
1	7582
2	8705
3	9809
4	9947
5	9995
6	10,020
7	10,037
8	10,049
9	10,058

Hill County Data (cont.)

Iteration for H_2

No*	Divergence D_B
1	269.00
2	280.48
3	293.32
4	300.68
5	304.07
6	306.19
7	307.74
8	308.95
9	309.93

Iteration for H_3

No*	Divergence D_B
1	312.18
2	344.52
3	380.83
4	387.20
5	391.70
6	392.96
7	394.58
8	399.47

Iteration for H_4

No*	Divergence D_B
1	371.12
2	394.75
3	398.62
4	400.69
5	402.03
6	402.98
7	403.74

*Iteration counter

BIBLIOGRAPHY

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, Inc.

Decell, H.P. and Quirein, J.A. (October 1973). "An Iterative Approach to the Feature Selection Problem". IEEE Cat. #CHO834-2, pp. 3B-1-3B-12.

Decell, H.P. and Smiley, W. (to appear). "Householder Transformations and Optimal Linear Combinations." Comm. in Stat.

Householder, A.S. (1968). "Unitary Triangularization of a Non-Symmetric Matrix." J. Assoc. Comput. Mach., pp. 339-342.

Kullback, S. (1968). Information Theory and Statistics. New York: Dover Publications.

Quirein, J.A. (November 1972). "Some Necessary Conditions for an Extremum." Report #12 NAS-9-12777. Dept. of Mathematics, Univ. of Houston, Texas

"Remote Multispectral Sensing in Agriculture." (1968). Report of the Laboratory for Agricultural Remote Sensing Vol. 3, Research Bulletin #844. Purdue Univ., Lafayette, Indiana.

Program Documentation

B-Average Bhattacharya Distance

by

Salma [illegible]. Marani

Department of Mathematics, University of Houston

Houston, Texas 77004

August, 1976

Report #57

NAS-9-1500C

DOCUMENTATION

Computation of the Total and the B-average Bhattacharya Distance:

(Univac 1108, Univ. of Houston).

This program consists of 3 subroutines to be executed in the following sequence:

(1) Subroutine BHATT

(2) Subroutine BHATB1

(3) Subroutine BHATB2

1. SUBROUTINE BHATT

ABSTRACT

This subroutine calculates the total Bhattacharyya Distance, BDIST, using all N channels. The output of this program, BDIST, will be used in comparing the difference $\delta_H = H_B - \text{BDIST}$ where H_B is the B-average Bhattacharyya Distance computed in the subroutines BHATB1, BHATB2.

User's Information:

(Double Precision Version Only).

In order to use this subroutine the following FORTRAN calling sequence must be given:

```
CALL BHATT(COVAR, XMEAN, M,N, BDIST)
```

where:

COVAR(input) is a real 3-dimensional array (M×N×N) and contains the M N×N class covariance matrices (positive definite symmetric) used as input.

XMEAN(input) is a real 2-dimensional array (M×N) and contains the M N-dimensional class mean vectors.

M(input) is the no. of classes under consideration i.e. the no. of covariance matrices and mean vectors.

N(input) is the dimension of the covariance matrices and the mean vectors.

BDIST (output) is the value of the total Bhattacharyya Distance computed by subroutine BHATT.

SUBROUTINES USED:

Subroutine BHATT in turn calls the following subroutines

1. Subroutine MATMUL. This subroutine computes the product of 2 matrices. It calls subroutines SUPSUM and ORDER.
2. Subroutine CHLSKY. This subroutine computes the inverse of a positive definite symmetric matrix.
3. Subroutine DET. This subroutine computes the determinant of a positive definite symmetric matrix.

NOTE: (1). The format statements for input, output are dependent upon the dimensions of the input data and corresponding adjustments have to be made to formats when different sets of data are run.

(2). The variables declared in the DIMFNSION statements have to similarly correspond to the dimensions of the input data.

ALGORITHM:

Subroutine BHATT computes the value of the total Bhattacharyya Distance using the covariance matrices and mean vectors as inputs.

The total Bhattacharyya Distance, BDIST, is computed by the formula

$$\mathrm{BDIST} = \frac{1}{m} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} H(i,j)$$

where $H(i,j)$, the interclass Bhattacharyya Distance between classes i and j is given by

$$H(i,j) = \exp[-\frac{1}{4} \delta_{ij}{}^{T}(\Sigma_i + \Sigma_j)^{-1}\delta_{ij} - \frac{1}{2} \ln \frac{|\Sigma_i + \Sigma_j|}{2^N|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}}]$$

where $\delta_{ij} = \mu_i - u_j$ and μ_i is the mean vector corresponding to class i and Σ_i is the covariance matrix corresponding to class i.

2. SUBROUTINE BHATB1:

ABSTRACT

This subroutine attempts to calculate the minimum B-average Bhattacharyya Distance using 1 Householder transformation to construct the B-matrix.

USER'S INFORMATION:

(Double Precision Version Only)

In order to use this subroutine the following FORTRAN calling sequence must be given:

CALL BHATB1 (COVAR, XMEAN, M,N, K, ITE, ALPHA)

where

COVAR(input) is a real 3-dimensional array (M N×N) containing the M N×N covariance matrices.

XMEAN(input) is a real 2-dimensional array (M×N) and contains the M N-dimensional mean vectors used as input.

M(input) is the number of classes under consideration (i.e. the no. of covariance matrices and mean vectors).

N(input) is the dimension of the covariance matrices and the mean vectors.

K(input) is the number of rows desired in the transformation matrix B (which is K×N)

ITE(input) is 1 + (the no. of iterations required)

ALPHA(input) is a varying parameter in the iteration formula.

OUTPUT OF SUBROUTINE BHATB1

This subroutine has the following output:

1. The transformation matrix B (which has dimension K×N) corresponding to a particular value of the Householder generator F.*
2. The value of the B-average interclass Bhattacharyya Distance $H_B(i,j)$, $i = 1,\ldots,M-1$; $j = i+1,\ldots,M$
3. The N-dimensional F-vector which is the generator of the Householder transformation $H = I-2FF^T$ used in constructing the B-matrix $B = (I_K|Z)H$.
4. The value of the B-average Bhattacharyya Distance, H_B corresponding to the matrix B.
5. The partial derivative vector $\frac{\partial H_B}{\partial F}$ which contains the partial derivatives of H_B with respect to the vector F.

*See 'ALGORITHM'

Subroutines Used

The following subroutines are in turn called by subroutine BHATB1:

1. Subroutine MATMUL - calls SUPSUM and ORDER.
2. Subroutine CHLSKY.
3. Subroutine DET.

ALGORITHM

Subroutine BHATB1 attempts to compute the minimum B-average Bhattacharyya Distance using one Householder transformation to compute the B-matrix. The B-average Bhattacharyya Distance is given by the formula

$$H_B = \frac{1}{m} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} H_B(i,j)$$

where

$$H_B(i,j) = \exp[-\frac{1}{4} \hat{\delta}_{ij}{}^T (\hat{\Sigma}_i + \hat{\Sigma}_j)^{-1} \hat{\delta}_{ij} - \frac{1}{2} \ell n(|\hat{\Sigma}_i + \hat{\Sigma}_j|/2^k |\hat{\Sigma}_i|^{1/2} |\hat{\Sigma}_j|^{1/2}]$$

where $\delta_{ij} = B(\mu_i - u_j)$ and $\hat{\Sigma}_i = B\Sigma_i B^T$ and B is a $K \times N$ matrix of rank K of the form $B = (I_K | Z)H$ where $H = I - 2FF^T$, $||F|| = 1$. An initial guess for F is taken to be $F_o{}^T = [\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}]^T$ and the corresponding matrix $B = (I_K | Z)(I - 2F_oF_o{}^T)$ is computed. The corresponding value of

$$H_B = \frac{1}{m} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} H_B(i,j)$$

is also computed.

The steepest descent iterator is then applied to alter the value of F

$$\text{i.e.} \qquad F_{p+1} = F_p - \alpha \frac{\partial H_B}{\partial F_p} \cdot H_B$$

where α is a varying parameter and is one of the inputs to the program. $\frac{\partial H_B}{\partial F_p}$ is the partial derivative vector (derived analytically). The value of F_{p+1} is then normalized so that $\|F_{p+1}\| = 1$. The B-matrix is recomputed with the new value of F. The corresponding value of H_B is computed. This procedure is repeated (ITE - 1) number of times (8 seems to be a good value for ITE). Two points should be noted:

(1). Whether $\frac{\partial H_B}{\partial F} \approx \Theta$.

(2). Whether $\delta_H = H_B$ - BDIST (the total Bhattacharyya Distance) is sufficiently small.

The values of α and ITE (which are both inputs to this subroutine) should be altered accordingly in order to achieve the above 2 objectives.

The value of F at which the <u>minimum</u> value of H_B occurs is saved. Call it F1.

3. <u>Subroutine BHATB2</u>

This subroutine attempts to compute the minimum B-average Bhattacharyya Distance using 2 Householder transformations.

<u>USER'S INFORMATION:</u>

(Double Precision Version)

(1) In order to use this subroutine the following FORTRAN calling sequence must be given:

```
CALL BHATB2(COVAR, XMEAN, M, N, K, ITE, ALPHA)
```

where

COVAR, XMEAN, M,N,K,ITE,ALPHA

have the same meanings as in SUBROUTINE BHATB1.

(2) This subroutine reads in the value of F1 computed in the previous program (subroutine BHATB1). The data cards for F1 should have the format 5F16.8 (e.g. if F1 is 12-dimensional then F1 is punched on 3 data cards; the first 2 cards contain 5 components of F1 and the last card contains 2 components of F1).

These data cards for F1 are placed following the data cards for the covariance matrices and the mean vectors.

(3) The value of F1 that is read in is then used to compute the Householder transformation $H_1 = I - 2F1F1^T$. The covariance matrices Σ_i and the mean vectors μ_i $i = 1,\ldots,m$ are transformed into $H_1\Sigma_iH_1$ and $H_1\mu_i$.

The number of Householder transformations by which the covariance matrices Σ_i and the mean vectors μ_i have to be transformed is denoted by the variable IJ.

For subroutine BHATB2 we require one Householder transformation to obtain $H_1\Sigma_iH_1$ and $H_1\mu_i$.

The FORTRAN statements "IJ = 1" appears after the comment:

"C-----------IJ Eq. No. of Householder Transformations Required---".

OUTPUT OF SUBROUTINE BHATB2

1. The vector F1 which is the generator of the Householder transformation $H_1 = I - 2E1F1^T$.
2. Same as subroutine BHATB1.

ALGORITHM:

Here each Σ_i is replaced by $H_1 \Sigma_i H_1$ and each μ_i is replaced by $H_1 u_i$. The B matrix is then taken to be $B = (I_K | Z)(I-2FF^T)$, F = 1. An initial guess for F, $F_o^T = [\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}]$ is made and the same procedure as in subroutine BHATB1 is applied. The value of F = F2 at which the minimum value of H_B occurs is saved.

USING MORE THAN 2 HOUSEHOLDER TRANSFORMATIONS TO CONSTRUCT THE B-MATRIX:

If more than 2 Householder transformations are required to compute the transformation matrix B i.e. if $\delta_H = H_B$ - BDIST is not small enough, then subroutine BHATB2 can be modified in the following way. For the B-matrix requiring 3 Householder transformations do the following:

(1) Place the data cards containing the vector F2 (computed in the previous program) following the data cards containing F1.

(2) The statement following the comment "C... Ij Eq. NO. OF HOUSEHOLDER TRANSFORMATIONS REQUIRED ..." should be "IJ = 2"

For $J \geq 4$ Householder transformations required in computing the B-matrix:

(1) the data cards for F1,...,F(J-1) should be placed after the data cards for the covariance matrices and mean vectors;

(2) the statement "IJ = 2" should be changed to "IJ = (J-1)".

References

1. H.P. Decell, Jr. and W.G. Smiley, III, "Householder Transformations and Optimal Linear Combinations", Dept. of Mathematics, University of Houston.

2. Salma K. Marani, Masters Thesis, "Bhattacharya Distance, Householder Transformations and Dimension Reduction in Pattern Recognition".

User's Guide: DATEXT

by

William A. Coberly, University of Tulsa, University of Houston
Jack D. Tubbs, NRC Postdoctoral Fellow-JSC/MPAD
Larry Hinman, Aeronutronic Ford, University of Houston

(OS/360 Dependent)

August, 1976
Report #58
NAS-9-15000

I. INTRODUCTION :

This program reads multispectal scanner data from a Universal format tape and outputs an intermediate data set in card image format for use as an input data set in various data analysis development programs. The general capabilities are summarized as follows:

1) decode the header record of the universal format tape.

2) extract all or part of the channels on the universal format tape. (The channel numbers are relative).

3) extract a rectangular region defined by first line (ISTART), last line (ISTOP), and a line skip factor (ISKIP) and analogous column or pixel values JSTART, JSTOP, AND JSKIP. (ISKIP or JSKIP = 1, means no lines are skipped.)

4) extract and label any region defined by a non-rectangular field or fields which is a subregion of .

5) randomly select a percentage SAMPCT of the regions or , which were defined in 3 or 4.

II. INPUT PARAMETERS :

SAMKEY	-1	-only header record is decoded
	0	-deterministic sample is extracted
	1	-random sample is extracted
SAMPCT		-if SAMKEY = 1, percent of data to be randomly sampled
SEED		-if SAMKEY = 1, initial seed for random number generator. (must be a positive odd integer)
ISTART		-beginning line for sample (absolute line number)
ISTOP		-last line for sample
ISKIP		-line skip factor (if ISKIP = 1, no lines are skipped)
JSTART		-beginning pixel for sample (relative pixel number)
JSTOP		-last pixel for sample
JSKIP		-pixel skip factor (if JSKIP = 1, no pixels are skipped)

NCHOUT -number of channels to be output

NCHLST -array of relative channel numbers of NCHOUT channels to be output

NFLDS -number of non-rectangular fields to be defined (if NFLDS = 0, then the rectangular region defined by ISTART etc, is output)

FID -array containing 8 character field ID for each field

NV -array containing number of vertices for each non-rectangular field (if the field is a quadralateral, then NV = 4)

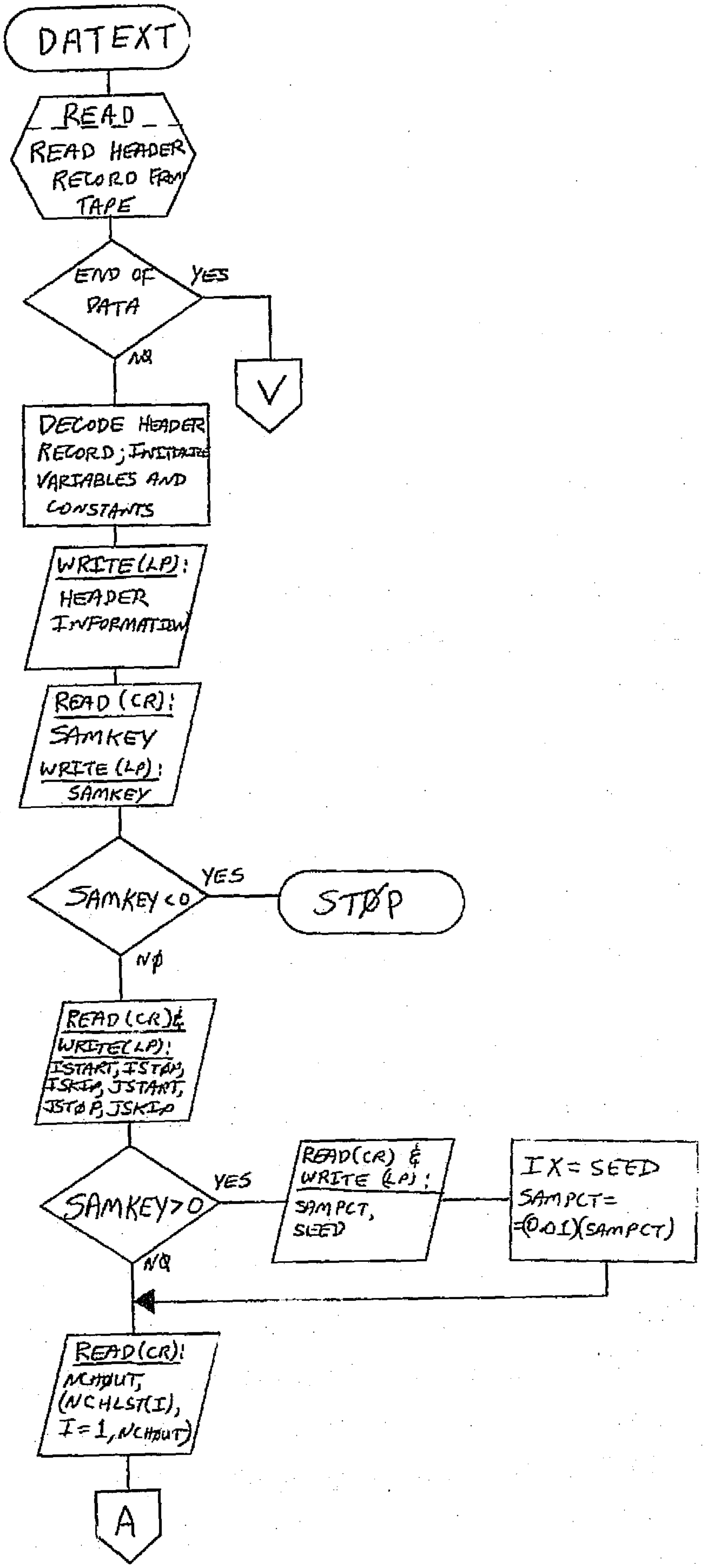
MINLIN -array containing the minimum line number for each field

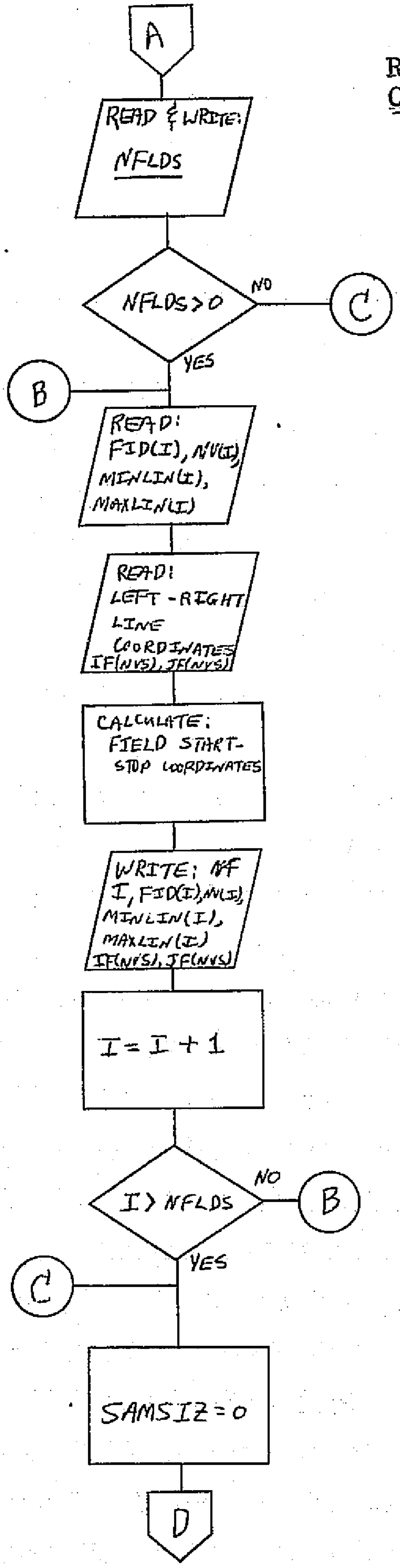
MAXLIN -array containing the maximum line number for each field

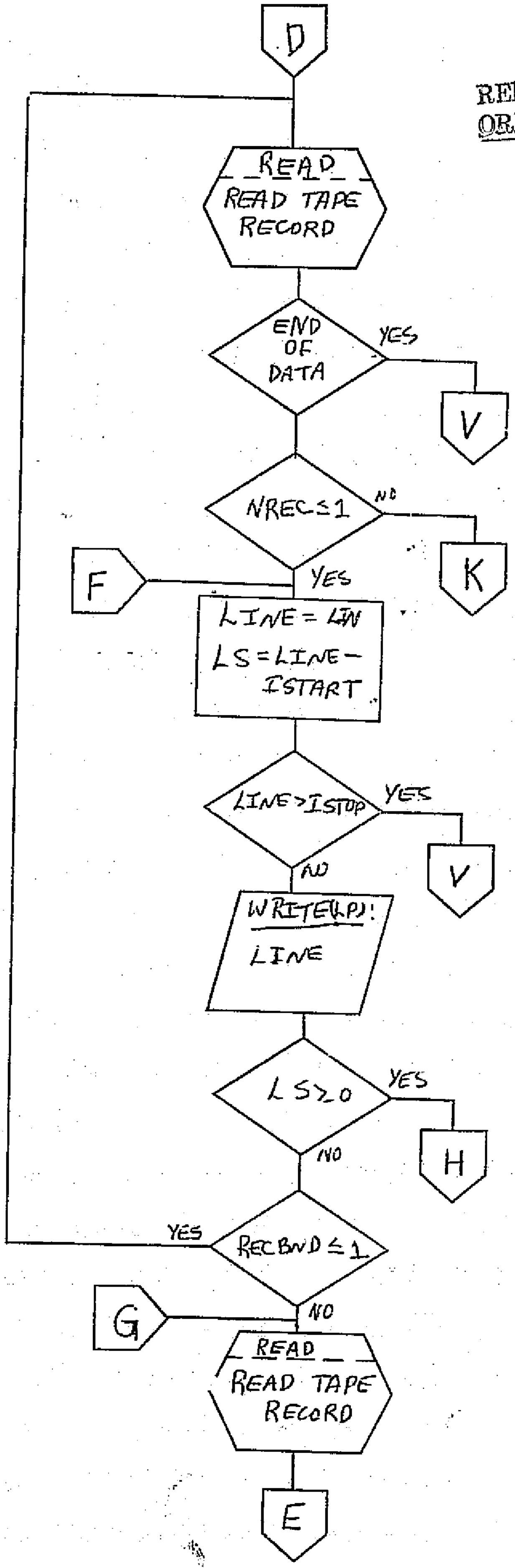
IF(J,I) -two dimensional array containing the line coordinates of the Jth vertex of the Ith field for J = 1, . . ., NV+1 (the first coordinate is repeated as the NV+1 coordinate a la ERIPS)

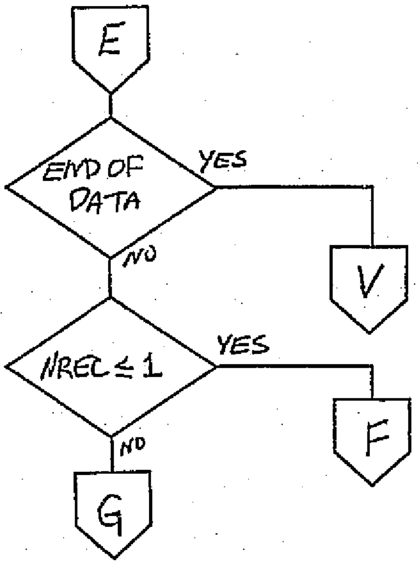
JF(J,I) -a two dimensional array containing the pixel coordinates of the Jth vertex of the Ith field for J = 1, . . ., NV+1 the first coordinate is repeated as the NV+1 coordinate a la ERIPS)

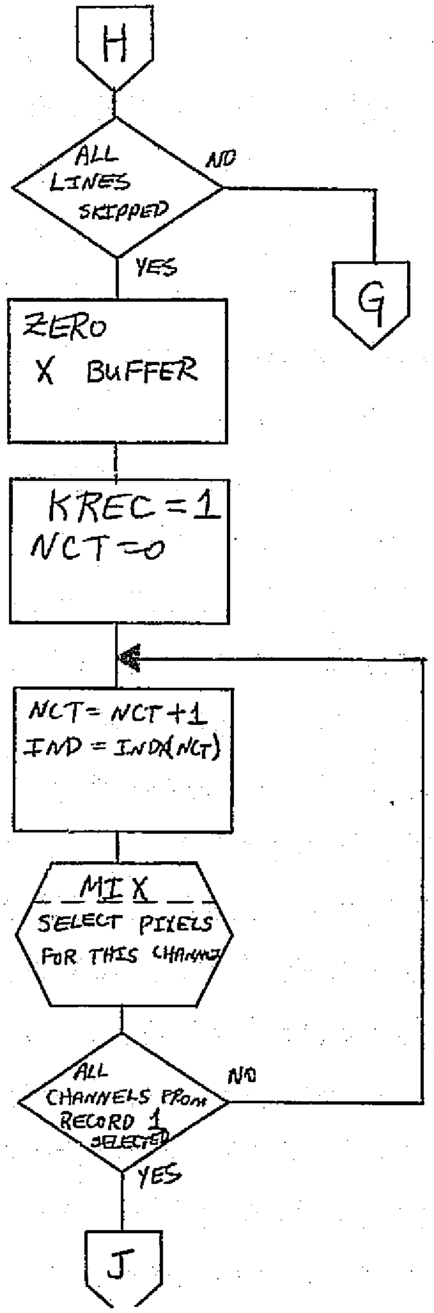
(the above vertices must be given in sequence such that the interior of the field lies to the right. See Appendix A for the ERIPS documentation for the FDLNIN routine)

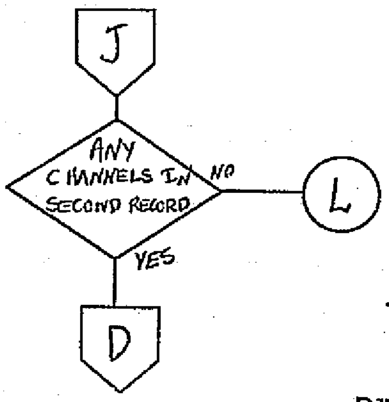

DATEXT
READ
READ HEADER RECORD FROM TAPE
END OF DATA
YES
V
NO
DECODE HEADER RECORD; INITIALIZE VARIABLES AND CONSTANTS
WRITE(LP): HEADER INFORMATION
READ (CR): SAMKEY
WRITE (LP): SAMKEY
SAMKEY < 0
YES
STØP
NØ
READ (CR) & WRITE(LP): ISTART, ISTØP, ISKIP, JSTART, JSTØP, JSKIP
SAMKEY > 0
YES
READ(CR) & WRITE (LP): SAMPCT, SEED
IX = SEED
SAMPCT = (0.01)(SAMPCT)
NO
READ(CR): NCHOUT, (NCHLST(I), I=1,NCHOUT)
A

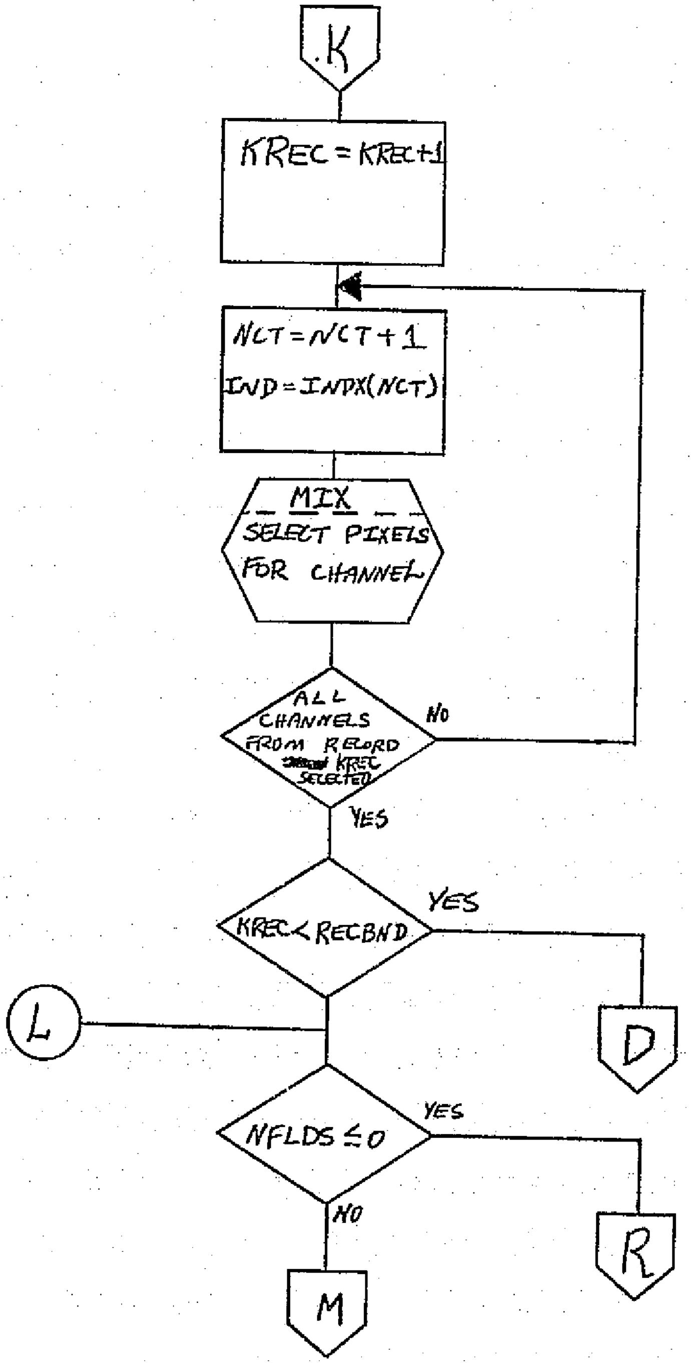

A
READ & WRITE:
NFLDS
NFLDS > 0
NO
C
YES
B
READ:
FID(I), NV(I),
MINLIN(I),
MAXLIN(I)
READ:
LEFT - RIGHT
LINE
COORDINATES
IF(NVS), JF(NVS)
CALCULATE:
FIELD START-
STOP COORDINATES
WRITE: NF
I, FID(I), NV(I),
MINLIN(I),
MAXLIN(I)
IF(NVS), JF(NVS)
I = I + 1
I > NFLDS
NO
B
YES
C
SAMSIZ = 0
D

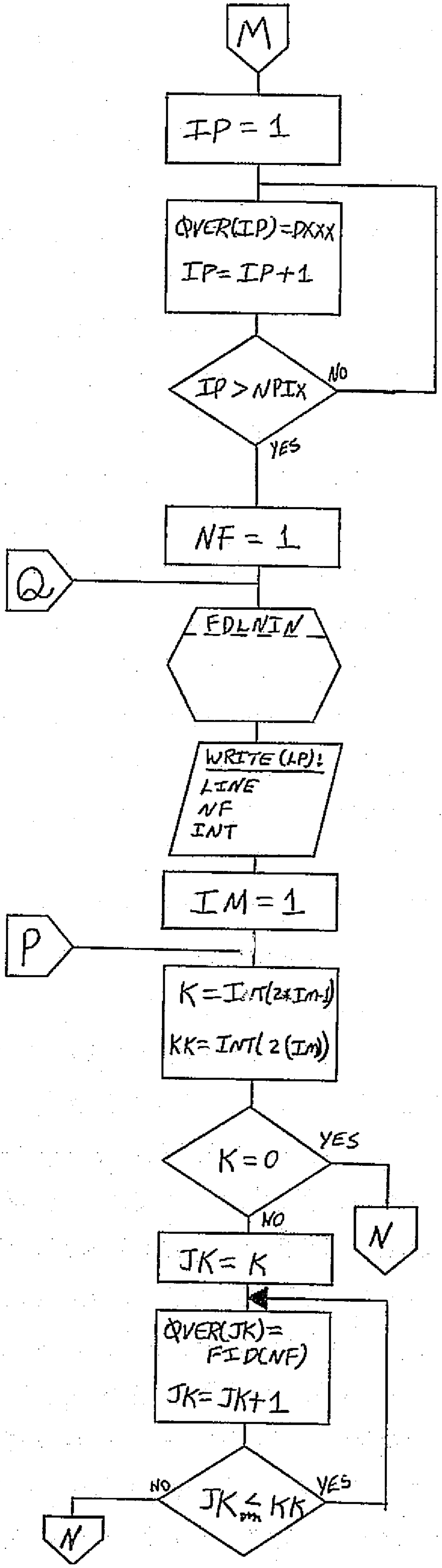

D
READ
READ TAPE RECORD
END OF DATA
YES
V
NREC≤1
NO
K
YES
F
LINE = LIN
LS = LINE − ISTART
LINE > ISTOP
YES
V
NO
WRITE(LP):
LINE
LS≥0
YES
H
NO
YES
RECBND≤1
G
NO
READ
READ TAPE RECORD
E

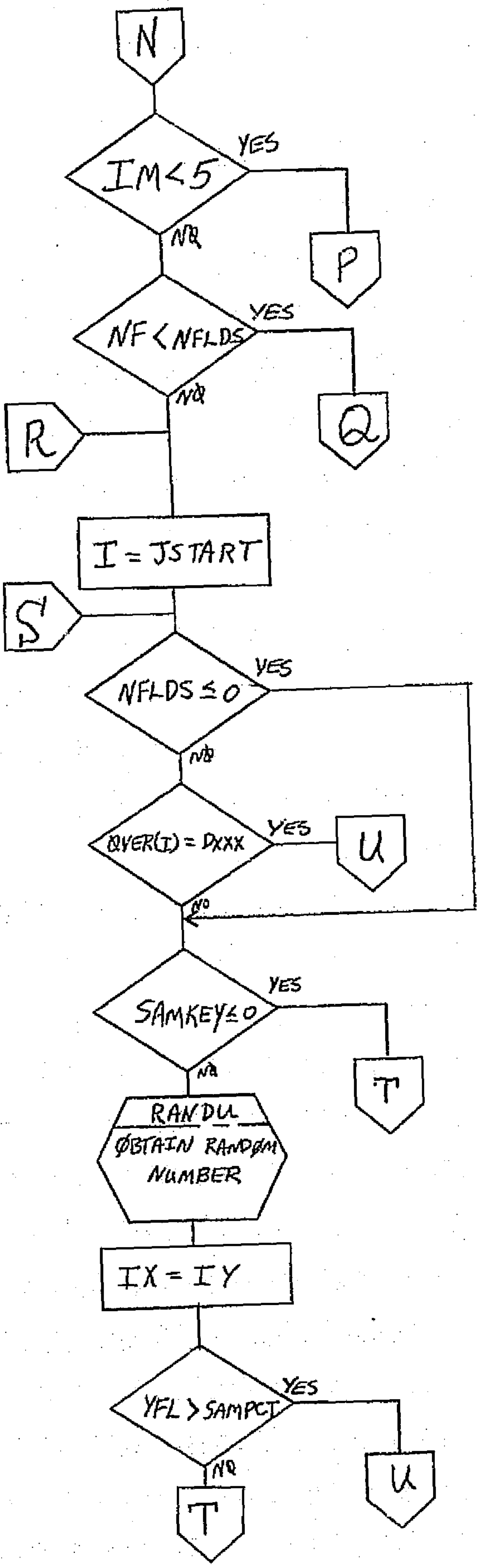

E
END OF DATA
YES
V
NO
NREC ≤ 1
YES
F
NO
G

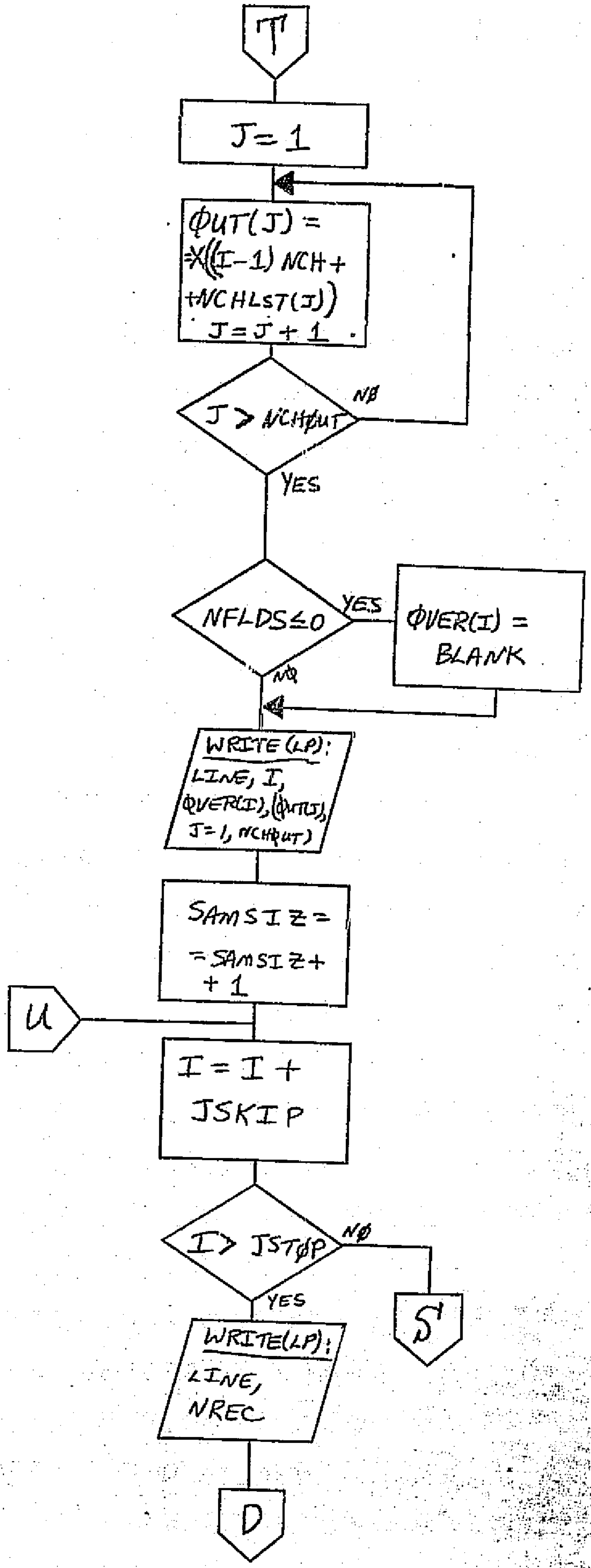

H
ALL LINES SKIPPED
NO
G
YES
ZERO X BUFFER
KREC = 1
NCT = 0
NCT = NCT + 1
IND = INDA(NCT)
MIX
SELECT PIXELS FOR THIS CHANNEL
ALL CHANNELS FROM RECORD 1 SELECTED
NO
YES
J

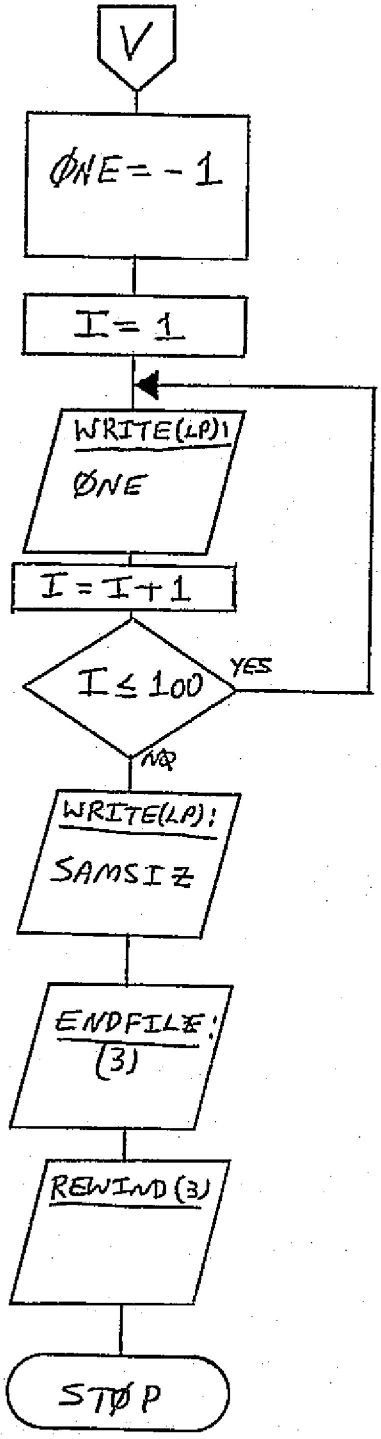

J
ANY CHANNELS IN SECOND RECORD
NO
L
YES
D

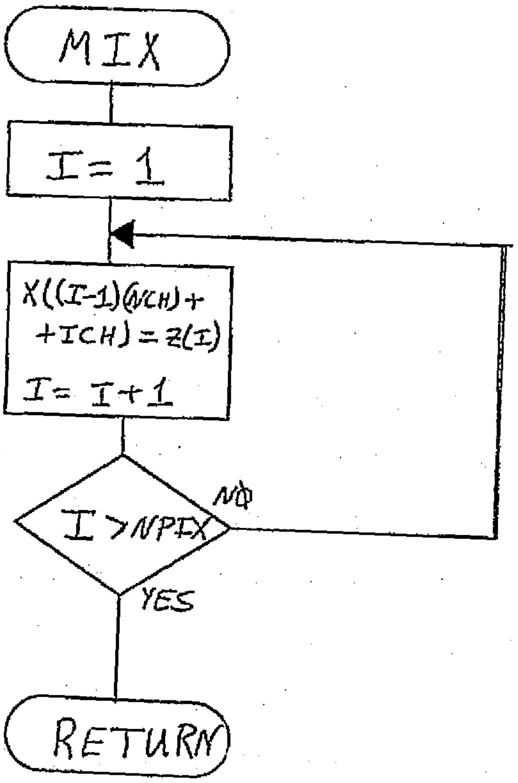

K
KREC = KREC+1
NCT = NCT + 1
IND = INDX(NCT)
MIX
SELECT PIXELS FOR CHANNEL
ALL CHANNELS FROM RECORD KREC SELECTED
NO
YES
KREC < RECBND
YES
D
L
NFLDS ≤ 0
YES
R
NO
M


M
IP = 1
QVER(IP)=DXXX
IP= IP+1
IP > NPIX
NO
YES
NF = 1
Q
FDLNIN
WRITE (LP):
LINE
NF
INT
IM = 1
P
K=INT(2*IM-1)
KK=INT(2(IM))
K=0
YES
NO
N
JK= K
QVER(JK)=
FIDCNF)
JK= JK+1
JK ≤ KK
NO
YES
N

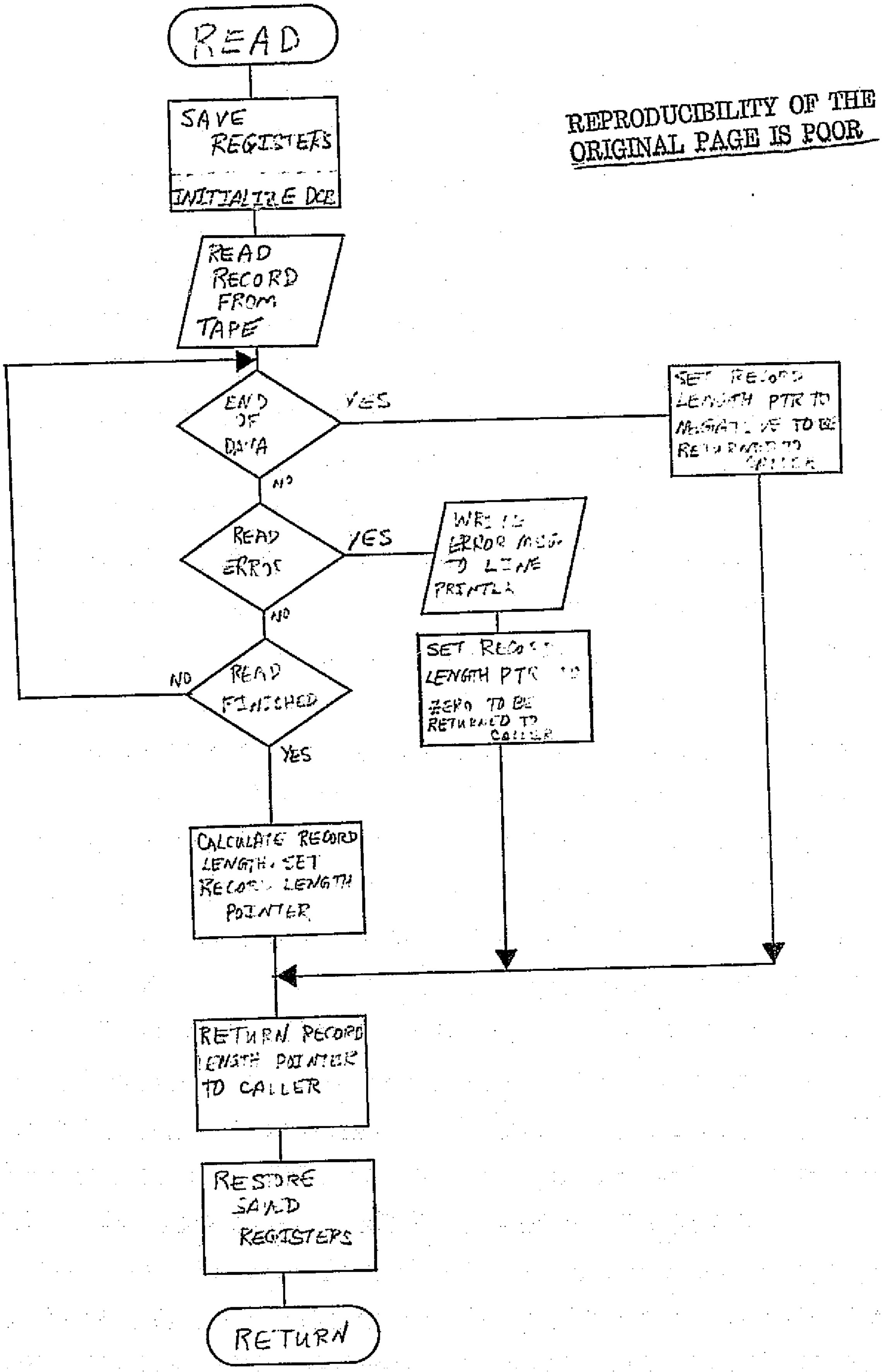

N
IM<5
YES
NØ
P
NF < NFLDS
YES
NØ
Q
R
I = JSTART
S
NFLDS ≤ 0
YES
NØ
QVER(I) = DXXX
YES
U
NO
SAMKEY≤0
YES
NØ
T
RANDU
ØBTAIN RANDØM NUMBER
IX = IY
YFL > SAMPCT
YES
NØ
T
U


T
J = 1
ØUT(J) =
=X((I-1) NCH +
+NCHLST(J))
J = J + 1
J > NCHØUT
NØ
YES
NFLDS≤0
YES
NØ
ØVER(I) =
BLANK
WRITE (LP):
LINE, I,
ØVER(I), (ØUT(J),
J=1, NCHØUT)
SAMSIZ =
= SAMSIZ +
+ 1
U
I = I +
JSKIP
I > JSTØP
NØ
YES
S
WRITE(LP):
LINE,
NREC
D


V
ØNE = -1
I = 1
WRITE(LP)1
ØNE
I = I+1
I ≤ 100
YES
NO
WRITE(LP):
SAMSIZ
ENDFILE:
(3)
REWIND (3)
STØP



REPRODUCIBILITY OF THE
ORIGINAL PAGE IS POOR


READ
SAVE REGISTERS
INITIALIZE DCB
READ RECORD FROM TAPE
END OF DATA
YES
SET RECORD LENGTH PTR TO NEGATIVE TO BE RETURNED TO CALLER
NO
READ ERROR
YES
WRITE ERROR MSG TO LINE PRINTER
SET RECORD LENGTH PTR TO ZERO TO BE RETURNED TO CALLER
NO
READ FINISHED
NO
YES
CALCULATE RECORD LENGTH, SET RECORD LENGTH POINTER
RETURN RECORD LENGTH POINTER TO CALLER
RESTORE SAVED REGISTERS
RETURN

IV. INPUT FORMAT FOR PARAMETERS

```
REQ:    SAMKEY        (10X, I10)
        ISTART
        ISTOP
        ISKIP
REQ:    JSTART        (10X, I10)
        JSTOP
        JSKIP

OPT:    SAMPCT        (10X, F10.0)
        SEED          (10X, I10)

REQ:    NCHOUT        (10X, I10)
        NCHLST        (10X, 16I2)

REQ:    NFLDS         (10X, I10)
      for I = 1, ....,NFLDS  (if NFLDS  0)
        FID(I)
        NV(I)
OPT:    MINLIN(I)     (A8, 2X, 3I5)
        MAXLIN(I)
        IF(J,I)       (11I5)
        JF(J,I)       (11I5)
```

V. FORMAT OF INPUT DATA SET

The Input Data Set is read from Fortran unit 1 (FT01F001) by the READ routine. The Input Data Set has the format of a Universal Format Image Data Tape described in NASA Earth Resources Data Format Control Book .(TR-543).

VI. FORMAT OF OUTPUT DATA SET

For each NCH dimensional pixel (X(I), I = 1, . . . , NCH) selected for output, the following record (80 bytes) is written onto Fortran unit 3 (FT03F001).

```
LINE number
PIXEL NUMBER
FID (if not applicable a blank is written)
X(NCHLST (1))
X(NCHLST (2))
    .
    .
    .
X(NCHLST (NCHOUT))
```

The format is (2I4, A8, 16I4). The logical record length is 80 bytes and the BLKSIZE is determined by the JCL card defining Fortran unit 3 (FT03F001).

VII. SUBROUTINES

MIX -arranges data by pixel rather than by channel
RANDU -random number generator (IBM SSP)
FDLNIN -determines intersection of a non-rectangular files for a scan line. (Fortran version of PL1 ERIPS utility routine)
READ -assembly language (360 OS) binary read routine (Hinman)

APPENDIX A

```
                 COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=50,SIZE=0000K,
                                    SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,NOID,NOXREF
      ISN 0002                INTEGER SEED
      ISN 0003                INTEGER BEGVID,RECLNG,RECBND,ANCLNG,INDX(16),XXXX(2500),
                             *     ONE,SAMKEY,SAMSIZ,NCHLST(16)
      ISN 0004                LOGICAL*1 Z(3060),Z2(2),X(10000),OUT(16)
      ISN 0005                INTEGER*2 ZINT2,NREC,LIN ,XX(5000)
      ISN 0006                DOUBLE PRECISION OVER,BLANK,DXXX,FID
      ISN 0007                DIMENSION FID(50),NV(50),MINLIN(50),MAXLIN(50),IF(12,50),
                             *        JF(12,50),INT(11),OVER(1000)
      ISN 0008                DATA BLANK/'        '/
      ISN 0009                DATA DXXX/'$$$$$$$$'/
      ISN 0010                DATA OUT/16*' '/,SAMSIZ/0/,LIN /0/
      ISN 0011                EQUIVALENCE (ZINT2,Z2(1)),(NREC,Z(1)),(LIN ,Z(71)),
                             *            (X(1),XX(1)),(X(1),XXXX(1))
                      C
                      C  READ HEADER RECORD AND DECODE THE FOLLOWING VARIABLES
                      C
                      C          NCH      -  NUMBER OF CHANNELS
                      C          NCH1     -  NUMBER OF CHANNELS ON FIRST RECORD OF BAND
                      C          NCH2     -  NUMBER OF CHANNELS ON OTHER RECORDS OF BAND
                      C          RECLNG   -  RECORD LENGTH
                      C          RECBND   -  NUMBER OF RECORDS PER BAND
                      C          NPIX     -  NUMBER OF PIXELS PER CHANNEL PER BAND
                      C          ANCLNG   -  LENGTH OF ANCILLARY BLOCK ON FIRST RECORD OF BAND
                      C          BEGVID   -  BEGIN VIDEO BYTE WITHIN SCAN
                      C          INDX     -  ARRAY OF INDICIES FOR BEGINNING BYTE OF EACH CHANNEL
                      C                      WITHIN TH APPROPRIATE RECORD
                      C
      ISN 0012                CALL READ(Z,LRCLG)
      ISN 0013                IF(LRCLG.LT.0) GO TO 999
      ISN 0015                ZINT2=0
      ISN 0016                Z2(2)=Z(90)
      ISN 0017                NCH=ZINT2
                      C
      ISN 0018                Z2(1)=Z(92)
      ISN 0019                Z2(2)=Z(93)
      ISN 0020                BEGVID=ZINT2
                      C
      ISN 0021                Z2(1)=Z(96)
      ISN 0022                Z2(2)=Z(97)
      ISN 0023                NPIX=ZINT2
                      C
      ISN 0024                Z2(1)=Z(100)
      ISN 0025                Z2(2)=Z(101)
      ISN 0026                RECLNG=ZINT2
                      C
      ISN 0027        10      ZINT2=0
```

IIAXPFLI-ICAXPFLI

REFERENCES

1. Program Name - FDLNINT
2. Programmer - R. J. Decker
3. Language - PL/1
4. LINKEDIT Attributes - NCAL
5. Inputs - Scan Line Number
6. Outputs - Intercepts (pixel numbers) of scan line and field sides
7. Special Items - Calling sequence:

   CALL FDLNINT(P,L);

   where P = pointer to field definition table

   L = 11 element vector declared

   FIXED BIN (15)

   L(11) should be loaded with the scan line number

   On return, the L vector will contain the ordered pixel intercepts. (e.g., a return of $\boxed{5}\boxed{7}\boxed{12}\boxed{20}\boxed{0 \longrightarrow 0}$ indicates pixels 5 through 7 and pixels 12 through 20 are contained in the field.)

FUNCTIONAL DESCRIPTION

This subroutine will return the pixel numbers of those pixels on a given line that are contained within the boundaries of a field.

DETAILED LOGIC DESCRIPTION

IIAXPFLI examines the number of vertices of the input field to determine if the field is a line-field or a polygon. If the input field is a line-field, then the intercepts are determined as follows:

The intercept of the line-field and L-0.5 is calculated as $P = (X_2-X_1)(L-0.5-Y_1) | (Y_2-Y_1) + X_1$. This calculation determines the projection of the intercept of the line-field and L+0.5 is calculated as $P = (X_2-X_1)(L+0.5-Y_1) | (Y_2-Y_1) + X_1$. This calculation determines the projection of the intercept of L+0.5 onto L. These projections are examined to determine which is the left one (P_L) and which is the right one (P_R). P_L is set to the integral value of P_L+0.5 and P_R is set to the integral value of P_R + 0.4999.

Approval	Approval
D. A. King 8/26/75	

If the field is a polygon, then IIAXPFLI finds the pixel intercepts of a scan line and the sides of the input field.

There are three distinct cases and each is handled separately; (1) the scan line intersects a side but not at the endpoints (i.e., vertices), (2) the scan line intersects a vertex that is not an end of a horizontal line, and (3) the scan line is concurrent with a horizontal side of the field.

FUNCTIONAL FLOWCHART

See Figure 1.

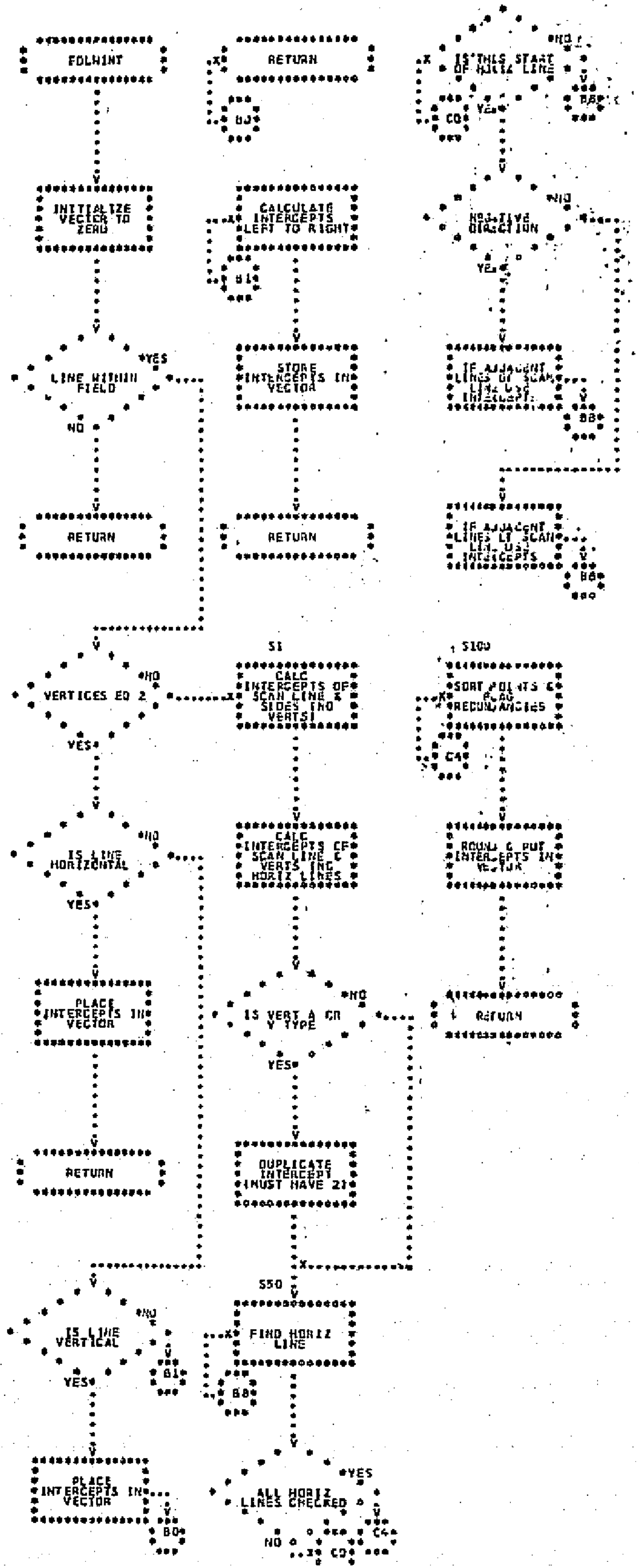

FOLNINT
INITIALIZE VECTOR TO ZERO
LINE WITHIN FIELD
YES
NO
RETURN
VERTICES EQ 2
NO
YES
IS LINE HORIZONTAL
NO
YES
PLACE INTERCEPTS IN VECTOR
RETURN
IS LINE VERTICAL
NO
YES
B1
PLACE INTERCEPTS IN VECTOR
B0
RETURN
B3
CALCULATE INTERCEPTS LEFT TO RIGHT
B1
STORE INTERCEPTS IN VECTOR
RETURN
S1
CALC INTERCEPTS OF SCAN LINE & SIDES (NO VERTS)
CALC INTERCEPTS OF SCAN LINE & VERTS INC HORIZ LINES
IS VERT A OR V TYPE
NO
YES
DUPLICATE INTERCEPT (MUST HAVE 2)
S50
FIND HORIZ LINE
B8
ALL HORIZ LINES CHECKED
YES
NO
C4
C0
IS THIS START OF HORIZ LINE
NO
B8
C0
YES
NEGATIVE DIRECTION
NO
YES
IF ADJACENT LINES GT SCAN LINE USE INTERCEPTS
B8
IF ADJACENT LINES LT SCAN LINE USE INTERCEPTS
B8
S100
SORT POINTS & FLAG REDUNDANCIES
C4
ROUND & PUT INTERCEPTS IN VECTOR
RETURN

FIGURE 1. FLOWCHART

APPENDIX B

```
ISN 0028              Z2(2)=Z(102)
ISN 0029              NCH2=ZINT2
              C
ISN 0030              ZINT2=0
ISN 0031              Z2(2)=Z(104)
ISN 0032              RECBND=ZINT2
              C
ISN 0033              Z2(1)=Z(105)
ISN 0034              Z2(2)=Z(106)
ISN 0035              ANCLNG=ZINT2
              C
ISN 0036              Z2(1)=Z(1785)
ISN 0037              Z2(2)=Z(1786)
ISN 0038              NCH1=ZINT2
              C
ISN 0039              ICT=0
ISN 0040              DO 20 I=1,NCH1
ISN 0041              ICT=ICT+1
ISN 0042         20   INDX(I)=ANCLNG+2+(I-1)*NPIX+1
ISN 0043              IF(RECBND.EQ.1) GO TO 40
ISN 0045              DO 30 I=2,RECBND
ISN 0046              DO 30 J=1,NCH2
ISN 0047              ICT=ICT+1
ISN 0048         30   INDX(ICT)=2+(J-1)*NPIX+1
ISN 0049         40   WRITE(6,200) NCH,NPIX,RECLNG,NCH1,NCH2,RECBND,ANCLNG,BEGVID
ISN 0050              WRITE(6,201) (I,INDX(I),I=1,NCH)
ISN 0051              WRITE(6,202) Z
ISN 0052          200 FORMAT(1H1 ,' NCH    = ',I6,/,
                     *          ' NPIX   = ',I6,/,
                     *          ' RECLNG = ',I6,/,
                     *          ' NCH1   = ',I6,/,
                     *          ' NCH2   = ',I6,/,
                     *          ' RECBND = ',I6,/,
                     *          ' ANCLNG = ',I6,/,
                     *          ' BEGVID = ',I6)
ISN 0053          201 FORMAT(1H ,'INDX(',I2,') = ',I8)
ISN 0054          202 FORMAT(100(/,5(2X,10Z2)))
              C
              C
              C
              C  READ SAMPLING PARAMETERS
              C
              C         SAMKEY = -1 - ONLY HEADER RECORD IS DECODED
              C                   0 - DETERMINISTIC SAMPLE
              C                   1 - RANDOM SAMPLE
              C         SAMPCT      - PERCENTAGE OF DATA TO BE SAMPLED RANDOMLY
              C         SEED        - SEED FOR RANDOM NUMBER GENERATOR
              C         ISTART      - BEGIN LINE FOR SAMPLE (ABSOLUTE LINE NUMBER)
```

```
               C          ISTOP        - LAST LINE FOR SAMPLE
               C          ISKIP        - LINE SKIP FACTOR (IF ISKIP=0, NO LINES ARE SKIPPED)
               C          JSTART       - BEGIN PIXEL FOR SAMPLE (RELATIVE PIXEL NUMBER)
               C          JSTOP        - LAST PIXEL FOR SAMPLE
               C          JSKIP        - PIXEL SKIP FACTOR (IF JSKIP=0, NO PIXELS ARE SKIPPED)
               C          NCHOUT       - NUMBER OF CHANNELS TO BE OUTPUT
               C          NCHLST       - ARRAY OF CHANNEL IDS TO BE OUTPUT (RELATIVE)
               C
               C
               C    ------------------------------------------------------------------
               C    ------------------------------------------------------------------
               C
ISN 0055                  READ(5,1000) SAMKEY
ISN 0056                  WRITE(6,1007) SAMKEY
ISN 0057                  IF(SAMKEY)41,42,42
ISN 0058               41 STOP
ISN 0059               42 READ(5,1000) ISTART,ISTOP,ISKIP,JSTART,JSTOP,JSKIP
ISN 0060                  WRITE(6,1008)ISTART,ISTOP,ISKIP,JSTART,JSTOP,JSKIP
ISN 0061                  IF(SAMKEY) 44,44,43
ISN 0062               43 READ(5,1002) SAMPCT,SEED
ISN 0063                  IX=SEED
ISN 0064                  WRITE(6,1009) SAMPCT,SEED
ISN 0065                  SAMPCT=SAMPCT/100.
ISN 0066               44 READ(5,1000) NCHOUT
ISN 0067                  READ(5,1003) (NCHLST(I),I=1,NCHOUT)
ISN 0068             1000 FORMAT(10X,I10)
ISN 0069             1002 FORMAT(10X,F10.0,/,10X,I10)
ISN 0070             1003 FORMAT(10X,16I2)
ISN 0071             1007 FORMAT(1H1,'SAMKEY     = ',I10)
ISN 0072             1008 FORMAT(1H ,'ISTART     = ',I10,/,
                         *           ' ISTOP      = ',I10,/,
                         *           ' ISKIP      = ',I10,/,
                         *           ' JSTART     = ',I10,/,
                         *           ' JSTOP      = ',I10,/,
                         *           ' JSKIP      = ',I10)
ISN 0073             1009 FORMAT(' SAMPCT     = ',F10.2/,' SEED       = ',I10 )
ISN 0074             1010 FORMAT(' NCHOUT     = ',I10)
ISN 0075             1011 FORMAT(' NCHLST     = ',16I5)
ISN 0076                  READ(5,2000) NFLDS
ISN 0077                  WRITE(6,2001) NFLDS
ISN 0078             2000 FORMAT(10X,I10)
ISN 0079             2001 FORMAT(1H ,'NFLDS      = ',I10)
ISN 0080                  IF(NFLDS) 440,440,438
ISN 0081              438 DO 439 NF=1,NFLDS
ISN 0082                  READ(5,2002)  FID(NF),NV(NF),MINLIN(NF),MAXLIN(NF)
ISN 0083                  NVS=NV(NF)   + 1
ISN 0084                  READ(5,2003)  (IF(J,NF),J=1,NVS)
ISN 0085                  READ(5,2003)  (JF(J,NF),J=1,NVS)
ISN 0086                  DO 605 II=1,NVS
ISN 0087              J=NVS-II+1
ISN 0088              J1=J+1
ISN 0089              IF(J1,NF)=IF(J,NF)
ISN 0090          605 JF(J1,NF)=JF(J,NF)
ISN 0091              IF(1,NF)=IF(NVS   ,NF)
ISN 0092              JF(1,NF)=JF(NVS   ,NF)
ISN 0093              IF(NVS+2,NF)=IF(3,NF)
ISN 0094              JF(NVS+2,NF)=JF(3,NF)
ISN 0095              NV3=NVS+2
ISN 0096              WRITE(6,2004) NF
ISN 0097              WRITE(6,2005) FID(NF),NV(NF),MINLIN(NF),MAXLIN(NF)
ISN 0098              WRITE(6,2006) (IF(J,NF),J=1,NV3)
ISN 0099          439 WRITE(6,2007) (JF(J,NF),J=1,NV3)
ISN 0100         2002 FORMAT(A8,2X,3I5)
ISN 0101         2003 FORMAT(11I5)
ISN 0102         2004 FORMAT(5X,'FIELD = ',I10)
ISN 0103         2005 FORMAT(5X,'FIELD ID = ''',A8,'''',/,
                *               5X,'NV     = ',I10,/,
                *               5X,'MINLIN = ',I10,/,
                *               5X,'MAXLIN = ',I10)
ISN 0104         2006 FORMAT(5X,'LINE   = ',12I5)
ISN 0105         2007 FORMAT(5X,'PIXEL  = ',12I5)
ISN 0106          440 CONTINUE
                C
                C    --------------------------------------------------------------
                C
                C    WRITE  DATA INTO CCB FORMAT
                C
                C    --------------------------------------------------------------
                C
ISN 0107              SAMSIZ = 0
ISN 0108           50 CALL READ(Z,LRCLG)
ISN 0109              IF(LRCLG.LT.0) GC TO 999
ISN 0111              IF(NREC-1) 55,55,60
                C
                C
ISN 0112           55 LINE=LIN
ISN 0113              IF(LINE.GT.ISTCP) GO TO 999
ISN 0115              LS=LINE-ISTART
ISN 0116              WRITE(6,307) LINE
ISN 0117          307 FORMAT(20X,I10)
ISN 0118              IF(LS.GE.0) GO TC 552
ISN 0120              IF(RECBND.LE.1) GO TO 50
                C
                C
ISN 0122          550 CALL READ(Z,LRCLG)
ISN 0123              IF(LRCLG.LT.0) GO TC 999
ISN 0125              IF(NREC-1) 55,55,550
```

REPRODUCIBILITY OF THE
ORIGINAL PAGE IS POOR

```
                C
ISN 0126           552 LSM=LS/ISKIP*ISKIP-LS
ISN 0127               IF(LSM.NE.0) GO TO 550
                C
ISN 0129           555 DO 56 I=1,2500
ISN 0130            56 XXXX(I)=0
                C
ISN 0131               KREC=1
ISN 0132               NCT=0
ISN 0133               DO 57 I=1,NCH1
ISN 0134               NCT=NCT+1
ISN 0135               IND=INDX(NCT)
ISN 0136            57 CALL MIX(Z(IND),NCT,NPIX,X,NCH)
ISN 0137               IF(NCH2.EQ.0) GO TO 7329
ISN 0139               GO TO 50
                C
                C
ISN 0140            60 KREC=KREC+1
ISN 0141               DO 61 I=1,NCH2
ISN 0142               NCT=NCT+1
ISN 0143               IND=INDX(NCT)
ISN 0144            61 CALL MIX(Z(IND),NCT,NPIX,X,NCH)
                C
ISN 0145               IF(KREC.LT.RECBND) GO TO 50
                C
                C   WRITE DATA TO OUTPUT DATA SET
                C
ISN 0147          7329 CONTINUE
ISN 0148               IF(NFLDS) 675,675,659
                C
ISN 0149           659 DO 660 IP=1,NPIX
ISN 0150           660 OVER(IP)=DXXX
ISN 0151               DO 665 NF=1,NFLDS
ISN 0152               CALL FDLNIN(LINE,NV(NF),IF(1,NF),JF(1,NF),INT,MINLIN(NF),
                     *                       MAXLIN(NF))
                C
                C      WRITE(6,6660) LINE,NF,INT
                C6660 FORMAT(30X,2I10,11I5)
                C
ISN 0153               DO 668 IM=1,5
ISN 0154               K=INT(2*IM-1)
ISN 0155               KK=INT(2*IM)
ISN 0156               IF(K.EQ.0) GO TO 670
ISN 0158               DO 669 JK=K,KK
ISN 0159           669 OVER(JK)=FID(NF)
ISN 0160           670 CONTINUE
ISN 0161           668 CONTINUE
ISN 0162           665 CONTINUE
```

```
                C
                C
ISN 0163         675 CONTINUE
ISN 0164             DO 80 I=JSTART,JSTOP,JSKIP
                C
ISN 0165             IF(NFLDS.LE.0) GO TO 680
                C
ISN 0167             IF(OVER(I).EQ.DXXX) GO TO 80
                C
ISN 0169         680 CONTINUE
ISN 0170             IF(SAMKEY) 75,75,70
ISN 0171          70 CALL RANDU(IX,IY,YFL)
ISN 0172             IX=IY
ISN 0173             IF(YFL.GT.SAMPCT) GO TO 80
                C
ISN 0175          75 DO 78 J=1,NCHOUT
ISN 0176          78 OUT(J)=X((I    -1)*NCH + NCHLST(J))
                C
ISN 0177             IF(NFLDS.LE.0) OVER(I)=BLANK
ISN 0179             WRITE(3,300) LINE,I,OVER(I),(OUT(J),J=1,NCHOUT)
ISN 0180         300 FORMAT(2I4,A8    ,16I4)
ISN 0181             SAMSIZ  = SAMSIZ + 1
ISN 0182          80 CONTINUE
                C
                C
ISN 0183             WRITE(6,301) LINE,NREC
ISN 0184         301 FORMAT(2X,2I5)
ISN 0185             GO TO 50
                C
ISN 0186         999 ONE=-1
ISN 0187             DO 90 I=1,100
ISN 0188          90 WRITE(3,400) ONE
ISN 0189         400 FORMAT(I4,76X)
ISN 0190             WRITE(6,405) SAMSIZ
ISN 0191         405 FORMAT(' SAMSIZE  = ',I10)
ISN 0192             ENDFILE 3
ISN 0193             REWIND 3
ISN 0194             STOP
ISN 0195             END
```

REPRODUCIBILITY OF THE ORIGINAL PAGE IS POOR

```
LEVEL  21.8 ( JUN 74 )                              OS/360  FORTRAN H                                      DATE

           COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=50,SIZE=0000K,
                              SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,NOID,NOXREF
    ISN 0002                  SUBROUTINE MIX(Z,ICH,NPIX,X,NCH)
    ISN 0003                  LOGICAL*1 Z(1),X(1)
    ISN 0004                  DO 1 I=1,NPIX
    ISN 0005                  LOC=(I-1)*NCH+ICH
    ISN 0006                1 X(LOC)=Z(I)
    ISN 0007                  RETURN
    ISN 0008                  END
```

```
LEVEL 21.8 ( JUN 74 )                               OS/360  FORTRAN H                                         DATE

          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=50,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,NOID,NOXREF
     ISN 0002                SUBROUTINE RANDU(IX,IY,YFL)
     ISN 0003                IY=IX*65539
     ISN 0004                IF(IY)5,6,6
     ISN 0005              5 IY=IY+2147483647 + 1
     ISN 0006              6 YFL=IY
     ISN 0007                YFL=YFL*.4656613E-9
     ISN 0008                RETURN
     ISN 0009                END
```

```
          COMPILER OPTIONS - NAME=  MAIN,OPT=02,LINECNT=50,SIZE=0000K,
                             SOURCE,EBCDIC,NOLIST,NODECK,LOAD,MAP,NOEDIT,NOID,NOXREF
   ISN 0002                SUBROUTINE FDLNIN (L,NV,Y,X,INT,MINLIN,MAXLIN)
   ISN 0003                INTEGER Y(12),X(12),INT(11),CUR
   ISN 0004                REAL PTS(10)
   ISN 0005                NV1=NV+1
   ISN 0006                DO 10 I=1,10
   ISN 0007             10 INT(I)=0
   ISN 0008                IF(L.LT.MINLIN.OR.L.GT.MAXLIN) RETURN
   ISN 0010          1     DO 15 I=1,10
   ISN 0011         15     PTS(I)=0.
   ISN 0012                IPT=0
   ISN 0013                DO 12 I=2,NV1
   ISN 0014                IF(.NOT.(L.GT.MIN0(Y(I),Y(I+1)).AND.L.LT.MAX0(Y(I),Y(I+1))))
                          *        GO TO 12
   ISN 0016                IPT=IPT+1
   ISN 0017                PTS(IPT)=(FLOAT((L-Y(I))*(X(I+1)-X(I))))/
                          *     (FLOAT(Y(I+1)-Y(I)))+FLOAT(X(I))
   ISN 0018         12     CONTINUE
   ISN 0019                DO 14 I=2,NV1
   ISN 0020                IF(.NOT.(L.EQ.Y(I).AND.L.NE.Y(I-1).AND.L.NE.Y(I+1)))GO TO 14
   ISN 0022                IPT=IPT+1
   ISN 0023                PTS(IPT)=FLOAT(X(I))
   ISN 0024                IF(.NOT.((L.LT.Y(I-1).AND.L.LT.Y(I+1)).OR.(L.GT.Y(I-1).AND.
                          *        L.GT.Y(I+1))))GO TO 14
   ISN 0026                IPT=IPT+1
   ISN 0027                PTS(IPT)=PTS(IPT-1)
   ISN 0028         14     CONTINUE
   ISN 0029                J=1
   ISN 0030         50     J=J+1
   ISN 0031                IF(J.GT.NV) GO TO 100
   ISN 0033                IF(Y(J).NE.L) GO TO 50
   ISN 0035                IF(Y(J+1).NE.L) GO TO 50
   ISN 0037                IF(X(J+1).LT.X(J)) GO TO 16
   ISN 0039                IF(Y(J-1).GE.L) GO TO 20
   ISN 0041                IPT=IPT+1
   ISN 0042                PTS(IPT)=X(J)
   ISN 0043         20     IF(Y(J+2).GE.L) GO TO 21
   ISN 0045                IPT=IPT+1
   ISN 0046                PTS(IPT)=X(J+1)
   ISN 0047         21     J=J+1
   ISN 0048                GO TO 50
   ISN 0049         16     IF(Y(J-1).LE.L) GO TO 17
   ISN 0051                IPT=IPT+1
   ISN 0052                PTS(IPT)=X(J)
   ISN 0053         17     IF(Y(J+2).LE.L) GO TO 18
   ISN 0055                IPT=IPT+1
   ISN 0056                PTS(IPT)=X(J+1)
```

```
ISN 0057        18      J=J+1
ISN 0058                GO TO 50
ISN 0059       100      CONTINUE
ISN 0060                IPT1=IPT-1
ISN 0061                DO 30 K=1,IPT1
ISN 0062                K1=K+1
ISN 0063                DO 30 I=K1,IPT
ISN 0064                IF(PTS(I).GE.PTS(K)) GO TO 30
ISN 0066                DUM=PTS(I)
ISN 0067                PTS(I)=PTS(K)
ISN 0068                PTS(K)=DUM
ISN 0069        30      CONTINUE
ISN 0070                IF(IPT.EQ.2) GO TO 103
ISN 0072                IPT2=IPT-2
ISN 0073                DO 40 I=2,IPT2,2
ISN 0074                IF(PTS(I).NE.PTS(I+1)) GO TO 40
ISN 0076                PTS(I)=-1
ISN 0077                PTS(I+1)=-1
ISN 0078        40      CONTINUE
ISN 0079       103      K=0
ISN 0080                DO 110 I=1,IPT,2
ISN 0081                IF(PTS(I).EQ.-1) GO TO 105
ISN 0083                K=K+1
ISN 0084                INT(K)=PTS(I)+.499
ISN 0085       105      CONTINUE
ISN 0086                IF(PTS(I+1).EQ.-1) GO TO 110
ISN 0088                K=K+1
ISN 0089                INT(K)=PTS(I+1) + .500
ISN 0090       110      CONTINUE
ISN 0091       120      IPT2=IPT-2
ISN 0092                DO 60 I=2,IPT2,2
ISN 0093                IF(INT(I).NE. INT(I+1)) GO TO 60
ISN 0095                INT(I)=0
ISN 0096                INT(I+1)=0
ISN 0097        60      CONTINUE
ISN 0098                IPT1=IPT-1
ISN 0099                DO 70 K=1,IPT1
ISN 0100                K1=K+1
ISN 0101                DO 65 I=K1,IPT
ISN 0102                IF(.NOT.(INT(I).NE.0.AND.INT(I).LT.INT(K).OR.INT(K).EQ.0))GO TO 65
ISN 0104                DUM=INT(I)
ISN 0105                INT(I)=INT(K)
ISN 0106                INT(K)=DUM
ISN 0107        65      CONTINUE
ISN 0108        70      CONTINUE
ISN 0109                RETURN
ISN 0110                END
```

```
*READ
*
*         READ ERIPS LOG TAPE
*
*
*         LARRY HINMAN, EARTH RESOURCES PROGRAM OFFICE, PHILCO-FORD
*
*         CALL RDLOGT(BUFADR, RCDLNG)
*
READ      CSECT
          SAVE  (14,12),T,*                 SAVE REGS
          LR    2,15                        SET BASE
          USING READ,2                      ASM BASE
          LA    3,SAVE                      NEW SAVE AREA ADDR
          ST    3,8(13)                     LSA
          ST    13,4(3)                     HSA
          LR    13,3                        SAVE AREA ADDR
*
          L     3,0(1)                      ADDR OF BUFFER
          L     5,4(1)                      ADDR OF WORD FOR RECORD LNGTH
          LA    7,TAPEDCB                   ADDR OF DCB
          USING IHADCB,7                    SECOND BASE
          TM    DCBOFLGS,X'10'              TEST FOR OPEN
          BO    INPUT                       DCB IS OPEN
*
          OPEN  (TAPEDCB,,LPDCB,OUTPUT) INIT DCB'S
*
INPUT     DS    0H                          READ RECORDS FROM DCB
          READ  INDECB,SF,TAPEDCB,(3),'S'      READ RECORD
*
          CHECK INDECB                      CHECK READ
*
          L     8,INDECB+16                 IOB ADDR
          LH    4,DCBBLKSI                  RECORD SIZE READ
          SH    4,14(8)                     LENGTH OF RECORD READ
*
RTNO      DS    0H                          SET RECORD LENGTH IN BYTES
          ST    4,0(5)                      RECORD LENGTH TO CALLER
*
*
RETURN    DS    0H                          RETURN LOGIC
          L     13,SAVE+4                   OLD SAVE AREA ADDR
          RETURN (14,12),T                  RETURN TO CALLER
*
*
ENDDATA   DS    0H                          END OF INPUT
          MVI   0(5),X'FF'                  SET RECORD LENGTH TO NEGATIVE
          B     RETURN                      RETURN TO CALLER
*
*
ERROR     DS    0H                          READ ERROR OCCURRED
          ST    0,FIELD                     DECB ADDR
          UNPK  TMP(9),FIELD(5)             CONVERT TO PSEUDO-EBCDIC
          TR    TMP(8),TABLE-240            CONVERT TO EBCDIC
          MVC   ERRMSG+40(8),TMP            MOVE TO OUTPUT BUFFER
*
          ST    1,FIELD                     ERROR BITS AND DCB ADDR
          UNPK  TMP(9),FIELD(5)             CONVERT TO PSEUDO-EBCDIC
          TR    TMP(8),TABLE+240            CONVERT TO EBCDIC
          MVC   ERRMSG+60(8),TMP            MOVE TO OUTPUT BUFFER
*
          PUT   LPDCB,ERRMSG                OUTPUT ERROR MESSAGE
          S     4                           ERROR OCCURRED
          BR    14                          RETURN TO SYSTEM
*
```

```
*
*
*
*   DATA
*
         DS    0F
FIELD    DS    CL5
TMP      DS    CL9
*
*
*
TAPEDCB  DCB   MACRF=R,RECFM=U,BLKSIZE=8800,EODAD=ENDDATA,                 X
               DSORG=PS,DDNAME=FT01F001,SYNAD=ERROR,DEVD=TA,EROPT=ACC
*
LPDCB    DCB   DSORG=PS,MACRF=PM,BLKSIZE=133,LRECL=133,RECFM=FBM,          X
               DDNAME=LP
*
         DS    0F
ERRMSG   DC    X'09',CL132'**READ ERROR, RECORD IGNORED**'
*
         DS    0F
TABLE    DC    C'0123456789ABCDEF'
*
SAVE     DS    18F
*
*
         DCBD  DSORG=PS
         END
```

Characterizations of Linear Sufficient Statistics

by

B. Charles Peters, Jr.[1], Richard Redner,[1]
and Henry P. Decell, Jr.[1]

University of Houston

August, 1976

Report #59
NAS-9-15000

Characterizations of Linear Sufficient Statistics

By B. Charles Peters, Jr.[1], Richard Redner,[1]
and Henry P. Decell, Jr.[1]

University of Houston

We develop a necessary and sufficient condition that there exist a continous linear sufficient statistic T for a dominated collection of totally finite measures defined on the Borel field generated by the open sets of a Banach space X. In particular, corollary necessary and sufficient conditions that there exist a rank k linear sufficient statistic T for any finite collection of probability measures having n-variate normal densitites are given. In this case a simple calculation, involving only the population means and covariances, determines the smallest integer k for which there exists a rank k linear sufficient statistic T (as well as an associated statistic T itself).

[1] This author was partially supported by NASA Contract NAS-9-15000 with the University of Houston during the preparation of this work.

1. Introduction. If W is a Banach space, $\mathcal{B}(W)$ will denote the Borel field generated by the open sets of W. The totally finite measures defined on $\mathcal{B}(W)$ will be denoted by $\mathcal{M}(W)$. For $\mu, \lambda \varepsilon \mathcal{M}(W)$ we will write $\mu << \lambda$ provided $B \varepsilon \mathcal{B}(W)$ and $\lambda(B) = 0$ implies $\mu(B) = 0$. Whenever $\mu << \lambda$, $[d\mu/d\lambda]$ will denote the equivalence class of Radon-Nikodym derivatives of μ with respect to [2] [3]. If $\mathcal{P} \subset \mathcal{M}(W)$, $\mathcal{P}$ will be called a dominated (by λ) set of measures provided there exists $\lambda \varepsilon \mathcal{M}(W)$ (λ not necessarily in $\mathcal{P}$) such that $\mu \varepsilon \mathcal{P}$ implies $\mu << \lambda$. We will call $\mathcal{P} \subset \mathcal{M}(W)$ equivalent to λ ($\mathcal{P} \equiv \lambda$) provided $\mathcal{P}$ is dominated by λ and $\mu(B) = 0$ for each $\mu \varepsilon \mathcal{P}$ implies $\lambda(B) = 0$.

If X and Y are Banach spaces and $T: X \to Y$ then, following the notation in [3], we write $f(\epsilon)T^{-1}(\mathcal{B}(Y))$ provided $f: X \to R$ (= Reals) and f is $(T^{-1}(\mathcal{B}(Y)), \mathcal{B}(R))$ - measurable (as well as $(\mathcal{B}(X), \mathcal{B}(R))$ - measurable).

In [3], Halmos and Savage develop an approach to sufficient statistics. Their results provide an alternate definition, within a very general mathematical framework, of statistical sufficiency for dominated sets of measures. This alternate definition is particularly suitable to the development of the results in this paper. We will require the statement (Theorem 1.) of the alternate definition in the setting of Banach spaces.

In all that follows X and Y will be Banach spaces, T a linear continuous mapping of X onto Y, and $\mathcal{P} \subset \mathcal{M}(X)$ a dominated set of measures.

Theorem 1. (Halmos-Savage [3]) A necessary and sufficient condition that T be a sufficient statistic for $\mathcal{P}$ is that there exist $\lambda \varepsilon \mathcal{M}(X)$ such that $\mathscr{P} \equiv \lambda$ and $g_\mu \in [d\mu/d\lambda]$ such that $g_\mu(\varepsilon)T^{-1}(\mathcal{B}(Y))$ for each $\mu \in \mathscr{P}$.

In this paper our particular concern will be that of developing necessary and sufficient conditions that a linear continuous mapping T of X onto Y be a sufficient statistic for a dominated set of measures $\mathscr{P} \subset \mathcal{M}(X)$.

In Theorem 2. we will require an additional condition on T which, to the best of our knowledge, is generally unavoidable . We will require that the kernel of T (= ker T) be <u>complemented</u>, in the sense that there exists a closed subspace S of X such that $X = \ker T \oplus S$ (e.g., if X is a Hilbert space, take $S = (\ker T)^{\perp}$).

In Theorem 4. we will show that the condition $X = \ker T \oplus S$ may be relaxed whenever $[d\mu/d\lambda]$ contains a continuous representative. The results we develop are finally used to establish necessary and sufficient conditions that a linear statistic $B:R^n \to R^k (k \leq n)$ be sufficient for a finite collection of probability measures having n-variate normal densities.

2. <u>Principal Results</u>. In all that follows we will assume that X and Y are Banach spaces, $T:X \to Y$ is a linear continuous mapping of X <u>onto</u> Y, and $\mathscr{P} \subset \mathcal{M}(X)$ is a dominated set of measures.

Theorem 2. Let $X = \ker T \oplus S$ for some closed subspace of X. A necessary and sufficient condition that T be a sufficient statistic for $\mathscr{P}$ is that there exist $\lambda \in \mathcal{M}(X)$ such that $\mathscr{P} \equiv \lambda$ and,

$$\ker T \subset \{y : g_\mu(x + y) = g_\mu(x),\ x \in X\}$$

for each $\mu \in \mathscr{P}$ and some $g_\mu \in [d\mu/d\lambda]$.

Proof. If T is a sufficient statistic for $\mathcal{P}$ and $\mu \varepsilon \mathcal{P}$ then there exists (Theorem 1 $\lambda \equiv \mathcal{P}$ and $g_\mu \varepsilon [d\mu/d\lambda]$ such that $g_\mu(\epsilon)T^{-1}(\mathcal{B}(Y)$. Suppose $y \varepsilon \ker T$ and, without loss of generality, there exists $x_0 \varepsilon X$ such that $g_\mu(x_0 + y) < g_\mu(x_0)$. Choose $r \varepsilon R$ such that $g_\mu(x_0 + y) < r < g_\mu(x_0)$. Since $g_\mu^{-1}(-\infty, r)$ and $g_\mu^{-1}(r,\infty)$ are elements of $\mathcal{B}(X)$ and $g_\mu(\epsilon)T^{-1}(\mathcal{B}(Y))$ it follows that there exist B_1 and $B_2 \varepsilon \mathcal{B}(Y)$ such that $x_0 + y \varepsilon g^{-1}(-\infty,r) = T^{-1}(B_1)$ and $x_0 \varepsilon g^{-1}(r,\infty) = T^{-1}(B_2)$. Now, since T is linear and $y \varepsilon \ker T$, $T(x_0) \varepsilon B_1 \cap B_2 = \phi$, which is absurd.

Conversely, suppose $\mathcal{P} \equiv \lambda$, $\mu \varepsilon \mathcal{P}$ and $\ker T \subset \{y: g_\mu(x + y) = g_\mu(x), x \varepsilon X\}$ for some $g_\mu \varepsilon [d\mu/d\lambda]$. We need only show (according to Theorem 1) that $g_\mu(\epsilon)T^{-1}(\mathcal{B}(Y)$. It will only be necessary to show that for $r \varepsilon R$ there exists $B_r \varepsilon \mathcal{B}(Y)$ such that $g_\mu^{-1}(-\infty,r) = T^{-1}(B_r)$. We will show first that $g_\mu^{-1}(-\infty,r) = T^{-1}\, T(g_\mu^{-1}(-\infty,r) \cap S)$ and then that $B_r \equiv T(g_\mu^{-1}(-\infty,r) \cap S) \varepsilon \mathcal{B}(Y)$.

If $x \varepsilon T^{-1}(T(g_\mu^{-1}(-\infty,r) \cap S)$ then $T(x) \varepsilon T(g_\mu^{-1}(-\infty,r) \cap S)$ and hence $T(x) = T(z)$ for some $z \varepsilon g_\mu^{-1}(-\infty,r) \cap S$. Since T is linear $x - z \varepsilon \ker T$ so that $g_\mu(x) = g_\mu(x - z + z) = g_\mu(z) < r$ and $x \varepsilon g_\mu^{-1}(-\infty,r)$.

If $x \varepsilon g_\mu^{-1}(-\infty,r)$ then, since $X = \ker T \oplus S$, $x = k + s$ for $k \varepsilon \ker T$ and $s \varepsilon S$. It follows that $T(x) = T(s)$, $s - x \varepsilon \ker T$, $g_\mu(s) = g_\mu(s - x + x) = g_\mu(x) < r$, $s \varepsilon g_\mu^{-1}(-\infty,r)$, $T(x) = T(s) \varepsilon T(g_\mu^{-1}(-\infty,r) \cap S)$ and, finally, that $x \varepsilon T^{-1}(T(g_\mu^{-1}(-\infty,r) \cap S))$.

We now show that $T(g_\mu^{-1}(-\infty,r) \cap S) \varepsilon \mathcal{B}(Y)$. Let $T_S: S \to Y$ be the restriction of T to S and observe that T_S is a one to one continuous mapping of the Banach space S onto the Banach space Y. Since T_S satisfies the hypothesis of the open mapping theorem T_S is a homeomorphism of S onto Y. Since such mappings take elements of $\mathcal{B}(S)$ into elements of $\mathcal{B}(Y)$ and g_μ is measurable, $g_\mu^{-1}(-\infty,r) \cap S \in \mathcal{B}(X) \cap S = \mathcal{B}(S)$. It follows that $T(g_\mu^{-1}(-\infty,r) \cap S) = T_S(g_\mu^{-1}(-\infty,r) \cap S) \in \mathcal{B}(Y)$ and the proof of the theorem is complete.

Theorem 3. Let $\mathcal{P} \equiv \lambda$, $\lambda(B) = \lambda(B - y)$ for each $y \in \ker T$ and $B \in \mathcal{B}(X)$ such that $\lambda(B) = 0$, $\lambda(C) > 0$ for each non-empty open subset C of X and let $[d\mu/d\lambda]$ contain a continuous representative element f_μ for each $\mu \in \mathcal{P}$.

A necessary and sufficient condition that T be a sufficient statistic for $\mathcal{P}$ is that

$$\ker T \subset \{y : f_\mu(y + x) = f_\mu(x),\ x \in X\}$$

Proof: In order to see that the condition is sufficient we need only show (according to Theorem 1.) that $f_\mu(\varepsilon)T^{-1}(\mathcal{B}(Y))$, or equivalently, if $r \in R$ that $f_\mu^{-1}(-\infty,r) = T^{-1}(B_r)$ for some $B_r \in \mathcal{B}(Y)$. In fact, since T is an open mapping and f_μ is continuous, $T(f^{-1}(-\infty,r)) \in \mathcal{B}(Y)$. We take $B_r \equiv T(f^{-1}(-\infty,r))$ and conclude the argument by showing that $f_\mu^{-1}(-\infty,r) = T^{-1}T(f_\mu^{-1}(-\infty,r))$. We clearly need only establish that $T^{-1}T(f_\mu^{-1}(-\infty,r)) \subset f_\mu^{-1}(-\infty,r)$. If $x \in T^{-1}T(f_\mu^{-1}(-\infty,r))$ then $T(x) = T(z)$ for some $z \in f_\mu^{-1}(-\infty,r)$. Since $x - z \in \ker T$ it follows that $f_\mu(x) = f_\mu(x - z + z) = f_\mu(z) < r$ and hence that $x \in f_\mu^{-1}(-\infty,r)$.

In order to prove the necessity of the condition, recall the proof of the necessity of the condition in Theorem 2. and observe that the hypothesis $X = \ker T \oplus S$ for some closed subspace S of X was not essential. We may conclude that if $\mu \in \mathcal{P}$ there exists $g_\mu \in [d\mu/d\lambda]$ such that $\ker T \subset \{y : g_\mu(y + x) = g_\mu(x),\ x \in X\}$ and $f_\mu = g_\mu$ except on a set $B \in \mathcal{B}(X)$ such that $\lambda(B) = 0$.

Fix $y \in \ker T$. Since $\{x : f_\mu(y + x) \neq g_\mu(y + x)\} = B - y$ and $\lambda(B - y) = \lambda(B) = 0$, we may conclude that $f_\mu(x) = f_\mu(y + x)$ except on $C = B \cup (B - y)$ and $\lambda(C) = 0$. Moreover, since the mapping $x \to y + x$ is a homeomorphism of X onto X and f_μ is continuous, C is an open subset of X. According to the hypothesis, $\lambda(C) = 0$ and C open imply C is empty so that $f_\mu(y + x) = f_\mu(x)$ for each $x \in X$.

3. Normal Families. In what follows we will assume that $\mathcal{P} = \{P_i\}_{i=0}^{m-1}$ is a family of m probability measures defined on $\mathcal{B}(R^n)$ having normal densities

$$p_i(x) = (2\pi)^{-n/2}|\Omega_i|^{-1/2} \exp[-\tfrac{1}{2}(x - \eta_i)^T\Omega_i^{-1}(x - \eta_i]; \ i = 0, 1, \ldots, m-1.$$

where η_i and Ω_i are known and Ω_i is symmetric and positive definite.

We will derive necessary and sufficient conditions that a $k \times n$ matrix B $(k \leq n)$ mapping R^n onto R^k (i.e., rank $(B) = k$) be a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$. We first prove a Lemma.

Lemma 1. If $1 \leq i \leq m - 1$ and $f_i(x) = p_i(x)/p_0(x)$ then

$$\{y : f_i(y + x) = f_i(x),\ x \in X\} = \ker(\Omega_i^{-1} - \Omega_0^{-1}) \cap \{\Omega_i^{-1}\eta_i - \Omega_0^{-1}\eta_0\}^\perp .$$

Proof: Fix $y \in R^n$. After a little matrix algebra (which we will omit) we find that $f_i(y + x) = f_i(x)$ for each $x \in R^n$ if and only if

$$2x^T(\Omega_i^{-1} - \Omega_0^{-1})y - 2y^T(\Omega_i^{-1}\eta_i - \Omega_0^{-1}\eta_0) + y^T(\Omega_i^{-1} - \Omega_0^{-1})y = 0$$

for each $x \in R^n$. For $x = -y/2$ we see that $y^T(\Omega_i^{-1}\eta_i - \Omega_0^{-1}\eta_0) = 0$ so that $y \in \{\Omega_i^{-1}\eta_0 - \Omega_0^{-1}\eta_0\}^\perp$. In addition, it follows that $2x^T(\Omega_i^{-1} - \Omega_0^{-1})y + y^T(\Omega_i^{-1} - \Omega_0^{-1})y = 0$ and, writing $x = (z - y)/2$, that $z^T(\Omega_i^{-1} - \Omega_0^{-1})y = 0$ for each $z \in X$. This clearly implies $(\Omega_i^{-1} - \Omega_0^{-1})y = \Theta$ so that $y \in \ker(\Omega_i^{-1} - \Omega_0^{-1})$. The remaining containment follows easily.

Theorem 4. A necessary and sufficient condition that a $k \times n$ rank k matrix B be a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$ is that

$$\ker B \subset \bigcap_{i=1}^{m-1} [\ker(\Omega_i^{-1} - \Omega_0^{-1}) \cap \{\Omega_i^{-1}\eta_i - \Omega_0^{-1}\eta_0\}^\perp] .$$

Proof: Since the preliminary conditions of Theorem 3 are clearly satisfied for $\lambda = P_0$, Lemma 1 insures the necessity and sufficiency of the condition.

Theorem 5. A necessary and sufficient condition that a $k \times n$ rank k matrix B be a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$ is that, for $j = 1, \ldots, m-1$,

(a) $\Omega_j B^T(B\Omega_j B^T)^{-1} = \Omega_0 B^T(B\Omega_0 B^T)^{-1}$

(b) $\eta_j - \Omega_j B^T(B\Omega_j B^T)^{-1}B\eta_j = \eta_0 - \Omega_0 B^T(B\Omega_0 B^T)^{-1}B\eta_0$

(c) $\Omega_j - \Omega_j B^T(B\Omega_j B^T)^{-1}B\Omega_j = \Omega_0 - \Omega_0 B^T(B\Omega_0 B^T)^{-1}B\Omega_0$.

Proof: Let $(x|y) = x^Ty$ and $(x|y)_i = x^T\Omega_i^{-1}y$ $i = 0, 1, \ldots, m-1$. For $S \subset R^n$, $S^\perp$ and $S^{\perp i}$ will denote, respectively, the orthogonal complements of S relative to the inner products $(\cdot|\cdot)$ and $(\cdot|\cdot)_i$.

If A is an $n \times n$ matrix A^{*i} will denote the adjoint of A relative to the inner product $(\cdot|\cdot)_i$ on R^n. If A is a $k \times n$ matrix A^{*i} will denote the adjoint of A relative to the inner products $(\cdot|\cdot)_i$ on R^n and $(\cdot|\cdot)$ on R^k. It follows that $B^{*i} = \Omega_iB^T$.

If B is a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$ then, according to Theorem 3., $\ker B \subset \ker(\Omega_j^{-1} - \Omega_0^{-1})$; $j = 1, \ldots, m-1$ and hence $(\ker B)^{\perp j} = (\ker B)^{\perp 0}$. Since this implies range (B^{*j}) = range (B^{*0}) we have that $B^{*0}(BB^{*0})^{-1}BB^{*j} = B^{*j}$ and hence that $\Omega_jB^T(B\Omega_jB^T)^{-1} = \Omega_0B^T(B\Omega_0B^T)^{-1}$ which is (a).

Now let $Q = \Omega_0B^T(B\Omega_0B^T)^{-1}B$ and observe that $Q^{*j} = Q = Q^2$ for $j = 1, \ldots, m-1$. It follows that $\ker Q = \ker B \subset \ker(\Omega_j^{-1} - \Omega_0^{-1})$ and that $Q(\Omega_j^{-1} - \Omega_0^{-1})^{*0} = (\Omega_j^{-1} - \Omega_0)^{*0}$ and hence that $Q(\Omega_j - \Omega_0) = \Omega_j - \Omega_0$ which, recalling the definition of Q, is equivalent to (c).

Since $\ker(\Omega_j^{-1} - \Omega_0^{-1}) \cap (\Omega_j^{-1}\eta_n - \Omega_0^{-1}\eta_0) \subset (\eta_j - \eta_0)^{\perp j}$ and $\eta_j - \eta_0 \in (\ker B)^{\perp j}$ = range (B^{*j}) = range (Q), it follows that $Q(\eta_j - \eta_0) = \eta_j - \eta_0$ which, recalling the definiton of Q, is equivalent to (b).

Since all of the preceeding arguments are reversible, (a), (b) and (c) imply B is a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$, completing the proof of the theorem.

In the next theorem we will use the fact that there exists a non singular matrix M such that $M\Omega_0M^T = I$ and hence that the affine transformation $x \to Mx - \eta_0$ provides a change of variables that allows (without loss of generality or the ability to recover the sufficient statistic relative to the original variables) one to assume that $\eta_0 = \Theta$ and $\Omega_0 = I$.

Theorem 6. If $\eta_0 = \Theta$ and $\Omega_0 = I$ then a necessary and sufficient condition that a $k \times n$ rank k matrix B be sufficient for $\{P_i\}_{i=0}^{m-1}$ is that there exist a rank k orthogonal projection Q such that, for $i = 1, \ldots, m-1$,

$$(I - Q)[\eta_1|\eta_2| \ldots |\eta_{m-1}| \Omega_1 - I |\Omega_2 - I| \ldots |\Omega_{m-1} - I] = Z$$

where Z is the $n \times (n+1)(m-1)$ zero matrix.

Proof: If B is a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$, we may assume without loss of generality that $BB^T = I$ since B is a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$ if and only if KB is a sufficient statistic for each nonsingular $k \times k$ matrix K. One may indeed choose K such that $KBB^TK^T = (KB)(KB)^T = I$.

For $i = 1, \ldots, m-1$ Theorem 5. implies that

$$\Omega_i B^T(B\Omega_i B^T)^{-1} = I\, B^T(B\, I\, B^T)^{-1} = B^T$$

so that

$$(B\Omega_i B^T)^{-1} = B\Omega_i^{-1} B^T \quad \text{and} \quad \Omega_i B^T(B\Omega_i B^T)^{-1} B = B^T B\ .$$

Right multiplication of the latter equation by $\Omega_i B^T B$ will establish that

$$\Omega_i B^T B = B^T B \Omega_i B^T B$$

from whence it follows, using symmetry, that

$$\Omega_i B^T B = B^T B \Omega_i\ .$$

Since $\eta_1 = \Theta$ and $\Omega_1 = I$, Theorem 5 further implies

$$\eta_i - B^T B = \Theta$$

and

$$\Omega_i - B^T B \Omega_i = I - B^T B$$

Since $BB^T = I$, it follows that $B^T = B^+$ (where $(\cdot)^+$ denotes the generalized inverse of $(\cdot)$) and hence that $Q \equiv B^T B = B^+ B$ is the orthogonal projection on the range of B^T [5]. Clearly Q has rank k and we conclude that

$$(I - Q)\eta_i = \Theta$$

and

$$(I - Q)(\Omega_i - I) = Z$$

and the condition follows. Conversely, if the conditon holds let B be any $k \times n$ rank k matrix such that range (B^T) = range (Q). Clearly $B^+ B = Q$, $BB^+ = I$ and $B^+ = B^T$. Using the symmetry of $I - Q$ and $\Omega_i - I$ we conclude that

$$\Omega_i B^T B = B^T B \Omega_i$$

and hence that

$$Q = B^+ B = B^+ B \Omega_i B^T (B \Omega_i B^T)^{-1} B = \Omega_i B^+ B B^T (B \Omega_i B^T)^{-1} B$$

$$= \Omega_i B^T (B \Omega_i B^T)^{-1} B \ .$$

In addition,

$$\Omega_i B^T (B \Omega_i B^T)^{-1} = B^T$$

The obvious substitution for Q guarantees the satisfaction of the conditions of Theorem 5.

Definition 1. We will say that a rank k orthogonal projection Q <u>generates a sufficient statistic for</u> $\{P_i\}_{i=0}^{m-1}$ provided Q satisfies the condition in Theorem 6.

Corollary 1. If $M = [\eta_1 | \eta_2 | \ldots | \eta_{m-1} | \Omega_1 - I | \ldots | \Omega_{m-1} - I]$ then

a) $Q = MM^+$ generates a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$

and

b) $k = \text{rank}\ (MM^+) \equiv \text{tr}\ (MM^+)$ is the smallest integer for which there exists a rank k orthogonal projection generating a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$.

Proof: Let k be the smallest integer for which there exists a rank k orthogonal projection P generating a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$.

According to the definition of M, $(I - P)M = Z$ so that $PM = M$ and $PMM^+ = MM^+$. Since $(I - MM^+)M = Z$, MM^+ generates a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$. However, $PMM^+ = MM^+$ implies that range $(MM^+) \subset$ range (P) so that the minimality of k and the fact that MM^+ is an orthogonal projection imply that range (MM^+) = range (P) and hence that $MM^+ = P$.

Corollary 2. If B is a sufficient statistic for $\{P_i\}_{i=0}^{m-1}$ then

$$(B\Omega_i B^T)^{-1} = B\Omega_i^{-1} B^T \quad i = 0, 1, \ldots, m - 1 \ .$$

Proof: The conclusion is an immediate consequence of line 6 in the proof of Theorem 6.

4. Concluding Remarks. Theorems 4 and 5, although not so stated, are valid for arbitrary families of n-variate normal probability measures. Corollary 1. formally gives the construction for a sufficient statistic for finite families of n-variate normal probability measures solely in terms of the known parameters that determine the densities. In fact, if k=rank (M) (=rank MM^{+}) then any rank k matrix B for which range (B)=range (M) is a sufficient statistic for the family. Moreover, in terms of the dimension of the range of a sufficient statistic, k=rank M is the smallest integer for which there exists a sufficient statistic.

Several open questions concerning the "appropriate" definition of a "almost" sufficient statistic using the characterizations given in Theorems 4. and 5. will be the subject of a later paper. In this connection the results of Le Cam [4], although the approach is different, should be of significant value.

5. Acknowledgement. The authors would like to express there sincere appreciation to Professor H. Elton Lacey for his comments.

REFERENCES

1. Anderson, T. W. (1958) An Introduction to Multivariate Statistical Analysis. Wiley, New York.

2. Bahadur, R. R. (1954) Sufficiency and statistical decision function. Ann. Mathe. Statist. 25 423-463.

3. Halmos, P. R. and Savage, L. J. (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann. Math. Statist. 20 225-241.

4. Le Cam, L. (1964) Sufficiency and approximate sufficiency. Ann. Math. Statist. 35 1419-1455.

5. Rao, C. R. and Mitra, S. K. (1971) Generalized Inverse of Matrices and its Applications. Wiley, New York.

A Stochastic Approximation Algorithm for Estimating Mixture Proportions

James Sparra

University of Houston
Department of Mathematics
Houston, Texas

Report #60
NASA Contract NAS-9-15000

November 1976

A Stochastic Approximation Algorithm for Estimating Mixture Proportions

by

James Sparra

1. Summary. A stochastic approximation algorithm for estimating the proportions in a mixture of normal densities is presented. The algorithm is shown to converge to the true proportions in the case of a mixcure of two normal densities.

2. Introduction. Let $A = \{\alpha \in R^m : \alpha_i > 0$ and $\sum_{i=1}^{m} \alpha_i = 1\}$. For each i, $i = 1,\ldots,m$, let μ_i be an element of R^n and Σ_i be a positive definite real symmetric $n \times n$ matrix. Let X be a random variable with values in R^n and with density function.

$$p(\hat{\alpha},x) = \sum_{i=1}^{m} \hat{\alpha}_i p_i(x), \quad \text{for} \quad x \in R^n$$

where $\hat{\alpha} \in A$ and

$$p_i(x) = (2\Pi)^{-n/2}|\Sigma_i|^{-1/2}\exp\{-\tfrac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\}$$

for each $i = 1,\ldots,m$.

We assume that $\hat{\alpha}$ is not known but that μ_i and Σ_i are known for $i = 1,\ldots,m$. An algorithm for estimating $\hat{\alpha}$ will be presented in part 3 of this paper and in part 4 the algorithm will be shown to converge to $\hat{\alpha}$ in mean square and with probability 1 in the case where $m = 2$.

3. The Algorithm. Let $\{x_k\}_{k=o}^{\infty}$ be a sequence of observations on X. Let $\alpha^o \varepsilon A$. For $n \geq 0$ define α^{n+1} by

$$\alpha_i^{n+1} = \alpha_i^n - c_n(\alpha_i^n - \frac{\alpha_i^n p_i(x_n)}{p_{\alpha^n}(x_n)}),$$

where

$$p_{\alpha^n}(x_n) = \sum_{i=1}^{m} \alpha_i^n p_i(x_n)$$

and $\{c_k\}_{k=o}^{\infty}$ is a sequence of positive numbers such that

$$\sum_{k=o}^{\infty} c_k = \infty \quad \text{and} \quad \sum_{k=o}^{\infty} c_k^2 < \infty .$$

We note that each iterate is in A and that, since X is a random variable, each iterate may itself be considered a random variable.

4. Convergence of the Algorithm.

Theorem: If $\hat{\alpha} \varepsilon R^2$ then the algorithm described in part 3 converges to $\hat{\alpha}$ in mean square and with probability 1.

Proof: We refer the reader to the algorithm described in [1,pp. 332-333] and to the proof of convergence given in [1,pp. 350-352]. The applicability of the theorem given there is clear if we let $f(\alpha) = E(Z_\alpha)$, for each $\alpha \varepsilon A$, where

$$(Z_\alpha)_i = \alpha_i - \frac{\alpha_i(p_i \circ X)}{p_\alpha \circ X} .$$

In order to show convergence we must show that conditions (A1)-(A3) in [1,pp. 332-333] are satisfied. First we note that

$$f(\alpha) = (\alpha_1 - \alpha_1 g_1(\alpha_1),\ \alpha_2 - \alpha_2 g_2(\alpha_2))$$

where

$$g_1(\alpha_1) = \int_{R^n} \frac{p_1(x)}{\alpha_1 p_1(x) + (1-\alpha_1) p_2(x)}\ p_{\hat{\alpha}}(x)dx$$

and

$$g_2(\alpha_2) = \int_{R^n} \frac{p_2(x)}{(1-\alpha_2) p_1(x) + \alpha_2 p_2(x)}\ p_{\hat{\alpha}}(x)dx.$$

Further, we note that

$$\frac{d^2 g_1(\alpha_1)}{d\alpha_1^2} = \int_{R^n} \frac{p_1(x)[p_1(x) - p_2(x)]^2}{[\alpha_1 p_1(x) + (1-\alpha_1) p_2(x)]^3} \cdot p_{\hat{\alpha}}(x)dx > 0$$

and

$$\frac{d^2 g_2(\alpha_2)}{d\alpha_2^2} = \int_{R^n} \frac{p_2(x)[p_2(x) - p_1(x)]^2}{[(1-\alpha_2) p_1(x) + \alpha_2 p_2(x)]^3} \cdot p_{\hat{\alpha}}(x)dx > 0.$$

Now, $g_1(\hat{\alpha}_1) = 1$ and $g_1(1) = 1$. So, since g_1 has positive second derivative we have that $g_1(\alpha_1) < 1$ if $\alpha_1 \varepsilon (\hat{\alpha}_1, 1)$ and $g_1(\alpha_1) > 1$ if $\alpha_1 \varepsilon (0, \hat{\alpha}_1)$.

Similarly, $g_2(\hat{\alpha}_2) = 1$ and $g_2(1) = 1$ and $g_2(\alpha_2) < 1$ if $\alpha_2 \varepsilon (\hat{\alpha}_2, 1)$ and $g_2(\alpha_2) > 1$ if $\alpha_2 \varepsilon (0, \hat{\alpha}_2)$.

We now show that (A1)-(A3) are satisfied: Let $\alpha \varepsilon A$. Then

(A1) $f(\alpha) = 0$ iff $g_1(\alpha_1) = 1 = g_2(\alpha_2)$ iff $\alpha = \hat{\alpha}$.

(A2) $(\alpha - \hat{\alpha})^T f(\alpha) = (\alpha_1 - \hat{\alpha}_1)(\alpha_1 - \alpha_1 g_1(\alpha_1)) + (\alpha_2 - \hat{\alpha}_2)(\alpha_2 - \alpha_2 g_2(\alpha_2))$.

If $\alpha_1 > \hat{\alpha}_1$ then $g_1(\alpha_1) < 1$ and $(\alpha_1 - \alpha_1 g_1(\alpha_1)) > 0$. Then also $\alpha_2 < \hat{\alpha}_2$ and $g_2(\alpha_2) > 1$ and $(\alpha_2 - \alpha_2 g_2(\alpha_2)) < 0$. Thus, if $\alpha_1 > \hat{\alpha}_1$ then $(\alpha - \hat{\alpha})^T f(\alpha) > 0$. Similarly, if $\alpha_1 < \hat{\alpha}_1$ then $(\alpha - \hat{\alpha})^T f(\alpha) > 0$. Thus, A2 is satisfied in any closed, convex subset of A.

(A3)

$$E(||Z_\alpha||^2) = \sum_{i=1}^{2} \left(\alpha_i^2 - 2 \int_{R^n} \frac{\alpha_i^2 p_i(x)}{p_\alpha(x)} \cdot p_{\hat{\alpha}}(x)dx + \int_{R^n} \left(\frac{\alpha_i p_i(x)}{p_\alpha(x)}\right)^2 \cdot p_{\hat{\alpha}}(x)dx\right)$$

Now, we note that each term in the ith summand, $i = 1,2$, is less than 1 so that there is an $h > 0$ such that $E(||Z_\alpha||^2) < h$ for all $\alpha \varepsilon A$ and A3 is satisfied.

Bibliography

1. C.C. Blaydon, K.S. Fu, and R.L. Kashyap, "Stochastic Approximation", Adaptive, Learning and Pattern Recognition Systems, Academic Press, New York and London, 1970, Edited by J.M. Mendel and K.S. Fu.

The Role of Eigenvalues in Linear Feature Selection Theory[1]

by

D. R. Brown and M. J. O'Malley

Department of Mathematics
University of Houston

December, 1976
Report 61
NAS-9-15000

The Role of Eigenvalues in Linear Feature Selection Theory

D. R. Brown and M. J. O'Malley

Department of Mathematics, University of Houston
Houston, Texas 77004

Introduction. Recent statistical work in feature selection for the multivariate normal pattern recognition problem has concentrated on linearly transforming pattern classes so that the transformed pattern classes are equivalently distinguishable. Since, in general, this is not possible, techniques have been developed to preserve the distinction of the transformed pattern classes using various measures of distinction. These measures of pattern class distinction are most often treated as eigenvalue problems ([1], [2], [5], [6], [7], [9], [13], [14], [15]). In this paper we consider a particular measure of pattern class distinction called the average interclass divergence, or more simply, divergence, ([1], [2], [4], [6], [7], [8], [9], [10], [11]), where divergence will be the pairwise average of the expected interclass divergence derived from Hajek's two-class divergence as defined, for example, in [9].

This work was supported in part by NASA under Contract JSC-NAS-15000.

It has been shown in [4] that there always exists a $k \times n$ real matrix B such that the transformation determined by B maximizes divergence in k-dimensional space, and, in fact, that B can be written in the form $(I_k|Z)U$, where U is an orthogonal $n \times n$ matrix. We will investigate the role of the eigenvalues of U in such problems, and give an example demonstrating that the divergence measure of pattern class distinction does not depend on these eigenvalues (Theorem 7).

Our example is derived from the family of examples constructed in [3]. This special class of examples permits analytical calculation of divergence, a task ordinarily eschewed as unrealistic, and yields a precise expression for divergence. The reader is cautioned, however, not to confuse the numerical simplicity of this example with impracticality, since, mathematically, the failure of the eigenvalues of U to affect divergence in the restricted case erases any hope that they might be meaningful in an arbitrary case, however applied.

1. Special divergence formulas. Let $\Omega_1,\ldots,\Omega_m$ and $\mu_1,\ldots,\mu_m$ be the covariance matrices and means for m classes, where for each $i = 1,\ldots,m$, Ω_i is an $n \times n$ positive definite matrix and μ_i is a column n vector. Let

$$S_i = \Sigma_{\substack{j=1\\ j\neq i}}^{m} (\Omega_j + \delta_{ij}\delta_{ij}^T), \text{ where } \delta_{ij} = \mu_i - \mu_j .$$

Then, assuming equal a priori probabilities, the average interclass divergence for these m classes is given by

$$D = \tfrac{1}{2}\,\mathrm{tr}\Big(\sum_{i=1}^{m} \Omega_i^{-1} S_i\Big) - \tfrac{1}{2}\, m(m-1)n \tag{1}$$

while, if B is a $k \times n$ matrix, the B-average interclass divergence is

$$D_B = \tfrac{1}{2}\,\mathrm{tr}\Big(\sum_{i=1}^{m} (B\Omega_i B^T)^{-1}(BS_i B^T)\Big) - \tfrac{1}{2}\, m(m-1)k \tag{2}$$

where tr represents the trace function.

Moreover, as observed in [3], if

$$\mathcal{G} = \{B \in M_{kn} \colon BB^T = I_k \ \text{ and } \ (B^TB)\Omega_i = \Omega_i(B^TB), \ \ i = 1,\dots,m\} \ ,$$

where I_k is the $k \times k$ identity matrix and M_{kn} is the set of all $k \times n$ real matrices, then, for any $B \in \mathcal{G}$, (2) may be rewritten as

$$D_B = \tfrac{1}{2}\,\mathrm{tr}\Big(B\Big(\sum_{i=1}^{m} \Omega_i^{-1} S_i\Big)B^T\Big) - \tfrac{1}{2}\, m(m-1)k \tag{3}$$

For the remainder of the paper we assume that each Ω_i is a diagonal matrix of the form: $\begin{pmatrix} x_i & \\ & I_{n-1}\end{pmatrix}$, where x_i is a positive real number, and $\mu_i = \mu_j$ for all i,j. Under these restrictions, $\sum_{i=1}^{m} \Omega_i^{-1} S_i$ is a diagonal matrix of the form $\begin{pmatrix} x & \\ & pI_{n-1}\end{pmatrix}$, where

$x = \sum_{i=1}^{m} \frac{1}{x_i}\Big(\sum_{\substack{j=1\\ j\neq i}}^{m} x_j\Big)$ and $p = m(m-1)$. It follows from (1) that the average interclass divergence for the m classes is given by

$$D = \tfrac{1}{2}(x - p) \tag{4}$$

As observed in the introduction, in seeking to maximize the B-average interclass divergence D_B, it suffices to consider those $k \times n$ matrices of the form $(I_k|Z)U$, where U is an $n \times n$ orthogonal matrix. In the sequel, when considering D_B, we shall always assume that B is of this form. For any such $k \times n$ matrix B, it is obvious that $BB^T = I_k$, and hence $B \in \mathcal{C}$ if and only if $(B^TB)\Omega_i = \Omega_i(B^TB)$ for $i = 1,\ldots,m$. We will derive necessary and sufficient conditions in order that $B \in \mathcal{C}$ (Theorem 2), but first we calculate D_B in the case that formula (3) is valid. Recall that all means are hereafter considered equal and all covariance matrices diagonal of the form stated above.

Theorem 1. Let $B = (I_k|Z)U$, where $U = (u_{ij})$ is an $n \times n$ orthogonal matrix, and suppose D_B is given as in (3) above. Then

$$D_B = (\sum_{i=1}^{k} u_{i1}^2)D \qquad (5)$$

Proof: Since $\operatorname{tr}(XY) = \operatorname{tr}(YX)$ whenever both products are defined, we have in this case $D_B = \frac{1}{2} \operatorname{tr}(B^TB(\sum_{i=1}^{m} \Omega_i^{-1} S_i)) - \frac{1}{2} pk$. If U is written in block form, $U = \begin{pmatrix} A & C \\ E & F \end{pmatrix}$, where A is $k \times k$, then $B^TB = U^T(I_k|Z)^T(I_k|Z)U = \begin{pmatrix} A^TA & A^TC \\ C^TA & C^TC \end{pmatrix}$. Since $\sum_{i=1}^{m} \Omega_i^{-1} S_i = \begin{pmatrix} x & \\ & pI_{n-1} \end{pmatrix} = p\cdot\begin{pmatrix} \frac{x}{p} & \\ & I_{n-1} \end{pmatrix} = p\begin{pmatrix} M & \\ & I_{n-k} \end{pmatrix}$, where M is the $k \times k$ matrix $\begin{pmatrix} \frac{x}{p} & \\ & I_{k-1} \end{pmatrix}$, then $B^TB(\sum_{i=1}^{m} \Omega_i^{-1} S_i) = p\cdot\begin{pmatrix} A^TAM & A^TC \\ C^TAM & C^TC \end{pmatrix}$. Therefore, $\operatorname{tr}(B^TB(\sum_{i=1}^{m} \Omega_i^{-1} S_i)) = p(\operatorname{tr}(A^TAM) + \operatorname{tr}(C^TC)) = p((\sum_{j=1}^{k} u_{j1}^2)\frac{x}{p} + \sum_{q=2}^{k} (\sum_{j=1}^{k} u_{jq}^2) + \sum_{q=k+1}^{n} (\sum_{j=1}^{k} u_{jq}^2)) = (\sum_{j=1}^{k} u_{j1}^2)x + p(\sum_{q=2}^{n} (\sum_{j=1}^{k} u_{jq}^2))$. Since U is orthogonal, $\sum_{q=2}^{n} (\sum_{j=1}^{k} u_{jq}^2) =$

$\sum_{j=1}^{k} (1 - u_{j1}^2) = k - \sum_{j=1}^{k} u_{j1}^2$, so that $D_B = \frac{1}{2}((\sum_{j=1}^{k} u_{j1}^2)x + p(k - \sum_{j=1}^{k} u_{j1}^2)) - \frac{1}{2} pk =$ $(\sum_{j=1}^{k} u_{j1}^2)(\frac{x-p}{2}) = (\sum_{j=1}^{k} u_{j1}^2)D$.

Our next result gives necessary and sufficient conditions in order that $B = (I_k|Z)U \varepsilon \mathcal{P}$. While the proof is rather tedious, these conditions are particularly easy to apply and hence useful in seeking examples.

Theorem 2. Let $B = (I_k|Z)U$, where $U = (u_{ij})$ is an $n \times n$ orthogonal matrix. If, for each $i = 1,\dots,m$, $\Omega_i = \begin{pmatrix} x_i & \\ & I_{n-1} \end{pmatrix}$, then:

(1) if $x_i = 1$ for all i, then $B \varepsilon \mathcal{P}$;

(2) if $x_i \neq 1$ for at least one i, then $B \varepsilon \mathcal{P}$ if and only if $\sum_{j=1}^{k} u_{j1}^2 = 1$ or $\sum_{j=1}^{k} u_{j1}^2 = 0$.

Proof: If $x_i = 1$, then $\Omega_i = I_n$ and $(B^TB)\Omega_i = \Omega_i(B^TB)$ for any $k \times n$ matrix B. Thus, if $x_i = 1$ for all i, then $B \varepsilon \mathcal{P}$ for any $k \times n$ matrix of the form $(I_k|Z)U$. We suppose that $x_i \neq 1$ for at least one i. As in the proof of Theorem 1, we decompose U into the block form $\begin{pmatrix} A & C \\ E & F \end{pmatrix}$, so that $B^TB = \begin{pmatrix} A^TA & A^TC \\ C^TA & C^TC \end{pmatrix}$, where A is again $k \times k$. For a fixed i such that $x_i \neq 1$, write Ω_i in block form $\begin{pmatrix} G_i & \\ & I_{n-k} \end{pmatrix}$, where G_i is the $k \times k$ matrix $\begin{pmatrix} x_i & \\ & I_{k-1} \end{pmatrix}$. Then $(B^TB)\Omega_i = \begin{pmatrix} A^TAG_i & A^TC \\ C^TAG_i & C^TC \end{pmatrix}$, while $\Omega_i(B^TB) = \begin{pmatrix} G_iA^TA & G_iA^TC \\ C^TA & C^TC \end{pmatrix}$. Thus, B^TB commutes with Ω_i if and only if

(1) $A^TAG_i = G_iA^TA$ and (2) $C^TAG_i = C^TA$. We write A^TA and C^TA in block form: $A^TA = \begin{pmatrix} L & M \\ N & W \end{pmatrix}$, $C^TA = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}$, where L and P are 1×1. Since A^TA is symmetric, $N = M^T$. Therefore, $A^TAG_i = \begin{pmatrix} Lx_i & M \\ M^Tx_i & W \end{pmatrix}$,

and $G_iA^TA = \begin{pmatrix} x_iL & x_iM \\ M^T & W \end{pmatrix}$. Thus $A^TAG_i = G_iA^TA$ if and only if $M = x_iM$ and similarly, $C^TAG_i = C^TA$ if and only if $Px_i = P$ and $Rx_i = R$. Since

$M = (\sum_{j=1}^{k} u_{j1}u_{j2}, \ldots, \sum_{j=1}^{k} u_{j1}u_{jk})$ and $\begin{pmatrix} P \\ R \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k} u_{jk+1}u_{j1} \\ \vdots \\ \sum_{j=1}^{k} u_{jn}u_{j1} \end{pmatrix}$, it

follows that $Mx_i = M$, $Px_i = P$, and $Rx_i = R$ if and only if $x_i(\sum_{j=1}^{k} u_{j1}u_{jq} = \sum_{j=1}^{k} u_{j1}u_{jq}$ for $q = 2,\ldots,n$. Thus, since $x_i \neq 1$, we have that $(B^TB)\Omega_i = \Omega_i(B^TB)$ if and only if $\sum_{j=1}^{k} u_{j1}u_{jq} = 0$ for $q = 2,\ldots,n$. Since the above argument is valid for any Ω_i for which $x_i \neq 1$, and since B^TB commutes with Ω_i for any i for which $x_i = 1$, it follows that $B \varepsilon \mathcal{G}$ if and only if $\sum_{j=1}^{k} u_{j1}u_{jq} = 0$ for $q = 2,\ldots,n$. We next show that $\sum_{j=1}^{k} u_{j1}u_{jq} = 0$ for $q = 2,\ldots,n$ if and only if $\sum_{j=1}^{k} u_{j1}^2 = 1$ or $\sum_{j=1}^{k} u_{j1}^2 = 0$. Since U is orthogonal, $\sum_{j=1}^{n} u_{j1}u_{jq} = \sum_{j=1}^{k} u_{j1}u_{jq} + \sum_{j=k+1}^{n} u_{j1}u_{jq} = 0$ for $q = 2,\ldots,n$, while $1 = \sum_{j=1}^{n} u_{j1}^2 = \sum_{j=1}^{k} u_{j1}^2 + \sum_{j=k+1}^{n} u_{j1}^2$. Thus, if $\sum_{j=1}^{k} u_{j1}^2 = 1$, then $u_{j1} = 0$ for $j = k+1,\ldots,n$, and $\sum_{j=1}^{n} u_{j1}u_{jq} = \sum_{j=1}^{k} u_{j1}u_{jq} = 0$ for $q = 2,\ldots,n$. If $\sum_{j=1}^{k} u_{j1}^2 = 0$, then $u_{j1} = 0$ for $j = 1,\ldots,k$ and, obviously $\sum_{j=1}^{k} u_{j1}u_{jq} = 0$ for $q = 2,\ldots,n$.

Conversely, suppose that $\sum_{j=1}^{k} u_{j1}u_{jq} = 0$ for $q = 2,\dots,n$. If $u_{11} = \dots = u_{k1} = 0$, then $\sum_{j=1}^{k} u_{j1}^{2} = 0$ and the proof is complete. Otherwise, let u_{r1} be the first non-zero element in the first column of U, where $r \le k$. Then $0 = \sum_{j=1}^{k} u_{j1}u_{jq} = u_{r1}u_{rq} + \sum_{j=r+1}^{k} u_{j1}u_{jq}$, so that $u_{rq} = \frac{-1}{u_{r1}} (\sum_{j=r+1}^{k} u_{j1}u_{jq})$ for $q = 2,\dots,n$. Thus, if $u_{r+11},\dots,u_{k1} = 0$, then $u_{rq} = 0$ for $q = 2,\dots,n$ and it follows that $1 = u_{r1}^{2} = \sum_{j=1}^{k} u_{j1}^{2}$. Suppose $u_{w1} \neq 0$ where $r < w \le k$. Since $u_{r1}u_{w1} + \sum_{q=2}^{n} u_{wq}u_{rq} = 0$, then substituting for u_{rq}, $q \ge 2$, we have

$$u_{r1}u_{w1} + \sum_{q=2}^{n} u_{wq}(\frac{-1}{u_{r1}} \sum_{j=r+1}^{k} u_{j1}u_{jq}) = u_{r1}u_{w1} + (\frac{-1}{u_{r1}}) \sum_{j=r+1}^{k} u_{j1}(\sum_{q=2}^{n} u_{wq}u_{jq}) = 0 \quad (6)$$

Since U is orthogonal, then for $j \neq w$, $\sum_{q=2}^{n} u_{wq}u_{jq} = -u_{w1}u_{j1}$ and for $j = w$, $\sum_{q=2}^{n} u_{wq}u_{jq} = \sum_{q=2}^{n} u_{wq}^{2} = 1 - u_{w1}^{2}$. It follows that $\sum_{j=r+1}^{k} u_{j1}(\sum_{q=2}^{n} u_{wq}u_{jq}) = u_{w1}(\sum_{j=r+1}^{k} (-u_{j1}^{2})) + u_{w1}$, and, substituting in (6), we have $u_{w1}(u_{r1} + (\frac{-1}{u_{r1}})(\sum_{j=r+1}^{k}(-u_{j1}^{2})) + (\frac{-1}{u_{r1}})) = 0$. Multiplying by u_{r1}, we have $u_{w1}(u_{r1}^{2} + \sum_{j=r+1}^{k} u_{j1}^{2} - 1) = u_{w1}(\sum_{j=r}^{k} u_{j1}^{2} - 1) = 0$. Since $u_{w1} \neq 0$, it now follows that $1 = \sum_{j=r}^{k} u_{j1}^{2} = \sum_{j=1}^{k} u_{j1}^{2}$.

We note that, if there exists at least one Ω_j which is not the identity matrix I_n, then the proof of Theorem 2 shows that B^TB commutes with all Ω_i's if and only if B^TB commutes with Ω_j. Moreover, in this case, the elements of $\mathcal{C}$ are precisely those $B = (I_k|Z)U$ for which the first column of

U is of the form $\begin{pmatrix} u_{11} \\ \vdots \\ u_{k1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ \vdots \\ u_{k+11} \\ \vdots \\ u_{n1} \end{pmatrix}$.

Hence, by Theorem 1, if $B \varepsilon \mathcal{C}$, then $D_B = D$ or $D_B = 0$. (Note that if $\Omega_i = I_n$ for all i, then $D = 0$.)

We close this section with a definition. If V denotes the set of all $n \times n$ orthogonal matrices, let $\mathcal{S} = \{U = (u_{ij}) \varepsilon V : \sum_{j=1}^{k} u_{j1}^2 = 1 \text{ or } 0\}$. Thus, if there exists $\Omega_j \neq I_n$, then $B = (I_k|Z)U \varepsilon \mathcal{C}$ if and only if $U \varepsilon \mathcal{S}$.

2. Eigenvalues of U . Let $U = (u_{ij})$ be an $n \times n$ orthogonal matrix. As is well known, [12] , the eigenvalues of U lie on the unit circle in the complex plane and non-real eigenvalues occur in conjugate pairs. Thus, if U has a real eigenvalue λ, then $\lambda = \pm 1$, and, if $\mu = a + bi$, $b \neq 0$ is an eigenvalue of U, then $\overline{\mu} = a - bi$ is also an eigenvalue of U . Clearly, $\det U = \pm 1$. Moreover, if 1 has multiplicity p as an eigenvalue of U, -1 multiplicity m, and $\{a_j + b_j i, a_j - b_j i\}_{j=1}^{q}$ $(b_j \neq 0)$ are the remaining eigenvalues of U, then U is similar to a block diagonal orthogonal matrix PUP^{-1} of the form:

$$PUP^{-1} = \begin{pmatrix} A_1 & & & & & & \\ & \ddots & & & & & \\ & & A_q & & & & \\ & & & 1 & & & \\ & & & & \ddots & & \\ & & & & & 1 & \\ & & & & & & -1 \\ & & & & & & & \ddots \\ & & & & & & & & -1 \end{pmatrix} \qquad (7)$$

where 1 appears on the diagonal p times, -1 appears m times, and each $A_j = \begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix}$ is a 2 × 2 orthogonal matrix with eigenvalues $a_j + b_j i$, $a_j - b_j i$. Furthermore, the order in which the A_j's, 1's, and -1's appear on the diagonal can be changed to any desired order by a similarity transformation. Thus, any two orthogonal $n \times n$ matrices with the same set of eigenvalues are similar. Finally, we observe that if U is a 2 × 2 orthogonal matrix, then

$$U = \begin{pmatrix} c & d \\ d & -c \end{pmatrix} \quad \text{or} \quad U = \begin{pmatrix} c & d \\ -d & c \end{pmatrix} \quad \text{where } c^2 + d^2 = 1 .$$

Let $B = (I_k|Z)U \in \mathcal{B}$. For the remainder of the paper we will be concerned with determining what role, if any, the eigenvalues of U play in determining D_B. If $\{\lambda_1,\ldots,\lambda_n\}$ is a set of n not necessarily distinct complex numbers for which there exists an $n \times n$ orthogonal matrix U with eigenvalues $\lambda_1,\ldots,\lambda_n$, then we will say that $\{\lambda_1,\ldots,\lambda_n\}$ is a (*) <u>set</u>. We note that if $T = \{\lambda_1,\ldots,\lambda_n\}$ is a set of n not necessarily distinct complex numbers such that T is closed under conjugation and every element of T has modulus 1, then T is a (*) <u>set</u>. Throughout the following, we assume that $1 \le k < n$, where k and n are positive integers, and we assume that at least one covariance matrix $\Omega_i \neq I_n$.

<u>Proposition 3</u>. Let $\{\lambda_1,\ldots,\lambda_n\}$ be a (*) set. Then there exists an orthogonal matrix U with eigenvalues $\lambda_1,\ldots,\lambda_n$ such that $B = (I_k|Z)U \in \mathcal{B}$ and $D_B = D$ if and only if one of the following conditions holds:

(i) λ_i is real for some i.

(ii) $k \ge 2$ and no λ_i is real.

Proof: Observe that if at least one λ_j is real, say λ_1, then by (7) there exists a block diagonal orthogonal matrix U of the form $U = \begin{pmatrix} \lambda_1 & \\ & C \end{pmatrix}$, where C is an $(n - 1) \times (n - 1)$ block diagonal orthogonal matrix with eigenvalues $\lambda_2,\ldots,\lambda_n$. Thus, if $U = (u_{ij})$, then $\sum_{j=1}^{k} u_{j1}^2 = u_{j1}^2 = \lambda_1^2 = 1$, so that $B = (I_k|Z)U \in \mathcal{E}$ and $D_B = D$ (Theorem 2). If no λ_j is real, then n is even, and by (7) there exists a block diagonal orthogonal matrix U with eigenvalues $\lambda_1,\ldots,\lambda_n$ such that $U = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_{\frac{n}{2}} \end{pmatrix}$, where each A_j is a 2×2 matrix of the form $\begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix}$, $b_j \neq 0$. Thus, the first column of U is $\begin{pmatrix} a_1 \\ -b_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, and hence, if $k \geq 2$, then $B = (I_k|Z)U \in \mathcal{E}$ and $D_B = D$.

Conversely, suppose that $k = 1$. If there exists an orthogonal matrix U with eigenvalues $\lambda_1,\ldots,\lambda_n$ such that $B = (I_k|Z)U \in \mathcal{E}$, then $U \in \mathcal{L}$. Thus, if $D_B = D$, then U is of the form $\begin{pmatrix} a & 0 & \ldots & 0 \\ 0 & & & \\ 0 & & C & \\ \vdots & & & \\ 0 & & & \end{pmatrix}$, where $a = \pm 1$ and C is an $(n - 1) \times (n - 1)$ orthogonal matrix. Therefore, a is an eigenvalue of U and $\lambda_i = a$ is real for some i.

It is natural to consider the analogous condition $D_B = 0$. That is, given a (*) set $\{\lambda_1,\ldots,\lambda_n\}$, does there exist an orthogonal matrix U with these eigenvalues such that $B = (I_k|Z)U \varepsilon \mathcal{C}$ and $D_B = 0$? The answer, as in the preceding case, is no in general, but it is true in some important cases.

Proposition 4. Let $T = \{\lambda_1,\ldots,\lambda_n\}$ be a (*) set. If either

(i) 1 and $-1 \varepsilon T$, or;

(ii) i and $-i \varepsilon T$,

then there exists an orthogonal matrix U with eigenvalues $\{\lambda_1,\ldots,\lambda_n\}$ such that $B = (I_k|Z)U \varepsilon \mathcal{C}$ and $D_B = 0$.

Proof. Let λ_1 and λ_2 denote the pair $1, -1$ or $i, -i$, let H be any $(n - 2) \times (n - 2)$ orthogonal matrix with eigenvalues $\lambda_3,\ldots,\lambda_n$, and let

$$U = \begin{pmatrix} 0 & Z & b_1 \\ Z & H & Z \\ b_2 & Z & 0 \end{pmatrix}$$

, where Z denotes an $(n - 2)$ row or column vector of zeros, and if $\{\lambda_1, \lambda_2\} = \{1, -1\}$, then $b_1 = b_2 = 1$, and if $\{\lambda_1, \lambda_2\} = \{i, -i\}$, then $b_1 = 1$, $b_2 = -1$.

Clearly, U is an orthogonal matrix. Moreover, the eigenvalues of U are $\{\lambda_1,\ldots,\lambda_n\}$, since $\det(xI_n - U) = (x^2 - b_1b_2) \det(xI_{n-2} - H)$ and hence the roots of $\det(xI_n - U) = 0$ are the roots of $\det(xI_{n-2} - H) = 0$, together with the roots of $x^2 - b_1b_2 = 0$. Since the roots of the former equation are the eigenvalues of H, its suffices to show that λ_1 and λ_2 are the roots of $x^2 - b_1b_2 = 0$. This follows immediately from the relationship defined between the values of λ_1 and λ_2 and the choices of b_1 and b_2. Thus, since we assume $k < n$, then Theorem 2 implies that $U \varepsilon \mathcal{S}$, so that $B = (I_k|Z)U \varepsilon \mathcal{C}$, and, by Theorem 1, $D_B = 0$.

Our next result shows that, if $n = 3$, then Proposition 4 does not characterize those (*) sets T for which there exists an orthogonal matrix U with set of eigenvalues T such that $B = (I_k|Z)U \varepsilon \mathcal{C}$ and $D_B = 0$. We will obtain a partial extension of this result to arbitrary n and we will make strong use of the extension in our main result, Theorem 7.

<u>Lemma 5</u>. Let $n = 3$, $k = 2$, and suppose that $\{\lambda_1, \lambda_2, \lambda_3\}$ is a (*) set, where $\lambda_1 = a + bi$, $\lambda_2 = a - bi$.

(1) If $\lambda_3 = 1$, then there exists a 3×3 orthogonal matrix U with eigenvalues $\lambda_1, \lambda_2, \lambda_3$ such that $U \varepsilon \mathcal{S}$ and $D_B = 0$, $B = (I_k|Z)U$, if and only if a, the real part of λ_1 and λ_2, is less than or equal to zero;

(2) if $\lambda_3 = -1$, then there exists a 3×3 orthogonal matrix U with eigenvalues $\lambda_1, \lambda_2, \lambda_3$ such that $U \varepsilon \mathcal{S}$ and $D_B = 0$, $B = (I_k|Z)U$, if and only if a, the real part of λ_1 and λ_2, is greater than or equal to zero.

<u>Proof</u>. Observe that if $U \varepsilon \mathcal{S}$ is such that $D_B = 0$, where $B = (I_k|Z)U$, then by Theorems 1 and 2, U is of the form $\begin{pmatrix} 0 & A & \\ 0 & & \\ v & 0 & 0 \end{pmatrix}$, where $v = \pm 1$ and A is a 2×2 orthogonal matrix. Moreover, if U has eigenvalues

$\lambda_1, \lambda_2, \lambda_3$, then $\det(U) = \lambda_1\lambda_2\lambda_3$. Thus, if $\lambda_3 = 1$, then $\det(U) = 1$, and if $\lambda_3 = -1$, then $\det(U) = -1$. We consider the case $\lambda_3 = 1$, the case $\lambda_3 = -1$ being similar.

If $v = 1$, then A is of the form $\begin{pmatrix} c & d \\ -d & c \end{pmatrix}$. Then $\det(xI_3 - U) = x^3 + dx^2 - dx - 1$, so that the eigenvalues of U are 1, $\frac{-(1+d) \pm i\sqrt{3-2d-d^2}}{2}$. Thus, there exists U with eigenvalues $\lambda_1, \lambda_2, 1$ if and only if there exists a real number d, $|d| \le 1$, such that

$$a = \frac{-(1+d)}{2} \quad , \quad b = \frac{\sqrt{3-2d-d^2}}{2} \quad . \qquad (8)$$

Since $|d| \le 1$, then $\frac{-(1+d)}{2} \le 0$, and thus, if U exists, then $a \le 0$. Conversely, if $a \le 0$, then $d = -(1+2a)$ satisfies both equations in (8) and $|d| \le 1$. If $v = -1$, then $A = \begin{pmatrix} c & d \\ d & -c \end{pmatrix}$, and the eigenvalues of U are 1, $\frac{(d-1) \pm i\sqrt{3+2d-d^2}}{2}$. An argument similar to the preceding one shows that there exists U with eigenvalues $\lambda_1, \lambda_2, 1$ if and only if $a \le 0$.

Corollary 6. Let n and k be positive integers, $1 \le k < n$, and suppose that $T = \{\lambda_1, \ldots, \lambda_n\}$ is a (*) set.

(1) If $1 \in T$ and if there exists $a + bi \in T$, with $a \le 0$, then there exists an $n \times n$ orthogonal matrix U with eigenvalues T such that $U \in \mathcal{A}$ and $D_B = 0$, where $B = (I_k|Z)U$.

(2) If $-1 \in T$ and if there exists $a + bi \in T$, with $a \ge 0$, then there exists an $n \times n$ orthogonal matrix U with eigenvalues T such that $U \in \mathcal{A}$ and $D_B = 0$, where $B = (I_k|Z)U$.

Proof. By Lemma 5 and its proof, if $a \leq 0$, then $A = \begin{pmatrix} 0 & c & d \\ 0 & -d & c \\ 1 & 0 & 0 \end{pmatrix}$, where $d = -(1 + 2a)$, is an orthogonal matrix with eigenvalues 1, $a \pm bi$. Thus, if $\overline{U}$ is the $n \times n$ block diagonal matrix $\begin{pmatrix} A & Z \\ Z & H \end{pmatrix}$, where H is an $(n - 3) \times (n - 3)$ orthogonal matrix with eigenvalues $T \backslash \{1, a \pm bi\}$, then $\overline{U}$ is an orthogonal matrix with eigenvalues the elements of T. Therefore, if U is the $n \times n$ matrix obtained from $\overline{U}$ by interchanging the third and $n^{\underline{th}}$ rows and columns of $\overline{U}$, then U is orthogonal, and, since U is similar to $\overline{U}$, the eigenvalues of U are also the elements of T. Finally, since the first column of U is $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$, we have $U \in \mathcal{L}$, and, by Theorems 1 and 2, $D_B = 0$, where $B = (I_k|Z)U$ and $k < n$. The proof of (2) is similar.

We make a few additional observations before stating our main result. Let U be an $n \times n$ orthogonal matrix with eigenvalues λ_1, $\{a_j + b_j i\}_{j=2}^{n}$, where b_j may be zero. Since $\mathrm{tr}(U)$ is the sum of the eigenvalues of U, it follows that if $\lambda_1 = 1$ and $a_j > 0$ for $j = 2,\ldots,n$, then $\mathrm{tr}(U) = 1 + \sum_{j=2}^{n} a_j > +1$, while if $\lambda_1 = -1$ and $a_j < 0$ for $j = 2,\ldots,n$ then $\mathrm{tr}(U) = -1 + \sum_{j=2}^{n} a_j < -1$. Also, if A is orthogonal and $\det(A) = -1$, then -1 is an eigenvalue of A. This follows immediately from the fact that $\det(A)$ is the product of the eigenvalues of A, repeated to their respective multiplicities. Finally, if A is orthogonal, $n \times n$, and n is even, then $\det(A) = -1$ implies that both -1 and 1 are eigenvalues of A.

Theorem 7. Let n and k be positive integers, $1 \leq k < n$, let U be an $n \times n$ orthogonal matrix, and let $B = (I_k|Z)U$ be such that $D_B = D$. If $\overline{U} = \begin{pmatrix} I_{n-1} & Z \\ Z & -1 \end{pmatrix} U$ and if $\overline{B} = (I_k|Z)\overline{U}$, then $B = \overline{B}$, so that $D_{\overline{B}} = D_B = D$. Either U or $\overline{U}$ is similar to an $n \times n$ orthogonal matrix $U_1 \varepsilon \mathscr{J}$ such that $D_{B_1} = 0$, where $B_1 = (I_k|Z)U_1$.

Proof. Note that the matrix $\overline{U}$ differs from U only in that the last row of $\overline{U}$ is the negative of the last row of U. Clearly, since $k < n$, we have $\overline{B} = B$.

Now suppose that n is even. If $\det(U) = -1$, then 1 and -1 are eigenvalues of U and thus, by Proposition 4, there exists an orthogonal matrix U_1 similar to U such that $B_1 = (I_k|Z)U_1 \varepsilon \mathscr{C}$ and $D_{B_1} = 0$. If $\det(U) = 1$, then $\det(\overline{U}) = -1$, and the above argument applied to $\overline{U}$ yields the same conclusion.

Suppose that n is odd. Then U must have at least one real eigenvalue, λ. If $\lambda = 1$ and if U has another eigenvalue $a + bi$, $a \leq 0$, then the conclusion follows from (1) of Corollary 6. Similarly, if $\lambda = -1$ and if U has another eigenvalue $a + bi$, $a \geq 0$, then the conclusion follows from (2) of Corollary 6. Suppose now that $\lambda = 1$ is an eigenvalue of U and that $a > 0$ for all other eigenvalues $a + bi$ of U. Then $\det(U) = 1$ and $\mathrm{tr}(U) > 1$. Since $\det(\overline{U}) = -1$, it follows that -1 is an eigenvalue of $\overline{U}$, and, since $\mathrm{tr}(\overline{U})$ can differ from $\mathrm{tr}(U)$ by at most 2, we have that $\mathrm{tr}(\overline{U}) > -1$. Thus, $\overline{U}$ must have an eigenvalue of the form $c + di$, where $c > 0$, and hence, by (2) of Corollary 6, there exists an orthogonal matrix

U_1, simi'ar to $\overline{U}$, such that $B_1 = (I_k|Z)U_1 \varepsilon \zeta$ and $D_{B_1} = 0$. The case in which $\lambda = -1$ is an eigenvalue of U and that $a < 0$ for all other eigenvalues $a + bi$ of U is handled in a similar manner, and we omit the proof.

3. Conclusion. This paper provides an example to show that, even under extremely strong conditions, the eigenvalues of U do not affect the value of divergence $D_{(I_k|Z)U}$ in the space of reduced dimension.

REFERENCES

[1] C. C. Babu, "On the application of divergence to feature selection in pattern recognition," IEEE Trans. Syst., Man, and Cybern., Vol. SMC-2, pp. 668-670, Nov. 1972.

[2] C. C. Babu and S. Kalra, "On feature extraction in multiclass pattern recognition," Int. J. Contr., Vol. 15, No. 3, pp. 595-601, 1972.

[3] D. R. Brown and M. J. O'Malley, "A counterexample in linear feature selection theory," IEEE Trans. Syst., Man, and Cybern., Vol. SMC-6, No. 1, pp. 59-61, Jan. 1976.

[4] H. P. Decell and J. A. Quircin, "An iterative approach to the feature selection problem," in Proc. IEEE Conf. Machine Processing of Remotely Sensed Data, Purdue Univ., Oct. 1973, IEEE Cat. #CM0834-2GE, pp. 3B1-3B12.

[5] K. Fukunaga and W. Koontz, "Application of the Karhunen-Loève expansion to feature selection and ordering," IEEE Trans. Comput., Vol. C-19, pp. 311-318, Apr. 1970.

[6] T. L. Henderson and D. G. Lainiotis, "Comments on linear feature extraction," IEEE Trans. Inform. Theory, Vol. IT-15, pp. 728-730, Nov. 1969.

[7] T. T. Kadota and I. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," IEEE Trans. Inform. Theory, Vol. IT-13, pp. 278-284, Apr. 1967.

[8] T. Kailath, "The divergence and Battacharyya distance measures in signal detection," IEEE Trans. Commun. Technol., Vol. COM-15, pp. 52-60, Feb. 1967.

[9] S. Kullback, Information Theory and Statistics. New York: Wiley, 1969.

[10] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition svstems," IEEE Trans. Inform. Theory, Vol. IT-9, pp. 11-17, Jan. 1963.

[11] J. Tou and R. Heydorn, "Some approaches to optimum feature extraction," in Computer and Information Sciences, Vol. 2, J. Tou, Ed. New York: Academic, 1967.

[12] F. W. Warner, Foundations of Differentiable Manifolds and Lie Groups. Glenview, Illinois: Scott, Foresman and Co., 1971.

[13] S. Watanabe, Ed., Methodologies of Pattern Recognition. New York: Academic, 1969.

[14] S. Watanabe, et al, "Evaluation and selection of variables in pattern recognition," in Computer and Information Sciences, Vol. 11, J. Tou, Ed., New York: Academic, 1967.

[15] T. Young, "Reliability of linear feature extractions," IEEE Trans. Comput., Vol. C-20, pp. 967-971, Sept. 1971.

A REVIEW OF THE LEC PERFORMANCE EVALUATION OF UHMLE

Henry P. Decell, Jr.
Department of Mathematics
University of Houston
Houston, Texas

&

William A. Coberly
University of Tulsa
Consultant to
Department of Mathematics
University of Houston
Houston, Texas

Report #62
NAS-9-15000

January 1977

A Review of the LEC Performance Evaluation of UHMLE

In March 1976, Lockheed was directed to submit a plan [1] for comparative evaluation of several candidate signature extensions algorithms. The results of that test [2], car ied out by LEC in April, were the basis for selection of two algorithms [3], OSCAR and ATCOR, for test and implementation in a sub-operational system by IBM. Four simulated (SIM) data sets and seven consecutive day (CD) data sets were used. In the following sections, two points will be addressed for each data set. 1) Analysis and evaluation of the UHMLE test. 2) Recommendations on changes in the UHMLE algorithm motivated by the test. The criterion for evaluation of each algorithm will be overall classification accuracy (Tables 8 and 9 of [2] are attached for convenience).

I. Simulated Data Test.

In previous tests carried out by the University of Houston consistently good results were observed using essentially the same data set. The poor performance of UHMLE on SIM1 and the marginal performance on SIM4 seems to contradict our previous experience. The following observation on the LEC test may explain this discrepency.

In SIM1 the iteration sequence seemed to converge before the signatures had moved into the unlabeled data region. A second run which first estimated an initial translation X + B and then applied the general UHMLE algorithm was successful. Even though translation was included in our operational algorithm delivered to JSC, the second run was not reported in the final LEC analysis.

Pass	Local Accuracy	1st LEC UHMLE TEST	2nd LEC UHMLE TEST w/translation option
SIM1	93.5	-21.7	-2.5
SIM2	98.6	-0.7	no trans.
SIM3	97.0	-1.0	" "
SIM4	92.8	-5.0	" "
Ave.	95.5	-7.1	-2.3
Std.		9.9	2.0

Table 1

Revised SIM test results.
Overall Accuracy Difference

The use of the translation in SIM1 would dramatically change the outlook of UHMLE in the SIM test.

The results do not suggest any modifications of the UHMLE algorithm except to re-state the need to apply the translation first.

II. Consecutive Day Test.

General: The consecutive day (CD) data set consisted of three Kansas Intensive Test Sites (ITS) outlined in [1]. From these a total of seven pairs of consecutive day passes were selected from 1973-74 LANDSAT-1 data acquisitions.

ITS	DATA SET ID	DATE TRAINING/RECOGNITION	SIZE ITS	HAZE	
				TRAINING	RECOGNITION
Finney	F1709-8	2/1 July 74	5 × 6		
"	F1673-2	27/26 May 74	"	X	
"	F1655-4	9/8 May 74	"		
"	F1726-7	19/20 July 74	"	X	
Saline	S1455-4	21/20 Oct 73	3 × 3		
"	S1725-4	18/17 July 74	"		X
Ellis	E1726-5	12/11 June 74	3 × 3		X

Table 2

Consecutive Day Data Sets

Two UHMLE tests were run on each data set. UH/ALL uses as its unlabeled sample the rectangular area containing the selected Test/Training fields. UH/FIELDS uses the test fields only as input. The following ground areas associated with each ITS are defined for further reference.

A0 - ITS ground truth site. (Not alligned with LANDSAT ground track.)

A1 - Smallest rectangular field containing selected training field. Used as input for UH/ALL.

A2 - A0 intersect A1 , used for classification area.

A3 - Designated test fields (≡ training fields within A2). Used for input to UH/FIELDS.

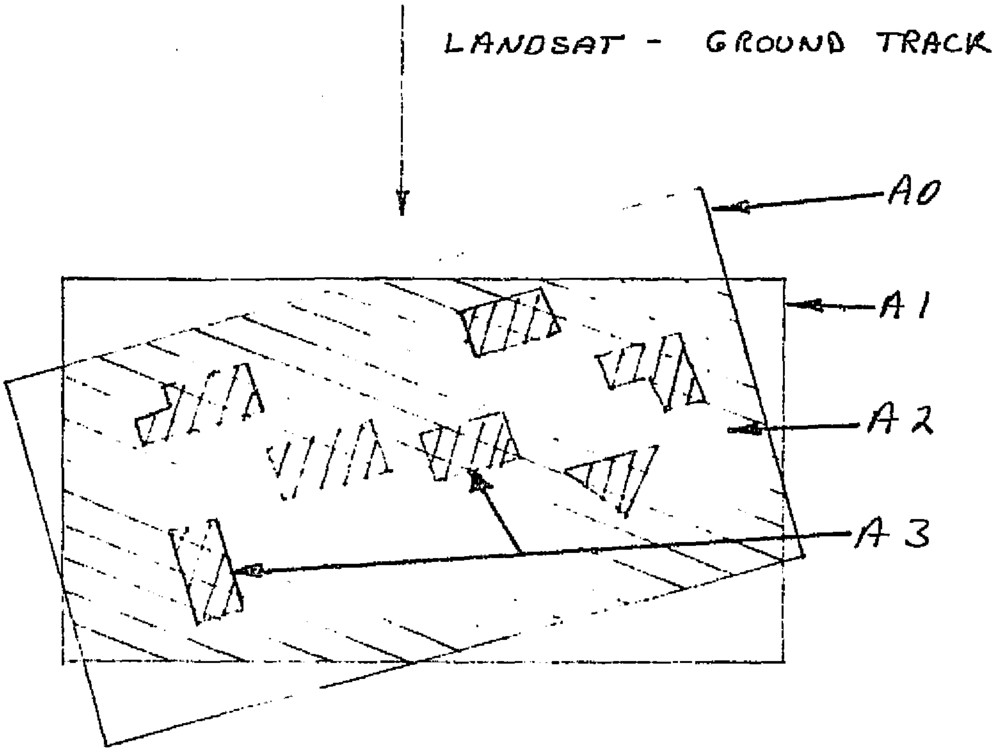


Figure 1

Ground area definitions

Proportion Estimates. UHMLE automatically estimates a proportion vector for the unlabeled input data set. These estimates are used in two ways in the Signature Extention (SE) test.

1) The UHMLE proportion estimates are used as *a priori* probabilities in the classification algorithm. Although this is not an unreasonable choice for the *a priori* probabilities, the UHMLE classification results are not comparable to those of the other candidate algorithms which used equally likely *a priori* probabilities. Moreover, in the UH/ALL test, the UHMLE proportion estimates correspond to Area A1. Area A2 was classified and only results from Area A3 were used for performance evaluation. In UH/FIELDS the unlabeled input data set and the classification region were equivalent.

2) In Tables 10-13 in [2], the estimated proportion of wheat for each algorithm is first compared to the local classification proportion estimate and then to the ground truth proportion estimate for both the SIM and CD data sets. In the CD test, the UH/ALL and UH/FIELDS are classification proportion estimates for area A2. The maximum-likelihood estimates from UHMLE (UH/ALL/MLE) correspond to area A1 . It is assumed here that the proportion estimate from local classification in Table 11 of [2] is based on A2. Hence UH/ALL/MLE is not comparable to the local standard. In Table 13 [2] the standard is ground truth. It is not clear whether or not the ground truth proportions correspond to A0 or A2 . In either case all proportion estimates listed in that table are not comparable.

Data Quality. This appears to be the most important factor in analyzing the UHMLE results. The CD data sets contained numerous data drops or "glitches." LEC was careful to choose training segments and fields so as to avoid this bad data in the computation of training statistics. However, several of the recognition segments used as input to UHMLE (in both UH/ALL and UH/FIELDS) were contaminated. This bad data effectively "captured" subclasses from both wheat and non-wheat categories and distorted means and particularly covariances in other subclasses. Only the data quality in Area A2 could be assessed from the available computer output. Further data drops, which may have been present in A1 (outside of A2), could also have an apparent degrading effect on UH/ALL test results. The implications and incidence of contaminated data is listed below in Table 3. We strongly recommend that this be the last time that this data set be used in any testing procedure.

Data Set	UH/FIELDS	UH/ALL
F 1709-8	Slight	Slight
F 1673-2	Bad	Bad
F 1655-4	Bad	Bad
F 1726-7	Bad	Bad
S 1455-4	Slight	Slight
S 1725-4	Good	Good
E 1726-5	Good	Good

Table 3

Incidence of Data Drops in CD Data Sets

Label Switching: In the UHMLE algorithm the various subclass statistics move in a quasi-independent manner to better "fit" the unlabeled data set. In this process a subclass component of the mixture model may seek out data in the unlabeled sample which is from a different category than the one assigned in the training segment. This poses no difficulty in terms of density estimation, however correct category labels are required for acreage proportion estimates. This phenomena is compounded by subclasses being "captured" by data drops, leaving unmodeled data free to be absorbed by an existing subclass. In a number of the CD tests substantially improved results are obtained if the label on a single subclass is reassigned. Interaction of the AI or DPA (at this point, prior to aggregation of acreage proportion estimates at the category level) with the view of detecting obvious category labeling errors, should be considered. This is a key point. We are simply saying that, when using UHMLE (or other algorithms), the spectral class identity extrapolated from the training segment may not be sufficient to establish crop category identity without AI interaction.

Individual CD Data Set Results. In this section each CD-data-set test is analyzed separately. Some revised results are reported along with supporting rationals.

F 1709-8 Two classes have inflated variances due to a data drop. However, both UH/ALL and UH/FIELDS do better than local classification.

F 1673-2 Very poor performance on both cases is observed. Two data drops have major effect on distorting variances and means on several subclasses. If one subclass, which is obviously mislabeled, is switched from wheat to non-wheat a substantial improvement is observed.

		LEC Test		Revised	
Local	UT	UH/FIELDS	UH/ALL	UH/FIELDS	UH/ALL
96.1	0.1	-23.7	-21.3	-3.1	-8.6

In Figure 2, the subclass means determined by UHMLE are plotted in the TACAP "brightness × green" coordinate system. Subclass W7 is clearly displaced from the other wheat subclasses. It is not unreasonable for mislabeling of this magnitude to be easily detected by an AI or DPA and corrected at the time of acreage estimation.

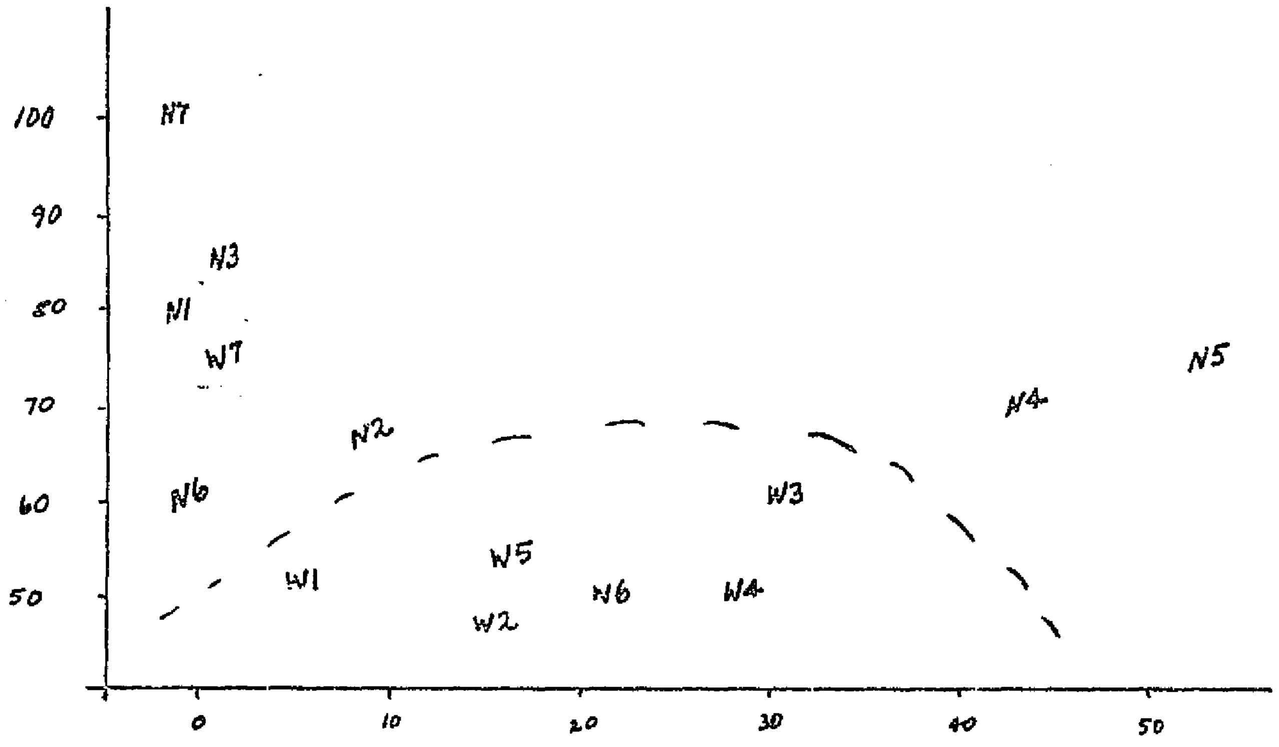


Figure 2. TACAP plot of class means.

FINNEY 1672 / UHMLE-FIELDS

F 1655-4   Again two data drops play a large role in distorting several subclass signatures in UH/ALL. One label switch again improves matters greatly. In UH/FIELDS the effects of

				Revised	
Local	UT	UH/FIELDS	UH/ALL	UH/FIELDS	UH/ALL
94.9	-3.8	-3.1	-15.0	not revised	-3.3

the data drops are not as apparent in the overall classification accuracy.

F 1726-7   Data drops substantially distort four subclasses in UH/ALL and to a lesser extent in UH/FIELDS. Even so, results are excellent (better than local classification) in UH/FIELDS. UH/ALL results are poor. No clear label switch is apparent.

S 1455-4   In this data set only four subclasses are modeled. Two subclasses are distorted by data drops, one severely in both cases. In the UH/ALL case the A1 area is much too large, introducing a large segment of extraneous data into the unlabeled sample. Further A2 is not contained in A1 (see Figure 3).

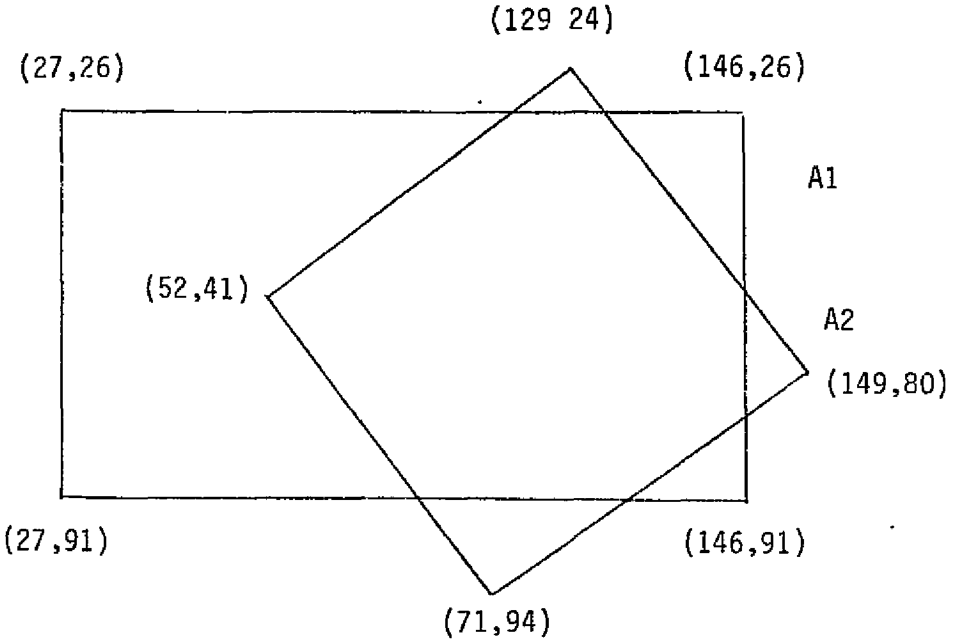


Figure 3.

Field Definition Errors in S 1455-4.

The poor data quality, errors in field definitions, and small number of subclasses render the interpretation of this test null and void. Inclusion of this test in the overall UHMLE evaluations is, therefore, meaningless.

S 1725-4 There are no data drops or anomolies in this test.

E 1726-5 There are no data drops. A reasonable case could be made for a label switch, however, the explanation is not as obvious as in the previous data sets and it will be omitted here. This case appears to be a reasonable test of the algorithm.

Summary of CD Test. If we introduce the three label changes (easily detected by an AI or DPA) suggested in F 1673-2 and F 1655-4 and omit the unacceptable test of S 1455-4, the performance of the algorithm is distinctly different than that reported in [2]. In light of the results presented here, the conclusions drawn by LEC in [2] concerning the relative performance of UHMLE are, at best, questionable. The original results along with the aforementioned revision and omission are listed in Table 4 below.

		LEC Original		Revised	
Data Set	Local	UH/FIELDS	UH/ALL	UH/FIELDS	UH/ALL
F 1709-8	79.5	2.7	7.3	same	same
F 1673-2	96.1	-21.3	-23.7	-3.1	-8.6
F 1655-4	94.9	-3.1	-15.0	same	-3.3
F 1726-7	80.0	0.9	-6.8	same	same
S 1455-4	86.5	-12.1	-29.5	OMIT	OMIT
S 1725-4	85.4	-4.3	0.9	same	same
E 1726-5	66.2	1.4	-7.3	same	same
Mean		-5.1	-10.6	-0.92	-2.97
Std. Dev.		8.7	13.1	2.9	6.1

Table 4.

Revised UHMLE Test Results.
Overall Classification Accuracy Differences.

We maintain that there is considerable evidence (provided, in part, by this analysis) for rejecting the original analysis and conclusions. If for no other reason, the poor data quality in five of the seven CD data sets chosen renders the LEC test results, as they pertain to UHMLE, invalid.

III. Conclusions.

Although the LANDSAT-2 data does not contain nearly the frequency of data drops observed in the LANDSAT-1 data used for this test, we clearly must incorporate a data editing scheme into the UHMLE algorithm or assume that preprocessing has deleted these pixels. There has been preliminary testing of a thresholding scheme which appears to be an adequate method when used in conjunction with an initial $X + B$ translation.

The reassessment of labels after signature extension remains a major priority in the UHMLE signature extension algorithm. This is a small task in terms of time compared to complete local training by the AI, and appears to be a necessary AI interaction function coupled with automatic processing of recognition segments.

SUMMARY

Our comments on the SD test and on the CD test suggest that the UHMLE algorithm _in particular_ and mixture density estimation _in general_ should still play an important role in the solution of the signature extension problem. In another paper [4], the signature (e.g., Procedure 1) extension problem, in the context of the LACIE training procedure is reformulated. Mixture density estimation (supervised or unsupervised) will certainly play a role in the exaction of the _Spectral Information Classes_ described in that paper. Additional work on the UHMLE algorithm, especially the details of incorporating it into the LACIE training procedure, we believe to be essential. These details are treated in the reformulation given in [4].

REFERENCES

1. Plan for Evaluating Several Signature Extension Algorithms, LEC Memorandum, April 1, 1976 Ref: 642-1877.

2. Performance Tests of Signature Extension Algorithms, LEC Ref: 642-2018 September 1976.

3. Selection of Signature Correction Algorithms for Implementation and Test by IBM. EOD Memorandum TF3/K. Baker: db: 4/26/76: 2071 April 30, 1976.

4. Henry P. Decell, Jr. and W. A. Coberly, On Signature Extension, Mathematics Department, University of Houston, December 1976 (in printing).

TABLE 8.– OVERALL ACCURACY FOR SIMULATED DATA*

[A minus sign means the algorithm was less accurate than local classification.]

Data	Local accuracy	Percentage difference between local accuracy and that obtained with various algorithms				
		R(S)	MLEST	UH fields	R(C)	UT
SIM1	93.5	0.0	-3.5	-21.7	-29.6	-99.3
SIM2	98.6	0.0	0.0	-0.7	0.0	-18.3
SIM3	97.0	0.1	0.0	-1.0	-5.2	-50.0
SIM4	92.8	-0.1	-3.2	-5.0	-2.9	-8.8
Mean	95.5	0.0	-1.7	-7.1	-9.4	-44.1
Std. dev.	2.8	0.1	1.9	9.9	13.6	40.8

*Prepared by LEC [2].

TABLE 9.— OVERALL ACCURACY FOR CONSECUTIVE DAY DATA*

[A minus sign means the algorithm was less accurate than local classification.]

Data	Local accuracy	Percentage difference between local accuracy and that obtained with various algorithms											
		R(S)	MLEST	OSCAR	REGRES	MOD R	R(C)	MOD OSCAR	ATCOR	UH fields	UT	R(S/C)	UH all
F1709-8	79.5	-5.8	-4.4	-7.0	-7.1	-7.6	-8.1	-7.8	-8.5	2.7	-8.2	-12.5	7.3
F1673-2	96.1	-2.0	-0.5	-3.2	-10.2	0.5	-1.7	-0.7	-5.0	-21.3	0.1	-1.7	-23.7
F1655-4	94.9	-3.3	-1.8	-2.1	-2.1	-2.7	-4.7	-3.0	-3.6	-3.1	-3.8	-3.8	-15.0
F1726-7	80.0	1.9	1.7	3.8	4.9	-1.9	-1.1	2.4	-5.9	0.9	-8.5	-7.1	-6.8
S1455-4	86.5	-0.2	-0.9	-3.5	-1.8	-3.2	-4.4	-2.5	0.1	-12.1	0.0	-3.5	-29.5
S1725-4	85.4	1.1	-0.5	-0.9	0.0	-3.2	-1.9	-5.0	-4.7	-4.3	-14.1	-11.0	0.9
E1726-5	66.2	-3.2	-6.0	-3.8	-3.5	-1.8	-4.1	-9.8	-2.7	1.4	-11.5	-9.8	-7.3
Mean	84.1	-1.6	-1.8	-2.4	-2.8	-2.8	-3.7	-3.8	-4.3	-5.1	-6.6	-7.1	-10.6
Std. dev.	10.2	2.7	2.6	3.3	4.9	2.5	2.4	4.2	2.7	8.7	5.5	4.2	13.1

*Prepared by LEC [2].

On the Convergence of

Optimal Linear Combination Procedures

William Tally

Report #63
NAS-9-15000

January 1977

Introduction:

The following algorithm has been suggested by Decell and Smiley in [1] for optimal linear combinations in the feature selection problem.

Let Ψ be a continuous function from M_n^k (see definition 1) into R^1 that is invariant under multiplication on the left by kxk invertible matrices. Then there exists $H_1 \in \mathcal{H}_n$ (see definition 2) such that

$$\Psi([I_k|Z]H_1) = \underset{H\in\mathcal{H}_n}{\text{l.u.b.}}\{\Psi([I_k|Z]H)\}.$$

Now for each positive integer i, let the element $H \in \mathcal{H}_n$ be chosen such that

$$\Psi([I_k|Z]H_iH_{i-1}\cdots H_1) = \underset{H\in\mathcal{H}_n}{\text{l.u.b.}}\ \Psi([I_k|Z]H\cdot H_{i-1}\cdots H_1)$$

The question of whether or not the above process terminates at an absolute Ψ-extremum (rank k maximal statistic) appeared in [1]. In this paper, we show that there exists a function Ψ as above for which the above process does not terminate at an absolute Ψ-extremum.

Let $H_1,\ldots,H_p$ be the matrices representing Householder transformations. Then for the matrix $[I_k|Z]H_1\cdots H_p$, let $\Theta([I_k|Z]H_1\cdots H_p)$ be the span in R^n of the k **row** vectors of that matrix. Suppose that $v_1,\ldots,v_k$ are linearly independent vectors in R^n. Then we show in this paper that there exists some integer $p \leq \min(n,n-k)$ and Householder transformations whose matrices are $H_1,\ldots,H_p$ for which

$\Theta([I_k|Z]H_1\cdots H_p) = \text{Span}\{v_1,\ldots,v_k\}$. We also determine the minimum integer p having the above property.

Preliminaries:

Definition 1. Let M_n^k be the set of all kxn rank k matrices.

Definition 2. Let $\mathcal{H}_n$ denote the set of all Householder transformations.

Definition 3. Let $\mathcal{S}_n^k$ denote the collection of all vector subspaces of R^n of dimension k.

Definition 4. Let $S^n = \{x \in R^n \mid \|x\| = 1\}$.

Definition 5. Let $\mathcal{C}$ be a closed subset of R^n and $x \notin \mathcal{C}$. Then there exists $c_x \in \mathcal{C}$ such that $\|x-c_x\| \leq \|x-c\|$ for any $c \in \mathcal{C}$. Let $\rho(x;\mathcal{C}) = \|x-c_x\|$.

Definition 6. Let A and B be elements of $\mathcal{S}_n^k$. Then there exists an element $a^* \in A \cap S^n$ having the property that $\rho(a^*; B\cap S^n) \geq \rho(a; B\cap S^n)$ for all $a \in A\cap S^n$. The number $\rho(a^*; B\cap S^n)$ will be called the distance from A to B and will be denoted by the symbol $d(A;B)$.

Proposition 1. For any elements A, B, and C in $\mathcal{S}_n^k$

i) $d(A;B) \geq 0$ and $d(A;B) = 0$ if and only if $A = B$.

ii) $d(A;C) \leq d(A;B) + d(B;C)$.

iii) For any $\xi > 0$ there exists a $\delta > 0$ such that whenever $d(A;B) < \delta$, then $d(B;A) < \xi$.

Definition 7. For any $P \in \mathcal{S}_n^k$ and $\xi > 0$, let

$$\mathcal{U}_\xi(P) = \{X \in \mathcal{S}_n^k \mid d(X;P) < \xi\}.$$

Definition 8. Let T be the topology on $\mathcal{S}_n^k$ determined by the subbasis $\{\mathcal{U}_\xi(P) \mid \xi > 0 \text{ and } P \in \mathcal{S}_n^k\}$.

Definition 9. Let $\mathcal{C}$ be a closed subset of $\mathcal{S}_n^k$ and let $P \in \mathcal{S}_n^k$. Let $D(P;\mathcal{C}) = \text{g.l.b.}\left\{d(P;C) \mid C \in \mathcal{C}\right\}$.

Proposition 2. $(\mathcal{S}_n^k, T)$ is normal.

Proof: Let $\mathcal{A}$ and $\mathcal{B}$ be two closed disjoint subsets of $\mathcal{S}_n^k$. Let $\mathcal{U}_1 = \left\{P \in \mathcal{S}_n^k \mid D(P;\mathcal{A}) < D(P;\mathcal{B})\right\}$ and $\mathcal{U}_2 = \left\{P \in \mathcal{S}_n^k \mid D(P;\mathcal{A}) > D(P;\mathcal{B})\right\}$. By Proposition 1, we can determine that $\mathcal{U}_1$ and $\mathcal{U}_2$ are both open and are disjoint. This completes the proof.

Definition 10. For any vector $w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$ in R^n, let $w^U = \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}$ and $w^L = \begin{pmatrix} w_{k+1} \\ \vdots \\ w_n \end{pmatrix}$.

Proposition 3. Suppose that $\left\{v_1, \ldots, v_k\right\}$ is a collection of linearly independent vectors in R^n. Let p be the dimension of $\text{Span}\left\{v_1^L, \ldots, v_k^L\right\}$ and assume $p > 0$. Then there exists a vector $x \in R^n$ such that $\|x\| = 1$, and if H_x is the Householder transformation determined by x, then the dimension of $\text{Span}\left\{H_x(v_1)^L, \ldots, H_x(v_k)^L\right\} = p-1$.

Proof: Case i) Dimension of $\text{Span}\left\{v_1^U, \ldots, v_k^U\right\}$ is less than k. We select a vector x^L in $\text{Span}\left\{v_1^L, \ldots, v_k^L\right\}$ such that $\|x^L\| = \sqrt{\tfrac{1}{2}}$. Since $\left[v_i^L - 2(v_i^L \cdot x^L)x^L\right] \cdot x^L = 0$ for $i=1,\ldots,k$. It follows that the dimension of $\text{Span}\left\{v_1^L - 2(v_1^L \cdot x^L)x^L, \ldots, v_k^L - 2(v_k^L \cdot x^L)x^L\right\}$ is $p-1$. Now by assumption there exists a vector x^U in R^k such that $\|x^U\| = \sqrt{\tfrac{1}{2}}$, and $v_i^U \cdot x^U = 0$ for $i=1,\ldots,k$. Since $v_i^L - 2(v_i \cdot x)x^L = v_i^L - 2(v_i^L \cdot x^L)x^L$, then the dimension of

Span $\left\{v^L-2(v_1^L\cdot x^L)x^L,\ldots,v_k^L-2(v_k^L\cdot x^L)x^L\right\}$ is p-1, for

$x = \begin{pmatrix} x^U \\ \vdots \\ x^U \end{pmatrix}^1$.

Case ii) The dimension of Span$\left\{v_1^U,\ldots,v_k^U\right\} = k$.
We select a vector x_o^L in Span $\left\{v_1^L,\ldots,v_k^L\right\}$ with $\|x_o^L\|=\sqrt{\tfrac{1}{2}}$.
Then we have that the dimension of
Span $\left\{v_1^L-2(v_1^L\cdot x_o^L)x_o^L,\ldots,v_k^L-2(v_k^L\cdot x_o^L)x_o^L\right\}$ is p-1. We assume then that $x^L = \lambda x_o^L$ for some $\lambda \angle 1$. We want a vector x^U in R^k such that if $x = \begin{pmatrix} x^U \\ x^L \end{pmatrix}$ then $\|x^U\|^2 + \|x^L\|^2 = 1$ and $v_i^L-2(v_i.x)x^L = v_i^L-2(v_i^L.x_o^L)x_o^L$ for i=1,...,k.
,

By substituting x_o^L into this equation in place of x^L we can determine that $v_i^U.x^U = (\frac{1-\lambda^2}{\lambda})v_i^L.x_o^L$ for i=1,...,k. By our assumption we can find a vector x^U satisfying the above equations whenever a choice of λ is made. We observe that if λ approaches 1, then $\|x^U\|$ must approach 0, and $\|x^L\|$ must approach $\sqrt{\tfrac{1}{2}}$ so that if λ approaches 1, then $\|x^U\|^2 + \|x^L\|^2$ must approach $\sqrt{\tfrac{1}{2}}$. If λ approaches 0, then $\|x^U\|$ approaches $+\infty$ and $\|x^L\|$ approaches 0 so $\|x^U\|^2 + \|x^L\|^2$ approaches $+\infty$ as λ approaches 0. It follows from this that there exists some λ for which $\|x^U\|^2 + \|x^L\|^2 = 1$. Thus we have the dimension of Span$\left\{v_1^L-2(v_1.x)x^L,\ldots,v_k^L-2(v_k.x)x^L\right\}$ is p-1 which is the required condition. This completes the proof of proposition 3.

/

Definition 11. For any $M \in M_n^k$ let $\Theta(M) = \text{Span}\{v_1, \ldots, v_k\}$ where $\{v_1, \ldots, v_k\}$ are the row vectors of M. Θ is easily seen to be continuous.

Proposition 4. Suppose that $\Theta([I_k | Z] H_1 \ldots H_p) = \text{Span}\{v_1, \ldots, v_k\}$ for Householder transformations $H_1, \ldots, H_p$. Then the dimension of $\text{Span}\{v_1^L, \ldots, v_k^L\}$ cannot exceed p.

Proof: We observe first of all that for any collection of vectors $\{y_1, \ldots, y_m\}$ and any Householder transformation H_x determined by the vector x that

$\text{Span}\{H_x(y_1), \ldots, H_x(y_m)\} \subset \text{Span}\{y_1, \ldots, y_m, x\}$ ..

Now $\Theta([I_k | Z] H_1 \ldots H_p) = \text{Span}\{H_p \ldots H_1(e_1), \ldots, H_p \ldots H_1(e_k)\}$ where e_i is the vector with 1 in the i^{th} place and 0 everywhere else. Thus by the above statements,

$\text{Span}\{v_1, \ldots, v_k\} \subset \text{Span}\{e_1, \ldots, e_k, x_1, \ldots, x_p\}$.

It follows that $\text{Span}\{v_1^L, \ldots, v_k^L\} \subset \text{Span}\{x_1^L, \ldots, x_p^L\}$.

Thus the dimension of $\text{Span}\{v_1^L, \ldots, v_k^L\}$ is less than or equal to p. This completes the proof of Proposition 4.

Proposition 5. For linearly independent vectors $\{v_1, \ldots, v_k\}$, if p is the dimension of $\text{Span}\{v_1^L, \ldots, v_k^L\}$ and $p > 0$, then there exists Householder transformations $H_1, \ldots, H_p$ such that $\Theta([I_k | Z] H_1 \ldots H_p) = \text{Span}\{v_1, \ldots, v_k\}$ and no fewer than p Householder transformations can have this property.

Proof: This is a consequence of Propositions 3 and 4 .

Construction of the map Ψ

Definition 12. For any $P \in \mathcal{S}_n^k$ let $P = \text{Span}\{v_1, \ldots, v_k\}$ and define $L(P)$ = the dimension of $\text{Span}\{v_1^L, \ldots, v_k^L\}$.

Definition 13. For $0 \le p \le n-k$ let $\mathcal{L}_p = \{A \in \mathcal{S}_n^k \mid L(A) \le p\}$.

Proposition 6. $\mathcal{L}_p$ is closed for $p=0,\ldots,n-k$.

Proof: This is a consequence of the fact that if $\{u_1, \ldots, u_m\}$ is a collection of vectors in R^{n-k} and q is the dimension of $\text{Span}\{u_1, \ldots, u_m\}$ then there exists a real number $\xi > 0$ such that if $\|u_i - u_i^*\|$ for $i=1,\ldots,m$, then the dimension of $\text{Span}\{u_1^*, \ldots, u_m^*\}$ is greater than or equal to q. This completes the proof of Proposition 6.

Now for some $P \in \mathcal{L}_1$ there exists $\xi > 0$ such that if $A \in \mathcal{L}_1$, then $\mathcal{U}_\xi(A)$ does not contain P. Let $\mathcal{A}$ be the closure in $\mathcal{S}_n^k$ of $\bigcup_{A \in \mathcal{L}_1}\{\mathcal{U}_\xi(A)\}$. By Urysohns lemma, [2] there exists a continuous function $\phi_1 : \mathcal{S}_n^k \to [0,1] \subset R^1$ such that $\phi_1(P) = 1$ and $\phi_1(A) = 0$ for any $A \in \mathcal{A}$. Let $I = \text{Span}\{e_1, \ldots, e_k\}$. Then $\mathcal{U}_\xi(I) \subset \mathcal{A}$ since $I \in \mathcal{L}_1$. Define a map $\phi_2 : \mathcal{S}_n^k \to [0, \frac{1}{2}]$ by $\phi_2(X) = 0$ if $X \notin \mathcal{U}_\xi(I)$ and $\phi_2(X) = \frac{\xi - d(X;I)}{2\xi}$ if $X \in \mathcal{U}_\xi(I)$. Let $\phi = \phi_1 + \phi_2$ and define $\Psi = \phi \circ \Theta$. We observe that $\mathcal{L}_1 = \Theta(\{[I_k|Z]H \mid H \in \mathcal{H}_n\})$. Also if $\Theta([I_k|Z]H_1) = I$ for some $H_1 \in \mathcal{H}_n$ then for any $H \in \mathcal{H}_n$, $\Theta([I_k|Z]H.H_1) \in \mathcal{L}_1$. That Ψ has the desired properties follows from the fact that the function ϕ has a maximum value of $\frac{1}{2}$ at I over the set $\mathcal{L}_1$ but ϕ has a maximum value of 1 at P over the entire space $\mathcal{S}_n^k$.

/

REFERENCES

1. Decell, H. P. and Smiley, W. G.III, Householder Transformations and Optimal Linear Combinations, 1974, Report #38, University of Houston Mathematics Department.

2. Royden, H. L., Real Analysis, page 148, 1970, Macmillan Company, London.

3. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York.

4. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

5. Quirein, J. A., "Divergence and Necessary Condition for Extremum" Report #12 NAS-9-12777 University of Houston, Department of Mathematics, Nov. 1972.

Sufficient Statistics for Mixtures of Measures in a Homogeneous Family

By

Charles Peters

Department of Mathematics

University of Houston

March, 1977
Report 64

Sufficient Statistics for Mixtures of Measures in a Homogeneous Family

by

Charles Peters
Department of Mathematics
University of Houston

1. Introduction:

Let $(X,\mathcal{A})$ and $(Y,\mathcal{B})$ be measureable spaces and let $T : X \to Y$ be surjective and measureable. Let $\mathcal{M}$ be a set of finite positive measures on $(X,\mathcal{A})$. For each $\mu \in \mathcal{M}$ there corresponds a measure μT^{-1} on $(Y,\mathcal{B})$ defined for $F \in \mathcal{B}$ by

$$\mu T^{-1}(F) = \mu(T^{-1}(F)).$$

If f is a μ-integrable real valued function on X, then as a consequence of the Radon Nikodym Theorem, there is a μT^{-1}- integrable function $e_\mu(f)$ on Y satisfying

$$\int_F e_\mu(f)d\mu T^{-1} = \int_{T^{-1}(F)} fd\mu$$

for each $F \in \mathcal{B}$. Clearly $e_\mu(f)$ is defined only up to sets in Y of μT^{-1} measure 0 and $f = g$ a.e. (μ) implies $e_\mu(f) = e_\mu(g)$ a.e. (μT^{-1}). The linear operator e_μ defined as above maps the space $\mathcal{L}^1(X,\mathcal{A},\mu)$ to the space $\mathcal{L}^1(Y,\mathcal{B},\mu T^{-1})$ and is called the conditional expectation operator. Its value

$e_\mu(f)$ at $f \varepsilon \mathcal{L}'(X, \mathcal{A}, \mu)$ is called the *conditional expectation of f given T*.

The conditional probability of an event $E \varepsilon \mathcal{A}$ is defined as

$$P_\mu(E) = e_\mu(\chi_E)$$

where χ_E is the indicator function of E. The conditional probability functions satisfy

(a) $$P_\mu : \mathcal{A} \to \mathcal{J}(Y, \mathcal{B}, \mu T^{-1}).$$

where $\mathcal{J}(Y, \mathcal{B}, \mu T^{-1})$ is the set of all real valued $\mathcal{B}$-measureable functions on Y, with equality defined as equality a.e. (μT^{-1}).

(b) For each $F \varepsilon \mathcal{B}, E \varepsilon \mathcal{A}$,

$$\mu(E \cap T^{-1}(F)) = \int_F P_\mu(E) d\mu T^{-1}$$

(c) $0 \leq P_\mu(E) \leq 1$ for each $E \varepsilon \mathcal{A}$ and $P_\mu(X) = 1$.

(d) If $\{E_n\}_{n=1}^{\infty}$ is a disjoint sequence of events in $\mathcal{A}$,

$$P_\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P_\mu(E_n) \quad \text{a.e. } (\mu T^{-1}).$$

It should be noted that P_μ satisfies property (c) even when μ is not a probability measure.

The transformation T is called a *sufficient statistic* for $\mathcal{M}$ if for each $E \varepsilon \mathcal{A}$ there is a $\mathcal{B}$-measureable function $P(E)$ on Y such that for each $\mu \varepsilon \mathcal{M}$, $P_\mu(E) = P(E)$ a.e., (μT^{-1}). The set $\mathcal{M}$ is *dominated* by a measure λ (perhaps not in $\mathcal{M}$) if for each $\mu \varepsilon \mathcal{M}$, μ is absolutely continuous with respect to λ,(written $\mu << \lambda$.) $\mathcal{M}$ is <u>homogeneous</u> if it is dominated by each of its members. A measure λ is <u>equivalent</u> to $\mathcal{M}$ if λ dominates $\mathcal{M}$ and $\mu(E) = 0$ for each $\mu \epsilon \mathcal{M}$ implies $\lambda(E) = 0$.

The notation and terminology used in this paper are taken from (Halmos and Savage; 1949), as are the following three theorems. The notation $\frac{d\mu}{d\lambda}(\epsilon)T^{-1}(\mathcal{B})$ means that there is an element of the equivalence class $\frac{d\mu}{d\lambda}$ of Radon-Nikodym derivatives which is $T^{-1}(\mathcal{B})$ measureable.

<u>Theorem</u> 1: If $\mathcal{M}$ is dominated, then a statistic T is sufficient for $\mathcal{M}$ if and only if there exists a measure λ equivalent to $\mathcal{M}$ such that for each $\mu \epsilon \mathcal{M}$, $\frac{d\mu}{d\lambda}(\epsilon)T^{-1}(\mathcal{B})$.

<u>Theorem 2</u>: If $\mathcal{M}$ is dominated, then a statistic T is sufficient for $\mathcal{M}$ if and only if T is sufficient for each pair $\{\mu,\nu\}$ of elements of $\mathcal{M}$.

<u>Theorem 3</u>: If $\mathcal{M}$ is homogeneous, then a statistic T is sufficient for $\mathcal{M}$ if and only if $\frac{d\mu}{d\nu}(\epsilon)T^{-1}(\mathcal{B})$ for each $\mu,\nu \epsilon \mathcal{M}$.

2. <u>Homogeneous</u> <u>Families</u>:

Henceforth, we will assume that $\mathcal{M}$ is homogeneous. Let $C(\mathcal{M})$ denote the cone generated by $\mathcal{M}$, excluding the zero measure. That is, $C(\mathcal{M})$ is the set of all finite linear combinations, with strictly positive coefficients, of elements of $\mathcal{M}$. Elements of $C(\mathcal{M})$ are termed <u>mixtures</u> of elements of $\mathcal{M}$. Clearly, $C(\mathcal{M})$ is also homogeneous; hence, the spaces $\mathcal{F}(Y,\mathcal{B},\mu T^{-1})$ are all the same for $\mu \epsilon C(\mathcal{M})$ and may be denoted simply by $\mathcal{F}$. For $\mu \epsilon C(\mathcal{M})$, P_μ maps $\mathcal{A}$ to $\mathcal{F}$ and it is clear from the definition of a sufficient statistic that T is sufficient for a subset $\mathcal{N}$ of $C(\mathcal{M})$ if and only if the conditional probability functions P_μ for $\mu \in \mathcal{N}$ are all equal.

Lemma 4: If $\mathcal{M}$ is dominated, $\mathcal{N} \subset C(\mathcal{M})$, and T is sufficient for $\mathcal{M}$, then T is sufficient for $\mathcal{N}$.

Proof: Let λ be that measure equivalent to $\mathcal{M}$ whose existence is assured by Theorem 1. If $\mu \in C(\mathcal{M})$, then μ can be written

$$\mu = \sum_{i=1}^{k} \beta_i \nu_i$$

with $\beta_i > 0$, $\nu_i \in \mathcal{M}$ for $i = 1,\ldots,k$. Hence,

$$\frac{d\mu}{d\lambda} = \sum_{i=1}^{k} \beta_i \frac{d\nu i}{d\lambda} \quad (\epsilon)\ T^{-1}(\mathcal{B}).$$

Thus T is sufficient for $C(\mathcal{M})$ and hence is sufficient for $\mathcal{N}$.

In order to characterize sufficient statistics for $\mathcal{N} \subset C(\mathcal{M})$, it suffices, by Theorem 2, to consider a pair

$$\mu_I = \sum_{i\epsilon I} \beta_i \mu_i$$

and

$$\mu_J = \sum_{j\epsilon J} \beta_j \mu_j$$

in $\mathcal{N}$, where I and J are finite sets; $\beta_k > 0$ for $k \in I \cup J$; and the measures $\{\mu_i\}_{i\epsilon I}$ are distinct members of $\mathcal{N}$, as are the measures $\{\mu_j\}_{j\epsilon J}$.

The set C( ) of all finite mixtures of elements of $\mathcal{M}$ is said to be identifiable (Teicher, 1960, 1961; Yakowitz 1969) if each element of $C(\mathcal{M})$ can be expressed in only one way as a linear combination with positive coefficients of elements of $\mathcal{M}$, except for the order of the summands. Equivalently, $C(\mathcal{M})$ is identifiable if the set $\mathcal{M}$ is linearly independent over the real numbers.

The concept of identifiability is very important in establishing the uniqueness and consistency of various estimators of the so called mixing parameters $\{\beta_i : i \varepsilon I\}$ in a mixture μ_I (Yakowitz, 1969).

Given a mixture μ_I in $C(\mathcal{M})$ we have for each $E \varepsilon \mathcal{A}$, $F \varepsilon \mathcal{B}$,

$$\int_F P_{\mu_I}(E)\, d\mu_I T^{-1} = \mu_I\, (E \cap T^{-1}(F))$$

$$= \sum_{i \varepsilon I} \beta_i \mu_i\, (E \cap T^{-1}(F))$$

$$= \sum_{i \varepsilon I} \beta_i \int_F P_{\mu_i}(E)\, d\mu_i\, T^{-1}$$

$$= \sum_{i \varepsilon I} \beta_i \int_F P_{\mu_i}\, (E)\, \frac{d\mu_i\, T^{-1}}{d\mu_I\, T^{-1}}\, d\mu_I\, T^{-1}.$$

Let $I_1, \ldots, I_r$ be the equivalence classes in I modulo the relation $i \equiv k$ if and only if $P_{\mu_i} = P_{\mu_k}$; that is, if and only if T is sufficient for the pair $\{\mu_i, \mu_k\}$. Then we have

$$\sum_{i \varepsilon I} \beta_i \int_F P_{\mu_i}(E)\, \frac{d\mu_i T^{-1}}{d\mu_I T^{-1}}\, d\mu_I T^{-1}$$

$$= \int_F \sum_{\ell=1}^{r} \sum_{i \varepsilon I_\ell} \beta_i \frac{d\mu_i T^{-1}}{d\mu_I T^{-1}} P_{\mu_{I_\ell}}(E)\, d\mu_I\, T^{-1},$$

where $P_{\mu_{I_\ell}}(E)$ is the common value of the $P_{\mu_i}(E)$ for $i \varepsilon I_\ell$. Thus,

$$P_{\mu_I} = \sum_{\ell=1}^{r} \frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} P_{\mu_{I_\ell}}$$

where μ_{I_ℓ} is the mixture

$$\mu_{I_\ell} = \sum_{i \varepsilon I_\ell} \beta_i \mu_i$$

Whenever the conditional probability function P_{μ_I} of a mixture μ_I is written in this fashion with $I_1, \ldots, I_r$ being equivalence classes modulo the relation $\equiv$, we will say that P_{μ_I} is written in normal form.

Definition 5: The set $C(\mathcal{M})$ is conditionally identifiable with respect to the statistic T if for each pair $\{\mu_I, \mu_J\}$ in $C(\mathcal{M})$, whenever $P_{\mu_I} = P_{\mu_J}$ and P_{μ_I}, P_{μ_J} are expressed in normal form

$$P_{\mu_I} = \sum_{\ell=1}^{r} \frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} P_{\mu_{I_\ell}}$$

$$P_{\mu_J} = \sum_{k=1}^{s} \frac{d\mu_{J_k} T^{-1}}{d\mu_J T^{-1}} P_{\mu_{J_k}} ,$$

then $r = s$ and for each $\ell = 1, \ldots, r$ there exists exactly one $k = 1, \ldots, r$ such that $\frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} = \frac{d\mu_{J_k} T^{-1}}{d\mu_J T^{-1}}$ and $P_{\mu_{I_\ell}} = P_{\mu_{J_k}}$. The set $C(\mathcal{M})$ is marginally identifiable with respect to T if the set $\{\mu T^{-1} | \mu \varepsilon \mathcal{M}\}$ is linearly independent over the real numbers.

Theorem 6: If $C(\mathcal{M})$ is both marginally identifiable and conditionally identifiable with respect to a statistic T, then $C(\mathcal{M})$ is identifiable.

Proof: Suppose $\mu_I = \sum_{i \varepsilon I} \beta_i \mu_i = \sum_{j \varepsilon J} \beta_j \mu_j = \mu_J$, where the measures in each sum are distinct members of $\mathcal{M}$. Then, expressed in normal form,

$$P_{\mu_I} = \sum_{\ell=1}^{r} \frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} P_{\mu_{I_\ell}} = \sum_{\ell=1}^{r} \frac{d\mu_{J_\ell} T^{-1}}{d\mu_J T^{-1}} P_{\mu_{J_\ell}} = P_{\mu_J},$$

and we may assume without loss of generality that

$$\frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} = \frac{d\mu_{J_\ell} T^{-1}}{d\mu_J T^{-1}}$$

and $$P\mu_{I_\ell} = P\mu_{J_\ell} \quad \text{for} \quad \ell = 1,\ldots,r.$$

Since $\mu_I T^{-1} = \mu_J T^{-1}$, it follows that $\mu_{I_\ell} T^{-1} = \mu_{J_\ell} T^{-1}$. For $i,k \in I_\ell$, $\mu_i T^{-1} \neq \mu_k T^{-1}$, for otherwise, since $P\mu_i = P\mu_k$, we would have $\mu_i = \mu_k$, contradicting the assumption that $\{\mu_i : i \in I\}$ are distinct. Similarly, the $\mu_j T^{-1}$ for $j \in J_\ell$ are all distinct. Since $C(\mathcal{M})$ is marginally identifiable, I_ℓ and J_ℓ have the same number of elements and for each $i \in I_\ell$ there is a unique $j(i) \in J_\ell$ such that $\beta_i = \beta_{j(i)}$ and $\mu_i T^{-1} = \mu_{j(i)} T^{-1}$. Since $P_{\mu_i} = P_{\mu_{j(i)}}$, it follows that $\mu_i = \mu_{j(i)}$ for each $i \in I_\ell$. Therefore, there is one to one map j from I onto J such that $\beta_{j(i)} = \beta_i$ and $\mu_{j(i)} = \mu_i$ for each $i \in 1$. Hence, $C(\mathcal{M})$ is identifiable, and the proof is complete.

For conditionally identifiable sets of measures, the following theorem and its corollary provide some characterizations of sufficient statistics.

Theorem 7: If $\mathcal{M}$ is homogeneous, $C(\mathcal{M})$ is conditionally identifiable with respect to a statistic T, and μ_I, μ_J are in $C(\mathcal{M})$, then T is sufficient for the pair μ_I, μ_J if and only if there exist partitions $I = I_1 \cup \ldots \cup I_r$ and $J = J_1 \cup \ldots \cup J_r$ such that for each $\ell = 1,\ldots,r$:

(a) $$d(\sum_{i\varepsilon I_\ell} \beta_i \mu_i)/ d(\sum_{j\varepsilon J_\ell} \beta_j \mu_j) = \frac{d\mu_{I_\ell}}{d\mu_{J_\ell}} = \frac{d\mu_I}{d\mu_J}$$

and

(b) T is sufficient for the set $N_\ell = \{\mu_k : k \varepsilon I_\ell \cup J_\ell\}$.

Proof: First suppose such partitions exist By (b) T is sufficient for the set N_1 and hence, by Lemma 4, it is sufficient for the pair $\{\mu_{I_1}, \mu_{J_1}\}$. It follows from (a) and Theorem 3 that T is sufficient for the pair $\{\mu_I, \mu_J\}$.

Suppose that T is sufficient for the pair $\{\mu_I, \mu_J\}$. Then, expressed in normal form,

$$\sum_{\ell=1}^{r} \frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} P_{\mu_{I_\ell}} = \sum_{\ell=1}^{r} \frac{d\mu_{J_\ell} T^{-1}}{d\mu_J T^{-1}} P_{\mu_{J_\ell}},$$

and we may assume without loss of generality that

$$\frac{d\mu_{I_\ell} T^{-1}}{d\mu_I T^{-1}} = \frac{d\mu_{J_\ell} T^{-1}}{d\mu_J T^{-1}} \text{ and } P_{\mu_{I_\ell}} = P_{\mu_{J_\ell}} \text{ for each } \ell.$$

The condition $P_{\mu_{I_\ell}} = P_{\mu_{J_\ell}}$ is equivalent to (b). By Theorem 3, there exists a representative $f \varepsilon \frac{d\mu_I}{d\mu_J}$ which is $T^{-1}(\mathcal{B})$ measureable. If $g \varepsilon \frac{d\mu_I T^{-1}}{d\mu_J T^{-1}}$, then $g\cdot T$ is $T^{-1}(\mathcal{B})$ measureable and for each $F \varepsilon \mathcal{B}$,

$$\int_{T^{-1}(F)} g\cdot T \; d\mu_J = \int_F g \; d\mu_J T^{-1} = \mu_I T^{-1}(F)$$

$$= \int_{T^{-1}(F)} f \; d\mu_J$$

It follows that $g \circ T = f$ a.e.(μ_J). Thus,

$$\frac{d\mu_I T^{-1}}{d\mu_J T^{-1}} \circ T = \{g \circ T \mid g \in \frac{d\mu_I T^{-1}}{d\mu_J T^{-1}}\} \subset \frac{d\mu_I}{d\mu_J} .$$

Since T is also sufficient for the pair $\{\mu_{I_\ell}, \mu_{J_\ell}\}$, a similar argument gives

$$\frac{d\mu_{I_\ell} T^{-1}}{d\mu_{J_\ell} T^{-1}} \circ T \subset \frac{d\mu_{I_\ell}}{d\mu_{J_\ell}}$$

for each ℓ. Since $\frac{d\mu_{I_\ell} T^{-1}}{d\mu_J T^{-1}} = \frac{d\mu_I T^{-1}}{d\mu_J T^{-1}}$ for each ℓ, it follows that (a) holds for each ℓ and the proof is complete.

Corollary 8: If $\mathcal{M}$ is homogeneous and $C(\mathcal{M})$ is conditionally identifiable with respect to a statistic T, then T is sufficient for a pair $\{\mu_I, \mu_J\}$ in $C(\mathcal{M})$ if and only if there exist subsets $I_1 \subset I$ and $J_1 \subset J$ such that:

(a) $$\frac{d\mu_{I_1}}{d\mu_{J_1}} = \frac{d\mu_I}{d\mu_J}$$

and

(b) T is sufficient for $N = \{\mu_k : k \in I_1 \cup J_1\}$.

Proof: That T sufficient implies the existence of I_1 and J_1 satisfying (a) and (b) is immediate from Theorem 7. Conversely if I_1 and J_1 satisfy

(a) and (b), then T is sufficient for μ_{I_1}, μ_{J_1} by (b) and hence, by (a), T is sufficient for μ_I, μ_J.

Given a pair of mixtures μ_I, μ_J in $C(\mathcal{M})$, we will call their likelihood ratio $\frac{d\mu_I}{d\mu_J}$ <u>indecomposable</u> if $I_1 \subset I$, $J_1 \subset J$ and $\frac{d\mu_{I_1}}{d\mu_{J_1}} = \frac{d\mu_I}{d\mu_J}$ imply $I_1 = I$ and $J_1 = J$. It is clear from Theorem 7 that if $C(\mathcal{M})$ is conditionally identifiable with respect to T and a pair of mixtures μ_I, μ_J in $C(\mathcal{M})$ have an indecomposable likelihood ratio, then T is sufficient for $\{\mu_I, \mu_J\}$ if and only if it is sufficient for $\{\mu_k : k \varepsilon I \cup J\}$. Also, it is not difficult to see that for each pair μ_I, μ_J in $C(\mathcal{M})$ there exist nonempty subsets $I_1 \subset I$ and $J_1 \subset J$ such that

$$\frac{d\mu_{I_1}}{d\mu_{J_1}} = \frac{d\mu_I}{d\mu_J}$$

and the likelihood ratio $\frac{d\mu_{I_1}}{d\mu_{J_1}}$ is indecomposable. If μ_I and μ_J represent the probability laws for two alternative hypotheses, then there would be two advantages in being able to identify subsets I_1 and J_1 satisfying these two criteria. First, the maximum likelihood decision procedure would be simplified, and second, the search for a statistic sufficient for deciding between the two hypotheses and having the property that $C(\mathcal{M})$ is conditionally identifiable could be restricted to those statistics sufficient for $\{\mu_k : I_1 \cup J_1\}$.

3. <u>Sufficient Linear Statistics for Mixtures of Normals</u>:

If $\mathcal{R}$ is a subring of the ring $\mathcal{J}$ introduced in Section 2, then with the usual definition of addition and multiplication by elements of $\mathcal{R}$ the set of all functions $\phi : \mathcal{A} \to \mathcal{J}$ is a module over $\mathcal{R}$. Thus, it is natural to consider *$\mathcal{R}$-independence* of a set $\mathcal{S}$ of such functions. To be precise, $\mathcal{S}$ is $\mathcal{R}$-independent if whenever $\phi_1,\dots,\phi_m$ is a finite set of distinct elements of $\mathcal{S}$ and $\gamma_1,\dots,y_m$ are elements of $\mathcal{R}$ such that

$$\gamma_1\phi_1(E) + \dots + \gamma_m\phi_m(E) = 0 \quad \text{for each } E \in \mathcal{A},$$

then $\gamma_1 = \dots = \gamma_m = 0$. If $\mathcal{R}'$ is a subring of $\mathcal{J}$ which contains all the bounded Radon-Nikodym derivatives $\frac{d\mu T^{-1}}{d\nu T^{-1}}$ for $\mu, \nu \in C(\mathcal{M})$, then it is clear that $\mathcal{R}$-independence of the set $\{P_\mu : \mu \in \mathcal{M}\}$ implies that $C(\mathcal{M})$ is conditionally identifiable with respect to T.

For the remainder of this section we will assume that X is $\mathbb{R}^n$, Y is $\mathbb{R}^k$ $(k \le n)$ and $T : X \to Y$ is linear and full rank. $\mathcal{A}$ and $\mathcal{B}$ are respectively, the Borel fields on $\mathbb{R}^n$ and $\mathbb{R}^k$. We also assume that each $\mu \in \mathcal{M}$ is described by a normal density function f_μ with mean m_μ and covariance Ω_μ. That is, for each $E \in \mathcal{A}$,

$$\mu(E) = \int_E f_\mu \, d\lambda_n,$$

where λ_n is Lebesgue measure on $\mathbb{R}^n$.

By a suitable choice of the coordinate system, we may represent the densities f_μ as joint density functions $f_\mu(y,z)$ on $\mathbb{R}^k \times \mathbb{R}^{n-k}$ while representing T as the projection $T(y,z) = y$. Then the marginal densities

$$g_\mu(y) = \int_{\mathbb{R}^{n-k}} f_\mu(y,z)dz$$

are normal with means Tm_μ and covariance matrices $T\Omega_\mu T^1$ (Anderson, 1958).

The conditional density functions

$$h_\mu(z \mid y) = \frac{f_\mu(y,z)}{g_\mu(y)}$$

are normal as functions of $z \in \mathbb{R}^{n-k}$ with means

$$(1) \qquad Sm_\mu + S\Omega_\mu T^1 (T\Omega_\mu T^1)^{-1} (y - Tm_\mu)$$

and covariances

$$(2) \qquad S\Omega_\mu S^1 - S\Omega_\mu T^1 (T\Omega_\mu T^1)^{-1} \; T\Omega_\mu S^1 .$$

where S is the linear operator $S(y,z) = z$. The conditional probabilities $P_\mu(E)$ are represented by

$$P_\mu(E \mid y) = \int_{S_y(E)} h_\mu(z|y)dz \quad .$$

where $S_y(E) = \{z \in \mathbb{R}^{n-k} \mid (y,z) \in E\}$.

<u>Theorem 9</u>: If $\mathcal{M}$ is a family of Borel measures on $\mathbb{R}^n$ given by n-variate normal density functions and $T : \mathbb{R}^n \to \mathbb{R}^k$ is linear of rank k, then $C(\mathcal{M})$ is conditionally identifiable with respect to T.

<u>Proof:</u> It can readily be verified that conditional identifiability of $C(\mathcal{M})$ is not affected by the change of variables just described. If μ_I and μ_J are in $C(\mathcal{M})$, then the Radon-Nikodym derivative $\frac{d\mu_I T^{-1}}{d\mu_J T^{-1}}$ is represented by a function of the form

$$\frac{g_I(y)}{g_J(y)} = \sum_{i \in I} \beta_i g_{\mu_i}(y) \Big/ \sum_{j \in J} \beta_j g_{\mu_j}(y);$$

i.e., a ratio of mixtures of k-variate normal density functions, which is continuous. Hence, by the remarks in the first paragraph of this section, it suffices to show that the set $\{P_\mu : \mu \in \mathcal{M}\}$ of conditional density functions is $\mathcal{R}$-independent, where $\mathcal{R}$ is the subring of $\mathcal{J}$ consisting of those elements of $\mathcal{J}$ which have a continuous representative. To this end, let $P_{\mu_1},\ldots,P_{\mu_r}$ be distinct and let $\gamma_1,\ldots,\gamma_r$ be continuous real valued functions on $\mathbb{R}^k$ such that for each $E \in \mathcal{A}$,

$$\gamma_1(y)P_{\mu_1}(E|y) +\ldots+ \gamma_r(y)P_{\mu_r}(E|y) = 0$$

for almost all y. In particular, choosing for E sets of the form $\mathbb{R}^k \times K$, where K is a borel set in $\mathbb{R}^{n-k}$, we have

$$\gamma_1(y) \int_K h_{\mu_1}(z|y)dz +\ldots+ \gamma_r(y) \int_K h_{\mu_r}(z|y)dz = 0$$

for almost all y. For each K, $\int_K h_{\mu_i}(z|y)dz$ is a continuous function of y. Hence,

$$\int_K (\gamma_1(y)h_{\mu_1}(z|y) +\ldots+ \gamma_r(y)h_{\mu_r}(z|y|)dz = 0$$

for each $y \in \mathbb{R}^k$. It follows that

$$\gamma_1(y)h_{\mu_1}(z|y) +\ldots+ \gamma_r(y)h_{\mu_r}(z|y) = 0$$

for each $y \in \mathbb{R}^k$, $z \in \mathbb{R}^{n-k}$. Let F be the set of $y \in \mathbb{R}^k$ where two or more of the conditional density functions $\bar{h}_{\mu_i}(z|y)$ are equal as functions of z. It is easily seen from (1) and (2) that the Lebesque measure of F is zero. For $y \notin F$, $\{h_{\mu_i}(\cdot|y),\ldots,h_{\mu_r}(\cdot|y)\}$ is a set of distinct normal density functions of z. Hence, (Yakowitz and Spragins; 1968), they are linearly independent over the real numbers. Therefore, for $y \notin F$, $\gamma_1(y) = \ldots = \gamma_r(y) = 0$. That is, $\gamma_1 = \ldots = \gamma_r = 0$ as elements of $\mathcal{F}$. Thus, $C(\mathcal{M})$ is conditionally identifiable.

If $\mu_I = \sum_{i\in I} \beta_i \mu_i$ is in $C(\mathcal{M})$, then μ_I has a density function

$$f_{\mu_I} = \sum_{i\in I} \beta_i f_{\mu_i}$$

which is a mixture of normal density functions. The following theorem is an immediate consequence of Theorems 7 and 9.

Theorem 10: Given the assumptions of Theorem 9, the statistic T is sufficient for a pair $\{\mu_I, \mu_J\}$ in $C(\mathcal{M})$ if and only if there exist partitions $I = I_1 \cup\ldots\cup I_r$ and $J = J_1 \cup\ldots\cup J_r$ such that for each $\ell = 1,\ldots,r$,

(a) $$\sum_{i\in I_\ell} \beta_i f_{\mu_i}(x)/ \sum_{j\in J_\ell} \beta_j f_{\mu_j}(x)$$

$$= \sum_{i\in I} \beta_i f_{\mu_i}(x)/ \sum_{j\in J} \beta_j f_{\mu_j}(x) \quad \text{for each } x \in \mathbb{R}^n,$$

and

(b) T is sufficient for the family $\{f_{\mu_k} : k \in I_\ell \cup J_\ell\}$ of normal density functions.

There is set of purely algebraic conditions which are equivalent to (b);

namely, that the expressions

$$\Omega_{\mu_k} - \Omega_{\mu_k} T^1 (T \Omega_{\mu_k} T^1)^{-1} T \Omega_{\mu_k}$$

$$m_{\mu_k} - \Omega_{\mu_k} T^1 (T \Omega_{\mu_k} T^1)^{-1} T m_{\mu_k}$$

$$\Omega_{\mu_k} T^1 (T \Omega_{\mu_k} T^1)^{-1} ,$$

are all independent of $k \in I_\ell \cup J_\ell$ (Peters, Redner, and Decell; 1976).

REFERENCES

1. Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons. New York.

2. Bahadur, R.R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* 25, 423-463.

3. Halmos, P.R. and Savage L.J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist. 20.* 225-241.

4. Peters, B.C., Redner R., and Decell, H.P. (1976). Characterizations of linear sufficient statistics. Tech. Report no. 59, Department of Mathematics , University of Houston.

5. Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* 32, 244-248.

6. Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist. 34*, 1265-1269.

7. Yakowitz, S. and Spragins, J. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* 39, 209-214.

8. Yakowitz, S. (1969). A consistent estimator for the identification of finite mixtures. *Ann. Math. Statist.* 4D, 1728-1735.

CHARACTERIZATIONS OF LINEAR SUFFICIENT STATISTICS

by

B. Charles Peters, Jr.,[1] Richard Redner,[1]
and Henry P. Decell, Jr.[1]

We develop necessary and sufficient conditions that a surjective bounded linear operator T from a Banach space X to a Banach space Y be a sufficient statistic for a dominated family of probability measures defined on the Borel sets of X. We give applications of these results that characterize linear sufficient statistics for families of the exponential type, including as special cases the Wishart and multivariate normal distributions. The latter result is used to establish precisely which procedures for sampling from a normal population have the property that the sample mean is a sufficient statistic.

[1] Author was partially supported by NASA/JSC Contract NAS-9-15000 with the University of Houston during the preparation of this work.

1. Introduction: Let T be a surjective measureable transformation from the measureable space $(X,\mathcal{A})$ to the measureable space $(Y,\mathcal{B})$, and let $\mathcal{D}$ be a set of totally finite measures on $\mathcal{A}$. Following Halmos and Savage [2], we say that T is a sufficient statistic relative to $\mathcal{D}$ if for each $E \in \mathcal{A}$ there exists a measureable function $P(E|\cdot) : (Y,\mathcal{B}) \to R$ (the real numbers) such that for each $F \in \mathcal{B}$, $\mu \in \mathcal{D}$

$$\mu(E \cap T^{-1}(F)) = \int_F P(E|y)d\mu T^{-1}(y) .$$

In another nonequivalent definition of a sufficient statistic given by Lehmann and Scheffe´[3], $\mathcal{B}$ is always taken to be $\mathcal{B}_T$, the largest σ-field on Y consistent with the measureability of T Bahadur [1] discusses the relationship between these two definitions at length.

In this paper our particular concern is that of developing necessary and sufficient conditions that a surjective bounded linear operator T from a Banach space X to a Banach space Y be a sufficient statistic, where $\mathcal{A}$ and $\mathcal{B}$ are the respective Borel fields of X and Y. Our first theorem shows that under a very natural condition the aforementioned definitions of sufficiency are equivalent. Specifically, the condition is that $\ker T = \{x \in X | Tx = \theta\}$ be complemented in X; that is, for some closed subspace S of X, $X = \ker T \oplus S$. (For example, if X is a Hilbert space, take $S = (\ker T)^{\perp}$.) As a corollary we obtain a simple characterization of sufficient linear statistics for dominated sets of measures. In Theorem 2 we replace the condition that ker T be complemented with conditions on the density functions corresponding to a dominated set $\mathcal{D}$. Finally, we give applications of these results that characterize linear sufficient statistics for families of the exponential type, including as special cases the Wishart and multivariate normal distributions. The latter result is used to establish precisely which procedures for sampling from a normal population have the property that the sample mean is a sufficient statistic. This generalizes the classical result that the sample mean is sufficient for independent samples. The final result deals with the connection between linear sufficient statistics and the Gauss-Markov theorem.

If W is a Banach space, $\mathcal{B}(W)$ will denote the Borel field generated by the open sets of W. The totally finite measures defined on $\mathcal{B}(W)$ will be denoted by $\mathcal{M}(W)$. We will write $\mu<<\nu$ for the relation of absolute continuity and $d\mu/d\nu$ for the equivalence class of Radon-Nikodym derivatives of μ with respect to ν. For the definitions of a dominated set of measures, equivalent sets of measures, and their connection with σ-finite measures defined on $\mathcal{B}(W)$, we refer the reader to Halmos and Savage [2].

2. Principal Results: Our first theorem shows that if ker T is complemented in S then, the two definitions of sufficiency described in the introduction are equivalent.

Theorem 1: Let X and Y be Banach spaces, let $A = \mathcal{B}(X)$ and let T be a surjective bounded linear operator from X to Y such that ker T is complemented in X . Then $\mathcal{B}_T + \mathcal{B}(Y)$.

Proof: Since T is Borel measureable, it suffices to show that $\mathcal{B}_T \subseteq \mathcal{B}(Y)$. Let S be a closed subspace of X such that $X = \ker T \oplus S$. If $F \in \mathcal{B}_T$, then $T^{-1}(F) \in \mathcal{B}(X)$ and if $\hat{T}$ denotes the restriction of T to S , then $\hat{T}^{-1}(F) = T^{-1}(F) \cap S \in \mathcal{B}(X)$. It follows that $\hat{T}^{-1}(F) \in \mathcal{B}(S)$, and since $\hat{T}$ is a topological isomorphism, $F = \hat{T}\hat{T}^{-1}(F) \in \mathcal{B}(Y)$.

Henceforth, we will assume that X and Y are Banach spaces; $\mathcal{A} = \mathcal{B}(X)$, $\mathcal{B} = \mathcal{B}(Y)$ and $T:(X,\mathcal{A}) \to (Y,\mathcal{B})$ is a surjective bounded linear operator. According to [2, Lemma 7], for a dominated collection of measures $\mathcal{D} \subset \mathcal{M}(X)$ a measure λ , equivalent to $\mathcal{D}$, can be defined by

$$\lambda(E) \equiv \sum_{i=1}^{\infty} a_i \mu_i(E)$$

where $\{\mu_i\}_{i=1}^{\infty}$ is a countable subset of $\mathcal{D}$ which is equivalent to $\mathcal{D}$ and $\sum_{i=1}^{\infty} a_i \mu_i(X) < \infty$. Obviously, if $\mathcal{D}$ is homogeneous, we can take $\lambda \in \mathcal{D}$. Combining the results of Theorem 1 with those of Lemma 2 and Theorem 1 of [2], we have:

Theorem 2: If ker T is complemented in X , then T is sufficient for $\mathcal{D}$ if and only if for each $\mu \in \mathcal{D}$ there exists a real valued function g_μ on Y such that $g_{\mu^\circ} T \in d\mu/d\lambda$.

Proof: By Theorem 1 of [2], T is sufficient if and only if for each $\mu \in \mathcal{D}$ there exists a real valued Borel measureable function g_μ on Y such that $g_{\mu^\circ} T \in d\mu/d\lambda$. Since ker T is complemented in X , $\mathcal{B}(Y) = \mathcal{B}_T$ and each real valued function g_μ such that

$g_\mu \circ T$ is Borel measureable on X must be Borel measureable on Y.

In all that follows $\delta g(x,z)$ will denote the Gateaux differential of the function g at x in the direction of z.

Corollary 1: If $\ker T$ is complemented in X, then T is sufficient for $\mathcal{D}$ if and only if for each $\mu \in \mathcal{D}$ there exists $f_\mu \in d\mu/d\lambda$ such that $x \in X$ and $y \in \ker T$ implies $\delta f_\mu(x;y) = 0$.

Proof: If T is sufficient, then for each $\mu \in \mathcal{D}$ there exists $g_\mu : Y \to R$ such that $f_\mu = g_\mu \circ T \in d\mu/d\lambda$. It follows immediately that $\delta f_\mu(x;y) = 0$ for each $x \in X$, $y \in \ker T$.

If $f_\mu \in d\mu/d\lambda$ and $\delta f_\mu(x;y) = 0$ for $\mu \in \mathcal{D}$, $x \in X$, $y \in \ker T$, then $f_\mu(x+y) = f_\mu(x)$ for each $x \in X$, $y \in \ker T$. For $z \in Y$ define $g_\mu(z) = f_\mu(x)$ where $z = Tx$. Then g_μ is well defined and $f_\mu = g_\mu \circ T$. Hence, T is sufficient.

The next theorem concerns a replacement of the complemented kernel condition whenever there is a continuous Radon-Nikodym derivative $f_\mu \in d\mu/d\lambda$ for each $\mu \in \mathcal{D}$.

Theorem 3: Let $V \subset X$ be an open set such that $\lambda(X \sim V) = 0$ and let $\lambda(U) > 0$ for each nonempty open subset U of V. Suppose $\lambda(B+y) = 0$ whenever $B \subset V$, $\lambda(B) = 0$ and $y \in \ker T$. For each $\mu \in \mathcal{D}$, let $f_\mu \in d\mu/d\lambda$ be continuous on V. Then T is sufficient if and only if $f_\mu(x) = f_\mu(z)$ whenever $x, z \in V$ and $Tx = Tz$.

Proof: If T is a sufficient statistic, then there exists $g_\mu \in d\mu/d\lambda$ such that $g_\mu(x) = g_\mu(z)$ whenever $x, z \in V$, $Tx = Tz$. Let $\mu \in \mathcal{D}$ and $y \in \ker T$ be fixed. The set

$$U = \{x \in V \cap (V-y) \mid f_\mu(x) \neq f_\mu(x+y)\}$$

is an open subset of V contained in $B \cup (B-y)$, where

$$B = \{x \in V \mid f_\mu(x) \neq g_\mu(x)\} \quad .$$

Since $\lambda(B) = 0$, it follows from the hypothesis that $\lambda(U) = 0$ and hence, $U = \emptyset$. Thus $f_\mu(x) = f_\mu(x+y)$ whenever $x, x+y \in V$.

Conversely, suppose $f_\mu(x) = f_\mu(z)$ for $\mu \in \mathcal{D}$, $x, z \in V$ whenever $Tx = Tz$. The function $g_\mu : T(V) \to R$ defined by $g_\mu(Tx) = f_\mu(x)$ for $x \in V$ is well defined on $T(V)$. Since f_μ is continuous on V, $f_\mu = g_\mu \circ T$ on V, and T is an open mapping, it follows that g_μ is continuous on the open set $T(V)$. For $y \notin T(V)$ define $g_\mu(y) = 0$. Then g_μ is Borel measureable on Y and $f_\mu = g_\mu \circ T$. Thus T is sufficient for $\mathcal{D}$.

The proof of the following corollary is clear and will be omitted.

Corollary 2: If, in addition to the hypotheses of Theorem 4, the set V is convex, then T is sufficient for $\mathcal{D}$ if and only if $\delta f_\mu(x;y) = 0$ for each $\mu \in \mathcal{D}$, $x \in V$, $y \in \ker T$.

3. Exponential Families: Let X and Y be Banach spaces, $(H, \langle\cdot|\cdot\rangle)$ a Hilbert space and ν a σ-finite measure on $\mathcal{B}(X)$ such that $\nu(X \sim V) = 0$ for some nonempty open convex set $V \subset X$ for which $\nu(U) > 0$ for each nonempty open set $U \subset V$. Let $\mathcal{D} = \{\mu_\gamma\}$, $\gamma \in \Gamma$ be a family of probability measures having exponential densities $f_\gamma(x) = c(\gamma)h(x) \exp \langle Q(\gamma)|t(x)\rangle \in d\mu_\gamma/d\nu$ where $c(\gamma) > 0$, $h(x) > 0$ on V a.e.(ν), $t : X \to H$ is continuous

and Gateaux differentiable on V , and $Q:\Gamma \to H$.

Theorem 4. Let $T:X \to Y$ be linear, bounded, surjective and $\nu(B+y) = 0$ whenever $B \varepsilon \mathcal{B}(X)$, $B \subset V$, $\nu(B) = 0$ and $y \varepsilon \ker T$. If $\beta \varepsilon \Gamma$, T is a sufficient statistic for the exponential family $\mathcal{D}$ if and only if $\langle Q(\gamma) - Q(\beta) | \delta t(x;y)\rangle = 0$ for each $\gamma \varepsilon \Gamma$, $x \varepsilon X$ and $y \varepsilon \ker T$.

Proof: Under the stated assumptions $\mathcal{D}$ is homogeneous and thus λ may be taken to be an arbitrary element, say μ_β , of $\mathcal{D}$. Applying Corollary 2, T is sufficient for $\mathcal{D}$ if and only if $\delta g_{\gamma,\beta}(x;y) = \theta$ for each $\gamma \varepsilon \Gamma$, $x \varepsilon V$ $y \varepsilon \ker T$, where

$$g_{\gamma,\beta}(x) = \frac{c(\gamma)}{c(\beta)} \exp \{\langle Q(\gamma) - Q(\beta) | t(x)\rangle\} .$$

This is equivalent to $\langle Q(\gamma) - Q(\beta) | \delta t(x;y)\rangle = 0$ for each $\gamma \varepsilon \Gamma$, $x \varepsilon V$, $y \varepsilon \ker T$.

4. Applications. Let $\mathcal{S}$ denote the symmetric $n \times n$ matrices, Γ the positive definite elements of $\mathcal{S}$ and $\mathcal{D}$ a family of Wishart probability measures with $m \geq n$ degrees of freedom having densities

$$f_\gamma(S) = c(\gamma)|S|^{(m-n-1)/2} \exp \{- \frac{1}{2} \operatorname{tr} (\gamma^{-1} S)\} .$$

Theorem 5. If $\beta \varepsilon \Gamma$ and $T:\mathcal{S} \to \text{range } (T)$ is linear, then T is a sufficient statistic for the Wishart family $\mathcal{D}$ if and only if $\operatorname{tr}[(\gamma^{-1} - \beta^{-1})K] = 0$ for each $\gamma \varepsilon \Gamma$ and $K \varepsilon \ker T$.

Proof. The preliminary conditions of Theorem 4. are satisfied with γ = Lebesgue measure on $\mathcal{S}$ and the obvious identifications of $c(\gamma)$

and $h(S)$. Let H equal $\mathcal{S}$ with $\langle A|B\rangle \equiv \operatorname{tr}(AB)$, $t(S) = S$ and $Q(\gamma)^* = -\gamma^{-1}/2$. Observe that $\delta t(S;F) = F$ and apply Theorem 4.

Remark: Theorem 5. implies that there is a nontrivial linear sufficient statistic if and only if there exists a linear manifold $\mathcal{M} \subsetneqq \mathcal{S}$ such that $\gamma^{-1} \in \mathcal{M}$ for each $\gamma \in \Gamma$.

We will now apply these results to normal families of probability measures. In Theorem 6. we will state set theoretical, algebraic and geometrical conditions, each equivalent to the condition that T be a linear sufficient statistic for a family $\mathcal{D} = \{P_\gamma\}$, $\gamma \in \Gamma$ of normal n-variate probability measures having densities, with respect to Lebesgue measure on R^n ,

$$p_\gamma(x) = (2\pi)^{-n/2}|\Omega_\gamma|^{-1/2} \exp\left[-\tfrac{1}{2}(x-\eta_\gamma)'\Omega_\gamma^{-1}(x-\eta_\gamma)\right]$$

We will assume that for some $\beta \in \Gamma$, $\eta_\beta = \theta$ and $\Omega_\beta = I$. This requirement imposes no loss of generality since for any $\beta \in \Gamma$ there exists a non singular matrix M_β for which $M_\beta \Omega_\beta M_\beta' = I$ and a change of coordinate system defined by the transformation $x \to M_\beta(x-\eta_\beta)$ allows one to recover the sufficient statistic in the original coordinate system.

Theorem 6. If $T:R^n \to R^k$ is a linear transformation of rank k and $\mathcal{D} = \{P_\gamma\}$, $\gamma \in \Gamma$ is an arbitrary family of n-variate normal probability measures such that for some $\beta \in \Gamma$, $\eta_\beta = \theta$ and $\Omega_\beta = I$ then the following conditions are equivalent:

(1) T is sufficient for $\mathcal{D} = \{P_\gamma\}$, $\gamma \varepsilon \Gamma$.

(2) $\ker T \subset \bigcap_{\gamma\varepsilon\Gamma} [\ker(\Omega_\gamma - I) \cap [\eta_\gamma]^\perp]$

(3) For each $\gamma \varepsilon \Gamma$,

(a) $T^+T\eta_\gamma = \eta_\gamma$

(b) $T^+T(\Omega_\gamma - I) = \Omega_\gamma - I$

where the notation $(\cdot)^+$ denotes the generalized inverse of $(\cdot)$.

Proof: To see that (1) → (2) observe that the preliminary conditions of Theorem 4. are satisfied with ν = Lebesgue measure on $X = R^n$. Make the obvious identifications for $c(\gamma)$ and $h(x)$. Let M_n denote the $n \times n$ real matrices and define $Q:\Gamma \to H = M_n \times R^n \times M_n$, $t:X \to H$ and $\langle\cdot|\cdot\rangle$ on H, respectively, by $Q(\gamma) = (-\Omega_\gamma^{-1}/2,\ \Omega_\gamma^{-1}\eta_\gamma,\ -\Omega_\gamma^{-1}\eta_\gamma\eta_\gamma/2)$, $t(x) = (xx', x, I)$ and $\langle(A_1,w_1,B_1)|(A_2,w_2,B_2)\rangle = \operatorname{tr}(A_1'A_2) + w_1'w_2 + \operatorname{tr}(B_1'B_2)$.

Since Q, t and $\langle\cdot|\cdot\rangle$ satisfy the remaining hypotheses of Theorem 4. and $\delta t(x,z) = (xz' + z'x, z, \theta)$ for each $x, z \varepsilon R^n$, it follows that for each $\gamma \varepsilon \Gamma$;

$$\begin{aligned}\ker T &\subset \{y\varepsilon R^n : x'(\Omega_\gamma^{-1} - I)y - y'\Omega_\gamma^{-1}\eta_\gamma = 0,\quad x \varepsilon R^n\}\\ &= \ker(\Omega_\gamma^{-1} - I) \cap [\Omega_\gamma^{-1}\eta_\gamma]^\perp = \ker(\Omega_\gamma - I) \cap [\eta_\gamma]^\perp.\end{aligned}$$

To see that (2) → (3) note that T^+T is the orthogonal projection on range $(T') = (\ker T)^\perp$. Since $\eta_\gamma \varepsilon (\ker T)^\perp$, (3a) holds. Furthermore, $\ker T^+T = \ker T \subset \ker(\Omega_\gamma - I)$ implies range $(\Omega_\gamma - I) \subset$ range (T^+T) and hence that $T^+T(\Omega_\gamma - I) = (\Omega_\gamma - I)$ which is (3b).

In order to see that (3) → (1) recall the definition of $Q(\gamma)$, $t(x)$ and the fact that $\delta t(x;z) = (xz' + z'x, z, \theta)$.

We need only show that $x'(\Omega_\gamma - I)y - \eta_\gamma' y = 0$ for each $\gamma \in \Gamma$, $x \in X$ and $y \in \ker T$. Using (3b) and symmetry together with (3a) it follows that

$$x'(\Omega_\gamma - I)y - \eta_\gamma' y = x'(\Omega_\gamma - I)T^+(Ty) - \eta_\gamma T^+(Ty) = 0 .$$

We state the following corollary without proof.

Corollary 3. Under the hypotheses of Theorem 6., there exists a $k \times n$ rank k sufficient statistic for $\{P_\gamma\}$, $\gamma \in \Gamma$ if and only if there exists a rank k orthogonal projection P on R^n such that (a) $P\eta_\gamma = \eta_\gamma$ and (b) $P(\Omega_\gamma - I) = \Omega_\gamma - I$ for each $\gamma \in \Gamma$. Moreover, any $k \times n$ rank k matrix such that $T^+T = P$ is a sufficient statistic for $\{P_\gamma\}$, $\gamma \in \Gamma$.

Corollary 4. If $\Gamma = \{0, 1, \cdots, m-1\}$, $\eta_0 = \theta$, $\Omega_0 = I$ and $B \equiv [\eta_1|\eta_2|\cdots|\eta_{m-1}|\Omega_1 - I|\Omega_2 - I|\cdots|\Omega_{m-1} - I]$ then T is a linear sufficient statistic for the finite family $\{P_\gamma\}$, $\gamma \in \Gamma$ of n-variate normal probability measures if and only if range (T') = range (B). Moreover, k = rank B is the smallest integer for which there exists a $k \times n$ sufficient statistic for $\{P_\gamma\}$, $\gamma \in \Gamma$.

Proof: The equivalent condition is an immediate consequence of Theorem 6. The minimality statement follows from the fact that if T is a $p \times n$ rank p sufficient statistic then $T^+TB = B$, hence, $T^+TBB^+ = BB^+$. It follows that range $(BB^+) \subset$ range (T^+T) and, since $(BB^+)B = B$, BB^+ satisfies Theorem 6.(3) so that $k = p$.

Example 1. Let $x_1, x_2, \cdots, x_n, \cdots$ be a sequence of univariate $N(\mu,\sigma)$ variables such that the joint density of $x_1, x_2, \cdots, x_n$ is $N(\mu\xi_n, \Omega_n)$ where $\xi_n' = (1,1,\cdots,1)$. Let $\{P_\mu\}$, $\mu \in R$ be the family of probability measures having densities $N(\mu\xi_n, \Omega_n)$ and $T \neq \theta$ a $1 \times n$ matrix.

Observe that T is sufficient for $\{P_\mu\}$, $\mu \in R$ if and only if $T\Omega_n^{1/2}$ is sufficient for the family of probability measures $\{\hat{P}_\mu\}$, $\mu \in R$ having densities $N(\mu\Omega_n^{-1/2}\xi_n, I)$ and, according to Theorem 6., that this is equivalent to the condition that $\ker T\,\Omega_n^{1/2} \subset [\Omega_n^{-1/2}\xi_n]^{\perp}$. This is equivalent to $\xi_n' = \alpha_n T\Omega_n$ for some scalar α_n . A simple calculation shows that $\alpha_n = n(T\Omega_n\xi_n)^{-1}$ so that the statistic T is sufficient for $\{P_\mu\}$, $\mu \in R$ if and only if $T = [(T\Omega_n\xi_n)^{-1}\xi_n'\Omega_n^{-1}]/n$. In particular, note that $T = \hat{T} \equiv (\xi_n'\Omega_n^{-1}\xi_n)^{-1}\xi_n'\Omega_n^{-1}$ is sufficient for $\{P_\mu\}$, $\mu \in R$ and th: $\hat{T}\,(x_1,\cdots,x_n)'$ is an unbiased estimate of μ for each integer n. This generalizes the classical result that the sample mean is a sufficient statistic for μ when the samples $x_1, x_2, \cdots$ are independent.

Further note that if $T \equiv \xi_n'/n$ (the statistic T for the sample mean) is a sufficient statistic for $\{P_\mu\}$, $\mu \in R$ for each integer n , the column sums (row sums) of Ω_n are identically $\alpha_n = (\xi_n'\Omega_n\xi)/n$. A routine induction argument shows that, in the latter case, Cov (x_i, x_j) = constant for i, j:1,2,$\cdots$, $i \neq j$.

Example 2. Let $y = W\gamma + \varepsilon$, where W is a fixed $m \times n$ matrix of rank n and $\varepsilon \sim N(\theta, I)$. According to the Gauss-Markov theorem, the minimum variance unbiased linear estimate of γ is $\hat{\gamma} = (W'W)^{-1}W'y$

Let $T = (W'W)^{-1}W'$ and observe that for $\gamma \varepsilon R^n$, $T'(TT')^{-1}T\,W\gamma = W\gamma$ and, since $T'(TT')^{-1}T=T^{+}T$, Theorem 6. implies T is a sufficient statistic for the set of probability measures $\{P_\gamma\}$, $\gamma \varepsilon R^n$ having densities $N(W\gamma,I)$.

On the other hand, if $\hat{T}$ is a sufficient linear statistic for $\{P_\gamma\}$, $\gamma \varepsilon R^n$ such that $\hat{T}y$ is an unbiased estimate of γ then, since $\hat{T}W = I$, $\hat{T}$ has rank n . Corollary 4 implies that n is the smallest integer for which there exists a linear $n \times m$ sufficient statistic for $\{P_\gamma\}$, $\gamma \varepsilon R^n$. Moreover, $\hat{T} = B(W'W)^{-1}W'$ for some nonsingular $n \times n$ matrix B . Since $\hat{T}W = I$, $\hat{T} = (W'W)^{-1}W'$.

Since $\hat{\gamma} = Ty$, the Gauss-Markov estimate of γ may be characterized as the unique linear sufficient statistic T for $\{P_\gamma\}$, $\gamma \varepsilon R^n$ for which Ty is an unbiased estimate of γ.

REFERENCES

1. Bahadur, R. R. (1954) Sufficiency and statistical decision function. Ann. Mathe. Statist. 25 423-463.

2. Halmos, P. R. and Savage, L. J. (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann. Mathe. Statist. 20 225-241.

3. Lehmann and Scheffe´ (1950) Completeness, similar regions and unbiased estimation. Sankhyã. Part I, 10 305-340.

LINEAR DIMENSION REDUCTION
AND BAYES CLASSIFICATION

Henry P. Decell, Jr.
Department of Mathematics
University of Houston
Houston, Texas

P.L. Odell
Programs in Mathematical Sciences
University of Texas at Dallas
Richardson, Texas

&

William A. Coberly
Department of Mathematics
Univeristy of Tulsa
Tulsa, Oklahoma

Report #66

February 1978

LINEAR DIMENSION REDUCTION

AND BAYES CLASSIFICATION

by

Henry P. Decell, Jr.[1], P. L. Odell[2] and William A. Coberly[3]

ABSTRACT

This paper develops an explicit expression for a compression matrix T of smallest possible left dimension k consistent with preserving the n-variate normal Bayes assignment of X to a given one of a finite number of populations and the k-variate Bayes assignment of TX to that population. The Bayes population assignment of X and TX are shown to be equivalent for a compression matrix T explicitly calculated as a function of the means and covariances (known) of the given populations.

1. Mathematics Department, University of Houston
2. Programs in Mathematical Sciences, Univ. of Texas at Dallas
3. Department of Mathematics, University of Tulsa

This work was partially supported by Johnson Space Center Contract NAS-9-15000.

INTRODUCTION

In this paper Π_i will denote an n-variate normal population having a priori probability $\pi_i>0$ and density $p_i(x)$; $i=0,1,\ldots,m$. Using recent results [1] that characterize linear sufficient statistics we will develop an explicit expression for a kxn compression ($k\leqq n$) matrix T for which, using the Bayes classification procedure [2] , in which costs of misclassification are tacitly assumed equal on all classes, X is assigned to Π_i if and only if TX is assigned to Π_i. We will further demonstrate that k is the smallest integer ($\leqq n$) for which the latter equivalence is valid and that T can be directly calculated in terms of the known population means and covariance matrices.

The applications which motivate the necessity for compressing or reducing the size of a data vector is summarized very well in a review paper by Laveen Kaval in [3]. Our own interest was motivated by a need to reduce computational requirements in a large area crop inventory project using multidimensional data taken remotely by near earth satellites [4].

In all that follows η_i and Σ_i will, respectively, denote the mean and covariance matrix of population Π_i, $i=0,1,\ldots,m$. It is well known that for each non-singular nxn matrix A and nx1 vector α, the Bayes assignment of x to Π_i is equivalent to the Bayes assignment of $A(x-\alpha)$ to Π_i. We will later assume that $\eta_0=\Theta$ and $\Sigma_0 = I$. This assumption will impose no loss of generality in the results that follow since we may set $\alpha\equiv\eta_0$ and choose A such that $A\Sigma_0 A^T=I$.
If the latter transformation of variables is necessary, we will not introduce new symbols for the variate $A(X-\eta_0)$, the densities $p_i(Ax-\eta_0)$

and their associated means and covariance matrices. Whenever Q is an sxn rank $(s\leq n)$ matrix, we will denote the s-variate normal density of Qx by (for population Π_i) $p_i(Qx)$.

PRINCIPAL RESULTS

According to [1], let $k(\leq n)$ be the smallest integer for which there exists a linear sufficient statistic (kxn matrix T) for the family of probability measures having densities $p_i(x)$; i=0,1, ..., m. The results in [1] demonstrate that the sufficiency of T is equivalent to the conditions:

$$\begin{aligned} &(1) \quad T^+T\eta_j = \eta_j \\ &(2) \quad T^+T(\Sigma_j - I) = \Sigma_j - I \end{aligned} \qquad j=0,1,\ldots, m$$

where $(\cdot)^+$ denotes the generalized inverse of $(\cdot)$.

Let M be the nx(n+1)m partitioned matrix

$$M \equiv [\eta_1|\eta_2|\cdots|\eta_m|\Sigma_1-I|\Sigma_2-I|\cdots|\Sigma_m-I]$$

and let M=FG be a full rank decomposition [5] of M, that is; F is nxk, G is kx(m+1)m and rank (F) = rank (G) = k. Again, according to [1] and the latter, k must be precisely the smallest integer $(\leq n)$ for which a kxn matrix T can be a sufficient statistic for the given family of probability measures.

It is well known [5] that $M^+=G^+F^+$ and hence that $MM^+=FF^+$. A simple computation reveals that $T\equiv F^T$ satisfies conditions (1) and (2) so that F^T is a sufficient statistic (of minimum left dimension) for the given family of probability measures. We have the following theorem.

Theorem 1. Let Π_i be an n-variate normal population with *a priori* probability $\pi_i>0$, mean η_i and covariance Σ_i; $i=0,1,\cdots,m$ (with $\eta_0=\theta$, $\Sigma_0=I$) and let $FG=M\equiv[\eta_1|\eta_2|\cdots|\eta_m|\Sigma_1-I|\Sigma_2-I|\cdots|\Sigma_m-I]$ be a full rank $(=k\leq n)$ decomposition of M. Then, the n-variate Bayes procedure assigns x to Π_i if and only if the k-variate Bayes procedure assigns F^Tx to Π_i. Moreover, k is the smallest integer for which there exists a kxn compression matrix T preserving the Bayes assignment of x and Tx to π_i; i=0, 1, ..., m

Proof: Recall that the n-variate Bayes procedure assigns x to π_j if and only if $\pi_j p_j(x)>\pi_i p_i(x)$; $i=0,1,\ldots,m$: $i\neq j$ (with arbitrary assignment of x to any of the populations $\Pi_k$for which $\pi_j p_j(x) = \pi_k p_k(x)$).

Let R be any (n-k) x n matrix such that $C = R(I-FF^+)$ has rank n-k and note that $\pi_j p_j(x) > \pi_i p_i(x)$; $i=0,1,\ldots,m$: $i\neq j$ is equivalent to

$$\pi_j p_j(\begin{bmatrix}F^T\\C\end{bmatrix} x) > \pi_i p_i(\begin{bmatrix}F^T\\C\end{bmatrix}x);\ i=0,1,\ldots,m:\ i\neq j$$

For any $q=0,1,\ldots,m$, the n-variate normal density $p_q(\begin{bmatrix}F^T\\C\end{bmatrix}x)$ has mean $\begin{bmatrix}F^T\eta_q\\C\ \eta_q\end{bmatrix}$ and covariance matrix:

$$\begin{bmatrix}F^T\Sigma_q F & F^T\Sigma_q C^T\\ C\Sigma_q F & C\Sigma_q C^T\end{bmatrix}$$

Condition (1) implies $C\eta_q=\theta$. Condition (2) implies that $I-FF^T$ commutes with Σ_q and it follows that $C\Sigma_q C^T=CC^T$ and $C\Sigma_q F = \theta$. We may therefore write $p_q(\begin{bmatrix}F^T\\C\end{bmatrix}x)$ as the product of the respective k-variate and (n-k)-variate densities $p_q(F^Tx)$ and $p_q(Cx|F^Tx)$, the conditional density of Cx given F^Tx. Since $p_q(Cx|F^Tx)>0$ does not depend upon q = 0, 1, ..., m; it follows that the n-variate Bayes assignment of x to Π_j; $j=0,1,\ldots, m$, implies the k-variate Bayes assignment F^Tx to Π_j. The foregoing arguments are reversible and hence the k-variate Bayes assignment of F^Ts to Π_j implies the n-variate Bayes assignment of x to Π_j, completing the proof of the equivalence. The minimality of k, in the sense that the n-variate and k-variate Bayes assignments of x and $F^T x$ are preserved, is a consequence of the developments preceding the theorem.

CONCLUDING REMARKS

Clearly the theorem is valid if there is at least one population with mean θ and covariance I, in which case we would label that population Π_0. If this is not the case, one would choose some population, say π_q, and perform the change of variables $x \rightarrow A(x-\eta_q)$ where $A\Sigma_q A^T=I$ prior to application of the theorem. The appropriate statistic for compression, in terms of the original variates, would then be $T=F^T A^{-1}$.

These results completely characterize the nature of data compression for the Bayes classification procedure in the sense that k is the smallest allowable data compression dimension consistent with preserving Bayes population assignment and, moreover, the theorem provides an explicit expression for the compression matrix T that depends only upon the known population means and covariances. The statistic $T=F^T$ given by the theorem is by no means unique (e.g., for any non singular kxk matrix B, $T\equiv BF^T$ will do! It is also true that there may be more efficient methods for calculating the statistic T (yet to be determined) than the method of full rank decomposition of M.

It should be noted that the matrix M has an "excellent chance" of having rank equal to n. Even in the case of two populations (m=2), there may well be n linearly independent columns among the 2(n+1) columns of M and, therefore, no integer $k<n$ and kxn rank k compression matrix T preserving the Bayes assignment of x and Tx.

There has been extensive work [6],[7],[8],[9],[10],[11],[12],[13], on determination of compression matrices (of a given rank) based upon criteria that, generally, attempt to describe the relative (to the variate x) "information content" in the variate Tx (e.g., divergence, Bhattacharyya distance, Chernoff bound, principal components, Wilks scatter, etc.) While these criteria provide bases for calculating compression matrices T, they provide little or no means for determining the degradation in probability of misclassification or sensitivity to population assignments.

In sampling situation one may choose to replace the columns of the matrix M by their estimates, that is η_j by $\bar{x}_j$ and Σ_j by S_j. The matrix defined by the estimate suggest a compression technique based on the selection of a k dimensional hyperplane which in some sense best fits the range space of matrix

$$\hat{M} = [\bar{x}_1|x_2|\cdots|\bar{x}_m|S_j-S_o|\cdots|S_m-S_o]$$

where

$$\bar{x}_o=\Theta \text{ and } S_o=I.$$

We feel that the results in this paper shed some light upon the subject. In future work we intend to extend these results and the results of [1] to a related concept of an "almost sufficient" statistic.

REFERENCES

[1] Peters, B.C., Redner, R., Decell, H.P. Jr., "Characterizations of Linear Sufficient Statistics," submitted to Sankyā, A, (1978).

[2] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc. (1958), pp. 126-152.

[3] Kanal, L., "Patterns in Pattern Recognition: 1968-1974, IEEE Transaction in Pattern Recognition, Vol. IT-20, Nov. 1974, pp. 697-722.

[4] Linz, J., Simonett, Ed., Remote Sensing of Environment, Addison-Wesley Inc. (1976).

[5] Boullion, T.L., and Odell, P.L., Generalized Inverse Matrices, Wiley-Interscience (1971), p. 11.

[6] deFigueiredo, R.J., "Optimal linear and nonlinear feature extraction based on the minimization of the increased risk of misclassification," in Proc. 2nd Int. Joint Conf. Pattern Recognition, 1974.

[7] Decell, H.P. Jr., and Quierein, J.A., "An iterative approach to the feature selection problem," in Proc. Purdue Univ. Conf. Machine Processing of Remotely Sensed Data, 1972, pp. 3B1-3B12.

[8] Cover, T.M., "Learning in pattern recognition," in Methodologies of Pattern Recognition, S. Watanabe, Ed. New York: Academic 1969, pp. 111-132.

[9] Whitney, A., "A direct method of nonparametric measurement selection," IEEE Trans. Comput. (Short Notes), vol. C-20, pp. 1100-1103, Sept. 1971.

[10] Simon, J.C., Roche, C., and Sabah, G., "On automatic generation of pattern recognition operators," in Proc. 1972 Int. Conf. Cybernetics and Soc., pp. 232-238, IEEE Publ. no. 72 CH06478SMC.

[11] Michael, M., and Lin, W.C., "Experimental study of information measure and inter-intra class distance ratios on feature selection and orderings," IEEE Trans. Syst., Man, Cybern., vol. SMC-3, pp. 172-181, Mar. 1973.

[12] Kailath, T., "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Commun. Technol., vol. COM-15, pp. 52-60, Feb. 1967.

[13] Mucciardi, A.N., and Gose, E.E., "A comparison of seven techniques for choosing subsets of pattern recognition properties," IEEE Trans. Comput., vol. C-20, pp. 1023-1031, Sept. 1971.

QUASI-NEWTON METHODS

by

Homer F. Walker
Department of Mathematics
University of Houston
Houston, Texas

Report #67
NASA Contract NAS-9-15000

QUASI-NEWTON METHODS

by

Homer F. Walker
Department of Mathematics
University of Houston
Houston, Texas

1. Introduction

Systems of nonlinear equations can seldom be solved exactly. Usually, one must obtain approximations to the solutions of such systems by iteration. Quasi-Newton methods (also known as variable metric, variance, secant, update, or modification methods) constitute a class of iterative procedures which may be regarded as generalizations of the secant method for solving a single equation in one unknown. Indeed, not only is the quasi-Newton equation (the equation characteristically satisfied by the iterates produced by these methods) a direct extension of the equation which defines the iterates of the secant method, but also these procedures share many of the computational advantages of the secant method over Newton's method.

Quasi-Newton methods were first introduced in the papers of Davidon [2], Fletcher and Powell [4], and Broyden [1]. In spite of their recent origins, these methods have proved themselves in dealing with practical problems and have become the subject of a large amount of research. The paper of Dennis and Moré [3] provides both an excellent in-depth survey and an elegant unified development of quasi-Newton methods and their theory as understood in the mid-1970's. The main body of this note is a rearrangement and condensation of material in [3].

In the following, we first formulate precisely the problem to be solved and motivate the introduction of quasi-Newton methods by considering the classical Newton and secant methods and their properties. We then survey three highly successful quasi-Newton methods: Broyden's method for the solution of general nonlinear equations, and the Davidon-Fletcher-Powell and Broyden-Fletcher-Goldfarb-Shanno procedures for unconstrained minimization. (The last two methods will henceforth be referred to as the DFP and BFGS methods, respectively.) Finally, we compare the properties of these methods to those of Newton's method and UHMLE in potential applications to maximum-likelihood estimation of parameters in mixture distributions.

2. The problem

We consider the problem of solving $F(x) = 0$ in an open convex subset D of R^n under the following assumptions on the mapping $F:D \rightarrow R^n$:

(a) F is continuously differentiable on D.

(b) There is an x^* in D such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular.

Newton's method for iteratively approximating the solution x^* begins with an initial approximation x_0 to x^* and attempts to obtain improved approximations by the iteration

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k) \qquad k = 0,1, \ldots .$$

The convergence properties of Newton's method which are important here are summarized in the following theorem.

Theorem: Whenever x_0 is sufficiently near x^*, there is a sequence $\{\alpha_k\}_{k=0,1,\ldots}$ of non-negative numbers which converges to zero and for which

(1) $$|x_{k+1} - x^*| \le \alpha_k |x_k - x^*| \qquad k = 0,1, \ldots \quad .$$

If, in addition to satisfying assumptions (a) and (b) above, F has a derivative which is Lipschitz continuous at x^*, i.e., there exists a κ for which $|F'(x) - F'(x^*)| \le \kappa |x - x^*|$ for all x sufficiently near x^*, then there exists a constant β such that

(2) $$|x_{k+1} - x^*| \le \beta |x_k - x^*|^2 \qquad k = 0,1, \ldots$$

whenever x_0 is sufficiently near x^*.

A sequence which satisfies an inequality of the form (1) with a sequence $\{\alpha_k\}_{k=0,1,\ldots}$ which converges to zero is said to converge superlinearly. If a sequence satisfies an inequality of the form (2), then it is said to converge quadratically. Superlinear convergence is fast; quadratic convergence is very fast. Since Lipschitz continuity is a very weak assumption, one might say that the theorem asserts that the convergence exhibited by the Newton iterates is always fast and almost always very fast.

The rapid convergence of the Newton iterates is the major advantage of Newton's method. Another advantage is that Newton's method is "self-corrective" in the sense that x_{k+1} depends only on F and x_k so that bad effects of previous iterations are not carried along. (Quasi-Newton methods are not self-corrective in this sense.) Balanced against these advantages is the fact that Newton's method often requires a great deal of computation at each iteration. Indeed, the determination of each iterate requires $O(n^2)$ function evaluations and $O(n^3)$ arithmetic operations. Thus one is led to ask whether there are methods which retain fast convergence while requiring fewer function evaluations and arithmetic operations at each iteration.

With this question in mind, consider the secant method in the case $n = 1$. This method begins with an initial approximation x_0 to x^* and defines successive approximations by the iteration

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{F(x_k) - F(x_{k-1})} F(x_k) .$$

One may regard the secant method as being obtained from Newton's method by replacing the derivative $F'(x_k)$ by a finite-difference approximation. A particular consequence is that the number of function evaluations per iteration is reduced from two for Newton's method to one for the secant method while the number of arithmetic operations per iteration is not significantly increased. It can be proved that, for x_0 sufficiently near x^*, the iterates produced by the secant method exhibit superlinear convergence rather than quadratic convergence as in the case of the Newton iterates. Nevertheless, superlinear convergence is still fast, and experience has shown that, as a general-purpose algorithm, the secant method is more efficient in total computation time than Newton's method. This suggests that generalizations of the secant method to higher dimensions might be similarly successful.

3. Quasi-Newton methods

Quasi-Newton methods are generalizations of the secant method which are applicable to problems of the type at hand involving an arbitrary number of independent variables. The key properties of these methods are that the iterates exhibit superlinear local convergence and that each iteration requires n function evaluations and $O(n^2)$ arithmetic operations. In spite of the fact that quasi-Newton methods do not have the quadratic convergence property of Newton's method, the comparatively small number of function evaluations and arithmetic operations make them preferable to Newton's method in many applications.

Quasi-Newton methods have the general form

$$x_{k+1} = x_k - B_k^{-1}F(x_k) \ ,$$

where B_k satisfies the quasi-Newton equation

$$B_k(x_k - x_{k-1}) = F(x_k) - F(x_{k-1}) \ . \tag{3}$$

Note that B_k has the action of a finite-difference approximation to $F'(x_{k-1})$ in the direction $(x_k - x_{k-1})$. Thus quasi-Newton methods in general bear the same relation to Newton's method as the secant method in the case $n = 1$.

It is clear that the secant method is a quasi-Newton method. In fact, if $n = 1$, then the quasi-Newton equation determines the scalar B_k exactly, and so the secant method is the only quasi-Newton method in this case. If $n > 1$, then the quasi-Newton equation alone does not determine B_k uniquely; hence, there is no unique natural extension of the secant method to the case of an arbitrary number of independent variables. This lack of uniqueness in the general case may be regarded as an advantage, for it allows a variety of quasi-Newton algorithms which may be drawn upon to take advantage of any special structure which may be present in specific problems of interest.

When $n > 1$, one must impose relations between successive matrices B_k and their predecessors which, together with the quasi-Newton equation, uniquely determine these matrices inductively. In general, those relations are chosen with an eye toward minimizing the computational complexity of the resulting update formula for determining B_{k+1} from B_k, x_k, and F while taking maximal advantage of whatever special structure may be shared by the particular problems under consideration. Of the three quasi-Newton methods presented below, the first (Broyden's method) is intended to be a general purpose algorithm which can be applied to all problems without regard to special structure. Consequently, in Broyden's method, B_{k+1} is obtained by adding a rank-one "correction term" to B_k in such a way that the quasi-Newton equation is satisfied and B_{k+1} agrees with B_k on the orthogonal complement of $(x_{k+1} - x_k)$. In a sense, this may be regarded as the "simplest" way to obtain B_{k+1} from B_k in such a way that the quasi-Newton equation is satisfied. On the other hand, the second two methods (the DFP and BFGS methods) are designed for unconstrained minimization problems, in which the Jacobian $F'(x)$ can be expected to be symmetric and positive-definite. Thus the update formulas for these methods are such that the successive B_k's "inherit" symmetry and positive-definiteness from the preceding ones. Not surprisingly, these formulas are more complex than the update formula of Broyden's method. In fact, in order to guarantee hereditary symmetry and positive-definiteness, it is necessary in these formulas to determine B_{k+1} from B_k with a correction term of rank two.

4. Broyden's method for general nonlinear equations

Broyden's method is, in a sense, the "simplest" of the most popular quasi-Newton methods and is intended to be a general-purpose algorithm for solving arbitrary nonlinear equations. To derive the formula used in Broyden's method to update the matrices B_k, suppose that, for some $k \geq 0$, one has arrived at x_k and B_k. Then x_{k+1} can be generated by the formula

$$x_{k+1} = x_k - B_k^{-1}F(x_k) .$$

Our objective is to use x_k, x_{k+1}, B_k and F to update B_k in the "simplest" way to obtain a matrix B_{k+1} which satisfies the quasi-Newton equation.

For convenience, we adopt the following notation:

$$x_k = x, \quad B_k = B, \quad B_{k+1} = \bar{B}, \quad x_{k+1} - x_k = s, \quad F(x_{k+1}) - F(x_k) = y.$$

In this notation, the quasi-Newton equation which we wish B_{k+1} to satisfy is $\bar{B}s = y$. This equation uniquely specifies the action of $\bar{B}$ in the direction of s. Since there is no apparent reason for $\bar{B}$ to differ from B on the orthogonal complement of s, it seems reasonable to impose on $\bar{B}$ the condition that $Bz = \bar{B}z$ for all z such that $z^T s = 0$. It is easily verified that there is a unique $\bar{B}$ which satisfies both this condition and the quasi-Newton equation. This $\bar{B}$ is given by the formula

$$\bar{B} = B + \frac{(y - Bs)s^T}{|s|^2} .$$

Note that $\bar{B}$ and B differ by a rank-one operator. Restoring subscripts, we obtain the iteration formulas for Broyden's method:

$$x_{k+1} = x_k - B_k^{-1}F(x_k)$$

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{|s_k|^2} \quad ,$$

where $y_k = F(x_{k+1}) - F(x_k)$ and $s_k = x_{k+1} - x_k$.

Does Broyden's method exhibit the key properties attributed to quasi-Newton methods in the preceding section? It can be shown that if x_0 and B_0 are sufficiently near x^* and $F'(x^*)$, respectively, then the Broyden iterates are well-defined and converge superlinearly to x^*. (The proof is very involved, and we omit it.) Also, it is clear that, for a given value of k, the determination of x_{k+1} and B_{k+1} requires only the n function evaluations necessary to specify $F(x_{k+1})$, assuming that $F(x_k)$ can be provided from storage. Finally, it is evident that, for a given k, x_{k+1} and B_{k+1} can be determined with $O(n^2)$ arithmetic operations if $B_k^{-1}F(x_k)$ can be evaluated with $O(n^2)$ arithmetic operations.

There are two ways of evaluating $B_k^{-1}F(x_k)$ with $O(n^2)$ arithmetic operations, both of which require information about B_{k-1}. The first way is based on the Sherman-Morrison formula [8] and produces $\bar{B}^{-1}$ from B^{-1} with $O(n^2)$ arithmetic operations in the following way: write

$$\bar{B} = B + \frac{(y - Bs)s^T}{|s|^2} = B + uv^T ,$$

where $u = (y - Bs)$, $v = \frac{s^T}{|s|^2}$; then

$$\bar{B}^{-1} = B^{-1} - \frac{1}{1 + <v, B^{-1}u>} B^{-1}uv^T B^{-1} .$$

The second way is based on a special factorization procedure due to Gill and Murray [5] which begins with a factorization $B = QR$ and yields a factorization $\overline{B} = \overline{Q}\,\overline{R}$ with $O(n^2)$ arithmetic operations. (Here, Q and $\overline{Q}$ are orthogonal and R and $\overline{R}$ are upper-triangular.) Since an n-dimensional linear system whose coefficient matrix is factored in this way can be solved with $O(n^2)$ arithmetic operations, this allows the evaluation of the terms $B_k^{-1}F(x_k)$ with $O(n^2)$ arithmetic operations as desired. For reasons of numerical stability, the Gill-Murray factorization procedure is generally preferable to the method using the Sherman-Morrison formula.

5. The DFP and BFGS methods for unconstrained minimization

For the purposes of this note, the basic problem of unconstrained minimization may be regarded as the problem of solving $\nabla f(x) = 0$ in an open convex subset D of R^n, where f is a nonlinear functional from D to R^1. Clearly, this problem is of the type introduced in Section 2, with ∇f playing the role of F. The special feature of this problem is that the Jacobian of the function whose zero is being sought is actually the Hessian $\nabla^2 f$, a matrix which is certainly symmetric. In fact, in most problems of practical interest, $\nabla^2 f$ is positive-definite near the minimum of f.

It seems reasonable to require that the matrices B_k appearing in a quasi-Newton method applied to an unconstrained minimization problem be symmetric and positive-definite. Since each B_k is to be determined from its predecessor by an update formula, it is reasonable to impose conditions on the update formula which guarantee that symmetry and positive-definiteness are inherited by the successive matrices B_k. Unfortunately, imposing hereditary symmetry as well as the quasi-Newton equation completely determines a rank-one update formula, and this formula does not guarantee hereditary positive-definiteness. Consequently, one is led to look for rank-two update formulas which insure that the successive matrices B_k inherit symmetry and positive-definiteness.

A general rank-two update formula which guarantees hereditary symmetry is the following:

$$\bar{B} = B + \frac{(y - Bs)c^T + c(y - Bs)^T}{<c,s>} - \frac{<y - Bs,s>}{<c,s>^2} cc^T ,$$

where c is any vector in R^n such that $<c,s> \neq 0$. A "natural" choice of c which insures hereditary positive-definiteness whenever $<y,s> > 0$ is $c = y$. (Since $<y,s> \approx <\nabla^2 f(x^*)s,s>$ near x^*, one expects $<y,s>$ to be positive near x^*.) The resulting update formula is that used in the Davidon-Fletcher-Powell (DFP) method. Denoting by $\bar{B}_{DFP}$ the updated matrix obtained from B by applying this formula, one has

$$\bar{B}_{DFP} = B + \frac{(y - Bs)y^T + y(y - Bs)^T}{<y,s>} - \frac{<y - Bs,s>yy^T}{<y,s>^2}$$

$$= (I - \frac{ys^T}{<y,s>})B(I - \frac{sy^T}{<y,s>}) + \frac{yy^T}{<y,s>} \quad .$$

As with Broyden's method, one can show that the DFP iterates converge superlinearly to x^* whenever x_0 and B_0 are sufficiently near x^* and $\nabla^2 f(x^*)$, respectively, and that each iteration requires n function evaluations and $O(n^2)$ arithmetic operations. Although the DFP update formula is a bit more complicated than the Broyden update formula, experience has shown that the DFP method is generally superior to Broyden's method for problems in unconstrained minimization.

At the k^{th} iteration, both Broyden's method and the DFP method require first the determination of $B_k^{-1}F(x_k)$ and then the updating of B_k. It is natural to ask whether a more efficient method might be obtained by applying an update formula directly to B_k^{-1}. If we denote B^{-1} by H and $\bar{B}^{-1}$ by $\bar{H}$, the quasi-Newton equation $\bar{B}s = y$ becomes $s = \bar{H}y$. Carrying out a development completely analogous to that leading to the DFP update formula yields the update formula of the Broyden-Fletcher-Shanno-Goldfarb (BFGS) method. Denoting by $\bar{H}_{BFGS}$ the updated matrix obtained from H by applying this formula, one has

$$\bar{H}_{BFGS} = (I - \frac{sy^T}{<y,s>})H(I - \frac{ys^T}{<y,s>}) + \frac{ss^T}{<y,s>} .$$

It is not difficult to see that, as in the case of the DFP update, this update adds a rank-two correction term to H and guarantees hereditary symmetry and, if $<y,s> > 0$, positive-definiteness. Again, it can be shown that the BFGS iterates converge superlinearly to x^* wherever x_0 and H_0 are sufficiently near x^* and $\nabla^2 f(x^*)^{-1}$, respectively. It is clear that each iteration requires n function evaluations and $O(n^2)$ arithmetic operations.

The BFGS method is not the same as the DFP method. In fact,

$$\bar{H}_{BFGS} = (\bar{B}_{DFP})^{-1} + vv^T$$

where $v = <y,Hy>^{1/2}[\frac{s}{<s,y>} - \frac{Hy}{<y,Hy>}]$. According to [3], there is "growing evidence that BFGS is the best current update formula for use in unconstrained minimization".

6. A potential application

We conclude this note by comparing the properties of quasi-Newton methods to those of Newton's method and UHMLE in a potential application to the problem of obtaining maximum-likelihood estimates of the parameters in mixture distributions. Such estimates, of course, play a fundamental role in certain approaches to signature extension, estimation of proportions, and clustering. For a description of the UHMLE algorithm, see [6] and [7].

Let X be an n-dimensional random variable with probability density function

$$p(x) = \sum_{i=1}^{m} \alpha_i^0 \, p_i(x) \ ,$$

where

$$p_i(x) = \frac{1}{(2\pi)^{n/2}|\Sigma_i^0|^{1/2}} \ e^{-1/2(x-\mu_i^0)^T \Sigma_i^{0^{-1}} (x-\mu_i^0)}$$

and the proportions α_i^0 are positive and sum to 1. Suppose that $\{x_k\}_{k=1,\ldots,N}$ is a sample of independent observations on X. By a <u>maximum-likelihood estimate</u> of the parameters $\{\alpha_i^0, \mu_i^0, \Sigma_i^0\}_{i=1,\ldots,m}$, we mean a choice of parameters $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$ which locally maximizes the <u>log-likelihood function</u>

$$L = \sum_{k=1}^{N} \log p(x_k) \ ,$$

regarded as a function of the parameters $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$. It is known that, loosely speaking, there is a unique strongly-consistent maximum-likelihood estimate. (See [7] for a clarification and proof of this statement.)

The problem which we consider here is to approximate numerically the strongly-consistent maximum-likelihood estimate. This is potentially a very difficult problem. Indeed, the number of independent variables is $(m - 1) + mn + m \frac{n(n+1)}{2}$, a number which may be very large. Furthermore, the evaluation of functions derived from the log-likelihood function usually involves summation over the entire sample of N observations and, hence, is a source of computational difficulty when the sample is large. In the table below, we list the key properties of UHMLE, Newton's method, and quasi-Newton methods when applied to solving likelihood equations obtained by differentiating the log-likelihood function. It should be noted that, in addition to the arithmetic operations listed in the table, each method requires at each iteration the evaluation of the functions $p_i(x_k)$, $i = 1,\ldots,m$, $k = 1,\ldots,N$.

METHOD	CONVERGENCE	ARITHMETIC OPERATIONS PER ITERATION
UHMLE	Linear	$0(mn^2N)$
Newton's Method	Quadratic	$0_1(m^2n^4N) + 0_2(m^3n^6)$
Quasi-Newton Methods	Superlinear	$0_1(mn^2N) + 0_2(m^2n^4)$

Of course, many factors must be considered in addition to convergence rates and the amount of arithmetic per iteration when deciding what sort of algorithm is best suited in a particular instance for application to the problem under consideration. For example, UHMLE is a type of gradient method; hence, one might expect UHMLE to enjoy the relatively good global convergence behavior usually associated with gradient methods. Furthermore, gradient methods are often competitive in speed of convergence to Newton's method and quasi-Newton methods when only "ball-park" approximations to the solution are desired. Since the nearness of the maximum-likelihood estimate to the true parameters will be limited by the variance of the sample observations, "ball-park" approximations will certainly suffice except, perhaps, in the case of a very large sample.

It is difficult to predict circumstances in which the advantage of fast convergence for Newton's method and quasi-Newton methods will outweigh the disadvantage of having to perform a great many arithmetic operations at each iteration with these methods. However, it should be noted that if N is very large relative to m and n, then the number of arithmetic operations per iteration required by quasi-Newton methods is comparable to the number required by UHMLE. Also, if N is very large, one might reasonably want to obtain very accurate approximations of the maximum-likelihood estimate, in which case the superlinear convergence of quasi-Newton methods is clearly preferable to the linear convergence of UHMLE. Consequently, if N is very large relative to m and n and if particularly accurate approximations of the maximum-likelihood estimate are desired, then quasi-Newton methods appear to have a clear-cut advantage over UHMLE. In such circumstances, one might retain the good global properties of UHMLE by employing a hybrid method which initially behaves like UHMLE and then behaves increasingly like a quasi-Newton method as the iteration proceeds.

BIBLIOGRAPHY

1. C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," Math. Comp. 19 (1965), pp. 577-593.

2. W. C. Davidon, "Variable metric method for minimization," Rep. ANL-5990 Rev. (1959), Argonne National Laboratories, Argonne, Illinois.

3. J. E. Dennis, Jr., and J. J. Moré, "Quasi-Newton methods, motivation and theory," SIAM Review 19 (1977), pp. 46-89.

4. R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," Comput. J. 6 (1963), pp. 163-168.

5. P. E. Gill and W. Murray, "Quasi-Newton methods for unconstrained minimization," J. Inst. Math. Appl. 9 (1972), pp. 91-108.

6. B. C. Peters, Jr. and H. F. Walker, "The numerical evaluation of the maximum-likelihood estimate of a subset of mixture proportions," University of Houston Math. Dept. Tech. Report No. 50, Contract NAS-9-12777 (1976).

7. B. C. Peters, Jr., and H. F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions, II," University of Houston Math. Dept. Tech. Report No. 51, Contract NAS-9-12777 (1976).

8. J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix," Ann. Math. Statist. 20 (1949), p. 621.

ON N$^{\underline{th}}$ ROOTS OF POSITIVE OPERATORS

D.R. Brown and M.J. O'Malley
Department of Mathematics
University of Houston
Houston, Texas 77004

Report #68

February 1978

ON N$^{\underline{th}}$ ROOTS OF POSITIVE OPERATORS

by D.R. Brown and M.J. O'Malley[1]

A bounded operator A on a Hilbert space H is positive provided $\langle Ax,x\rangle \geq 0$ for all $x \in H$. These operators are symmetric, and as such constitute a natural generalization of non-negative real diagonal matrices. The following result is thus both well known and not surprising:

Theorem: A positive operator has a unique positive square root (under operator composition).

This may be established by integration of the correct function, invoking the spectral theorem for self-adjoint operators. A more accessible argument for those not acquainted with the mysteries of spectral measures may be found in [1,p.317].

While square roots and their iterates seem to provide a sufficient analytic tool for most purposes, it is also a (folk) theorem that positive operators possess unique positive $n^{\underline{th}}$ roots for every positive integer n. As in the $n = 2$ case, existence follows from an application of the spectral theorem; however, we give an argument in the spirit of [1]. The purpose in so doing is not to exercise the reader's knowledge of induction, but rather to illustrate another use of the Law of the Mean as a motivational instrument.

[1] Both authors received partial support under NASA contract NAS-9-15000.

Let I be the identity operator on H, and let $B(H)$ denote the set of bounded operators on H. We will need the following properties of positive operators:

(1) the relation on positive operators defined by $A \leq B$ if and only if $B - A$ is positive, is reflexive, transitive, and consistent with the notation $0 \leq A$ for any positive A; moreover, this relation is preserved by operator addition and positive real scalar multiplication, and reversed by negative scalar multiplication.

(2) If A and B are positive and if $AB = BA$, then AB is positive.

(3) If $0 \leq A \leq I$, then $0 \leq I-A \leq I$.

(4) If $0 \leq A$, then $A \leq ||A||I$, so that $(||A||)^{-1}A \leq I$, if $A \neq 0$.

(5) If $0 \leq A \leq I$, then $A^n \leq A$ for all positive integers n.

We also require:

<u>Lemma</u>. If $\{S_n\}$ is a sequence in $B(H)$ such that $0 \leq S_n \leq S_{n+1} \leq I$, then there exists $S \in B(H)$ such that $\{S_n u\} \to Su$ for all $u \in H$.

All of the conclusions above are verified by straightforward arguments in [1,pp. 317-320].

<u>Theorem</u>: Let $A \in B(H)$, $0 \leq A$, and let k be a positive integer. Then there exists a unique positive operator B such that $B^k = A$.

<u>Proof</u>: By (4) above, we need only consider the case in which $A \leq I$.

We first prove the existence of B. Since the theorem is a tautology for all operators when $k = 1$, we assume the existence of positive $(k-1)$-st roots for all positive operators.

Under the momentary supposition that B exists, let $R = I - A$ and $S = I - B$. Then $(I - S)^k = I - R$, so that

$$(*) \qquad S = (1/k) [R + \sum_{r=2}^{k} \binom{k}{r} (-1)^r S^r] .$$

Clearly the existence of a positive operator satisfying this implicit relation is necessary and sufficient to establish the existence of the desired operator B. To this end, we define a sequence of operators by $S_o = 0$, $S_{n+1} = (1/k) [R + \sum_{r=2}^{k} \binom{k}{r} (-1)^r S_n^r]$. In order to show $S_n \leq S_{n+1}$ it suffices to show, under the assumption $0 \leq S_{n-1} \leq S_n \leq I$, that $0 \leq S_{n+1} - S_n =$ $(1/k) [\sum_{r=2}^{k} \binom{k}{r} (-1)^r (S_n^r - S_{n-1}^r)]$.

To accomplish this, we digress to a consideration of the polynomial $f(x) = \sum_{r=2}^{k} \binom{k}{r} (-1)^r x^r = (1-x)^k + kx - 1$. Since Since $f'(x) = k [1 - (1 - x)^{k-1}] \geq 0$ on $[0,1]$, clearly f is increasing on this interval. To translate this to operators, it is necessary to examine the situation more carefully. By the Mean Value Theorem, given $0 \leq y < z \leq 1$, there exists a (unique) number $c \in (y,z)$ such that

$$(**) \qquad f(z) - f(y) = f'(c)(z - y) \quad .$$

Upon solving, $c = 1 - [(1/k) \sum_{r=0}^{k-1} (1 - y)^{k-r-1}(1 - z)^r]^{1/(k-1)}$

Returning to our operator problem, we wish to apply this information to the sequence $\{S_n\}$. Since all members of this family are polynomials in $R = I - A$, any two of them commute. This is a property sufficient to permit imitation of equation (**) with operators; let $z = S_n$, $y = S_{n-1}$. In this format, we use C to represent the operator $I - J$, where J is (any) positive $(k-1)$st root of the operator $(1/k) \sum_{r=o}^{k-1} (I - S_{n-1})^{k-r-1}(I - S_n)^r$. The following chain of equalities is easily calculated:

$$S_{n+1} - S_n = (1/k)\cdot(f(S_n) - f(S_{n-1}))$$

$$= (1/k)\{k[I - (I - C)^{k-1}]\}\cdot(S_n - S_{n-1})$$

$$= [I - (I - C)^{k-1}]\cdot(S_n - S_{n-1})$$

$$= [I - J^{k-1}]\cdot(S_n - S_{n-1})$$

$$= [I - \{(1/k) \sum_{r=0}^{k-1} (I-S_{n-1})^{k-r-1}(I-S_n)^r\}]\cdot(S_n - S_{n-1})$$

By application of remarks (2), (3) and (5), the assumption of existence of $(k-1)$st roots, and the inductive hypothesis $S_{n-1} \leqq S_n$, the latter operator product exists and is positive. Hence $S_n \leqq S_{n+1}$, and the sequence $\{S_n\}$ is increasing. Of course, the Law of the Mean is not applicable in this setting, nor is it used other than to motivate the choice of C. Indeed, the discerning reader will note that the extremes of the chain above may be shown to be equal without the introduction of C. However, the rather unusual factorization of $S_{n+1} - S_n$ would be more difficult to discover without the example furnished by the derivative in the real function situation.

To invoke the Lemma and complete the proof of existence of $k^{\underline{th}}$ roots, it remains to show $S_n \leq I$ for all n. Assuming $0 \leqq S_m \leqq I$, we have $kS_{m+1} = R + \sum_{r=2}^{k} \binom{k}{r}(-1)^r S_m^r = R - I + kS_m + (I - S_m)^k$. By remark (5), $(I - S\ )^k \leq I - S_m$; therefore

$$R + kS_m - I + (I - S_m)^k \leq R + kS_m - I + I - S_m$$
$$\leq I + (k-1)S_m \leqq kI.$$

Hence $kS_{m+1} \leqq kI$ and $S_{m+1} \leqq I$, as desired. Thus, the Lemma gives an operator as in (*), and $I - S = B$ is a $k^{\underline{th}}$ root of A.

In order to prove the uniqueness of a positive $k^{\underline{th}}$ root of A, we first observe that if T is any positive $k^{\underline{th}}$ root of A, then T must perforce commute with A, hence with $I - A = R$, hence with each S_n, and thus with S and $I - S = B$. Let $u \in H$, $v = (B-T)u$. Then $0 = \langle (B^k - T^k)u, v \rangle = \langle (\sum_{r=0}^{k-1} B^{k-r-1}T^r)(B-T)u, v \rangle = \sum_{r=0}^{k-1} \langle B^{k-r-1}T^r v, v \rangle$. Since B and T commute, $0 \leq B^{k-r-1}T^r$, whence $\langle B^{k-r-1}T^r v, v \rangle = 0$, $r = 0,1,\ldots,k-1$. Let F_r be any positive (hence symmetric) square root of $B^{k-r-1}T^r$. Then $||F_r v||^2 = \langle F_r v, F_r v \rangle = \langle F_r^2 v, v \rangle = 0$, so that $F_r v = 0$ and $B^{k-r-1}T^r v = F_r^2 v = 0$. Therefore $B^{k-r-1}T^r(B-T)u = 0$. or $B^{k-r}T^r u = B^{k-r-1}T^{r+1}u$, $r = 0,1,\ldots k-1$. In particular, for $r = k-1$, $BT^{k-1} = T^k$. Multiplying by T, we have $B^{k+1} = BA = BT^k = T^{k+1}$.

If $k = 2$, the argument above shows $Bv = 0 = Tv$, whence $||(B-T)u||^2 = \langle (B-T)^2 u, u \rangle = \langle (B-T)v, u \rangle = 0$. Hence $Bu = Tu$ for all $u \in H$, and B is thus unique. Now assume all positive roots, of order less than k, for positive operators are unique. If $k = 2j$, then $(B^j)^2 = B^{2j} = B^k = T^k = (T^j)^2$, whence $B^j = T^j$ and thus $B = T$. If k is odd, we have shown above that $B^{k+1} = T^{k+1}$, so, by the even exponent argument, again B = T. This completes the proof.

REFERENCE

1. Schechter, Martin, Principles of Functional Analysis, Academic Press, New York, 1971.

University of Houston
Houston, Texas, 77004

A FIXED POINT THEOREM FOR CERTAIN OPERATOR VALUED MAPS

D.R. Brown & M.J. O'Malley
Department of Mathematics
University of Houston
Houston, Texas

Report #69

March 1978

A FIXED POINT THEOREM FOR CERTAIN OPERATOR VALUED MAPS

by D.R. Brown and M.J. O'Malley[1]

1. Introduction. Let H be a real Hilbert space, and let $B_1(H)$ denote the space of symmetric, bounded operators on H which have numerical range in $[0,1]$, topologized by the strong operator topology (that is, the topology of point-wise convergence). It is well known [3], that if $T \in B_1(H)$, then there exists a unique $S \in B_1(H)$ such that $S^2 = T$. We represent S by $T^{\frac{1}{2}}$. The following theorem is due to John Neuberger [2].

Theorem A: Suppose $w \in H$, P is an orthogonal projection on H, and L is a (strongly) continuous function from H into $B_1(H)$. Let $Q_o = P$, and set $Q_{n+1} = Q_n^{\frac{1}{2}} L(Q_n^{\frac{1}{2}} w) Q_n^{\frac{1}{2}}$, $n = 0,1,2,\ldots$. Then $\{Q_n\}_{n=o}^{\infty}$ converges to an element $Q \in B_1(H)$ for which $z = Q^{\frac{1}{2}} w$ is a fixed point of P and a fixed point of L in the sense that $L(z)z = z$.

In this paper, under the same hypotheses as Theorem A, we develop a family of Neuberger-like results to find points $z \in H$ satisfying $L(z)z = z$ and $P(z) = z$. This family includes Neuberger's theorem and has the additional property that "most" of the sequences $\{Q_n\}$ converge to idempotent elements of $B_1(H)$. The limit operator of Theorem A need not be idempotent.

Such theorems as those above not only play a valuable role in the search for numerical solutions of partial differential equations, but are also useful, in the finite-dimensional case, in attacking the problem of determining the nonzero

[1] Both authors received partial support under NASA contract NAS-9-15000.

fixed points of a function $\emptyset:R^n \longrightarrow R^n$. In particular, if $x \in R^n-\{0\}$, then x is a fixed point of $\emptyset$ if and only if $A(x)x = x$, where A is the matrix valued function defined by $A(x) = (||x||^{-2})\cdot\emptyset(x)\cdot(x^T)$. In fact, it follows that this can occur if and only if $A(x)$ is a nonzero symmetric idempotent.

It is a pleasure to record our indebtedness to H.P. Decell for the remark immediately above, and to several other members of the University of Houston Mathematics Department, particularly Phillip Walker, for helpful conversations regarding the preparation of this paper.

2. Fixed Points of $L(z)$. Recall that an operator is positive if $\langle Ax,x\rangle \geqq 0$ for all $x \in H$, where $\langle\ ,\ \rangle$ is the inner product of H. We presume familiarity with the standard properties of positive operators as set forth, for example, in [3]. By invocation of the Spectral Theorem, or, alternately, by a sequential construction, it is possible to provide, for any $T \in B_1(H)$ and any positive integer n, a unique operator $T^{1/n} \in B_1(H)$ such that $(T^{1/n})^n = T$. This notion extends immediately to arbitrary positive rational powers of T by defining $T^{r/s} = (T^{1/s})^r$. Moreover, by again appealing to the Spectral Theorem, it follows that if $\{Q_j\}$ is a sequence in $B_1(H)$ converging strongly to Q, and t is an arbitrary positive rational number, then $\{Q_j^t\}$ converges strongly to Q^t. Finally, recall that the usual quasi-order defined for positive operators by $A \leqq B$ if and only if $B - A$ is positive satisfies an additional anti-symmetry condition, to wit: if A and B are positive and commute, then $A \leqq B$ and $B \leqq A$ forces $A = B$.

Lemma 1. Let $Q \in B_1(H)$ and let α be a positive rational number other than 1. If $Q^\alpha = Q$, then $Q = Q^2$; that is, Q is an idempotent.

Proof: Let $\alpha = r/s$; the presumed equality is equivalent to $Q^r = Q^s$. Without loss of generality, assume $r < s$ and that r is the minimal positive power of Q which reoccurs in the sequence $\{Q^n\}$. From the fact that powers of an operator descend in the quasi-order mentioned above, together with the limited anti-symmetry of this relation, it follows that $Q^t = Q^r$ for all integral t between r and s. From $Q^r = Q^{r+1}$, it follows that $Q^t = Q^r$ for all $t \geq r$. If r is odd, then $(Q^{(r+1)/2})^2 = Q^{r+1} = Q^{2r} = (Q^r)^2$. By uniqueness of square roots, $Q^r = Q^{(r+1)/2}$, whence $r = (r+1)/2$ and $r = 1$. If r is even, then $(Q^{r/2})^2 = Q^r = (Q^r)^2$, whence $r = r/2$, which is impossible for positive r. Thus $r = 1$ and $Q = Q^2$.

We are now ready to prove our

Theorem 2. Let $w \in H$, let P be an orthogonal projection on H, and let $L: H \longrightarrow B_1(H)$ be strongly continuous. Let α, β be positive rational numbers with $\alpha \in [\frac{1}{2}, \infty)$. Set $Q_o = P$, and let $Q^\alpha_{n+1} = Q^\alpha_n L(Q^\beta_n w) Q^\alpha_n$, $n = 0, 1, 2, \ldots$. Then $\{Q_n\}_{n=o}^{\infty}$ is a decreasing sequence of elements of $B_1(H)$ which converge to an element $Q \in B_1(H)$ such that

(1) if $\alpha > \frac{1}{2}$, then Q is idempotent and $z = Qw$ satisfies $L(z)z = z$, and $Pz = z$, and

(2) if $\alpha = \frac{1}{2}$ and $\beta \geq \frac{1}{2}$, then $z = Q^\beta w$ satisfies $L(z)z = z$ and $Pz = z$.

Proof: Fix $\alpha \geq \frac{1}{2}$ and $\beta > 0$. Since $Q_o = P \in B_1(H)$ and the range of L

is in $B_1(H)$, it follows inductively that $Q_n \in B_1(H)$ for all n. Since $2\alpha \geqq 1$, $Q_n^{2\alpha} \leqq Q_n$; moreover, $\langle (Q_n^{2\alpha} - Q_{n+1})x,x\rangle = \langle (Q_n^{2\alpha} - Q_n^{\alpha}L(Q_n^{\beta}w)Q_n^{\alpha})x,x\rangle =$ $\langle Q_n^{\alpha}(I - L(Q_n^{\beta}w)Q_n^{\alpha}x,x\rangle = \langle (I - L(Q_n^{\beta}w))Q_n^{\alpha}x,Q_n^{\alpha}x\rangle$. Thus, since $I - L(Q_n^{\beta}w) \geqq 0$, it follows that $Q_{n+1} \leqq Q_n^{2\alpha}$. Hence we have

(*) $$Q_{n+1} \leqq Q_n^{2\alpha} \leqq Q_n,\ n = 0,1,2,\dots .$$

In particular, the sequence $\{Q_n\}$ is monotonically decreasing in the (operator) interval from 0 to I. Thus we have by [3, p.318] that the sequence $\{Q_n\}$ converges strongly to an element $Q \in B_1(H)$, whence $\{Q_n^{\alpha}\}$ converges to Q^{α} and $\{Q_n^{\beta}\}$ converges to Q^{β}. Since L is continuous and operator multiplication is jointly continuous in the strong topology on $B_1(H)$, we have by uniqueness of limits that $Q = Q^{\alpha}L(Q^{\beta}w)Q^{\alpha}$. Also, from (*) and the closed graph of the relation $\leqq$, we have $Q \leqq Q^{2\alpha} \leqq Q$. Thus, since Q and $Q^{2\alpha}$ commute, we have that $Q = Q^{2\alpha}$. Moreover, since $P = Q_o$, we have $PQ_n = Q_n$, whence $PQ^{\gamma} = Q^{\gamma}$ for all positive rational γ.

(i) Suppose $\alpha > \frac{1}{2}$. By lemma 1, $Q = Q^2$, from which it follows that $Q = Q^{\gamma}$ for all positive rational γ, and, in particular, $Q = QL(Qw)Q$.

Let $z = Qw$, and fix $x \in H$. Then $\langle Qx,x\rangle = \langle QL(z)Qx,x\rangle = \langle L(z)Qx,Qx\rangle$, and since $Q^2 = Q$, it follows that $0 = \langle Qx,Qx\rangle - \langle L(z)Qx,Qx\rangle = \langle (I - L(z))Qx,Qx\rangle$. Therefore, since $I{-}L(z)$ and hence $(I{-}L(z))^{\frac{1}{2}}$ belong to $B_1(H)$, we have that $Q = L(z)Q$. In particular, $z = Qw = L(z)Qw = L(z)z$.

(ii) Suppose $\alpha = \frac{1}{2}$, $\beta \geqq \frac{1}{2}$. Let $z = Q^{\beta}w$; then $Q = Q^{\frac{1}{2}}L(z)Q^{\frac{1}{2}}$ from which $\langle Qx,x\rangle = \langle Q^{\frac{1}{2}}L(z)Q^{\frac{1}{2}}x,x\rangle = \langle L(z)Q^{\frac{1}{2}}x,Q^{\frac{1}{2}}x\rangle$. Since $\langle Qx,x\rangle = \langle Q^{\frac{1}{2}}x,Q^{\frac{1}{2}}x\rangle$ also, we have $0 = \langle Q^{\frac{1}{2}}x{-}L(z)Q^{\frac{1}{2}}x,Q^{\frac{1}{2}}x\rangle = \langle (I{-}L(z))Q^{\frac{1}{2}}x,Q^{\frac{1}{2}}x\rangle$. Now, as in (i), it follows that $Q^{\frac{1}{2}} = L(z)Q^{\frac{1}{2}}$. In particular, $z = Q^{\beta}w = Q^{\frac{1}{2}}Q^{\beta-\frac{1}{2}}w = L(z)Q^{\frac{1}{2}}Q^{\beta-\frac{1}{2}}w = L(z)Q^{\beta}w = L(z)z$. That $Pz = z$ in both cases is obvious from the fact that $PQ^{\gamma} = Q^{\gamma}$ for all positive rational γ. This completes the proof.

Given a nonzero element $z \in H$ such that $L(z)z = z$, it is reasonable to ask if our sequences are able to produce z. We note now that, by proper selection of w and P, z is attainable from each of our sequences. Specifically, if α and β are fixed as in the theorem, then let $w = z$ and let P be the orthogonal projection of H onto the line through z. From the construction of the sequence $\{Q_n\}$, $Q_1 = PL(z)P$, whence $Q_1 = P$. If follows immediately that $Q_n = P$ for all n and thus $Q = P$. Hence $z = Qw = Pw$ (or $z = Q^{\beta}w = P^{\beta}w = Pw$) is the fixed point yielded by our theorem.

While it is not reasonable to expect the praticioner to guess P so accurately, these remarks do attach the virtue of theoretical completeness to these processes.

3. Examples. (1) Suppose that $\alpha = \frac{1}{2}$ and that $\gamma, \delta \in [\frac{1}{2},\infty)$ such that neither of γ, δ is an integral multiple of the other. We show that for fixed $w \in H$ and P, the Q and z obtained by using γ for β need not be the same as those obtained by using δ for β. Moreover, the limit operator Q in this case need not be an idempotent, although it can be one. Assume $\delta < \gamma$. Let k be the least positive integer such that $\gamma < k\delta$. Note $2 \leqq k$ and $(k-1)\delta < \gamma$. Let a be any number in the interval $(0,1)$. Then $a^{k\delta} < a^{\gamma} < a^{(k-1)\delta} \leqq a^{\delta}$.

Define $L:R \longrightarrow [0,1]$ by

$$L(x) = \begin{cases} 1 , & x \leq a^{\gamma} \\ [(1-a)/(a^{\gamma}-a^{(k-1)\delta})]\cdot(x-a^{\gamma}) + 1 , & a^{\gamma} \leq x \leq a^{(k-1)\delta} \\ a , & a^{(k-1)\delta} \leq x. \end{cases}$$

Set $P = 1$, $w = 1$. Using γ for β in the theorem yields $Q_o = 1$ and $Q_1 = a$. Inductively, $Q_n = a$, so that $Q = a$. Hence $z = Q^{\gamma}w = a^{\gamma}\cdot 1 = a^{\gamma}$ in this case. On the other hand, using δ for β gives $Q_o = 1$, $Q_1 = a$, but $Q_2 = a^2, \ldots, Q_k = a^k$. Moreover, $Q_n = a^k$ for $n \geq k$, hence $Q = a^k$ and $z = Q^{\delta}w = a^{k\delta}\cdot 1 = a^{k\delta}$. By the choices of a and k, the exponents γ and δ yield distinct operators and distinct fixed points. Moreover, neither of the limit operators determined by γ and δ is idempotent.

(2) Suppose that $\alpha > \frac{1}{2}$, so that any limiting Q obtained through the theorem is idempotent. We show for fixed $w \in H$ and P, that the resulting limit idempotents may vary with the choice of β, as may the fixed points determined in this manner. To this end, let $\alpha = 1$ in the theorem. Let $L:R^3 \longrightarrow B_1(R^3)$ be as follows: all image matrices are diagonal, where $\begin{pmatrix} x & 0 & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{pmatrix}$ will be represented as $\text{diag}(x,y,z)$. We require $L(1,1,1) = \text{diag}(1,\frac{1}{4},1)$, $L(1,\frac{1}{2},1) = \text{diag}(1,\frac{1}{4},\frac{1}{4})$, $L(1,\frac{1}{4},1) = \text{diag}(\frac{1}{4},\frac{1}{4},1)$, $L(1,y,z) = \text{diag}(1,y,z)$ for $(y,z) \in [0,\frac{1}{2}] \times [0,\frac{1}{2}]$, and $L(x,y,1) = \text{diag}(x,y,1)$ for $(x,y) \in [0,\frac{1}{2}] \times [0,\frac{1}{2}]$. The extension theorem of Tietze (c.f. [1]) permits a continuous extension of L to all of R^3 into the diagonal matrices whose entries are in the interval $[0,1]$. Let $P = I_3$, the identity operator, and let w be the vector $(1,1,1)$. If $\beta = \frac{1}{2}$, a brief examination of the defining sequence of Q_n's in Theorem 2

shows that the limit idempotent $Q = \mathrm{diag}(1,0,0)$, and $z = Qw = (1,0,0)$. On the other hand, if $\beta = 1$, then limit $Q = \mathrm{diag}(0,0,1)$, and $z = (0,0,1)$.

(3) With notation as in (2), suppose $\beta = 1$ is fixed. We show for fixed $w \in H$ and P, that the resulting limit idempotents may vary with α, as may the fixed points determined in this manner. Letting $P = I_3$ and $w = (1,1,1)$ as in (2), we require this time that $L(1,1,1) = L(1,\frac{1}{2},1) = \mathrm{diag}(1,\frac{1}{2},1)$, $L(1,1/8,1) = L(1,0,0) = \mathrm{diag}(1,0,0)$, and $L(1,1/32,1) = L(0,0,1) = \mathrm{diag}(0,0,1)$. Extending as before, we have a continuous L defined on R^3 into the diagonal matrices with entries in $[0,1]$. For any choice of α, $Q_1 = \mathrm{diag}(1,\frac{1}{2},1)$. If $\alpha = 1$, $Q_2 = \mathrm{diag}(1,1/8,1)$, $Q_3 = Q_n = Q = \mathrm{diag}(1,0,0)$, $z = (1,0,0)$. On the other hand, if $\alpha = 2$, then $Q_2 = \mathrm{diag}(1,1/32,1)$, $Q_3 = Q_n = Q = \mathrm{diag}(0,0,1)$, $z = (0,0,1)$.

It is easy to see that a slightly more complicated definition of L would yield a single example incorporating the features of all three prior illustrations.

References

1. Kelley, John L., General Topology, Van Nostrand, New York, 1955.

2. Neuberger, John, Projection Methods for Linear and Nonlinear Systems of Partial Differential Equations, Springer-Verlag Lecture Notes (to appear).

3. Schechter, Martin, Principles of Functional Analysis, Academic Press, New York, 1971.

University of Houston
Houston, Texas 77004