DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR-141880

(NASA-CR-141880) ON THE VARIATIONAL EQUATIONS FOR HOUSEHOLDER TRANSFORMATIONS IN FEATURE SELECTION (Houston Univ.) 8 p HC $3.25 CSCL 12A

N75-26743

Unclas
G3/64 27285

ON THE VARIATIONAL EQUATIONS
FOR HOUSEHOLDER TRANSFORMATIONS
IN FEATURE SELECTION

BY H. P. DECELL, JR. & M. MAYEKAR

JUNE 1974 REPORT #39

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

On The Variational Equations For Householder Transformations in Feature Selection -- Divergence

March, 1975

by

Henry P. Decell, Jr. and Mike Mayekar
Department of Mathematics

Report 39
NAS-9-12777

Introduction

In [9] Decell and Smiley and in [2] Decell and Quirein have results that suggest the possibility of using a sequential monotone process for solving the feature selection problem (multivariate normal populations and best k linear combinations) using Householder transformations. The results are general in that they apply to a large class of separability criteria [9].

In this report these results will be applied to the divergence separability criterion and an expression for the gradient of the divergence (in the reduced feature space) with respect to the generator of a single Householder transformation will be developed. This expression for the gradient can be used in any number of differential correction schemes (iterators) that attempt to extremize the divergence (in the reduced feature space).

Two data sets provided by the Earth Observations Division-JSC are used to demonstrate selecting the Householder transformations that generate the $k \times n$ matrix defining the "best" (in the sense of extremizing the divergence) k linear combinations of features. The tests allow initial comparisons to be made with results obtained in [2]. In particular, this new technique does not appear to require initial guesses for the iterator to be generated by the without replacement, exhaustive search, or other similar schemes.

An Expression for the Gradient

Using the results in [2] and [9] we need only calculate the gradient of the function

$$Q(U,\lambda) = D_B + \lambda(U^T U - 1)$$

where $B = (I_K/Z)(I - 2UU^T)$, λ a Lagrange multiplier and

$$D_B \equiv \frac{1}{2} \operatorname{tr} \sum_{i}^{m} \{(BV_i B^T)^{-1}(BS_i B^T)\} - \frac{m(m-1)K}{2}$$

As usual: V_i ; $i=1,\ldots,m$ are the class covariances

δ_{ij}; $i,j=1,\ldots,m$ are the difference in class means

and

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

First Taking differential of D_B, it is easily verified $dD_B = F + G$

where
$$F = \tfrac{1}{2}\,\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(BV_iB^T)^{-1}\,(dBS_iB^T + BS_i\,dB^T)\right\}\right]$$

$$= \tfrac{1}{2}\,\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(dBS_iB^T)\,(BV_iB^T)^{-1}\right]\right. + \tfrac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(BV_iB^T)^{-1}(BS_i\,dB^T)\right\}\right]$$

$$= \tfrac{1}{2}\,\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(dBS_iB^T)\,(BV_iB^T)^{-1}\right]\right. + \tfrac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(dBS_iB^T)\,(BV_iB^T)^{-1}\right\}^T\right]$$

$$= \mathrm{tr}\left\{\sum_{i=1}^{m}(dBS_iB^T)\,(BV_iB^T)^{-1}\right\}$$

$$G = -\tfrac{1}{2}\,\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(BV_iB^T)^{-1}\,(dBV_iB^T + BV_i\,dB^T)\,(BV_iB^T)^{-1}\,(BS_iB^T)\right\}\right]$$

$$= -\tfrac{1}{2}\,\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(dBV_iB^T)\,(BV_iB^T)^{-1}\,(BS_iB^T)\,(BV_iB^T)^{-1}\right\}\right]$$

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

$$= -\,\mathrm{tr}\left[\sum_{i=1}^{m}\left\{(dBV_iB^T)\,(BV_iB^T)^{-1}\,(BS_iB^T)\,(BV_iB^T)^{-1}\right\}\right]$$

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

Now we define
$$H_i = \left[\left\{S_iB^T - V_iB^T\,(BV_iB^T)^{-1}\,(BS_iB^T)\right\}\,(BV_iB^T)^{-1}\right]$$

Hence $$dD_B = \mathrm{tr} \sum_{i=1}^{m} [dBH_i]$$

$$= \mathrm{tr} \sum_{i=1}^{m} [d\{(I_K/Z)(I - 2UU^T)\}H_i]$$

$$= -2\mathrm{tr} \sum_{i=1}^{m} [(I_K/Z)(dUU^T + UdU^T)H_i]$$

$$= -2\mathrm{tr} \sum_{i=1}^{m} [(I_K/Z)dUU^T H_i] - [(I_K/Z)UdU^T H_i]$$

$$= -2\mathrm{tr} \sum_{i=1}^{m} [H_i{}^T UdU^T \left(\frac{I_K}{Z}\right)]^T - [H_i(I_K/Z)UdU^T]$$

$$= -2\mathrm{tr} \sum_{i=1}^{m} [\left(\frac{I_K}{Z}\right)H_i^T UdU^T] - [H_i(I_K/Z)UdU^T]$$

$$= -2\mathrm{tr} \sum_{i=1}^{m} [\{H_i(I_K/Z)\}^T U + \{H_i(I_K/Z)\}U]dU^T$$

If we calculate the differential of $\lambda(U^TU - 1)$ with respect to U we have

$$d(U^TU - 1) = \lambda(U^TdU + dU^TU) = \lambda\{(dU^TU)^T + dU^TU\}$$

$$= \lambda\{\mathrm{tr}(dU^TU)^T + \mathrm{tr}(dU^TU)\}$$

$$= 2\lambda \mathrm{tr}(dU^TU) = 2\lambda \mathrm{tr}(UdU^T)$$

Clearly the differential of $\lambda(U^TU - 1)$ with respect to λ is $d\lambda(U^TU - 1)$ so that if we define the matrix

$$P(U) = \sum_{i=1}^{m} \{[H_i(I_K/Z)]^T + H_i(I_K/Z)\}$$

it follows that

$$\text{Grad } Q(U,\lambda) = \begin{pmatrix} -2P(U)U + 2\lambda U \\ \hdashline U^T U - 1 \end{pmatrix}$$

$$= -2 \begin{pmatrix} P(U)U - \lambda U \\ \hdashline \dfrac{U^T U - 1}{-2} \end{pmatrix}$$

Routine to find Maximum Average Divergence D_B

I. Take the starting value $U_o = \begin{pmatrix} \frac{1}{N} \\ \cdot \\ \cdot \\ \cdot \\ \frac{1}{N} \end{pmatrix}$

Compute initial B matrix $B(U_o) = (I_K/Z)(I - 2U_oU_o^T)$ and the value of $D_B(U_o)$

$$D_{B(U_o)} = \frac{1}{2} \text{trace}\{\sum_{i=1}^{m} (BV_iB^T)^{-1}(BS_IB^T)\} - \frac{m(m-1)}{2} K$$

Use a crude variation of the Steepest Descent Method to extremize D_B.

$$\text{Max } D_B \Longleftrightarrow \text{Min}(-D_B)$$

$$U_{p+1} = \{U_p - \alpha \text{ Grad}(-D_{B(U_p)})\} \Big/ ||U_p - \alpha \text{ Grad}(-D_{B(U_p)})|| \cdot D_{B(U_p)}$$

where

$$\mathrm{Grad}(-D_{B(U_p)}) = 2\sum_{i=1}^{m} \{H_i(I_K/Z)\}^T + H_i(I_K/Z)\ U_p$$

and

$$B(U_p) = (I_K/Z)(I - 2U_pU_p^T)$$

Compute the B-matrix with the new value of V and also the corresponding value of D_B. Repeat the procedure until $D_{B(U_p)}$ begins to stabilize.

II. The same procedure as in I except V_i is replaced by $H_1V_iH_1$ and S_i by $H_1S_iH_1$ where $H_1 = (I - 2UU^T)$, U is the value obtained at max D_B in I.

III. The same procedure as in II except $H_1V_iH_1$ is replaced by $H_2H_1V_iH_1H_2$ and $H_1S_iH_1$ is replaced by $H_2H_1S_iH_1H_2$.

IV. Continue, V continue... etc. until D_B does not increase as a function of Roman numeral steps. Note that the iteration in each phase (i.e. I, II, III, etc) uses the same arbitrary initial guess $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^T$. In addition, an attempt to satisfy the constraint $U^TU = 1$ is <u>forced</u> <u>arbitrarily</u> on the steepest decent procedure. This is a very crude scheme and potentially generates error. Moreover, the step size α is taken to be constant in all phases and is obviously inefficient (except in the decrease in step size caused by the division

The following test cases seem to indicate relative insensitivity to these crude iteration adjustments. More sophisticated, careful computations are being implemented to further refine the technique and eliminate these inefficiencies. The technique will be available on the LARS terminal shortly.

Results: Data Set I (210 Flight Line)

N = 12, m = 9, K = 6, B is 6 by 12 matrix.

Total Divergence D = 10660

D_{B11}	3686	D_{B21}	8221	D_{B31}	8697
D_{B12}	6639	D_{B22}	9248	D_{B32}	9730
D_{B13}	7769	D_{B23}	9786	D_{B33}	9940
D_{B14}	7843	D_{B24}	9944	D_{B34}	9994
D_{B15}	7605	D_{B25}	9987	D_{B35}	10018
D_{B16}	6093	D_{B26}	10020	D_{B36}	10035
D_{B17}	5825	D_{B27}	10028	D_{B37}	10047
D_{B18}	7279	D_{B28}	10032	D_{B38}	10056

Data Set II (Hill County)

N = 16, m = 5, K = 6

Total Divergence D = 636

D_{B11}	93	D_{B21}	227	D_{B31}	228
D_{B12}	106	D_{B22}	274	D_{B32}	275
D_{B13}	113	D_{B23}	275	D_{B33}	276
D_{B14}	129	D_{B24}	260	D_{B34}	280
D_{B15}	153	D_{B25}	287	D_{B35}	288
D_{B16}	183	D_{B26}	290	D_{B36}	290
D_{B17}	220	D_{B27}	293	D_{B37}	294
D_{B18}	223	D_{B28}	298	D_{B38}	300

RE.:ERENCES

1. Quirein, J. A., "Som. Necessary Conditions For An Extreme."

2. Decell, H. P., and Quirein, J. A., "An Iterative Approach to the Feature Selection Problem," Report #26 NAS-9-12777, Department of Mathematics, University of Houston, March, 1973.

3. Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic Press, New York and London, 1972.

4. Toy, J. T., Computer and Information Sciences, II, Academic Press, New York and London, 1967.

5. Householder, Alston S., Unitary Triangularization of a Non-Symmetric Matrix, J. Assoc. Comput. Mach., 5 (1958), 339-342.

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

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

8. Marani, Salma, "Routine to Find the Maximum Average Divergence." Report #36 NAS-9-12777, Department of Mathematics, University of Houston, August 1974.

9. Decell, H. P., and Smiley, William, "Householder Transformations and Optimal Linear Combinations," Report #38 NAS-9-12777, Department of Mathematics, University of Houston.