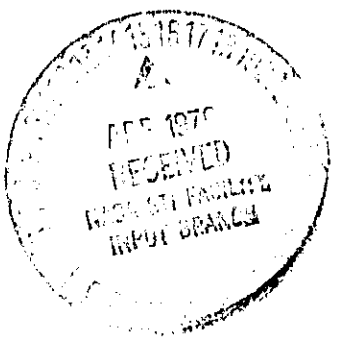
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

APPLICATION OF MATRIX DERIVATIVES TO OPTIMIZATION PROBLEMS IN STATISTICAL PATTERN RECOGNITION

BY JOSEPH S. MORRELL

REPORT #45 AUGUST 1975



PREPARED FOR
EARTH RESOURCES LABORATORY
UNDER
CONTRACT NAS-12777

HOUSTON, TEXAS 77004

Applications of Matrix Derivatives To Optimization Problems in Statistical Pattern Recognition

August, 1975

Joseph S. Morrell
Department of Mathematics
University of Southern Mississippi

Report #45

Abstract

A necessary condition for a real valued Frechet differentiable function of a vector variable to have an extremum at a vector x_o is that the Frechet derivative vanishes at x_o. This paper establishes a relationship between Frechet differentials and matrix derivatives obtaining a necessary condition on the matrix derivative at an extrema. These results are applied to various scalar functions of matrix variables which occur in statistical pattern recognition.

Applications of Matrix Derivatives To Optimization Problems in Statistical Pattern Recognition

1. Introduction.

Let S be a transformation on a normed space X to a normed space Y. If for, $x \in X$, there is a bounded linear operator $A \in B(X,Y)$ such that

$$1.1) \qquad \lim_{||h||\to 0} \frac{||S(x+h) - S(x) - A(h)||}{||h||} = 0$$

then S is Frechet differential at x. The vector $A(h)$ is referred to as the Frechet differential of S at x with increment h and A is denoted by $\delta S(x,)$ or $S'(x)$.

We list below some important properties of the Frechet differential. For proofs and a detailed treatment of Frechet differentials see [6 ,pp.175-178]

Theorem 1.1: If S has a Frechet differential, it is unique

Theorem 1.2: If S has a Frechet differential at x, then S is continuous at x.

Theorem 1.3: Let f be a real valued function which is Frechet differentiable at $x_o \in X$. If f has an extrema at x_o, then $\delta f(x_o,h) = 0$ for all $h \in X$.

Example 1.4: Let $X = R^n$ where R is the field of real numbers and let $f(x) = f(x_1,\ldots,x_n)$ be a real valued function on X having continuous partial derivatives existing with respect to each variable x_i. Then the Frechet

differential is

$$\delta f(x,h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}_x h_i.$$

We denote by $M_{r\times s}$ the vector space of real $r\times s$ matrices. For $A \in M_{r\times s}$ we denote the element in the ith row and jth column of A by $\langle A\rangle_{ij}$. Let $tr(A) = \sum_{i=1}^{n}\langle A\rangle_{ii}$, the usual trace of A, and let A^T denote the transpose of A.

The set $M_{r\times s}$ is a normed space with

$$||A|| = [tr(AA^T)]^{1/2} = [\sum_{i=1}^{r}\sum_{j=1}^{s}(\langle A\rangle_{ij})^2]^{1/2} \quad \text{for}$$

$A \in M_{rs}$. An inner product compatible with this norm is given by

$$(A,B) = \sum_{i=1}^{r}\sum_{j=1}^{s}\langle A\rangle_{ij}\cdot\langle B\rangle_{ij} = tr(A\cdot B^T).$$

Let $A \in M_{p\times q}$ have each entry a function of the entries of $X \in M_{m\times}$ and let $\frac{\partial\langle A\rangle_{ij}}{\partial\langle X\rangle_{\gamma\delta}}$ exist for all $1 \le i \le p$, $i \le j \le q$, $1 \le \gamma \le m$, $1 \le \delta \le n$.

We define $\frac{\partial A}{\partial\langle X\rangle_{\gamma\delta}} \in M_{p\times q}$ and $\frac{\partial\langle A\rangle_{ij}}{\partial X} \in M_{m\times n}$ by

$$\left\langle\frac{\partial A}{\partial\langle X\rangle_{\gamma\delta}}\right\rangle_{ij} = \frac{\partial\langle A\rangle_{ij}}{\partial\langle X\rangle_{\gamma\delta}} = \left\langle\frac{\partial\langle A\rangle_{ij}}{\partial X}\right\rangle_{\gamma\delta}.$$

We make the convention that all partials are taken considering the entries of X

as being independent unless otherwise specified. For example

Example 1.5: $\frac{\partial X}{\partial \langle X \rangle_{\gamma\delta}} = J_{\gamma\delta}$, the m n matrix with

$$\langle J_{\gamma\delta} \rangle_{ij} = \begin{cases} 1 & \text{if } \gamma = i \text{ and } \delta = j \\ 0 & \text{otherwise} \end{cases}$$

. The above holds even in the symmetric case due to our convention. For future reference we denote by K_{ij} the $p \times q$ matrix with $\langle K_{ij} \rangle_{\gamma\delta} = \begin{cases} 1 & \text{if } i = \gamma \text{ and } j = \delta \\ 0 & \text{otherwise} \end{cases}$ and by J_j the $n \times 1$ vector with 1 in the jth component and zero elsewhere.

Example 1.6: Let $Y = tr(X)$ and $X \in M_{m \times n}$. Then $\frac{\partial Y}{\partial X} = I_{m \times n}$, the identity in $M_{m \times n}$.

One writes $\frac{\partial Y}{\partial X}$ instead of $\frac{\partial \langle Y \rangle_{ij}}{\partial X}$ or $\frac{\partial Y}{\partial \langle X \rangle_{\gamma\delta}}$ if and only if X or Y is a scalar.

Example 1.7: Let $|X|$ denote the determinant of X. Then $\frac{\partial |X|}{\partial \langle X \rangle_{\gamma\delta}} = cof(\langle X \rangle_{\gamma\delta})$. Thus $\frac{\partial |X|}{\partial X} = cof(X)$ and if, X is full rank, $\frac{\partial |X|}{\partial X} = |X| X^{-T}$, where $X^{-T} = (X^{-1})^T$.

Several equations are listed below which are easily verified using component-wise arguments.

Let $y = f(X)$ be a scalar function of $X \in M_{m \times n}$ and $u(y)$ a scalar function of y.

1.2) $$\frac{\partial u}{\partial X} = \frac{\partial u}{\partial y} \cdot \frac{\partial y}{\partial X} \quad \text{if} \quad \frac{\partial u}{\partial y} \quad \text{and} \quad \frac{\partial y}{\partial X} \quad \text{exist}$$

Example 1.8: If X is full rank and $t \in R$, then

$$\frac{\partial |X|^t}{\partial X} = t|X|^{t-1} \frac{\partial |X|}{\partial X} = t|X|^t X^{-T}.$$

Let $Y(X) \in M_{r\times s}$ and $W(X) \in$. Then

1.3) $$\frac{\partial(W\cdot Y)}{\partial\langle X\rangle_{\gamma\delta}} = W\frac{\partial Y}{\partial\langle X\rangle_{\gamma\delta}} + \frac{\partial W}{\partial\langle X\rangle_{\gamma\delta}}Y$$

if $\frac{\partial Y}{\partial\langle X\rangle_{\gamma\delta}}$ and $\frac{\partial W}{\partial\langle X\rangle_{\gamma\delta}}$ exist .

If $f(X)$ is full rank and $\frac{\partial f(X)}{\partial\langle X\rangle_{\gamma\delta}}$ exist, then by equation 1.3) we have

1.4) $$\frac{\partial f^{-1}(X)}{\partial\langle X\rangle_{\gamma\delta}} = -f^{-1}(X)\frac{\partial f(X)}{\partial\langle X\rangle_{\gamma\delta}}f^{-1}(X)$$

Example 1.9: If X is full rank, then

$$\frac{\partial(X^{-1})}{\partial\langle X\rangle_{\gamma\delta}} = -X^{-1}J_{\gamma\delta}X^{-1}.$$

Let $Y = Y(X) \in M_{p\times q}$ and $w = w(Y) \in M_{r\times s}$ with $X \in M_{m\times n}$. If the indicated derivatives exist , then

1.5) $$\frac{\partial w}{\partial\langle X\rangle_{\gamma\delta}} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\partial w}{\partial\langle Y\rangle_{ij}}\cdot\frac{\partial\langle Y\rangle_{ij}}{\partial\langle X\rangle_{\gamma\delta}}$$

and, if w is scalar valued,

$$\frac{\partial w}{\partial X} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\partial w}{\partial\langle Y\rangle_{ij}}\frac{\partial\langle Y\rangle_{ij}}{\partial X}.$$

In his excellent paper on matrix derivatives Dwyer proves an extremely useful theorem establishing a procedure for calculating $\frac{\partial \langle Y \rangle_{ij}}{\partial X}$ when $\frac{\partial Y}{\partial \langle X \rangle_{\gamma\delta}}$ is of a certain form. We state the theorem without proof. For a proof of this theorem see Dwyer [4, p 612].

Theorem 1.10: Let $X \in M_{m\times n}$ and $Y \in M_{p\times \ell}$. Then

$$\frac{\partial Y}{\partial \langle X \rangle_{\gamma\delta}} = \sum_q A_q J_{\gamma\delta} B_q + \sum_h C_h J_{\gamma\delta}^T D_h$$

if and only if

$$\frac{\partial \langle Y \rangle_{ij}}{\partial X} = \sum_q A_q^T \cdot K_{ij} \cdot B_q^T + \sum_h D_h \cdot K_{ij}^T \cdot C_h .$$

All matrix multiplications must be defined when applying this theorem. This condition must also be observed in applying the following corollary which is a restatement of the theorem for the scalar valued case.

Corollary 1.11: Let f be scalar valued. Then

$$\frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\delta}} = \sum_q A_q \cdot J_{\gamma\delta} \cdot B_q + \sum_h C_h \cdot J_{\gamma\delta}^T \cdot D_h$$

if and only if

$$\frac{\partial f(X)}{\partial X} = \sum_q A_q^T \cdot B_q^T + \sum_h D_h \cdot C_h .$$

2. The Frechet Differential of Matrix Valued Functions of a Matrix Variable.

The following theorem establishes a relationship between the Frechet differential and the matrix derivatives.

Theorem 2.1: Let f be an operator from $M_{m\times n}$ to $M_{p\times q}$ with $\frac{\partial \langle f(X)\rangle_{ij}}{\partial \langle X\rangle_{\gamma\delta}}$ continuous for $1 \le i \le p$, $1 \le j \le q$, $1 \le \gamma \le m$, $1 \le \delta \le n$, with all entries of X independent. Then f is Frechet differentiable and

$$\delta f(X,H) = \sum_{\gamma,\delta} \frac{\partial f(X)}{\partial \langle X\rangle_{\gamma\delta}} \langle H\rangle_{\gamma\delta}.$$

Proof: $\delta f(X,H)$ is obviously linear in H.

$$||\delta f(X,H)|| = \Big\| \sum_{\gamma,\delta} \frac{\partial f}{\partial \langle X\rangle_{\gamma\delta}} \langle H\rangle_{\gamma\delta} \Big\| \le \sum_{\gamma,\delta} |\langle H\rangle_{\gamma\delta}| \, \Big\| \frac{\partial f}{\partial \langle X\rangle_{\gamma\delta}} \Big\| .$$

Applying the Cauchy inequality,

$$||\delta f(X,H)|| \le \Big(\sum_{\gamma,\delta} |\langle \rangle_{\gamma\delta}|^2\Big)^{1/2} \Big(\sum_{\gamma,\delta} \Big\| \frac{\partial f}{\partial \langle X\rangle_{\gamma\delta}} \Big\|^2\Big)^{\frac{1}{2}} = ||H|| \sum_{\gamma,\delta} \Big\| \frac{\partial f}{\partial \langle X\rangle_{\gamma\delta}} \Big\|^2\Big)^{\frac{1}{2}} .$$

Thus $\delta f(X,\cdot)$ is bounded with operator norm

$$||\delta f(X,\cdot)|| \le \Big(\sum_{\gamma,\delta} \Big\| \frac{\partial f}{\partial \langle X\rangle_{\gamma\delta}} \Big\|^2\Big)^{1/2}.$$

Let $H \in M_{m\times n}$. For $1 \le i \le m$, $i \le j \le n$ let

$G_{i,j} = \sum_{\ell<i} \sum_{k=1}^{n} \langle H\rangle_{\ell k} K_{\ell k} + \sum_{k\le j} \langle H\rangle_{ik} K_{ik}$. Let $G_{1,0} = \Theta$ and $G_{i,0} = G_{i-1,n}$ for $1 < i \le r$. For $X \in M_{m\times n}$, $f(X+H) - f(X) = \sum_{i=1}^{m} \sum_{j=1}^{m} f(X+G_{i,j}) - f(X+G_{i,j-1})$ and $(X+G_{i,j}) - (X+G_{i,j-1}) = \langle H\rangle_{ij} J_{ij}$.

Using the mean-value theorem and the continuity of $\frac{\partial \langle f(X)\rangle_{k\ell}}{\partial \langle X\rangle_{ij}}$, we have for each $\varepsilon > 0$, a $\delta > 0$ such that $||H|| < \delta$ implies

$$|\langle f(X+G_{i,j})\rangle_{k\ell} - \langle f(X+G_{i,j-1})\rangle_{k\ell} - \frac{\partial \langle f(X)\rangle_{k\ell}}{\partial \langle X\rangle_{ij}} \langle H\rangle_{ij}| < \frac{\varepsilon\,||H||}{4pqmn}$$

for all $1 \le i \le m$, $1 \le j \le n$, $1 \le k \le p$, $1 \le \ell \le q$.

Let $\varepsilon > 0$ and $||H|| < \delta$ for $\delta > 0$ described above. Then

$$\frac{||f(X+H) - f(X) - \delta f(X,H)||}{||H||} \le \frac{1}{||H||} \Big|\Big| \sum_{i,j} f(X+G_{i,j}) - f(X+G_{i,j-1}) - \frac{\partial f(X)}{\partial \langle X\rangle_{ij}} \langle H\rangle_{ij} \Big|\Big|$$

$$\le \sum_{i,j} ||f(X,G_{i,j}) - f(X+G_{i,j-1}) - \frac{\partial f(X)}{\partial \langle H\rangle_{ij}} \langle H\rangle_{ij}||$$

$$= \sum_{i,j} \Big[\sum_{k,\ell} \Big(\langle f(X+G_{i,j}) - f(X+G_{i,j-1})\rangle_{k\ell} - \frac{\partial \langle f(X)\rangle_{k\ell}}{\partial \langle X\rangle_{ij}} \langle H\rangle_{ij}\Big)^2\Big]^{\frac{1}{2}}$$

$$< \sum_{i,j} \Big[\sum_{k,\ell} \Big(\frac{\varepsilon}{4pqmn}\Big)^2\Big]^{1/2} < \varepsilon. \quad \text{Thus}$$

$\lim_{||H||\to 0} \frac{||f(X+H) - f(X) - \delta f(X,H)||}{||H||} = 0$ and the theorem is proved.

If the entries of X are not independent then a parallel argument will show that $\delta f(X,H) = \sum \frac{\partial \bar{f}(X)}{\partial \langle X \rangle_{\gamma\delta}} \langle H \rangle_{\gamma\delta}$ where the summation is over all γ,δ for which $\langle X \rangle_{\gamma\delta}$ is independent and $\frac{\partial \bar{f}(X)}{\partial \langle X \rangle}$ is taken with the entries of X not being considered as independent.

If f is a scalar valued function of a matrix X and is Frechet differentiable at X, where all entries of X are independent the conclusion of the theorem can be stated as $\delta f(X,H) = (\frac{\partial f(X)}{\partial X}, H)$, the inner product of H with the matrix derivative of f at X. It is shown below that this relationship holds in certain cases in which the entries of X are not independent.

Theorem 2.2: Let f be a scalar valued Frechet differentiable function of a symmetric matrix. Then for symmetric H,

$$\delta f(X,H) = (\frac{\partial f(X)}{\partial X}, H).$$

Proof. From the statement following Theorem 2.1 we have

$$\delta f(X,H) = \sum_{i \leq j} \frac{\partial f}{\partial \langle X \rangle_{ij}} \cdot \langle H \rangle_{ij}, \text{ where } \frac{\partial \bar{f}}{\partial \langle X \rangle ij}$$

is taken without the usual convention of treating the entries of X as being independent. From elementary calculus, $\frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\delta}} = \sum_{i,j} \frac{\partial f(X)}{\partial \langle X \rangle_{ij}} \frac{\partial \langle X \rangle_{ij}}{\partial \langle X \rangle_{\gamma\delta}}$ when $\frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\delta}}$ is taken by treating all entries as being independent. When X is symmetric and $\delta \neq \gamma$,

$$\frac{\partial \bar{f}(X)}{\partial \langle X \rangle_{\gamma\delta}} = \frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\delta}} + \frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\delta}} \quad \text{and}$$

$$\frac{\partial \bar{f}(X)}{\partial \langle X \rangle_{\gamma\gamma}} = \frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\gamma}}$$

since it is assumed that symetery is the only dependency condition. Thus

$$\delta f(X,H) = \sum_{i=1}^{n} \frac{\partial f(X)}{\partial \langle X \rangle_{ii}} \langle H \rangle_{ii} + \sum_{i<j} \left[\frac{\partial f(X)}{\partial \langle X \rangle_{ij}} + \frac{\partial f(X)}{\partial \langle X \rangle_{ji}}\right] \langle H \rangle_{ij}$$

$$= \sum_{i=1}^{n} \frac{\partial f(X)}{\partial \langle X \rangle_{ii}} \langle H \rangle_{ii} + \sum_{i<j} \frac{\partial f(X)}{\partial \langle X \rangle_{ij}} \langle H \rangle_{ij} + \sum_{i>j} \frac{\partial f(X)}{\partial \langle X \rangle_{ij}} \langle H \rangle_{ij}$$

$$= \sum_{i,j} \frac{\partial f(X)}{\partial \langle X \rangle_{ij}} \langle H \rangle_{ij}$$

$$= \left(\frac{\partial f(X)}{\partial X}, H\right) \quad \text{and the theorem is proved.}$$

If f satisfies the hypothesis of Theorem 2.2 and X and H are diagonal, then a similar argument proves that $\delta f(X,H) = \left(\frac{\partial f}{\partial X}, H\right)$.

The matrices $A,B \in M_{m\times n}$ are <u>orthogonal</u> if $(A,B) = \theta$. The only matrix $M_{m\times n}$ which is orthogonal to every matrix in $M_{m\times n}$ is θ, the zero matrix. The only symmetric matrix orthogonal to every orthogonal matrix is θ, and the only diagonal matrix orthogonal to every diagonal matrix is θ. An immediate consequence of the above facts and Theorem 2.2 is

<u>Corollary</u> 2.3: Let f be a scalar valued function of a matrix X which has all entries independent, is symmetric, or is diagonal and satisfies the hypothesis

of Theorem 2.2. If f has an extreme value at X_0, then the matrix derivative of f vanishes at X_0.

3. Applications to MLSE, MLEST, and PMC.

In the remainder of this paper the following notation will be used. Let $\{x_k\}_{k=1,\ldots,N}$ be a collection of N samples each having n features ($x_k \in R^n$), with the samples taken from a mixture of m classes. Let $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$ be signature parameters where $\alpha_i \in R$, $\mu_i \in R^n$ and $\Sigma_i \in M_{m\,n}$ are respectively, the a priori probability, the mean vector, and the covariance matrix for observations from the ith class. The distribution $p(x_k)$ is normal with

$$p(x_k) = \sum_{i=1}^{m} \alpha_i \, p_i(x_k)$$

where the conditional density function, $p_i(x_k)$, is given by

$$p_i(x_k) = \frac{1}{(2\Pi)^{n/2}|\Sigma_i|^{1/2}} \; e^{-1/2 \, \Gamma_i}$$

where

$$\Gamma_i = (x_k - \mu_i)^T \Sigma_i^{-1} (x_k - \mu_i).$$

If some of the signature parameters are known, a maximum likelihood signature estimate (MLSE) is a choice of the remaining parameters which maximize the log-likelihood function

$$L = \sum_{k=1}^{N} \log(p(x_k)).$$

We now obtain the likelihood equations by taking the matrix derivative of L with respect to the means and covariance in each class and applying Corollary 2.3.

$$\frac{\partial L}{\partial \Sigma_i} = \sum_{k=1}^{N} \frac{\alpha_i}{p(x_k)} [e^{-1/2\Gamma_i} \frac{\partial |\Sigma_i|^{-1/2}}{\partial \Sigma_i} + \frac{1}{|\Sigma_i|^{1/2}} e^{-1/2\Gamma_i} (-\frac{1}{2} \frac{\partial \Gamma_i}{\partial \Sigma_i})]$$

From Example 1.8,

$$\frac{\partial |\Sigma_i|^{-1/2}}{\partial \Sigma_i} = -\frac{1}{2} |\Sigma_i|^{-1/2} \Sigma_i^{-T}.$$

Since

$$\frac{\partial \Gamma_i}{\partial <\Sigma_i>_{\gamma\delta}} = (x_k - \mu_i)^T (-\Sigma_i^{-1} J_{\gamma\delta} \Sigma_i^{-1})(x_k - \mu_i),$$

then by Corollary 1.11,

$$\frac{\partial \Gamma_i}{\partial \Sigma_i} = -\Sigma_i^{-T}(x_k - \mu_i)(x_k - \mu_i)^T \Sigma_i^{-T}. \quad \text{Thus}$$

$$\frac{\partial L}{\partial \Sigma_i} = \sum_{k=1}^{N} \alpha_i \frac{p_i(x_k)}{p(x_k)} [-\frac{1}{2} \Sigma_i^{-T} + \frac{1}{2} \Sigma_i^{-T}(x_k - \mu_i)(x_k - \mu_i)^T \Sigma_i^{-T}].$$

Applying Corollary 2.3 and the symmetry of Σ_i, we have

$$(\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}) \Sigma_i^{-1} = \Sigma_i^{-1} (\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} (x_k - \mu_i)(x_k - \mu_i)^T] \Sigma_i^{-1},$$

and hence

$$3.1) \qquad \Sigma_i = \left(\frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}(x_k - \mu_i)(x_k - \mu_i)^T\right) \Big/ \left(\frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right)$$

at extrema.

Next $$\frac{\partial L}{\partial \mu_i} = \sum_{k=1}^{N} \frac{\alpha_i}{p(x_k)} p_i(x_k) \left(-\frac{1}{2}\frac{\partial \Gamma_i}{\partial \mu_i}\right),$$

where

$$\frac{\partial \Gamma_i}{\partial \langle \mu_i \rangle_j} = (x_k - \mu_i)^T \Sigma_i^{-1}(-J_j) + (-J_j^T)\Sigma_i^{-1}(x_k - \mu_i).$$

Hence

$$\frac{\partial \Gamma_i}{\partial \mu_i} = 2\Sigma_i^{-1}(x_k - \mu_i),$$

and

$$\frac{\partial L}{\partial \mu_i} = \sum_{k=1}^{N} \alpha_i \frac{p_i(x_k)}{p(x_k)}[\Sigma_i^{-1}(x_k - \mu_i)].$$

At extrema we obtain

$$3.2) \qquad \mu_i = \left(\frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} x_k\right) \Big/ \left(\frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right).$$

By use of a Lagrange multiplier to enforce the constraints $\sum_{i=1}^{m}\alpha_i = 1$ and $\alpha_i \geq 0$, the following expression is obtained at extrema:

$$3.3) \qquad \alpha_i = \frac{\alpha_i}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} .$$

Equations 3.1), 3.2), 3.3) are the <u>likelihood equations</u> which serve as a point of departure for the results on MLSE by Walker and Peters in [8].

Maximum likelihood estimation of signature transformation (MLEST) is a procedure that adjusts signatures from a training segment to compensate for haze and sun angle, assuming that the adjustments are given by an affine transformation $x_k = Ay_k + b$, where A is an $n \times n$ non-singular matrix and b is an n-vector which transforms y_k, the kth pixel from the training segment to x_k, the kth pixel in the recognition segment. Using this transformation, a set of parameters for the recognition setment is obtained as follows:

$$\mu_i' = A\mu_i + b$$

$$\Sigma_i' = A\Sigma_i A^T$$

The conditional density function of the transformed ith class is

$$p_i(x_k) = \frac{1}{(2\Pi)^{n/2}|A\Sigma_i A^T|^{1/2}} e^{-1/2\Gamma_i'},$$

where

$$\Gamma_i' = (x_k - A\mu_i - b)^T (A\Sigma_i A^T)^{-1} (x_k - A\mu_i - b).$$

The transformed likelihood function is

$$L' = \sum_{k=1}^{N} \log(p'(x_k)),$$

where

$$p'(x_k) = \sum_{i=1}^{m} \alpha_i \, p_i'(x_k).$$

Throughout the remainder of this discussion the primes will be suppressed.

Now $$\frac{\partial L}{\partial A} = \sum_{k=1}^{N} \sum_{i=1}^{m} \frac{\alpha_i}{p(x_k)} \frac{\partial p_i(x_k)}{\partial A},$$

where

$$\frac{\partial p_i(x_k)}{\partial A} = \frac{1}{(2\Pi)^{n/2}} \left[e^{-1/2\Gamma_i} \frac{\partial |A\Sigma_i A^T|^{-1/2}}{\partial A} + \frac{1}{|A\Sigma_i A^T|^{1/2}} e^{-1/2\Gamma_i} \left(-\frac{1}{2} \frac{\partial \Gamma_i}{\partial A}\right)\right].$$

Also,

$$\frac{\partial \Gamma_i}{\partial \langle A \rangle_{\gamma\delta}} = -\mu_i^T J_{\gamma\delta}^T (A\Sigma_i A^T)^{-1} (x_k - A\mu_i - b) + (x_k - A\mu_i - b)^T (A\Sigma_i A^T)^{-1} (-J_{\gamma\delta}\mu_i)$$

$$+ (x_k - A\mu_i - b)^T (-A\Sigma_i A^T)^{-1} (J_{\gamma\delta}\Sigma_i A^T + A\Sigma_i J_{\gamma\delta}^T)(A\Sigma_i A^T)^{-1} (x_k - A\mu_i - b).$$

Thus

$$\frac{\partial \Gamma_i}{\partial A} = -2(A\Sigma_i A^T)^{-1}(x_k - A\mu_i - b)[\mu_i^T + (x_k - A\mu_i - b)^T A^{-T}]$$

$$= -2(A\Sigma_i A^T)^{-1}(x_k - A\mu_i - b)(x_k - b)^T A^{-T}.$$

$$\frac{\partial |A\Sigma_i A^T|^{-1/2}}{\partial \langle A \rangle_{\gamma\delta}} = -\frac{1}{2}|A\Sigma_i A^T|^{-1/2}(A\Sigma_i A^T)^{-1}[A\Sigma_i J_{\gamma\delta}^T + J_{\gamma\delta}\Sigma_i A^T].$$

Thus

$$\frac{\partial |A\Sigma_i A^T|^{-1/2}}{\partial A} = -|A\Sigma_i A^T|^{-1/2}(A\Sigma_i A^T)^{-1}A\Sigma_i = -|A\Sigma_i A^T|^{-1/2}A^{-T},$$

since A is invertible. Thus

$$\frac{\partial L}{\partial A} = \sum_{k=1}^{N}\sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)} A^{-T}[-I + \Sigma_i^{-1}A^{-1}(x_k - A\mu_i - b)(x_k - b)^T A^{-T}].$$

At extrema we have

$$[\sum_{k=1}^{N}\sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)}]I = [\sum_{k=1}^{N}\sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)} \Sigma_i^{-1}A^{-1}(x_k - A\mu_i - b)(x_k - b)^T]A^{-T}.$$

Hence

3.4) $$A = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)}(x_k - b)(x_k - A\mu_i - b)^T A^{-T}\Sigma_i^{-1}.$$

Next, $$\frac{\partial L}{\partial b} = \sum_{k=1}^{N} \sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)} \left(-\frac{1}{2} \frac{\partial \Gamma_i}{\partial b}\right),$$

where

$$\frac{\partial \Gamma_i}{\partial \langle b \rangle_j} = (x_k - A\mu_i - b)^T (A\Sigma_i A^T)^{-1} (J_j) + (-J_j^T)(A\Sigma_i A^T)^{-1}(x_k - A\mu_i - b)$$

and hence

$$\frac{\partial \Gamma_i}{\partial b} = -2(A\Sigma_i A^T)^{-1}(x_k - A\mu_i - b).$$

Thus

$$\left[\sum_{k=1}^{N} \sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)} (A\Sigma_i A^T)^{-1}\right] b = \sum_{k=1}^{N} \sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)} (A\Sigma_i A^T)^{-1}(x_k - A\mu_i)]$$

at extrema and

3.5) $$b = A\left[\sum_{k=1}^{N} \sum_{i=1}^{m} \alpha_i \frac{p_i(x_k)}{p(x_k)} \Sigma_i^{-1}\right]^{-1} \left[\sum_{k=1}^{N} \sum_{i=1}^{m} \Sigma_i^{-1} A^{-1}(x_k - A\mu_i)\right].$$

Equations 3.3), 3.4), and 3.5) are the transformed likelihood equations which serve as a starting point for recent results on MLEST obtained by McCabe and Solomon in [7].

In [5] Walker and Guseman show that probability of misclassification (PMC) of a transformed observation is a differentible function of the kxn feature selection matrix B which transforms normally distributed observations in R^n to normally distributed observations in R^k. Calculating the differentials of the PMC is a direct result of calculating the differentials of the transformed density function. By making use of previous calculations we give an abbreviated

version of these calculations.

The transformed density function is

$$p(x,B) = \frac{1}{(2\Pi)^{n/2}|B\Sigma B^T|^{1/2}} \; e^{-1/2\Gamma},$$

where

$$\Gamma = (x - B\mu)^T(B\Sigma B^T)^{-1}(x - B\mu).$$

From the calculations in the MLEST problem,

$$\frac{\partial|B\Sigma B^T|^{-1/2}}{\partial B} = - |B\Sigma B^T|^{-1/2}(B\Sigma B^T)^{-1}B\Sigma$$

and

$$\frac{\partial\Gamma}{\partial B} = -2(B\Sigma B^T)^{-1}(x - B\mu)(\mu^T) - 2(B\Sigma B^T)^{-1}(x - B\mu)(x - B\mu)^T(B\Sigma B^T)^{-1}B\Sigma.$$

Hence

$$\frac{\partial p(x,B)}{\partial B} = - p(x,B)\{(B\Sigma B^T)^{-1}(B\Sigma) - (B\Sigma B^T)^{-1}(x - B\mu)[\mu^T + (x - B\mu)^T(B\Sigma B^T)^{-1}B\Sigma]\}.$$

The Frechet differential of p at (x,B) with increment C is $(\frac{\partial p(x,B)}{\partial B},C)$ and using the above expression for $\frac{\partial p(x,B)}{\partial B}$, it is easy to show that

$$(\delta p(x,B),C) = -p(x,B)\{\mathrm{tr}[C\Sigma B^T(B\Sigma B^T)^{-1} - (x-B)^T(B\Sigma B^T)^{-1}[C\mu + \tfrac{1}{2}(C\Sigma B^T + B\Sigma C^T)\ (B\Sigma B^T)^{-1}(x-B\mu)]\},$$

which is the form of the differential obtained in [5].

4. Matrix Derivatives of traces and applications.

We begin this section with the main theorem.

Theorem 4.1: Let $u = u(Y)$ be a scalar valued function of $Y = Y(X)$ where $Y \in M_{p\times r}$ and $X \in M_{m\times n}$. If $\frac{\partial u}{\partial Y}$ and $\frac{\partial Y}{\partial \langle X\rangle_{\gamma\delta}}$ exist and

$$\frac{\partial Y}{\partial \langle X\rangle_{\gamma\delta}} = \sum_q A_q\, J_{\gamma\delta}\, B_q + \sum_h C_h\, J^T_{\gamma\delta}\, D_h,$$

then

$$\frac{\partial u}{\partial X} = \sum_q A_q^T \left(\frac{\partial u}{\partial Y}\right) B_q^T + \sum_h D_h \left(\frac{\partial u}{\partial Y}\right)^T C_h.$$

(Remark: As in previous theorems on matrix derivatives, it is essential to determine if the aforementioned side condition concerning matrix multiplication is satisfied.)

Proof: From equation 1.5,

$$\frac{\partial u}{\partial X} = \sum_{i,j} \frac{\partial u}{\partial \langle Y\rangle_{ij}} \frac{\partial \langle Y\rangle_{ij}}{\partial X} = \sum_{i,j} \left\langle \frac{\partial u}{\partial Y} \right\rangle_{ij} \frac{\partial \langle Y\rangle_{ij}}{\partial X}.$$

By applying Theorem 1.10 to the hypothesis, we have

$$\frac{\partial \langle Y \rangle_{ij}}{\partial X} = \sum_q A_q^T K_{ij} B_q^T + \sum_h D_h K_{ij}^T C_h.$$

Thus

$$\frac{\partial u}{\partial X} = \sum_{i,j} \frac{\partial u}{\partial Y}_{ij} \left(\sum_q A_q^T K_{ij} B_q^T + \sum_h D_h K_{ij}^T C_h \right)$$

$$= \sum_q A_q^T \left(\sum_{i,j} \frac{\partial u}{\partial Y}_{ij} K_{ij} \right) B_q^T + \sum_h D_h \left(\sum_{i,j} \cdot \frac{\partial u}{\partial Y}_{ij} K_{ij}^T \right) C_h$$

$$= \sum_q A_q^T \left(\frac{\partial u}{\partial Y} \right) B_q^T + \sum_h D_h \left(\frac{\partial u}{\partial Y} \right)^T C_h$$

and the theorem is proved.

The following result (due to Dwyer in [2]) is the most useful form of the above theorem, especially in applications to multivariate statistical analysis.

Corollary 4.2: Let $f(X)$ be a matrix valued function of a matrix X. If

$$\frac{\partial f(X)}{\partial \langle X \rangle_{\gamma\delta}} = \sum_q A_q J_{\gamma\delta} B_q + \sum_h C_h J_{\gamma\delta}^T D_h,$$

then

$$\frac{\partial \mathrm{tr}(f(X))}{\partial X} = \sum_q A_q^T B_q^T + \sum_h D_h C_h.$$

Proof: From Example 1.6 we have $\frac{\partial \mathrm{tr}(Y)}{\partial Y} = I$. Now apply the theorem.

The Corollary above provides an effective tool for solving optimization problems dealing with the trace function. To illustrate the strength of the corollary, we obtain some short cuts to finding necessary conditions for extrema for some important trace functions.

In [10] Quirein, using extensive and tedious calculations, obtained the derivatives of

$$\phi = \mathrm{tr}(B\Lambda B^T) - \mathrm{tr}\{M^T[(B\Lambda B^T)^{-1} - (BDB^T)^{-1} - I_k]\}$$

with respect to B and D, where B is a k×n matrix of rank k, Λ is an n×n positive definite, symmetric matrix, D is a positive definite diagonal matrix, and I_k is the k×k identity matrix. We present a less painful method of calculating these derivatives.

Since

$$\frac{\partial(B\Lambda B^T)}{\partial \langle B \rangle_{\gamma\delta}} = J_{\gamma\delta}\,\Lambda B^T + B\Lambda\, J_{\gamma\delta}^T,$$

and

$$\frac{\partial\, M^T[(B\Lambda B^T)^{-1} - (BDB^T)^{-1} - I_k]}{\partial \langle B \rangle_{\gamma\delta}} = M^T\{-(B\Lambda B^T)^{-1}[J_{\gamma\delta}\Lambda B^T + B\Lambda J_{\gamma\delta}^T](B\Lambda B^T)^{-1} + (BDB^T)^{-1}[J_{\gamma\delta}DB^T + BDJ_{\gamma\delta}^T](BDB^T)^{-1}\},$$

then by Corollary 4.2,

$$\frac{\partial \phi}{\partial B} = 2B\Lambda + 2(BDB^T)^{-1}M^T(BDB^T)B\Lambda - 2(BDB^T)^{-1}M^T(BDB^T)^{-1}BD$$

$$= 2[I_k + (BDB^T)^{-1}M^T(BDB^T)^{-1}]B\Lambda - 2(BDB^T)^{-1}M^T(BDB^T)^{-1}BD.$$

Since

$$\frac{\partial M^T[(B\Lambda B^T)^{-1} - (BDB^T)^{-1} - I_k]}{\partial <D>_{\gamma\delta}} = M^T(BDB^T)^{-1}B\, J_{\gamma\delta}\, B^T(BDB^T)^{-1},$$

then

$$\frac{\partial \phi}{\partial D} = -B^T(BDB^T)^{-1}M(BDB^T)^{-1}B.$$

In [9] Quirein extremizes the B-average interclass divergence

$$D_B = \frac{1}{2}\,\mathrm{tr}(Q) - \frac{m(m-1)}{2}\,k, \text{ with}$$

$$Q = \sum_{i=1}^{m}(B\Lambda_i B^T)^{-1}(BS_i B^T)$$

where B is $k\times n$ of rank k, Λ_i is symmetric $n\times n$ of full rank, and S_i is symmetric for $i=1,\ldots,m$. (Actually, Λ_i is the covariance matrix for the ith class and $S_i = \sum_{\substack{j=1 \\ j\neq i}}^{m}[\Lambda_j + \delta_{ij}\delta_{ij}^T]$ where $\delta_{ij} = \mu_i - \mu_j$, the difference the ith and jth class means.)

We again present an abbreviated version of the calculation.

Since

$$\frac{\partial Q}{\partial \langle B \rangle_{\gamma\delta}} = \sum_{i=1}^{m} (B\Lambda_i B^T)^{-1}[J_{\gamma\delta} S_i B^T + B S_i J_{\gamma\delta}^T] -$$

$$(B\Lambda_i B^T)^{-1}[J_{\gamma\delta}\Lambda_i B^T + B\Lambda_i J_{\gamma\delta}^T](B\Lambda_i B^T)^{-1}(BDB^T),$$

then

$$\frac{\partial D_B}{\partial B} = \sum_{i=1}^{m} (B\Lambda_i B^T)^{-1}[BS_i - (BS_i B^T)(B\Lambda_i B^T)^{-1} B\Lambda_i].$$

The next application will be in the problem of extremizing B-average interclass divergence in a reduced feature space with respect to the generator of a single Householder transformation, B, which compresses n-feature data to k features. In [2] Decell and Mayekar addressed this problem in modified form and Decell and others (see [1] and [3]) have made significant progress in the area of feature selection. The result below suggest another possible approach to the feature selection problem.

Let D_B be as in previous example and $B = (I_k | Z)(I - 2\frac{UU^T}{U^T U})$, where $(I_k|Z)$ is $k \times n$ with I_k, the $k \times k$ identity, in the first $k \times k$ block and zeroes elsewhere, and U is a non-zero $n \times 1$ vector.

First,

$$\frac{\partial(\frac{UU^T}{U^T U})}{\partial \langle U \rangle_j} = -\frac{1}{(U^T U)^2}\{U^T U[UJ_j^T + J_j U^T] - UU^T[J_j^T U + U^T J_j]\}.$$

Using the fact that U^TU, J_j^TU, and U^TJ_j are scalars, we can write

$$\frac{\partial B}{\partial \langle U\rangle_j} = -\frac{2(I_k|Z)}{(U^TU)^2}\{[UJ_j^TU^TU + U^TUJ_jU^T] - [UJ_j^TUU^T + UU^TJ_jU^T]\}.$$

Thus by Theorem 4.1,

$$\frac{\partial D_B}{\partial U} = -\frac{2}{(U^TU)^2}\{[U^TU(\frac{\partial D_B}{\partial B})^T(I_k|Z)U + U^TU(I_k|Z)^T(\frac{\partial D_B}{\partial B})U]$$

$$-[UU^T(\frac{\partial D_B}{B})^T(I_k|Z)U + UU^T(I_k|Z)^T(\frac{\partial D_B}{B})U]$$

Thus at extrema we have

$$\frac{-2}{(U^TU)}[I - \frac{UU^T}{U^TU}][(\frac{\partial D_B}{B})^T(I_k|Z) + (I_k|Z)^T(\frac{\partial D_B}{B})]U = \theta.$$

Thus, to extremize D_B, it is necessary to solve the equation $(I - \frac{UU^T}{U^TU})\ A(U)\ U = \theta$, where $A(U) = [(\frac{\partial D_B}{B})^T(I_k|Z) + (I_k|Z)^T(\frac{\partial D_B}{B})]$, the $n\times n$ matrix of rank k. The above equality is equivalent to

$$A(U)\ U = \lambda U \quad \text{for some} \quad \lambda \in R$$

since $\frac{UU^T}{U^TU}$ projects in the U direction.

The eignevalue problem $A(U)\ U = \lambda U$ suggest possible iteration schemes, but a scheme with good convergence properties is yet to be determined and the question remains open.

References

1. H.P. Decell, Jr. and J.A. Quirein, An Iterative Approach to the Feature Selection Problem, University of Houston, Department of Mathematics Report #26, March, 1973.
2. H.P. Decell, Jr. and M. Mayekar, On the Variational Equations for Householder Transformations in Feature Selection, University of Houston, Department of Mathematics Report #39, June, 1974.
3. H.P. Decell, Jr. and W.G. Smiley, III, Householder Transformations and Optimal Linear Combinations, University of Houston, Department of Mathematics Report #38, June, 1974.
4. Paul S. Dwyer, Some Applications of Matrix Derivatives in Multivariate Analysis, JASA, June, 1967, pp. 607-625.
5. L.F. Guseman, Jr. and Homer F. Walker, The Differentibility of the Probability of Misclassification, University of Houston, Mathematics Department Report #28, August, 1973.
6. David G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, Inc. New York, 1969.
7. Tom McCabe and Jimmy Solomon, An Iterative Scheme for Computing an Affine Transformation for Signature Extension, to appear.
8. B. Charles Peters, Jr. and Homer F. Walker, An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions, to appear.
9. J.A. Quirein, Divergence-Some Necessary Condition for an Extremum, University of Houston, Mathematics Department Report #12 November 1972.
10. J.A. Quirein, Partial Derivatives for Various Scalar Functions of Matrices, University of Houston, Department of Mathematics Report #29, July, 1973.