Construction of a Vector Equivalent to a Given Vector from the Point of View of the Analysis of Principal Components

Yves Escoufier, P. Robert, and J. Cambon, Montpellier

I. INTRODUCTION - All random vectors considered in this paper are assumed to be centered, with elements belonging to the set $L_2(\Omega, \mathcal{A}, \mathcal{P})$ of random variables with finite variances on $(\Omega, \mathcal{A}, \mathcal{P})$.

Let $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ be two such vectors ; let $\underset{\sim}{Z}_1$ and $\underset{\sim}{Z}_2$ be the random vectors of their principal components, respectively. In section III we study the problem of attaching weights to the elements of $\underset{\sim}{X}_2$ in such a way that the vector of principal components Z_2^* of this weighted vector be as closed as possible to $\underset{\sim}{Z}_1$. In section IV we study ways of choosing a subvector $\underset{\sim}{X}_2^*$ of $\underset{\sim}{X}_2$, simultaneously attaching weights to the selected elements, so that the vector Z_2^* of principal components of the resulting weighted $\underset{\sim}{X}_2^*$ will best approximate $\underset{\sim}{Z}_1$, over all choices of $\underset{\sim}{X}_2^*$ and all weighting rules.

When conducting a study based on principal components of a vector difficult or costly to observe, one would welcome a theoritically sound method of substituting to it a vector easier to measure and having approximately the same principal components. At this time not much is available. On the one hand, it is known that the simplest non-singular transformation of the original vector may perturb deeply the principal components ; on the other hand, the results to be presented here indicate that the currently proposed techniques ([BEALE et alt 1967], [JOLIFFE 1973]) may suffer from insufficient theoritical foundations.

In section II we recall essential results obtained by ESCOUFIER ([1970], [1973-a], [1973-b]) which are the basis of the solutions to the problems investigated in this paper. To comply with printing space regulations, proofs are omitted (see quoted references).

II. METHEMATICAL BASES - Let $\underset{\sim}{X}$ be a pxl random vector and $k \leq p$ be the rank of $E(\underset{\sim}{X}\underset{\sim}{X}')$. It is known that there exists a pxp orthogonal matrix $\underset{\sim}{H}^*$ such that

$$\underset{\sim}{H}^{*\prime}\, E(\underset{\sim}{X}\underset{\sim}{X}')\, \underset{\sim}{H}^* = \left[\begin{array}{c|c} \underset{\sim}{D} & 0 \\ \hline 0 & 0 \end{array}\right]$$

where $\underset{\sim}{D}$ is a kxk diagonal matrix with positive diagonal elements. Let $\underset{\sim}{H}$ be the pxk submatrix of $\underset{\sim}{H}^*$ such that $\underset{\sim}{H}'\, E(\underset{\sim}{X}\underset{\sim}{X}')\, \underset{\sim}{H} = \underset{\sim}{D}$.

<u>Definition 1</u>. The kxl random vector $\underset{\sim}{Z} = \underset{\sim}{H}'\underset{\sim}{X}$ is called the vector of principal components of $\underset{\sim}{X}$.

<u>Definition 2</u>. The pxl random vector $\underset{\sim}{X}_1$, with the vector of principal components $\underset{\sim}{Z}_1$, and the gxl random vector $\underset{\sim}{X}_2$, with the vector of principal components $\underset{\sim}{Z}_2$, are said to be equivalent if there exist a positive integer $k \leqslant \min(p,g)$, a constant α and a kxk matrix $\underset{\sim}{C}$ such that :

i) $\underset{\sim}{C}\underset{\sim}{C}' = \underset{\sim}{C}'\underset{\sim}{C} = \alpha \underset{\sim}{I}_k$, ii) $\underset{\sim}{Z}_1 = \underset{\sim}{C}'\underset{\sim}{Z}_2$.

($\underset{\sim}{I}_k$ is the kxk identity matrix).

It is easily verified that the relation established in Definition 2 is an equivalence relation. Its value, in our context, stems from the following properties [ESCOUFIER, 1973-b] :

<u>Proposition 1</u>. If $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ are equivalent vectors, then

i) for each diagonal element $\delta_i^{(1)}$ of $E(\underset{\sim}{Z}_1\underset{\sim}{Z}'_1)$ there is a diagonal element $\delta_j^{(2)}$ of $E(\underset{\sim}{Z}_2\underset{\sim}{Z}'_2)$ such that $\delta_j^{(2)} = \alpha\delta_i^{(1)}$;

ii) there exist permutation matrices $\underset{\sim}{P}$ and $\underset{\sim}{Q}$ such that

$$\underset{\sim}{Q}\,\underset{\sim}{C}\,\underset{\sim}{P} = \left|\begin{array}{c:c:c:c} \underset{\sim}{A}_1 & 0 & 0 \cdots & 0 \\ \hdashline 0 & \underset{\sim}{A}_2 & 0 & 0 \\ \hdashline \vdots & & \ddots & \vdots \\ \hdashline 0 & 0 & \cdots & \underset{\sim}{A}_r \end{array}\right|$$

where r is the number of distinct diagonal elements of $E(\underset{\sim}{Z}_2\underset{\sim}{Z}'_2)$;

iii) if all the diagonal elements of $E(\underset{\sim}{Z}_2\underset{\sim}{Z}'_2)$ are distinct, there is a permutation σ of the indices 1,2,...,k for which $Z_{1i} = \pm\sqrt{\alpha}\, Z_{2,\sigma(i)}$.

Proposition 1 should make clear what is to be understood by the "equivalence" of $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$.

The next step is to find an easily computable criterion for the equivalence of two vectors. ESCOUFIER [1973-a] associates to a vector $\underset{\sim}{X}$ a linear transformation $U_{\underset{\sim}{X}}$ defined on $L_2(\Omega, \mathcal{A}, \mathcal{P})$ by

$$U_{\underset{\sim}{X}}(Y) = \sum_{i=1}^{p} [\, E(X_i Y) \cdot X_i] \text{ for all } Y \in L_2(\Omega, \mathcal{A}, \mathcal{P}).$$

The following properties are then established :

a)- the set of operators $U_{\underset{\sim}{X}}$ is a subset of the class of Hilbert-Schmidt operators on L_2 and so is a Hilbert space ;

b)- if $\underset{\sim}{\Sigma} = \begin{bmatrix} \underset{\sim}{\Sigma}_{11} & \underset{\sim}{\Sigma}_{12} \\ \underset{\sim}{\Sigma}_{21} & \underset{\sim}{\Sigma}_{22} \end{bmatrix}$ is the covariance matrix of the vector $\underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{bmatrix}$, the scalar product for the Hilbert space considered is given by

$$\left\langle U_{\underset{\sim}{X}_1}, U_{\underset{\sim}{X}_2} \right\rangle = \operatorname{Tr}\,(\underset{\sim}{\Sigma}_{12}\, \underset{\sim}{\Sigma}_{21}) \;;$$

c)- if Z_k is a principal component of $\underset{\sim}{X}$ associated with the eigenvalue λ_k, then

$$U_{\underset{\sim}{X}}(Z_k) = \lambda_k Z_k .$$

From the above, the following practical criterion can be deduced [ESCOUFIER, 1973-b] :

<u>Proposition 2</u>. The vectors $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ are equivalent in the sense of Definition 2 if and only if the coefficient

$$RV(\underset{\sim}{X}_1, \underset{\sim}{X}_2) = \operatorname{Tr}\,(\Sigma_{12}\,\Sigma_{21}) \Big/ \left[\operatorname{Tr}\,(\underset{\sim}{\Sigma}^2_{11}).\operatorname{Tr}\,(\underset{\sim}{\Sigma}^2_{22})\right]^{1/2}$$

is equal to 1.

III. <u>THE FIRST PROBLEM</u> - From the last proposition it can be seen that the first problem can be restated as :

- Given vectors $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ find a diagonal matrix $\underset{\sim}{\Delta}$ so as to maximize $RV(\underset{\sim}{Y}, \underset{\sim}{\Delta}\underset{\sim}{X})$, where

$$RV(\underset{\sim}{Y}, \underset{\sim}{\Delta}\underset{\sim}{X}) = \operatorname{Tr}\left[\underset{\sim}{\Delta}E(\underset{\sim}{X}\underset{\sim}{Y}')E(\underset{\sim}{Y}\underset{\sim}{X}')\underset{\sim}{\Delta}\right] \Big/ \left\{\operatorname{Tr}\left[E(\underset{\sim}{Y}\underset{\sim}{Y}')\right]^2.\operatorname{Tr}\left[\underset{\sim}{\Delta}E(\underset{\sim}{X}\underset{\sim}{X}')\underset{\sim}{\Delta}\right]^2\right\}^{1/2};$$

Hence look for $\underset{\sim}{\Delta}$ to maximize

$$\operatorname{Tr}\left[\underset{\sim}{\Delta}E(\underset{\sim}{X}\underset{\sim}{Y}')\,E(\underset{\sim}{Y}\underset{\sim}{X}')\underset{\sim}{\Delta}\right] \Big/ \left\{\operatorname{Tr}\left[\underset{\sim}{\Delta}E(\underset{\sim}{X}\underset{\sim}{X}')\underset{\sim}{\Delta}\right]^2\right\}^{1/2}.$$

Denote by m_j the j-th diagonal element of $\underset{\sim}{\Delta}$, by v_{ij} the elements of $E(\underset{\sim}{X}\underset{\sim}{X}')$, by u_{ij} the elements of $E(\underset{\sim}{Y}\underset{\sim}{X}')$ and define the vectors $\underset{\sim}{U}$, $\underset{\sim}{M}$ and the matrix $\underset{\sim}{A}$ by

$$\underset{\sim}{U} = (U_j) \text{ where } U_j = \sum_i u_{ij},$$

$$\underset{\sim}{M} = (M_j) \text{ where } M_j = m_j^2,$$

$$\underset{\sim}{A} = (A_{ij}) \text{ where } A_{ij} = v_{ij}^2 .$$

By simply expanding the traces it can be verified that

$$\operatorname{Tr}\left[\underset{\sim}{\Delta}E(\underset{\sim}{X}\underset{\sim}{Y}')\,E(\underset{\sim}{Y}\underset{\sim}{X}')\underset{\sim}{\Delta}\right] = \underset{\sim}{U}'\underset{\sim}{M},$$

$$\operatorname{Tr}\left[\underset{\sim}{\Delta}E(\underset{\sim}{X}\underset{\sim}{X}')\underset{\sim}{\Delta}\right]^2 = \underset{\sim}{M}'\underset{\sim}{A}\underset{\sim}{M} .$$

Thus we are lead to the quadratic programming problem for $\underset{\sim}{M}$:

<u>Problem P1</u> $\left\{ \begin{array}{l} \text{Minimize } \underset{\sim}{M}'\underset{\sim}{A}\underset{\sim}{M} \\ \text{Subject to } \underset{\sim}{U}'\underset{\sim}{M} = 1 \text{ and } \underset{\sim}{M} \geqslant 0. \end{array} \right.$

Note that if $E(\underset{\sim}{X}\underset{\sim}{X}')$ is positive semi-definite, so is A [GOWER, 1971]. In such a case, the function $(\underset{\sim}{M}'\underset{\sim}{A}\underset{\sim}{M})$ is convex on the convex set ($\underset{\sim}{U}'\underset{\sim}{M} = 1$ and $\underset{\sim}{M} \geqslant 0$).

It was found convenient to substitute to Problem P1 on the unknown $\underset{\sim}{M}$ the following equivalent problem on the unknown vector $\underset{\sim}{N}$:

<u>Problem P2</u> $\left\{ \begin{array}{l} \text{Minimize } \underset{\sim}{N}'\ \underset{\sim}{C}\ \underset{\sim}{N}, \\ \text{Subject to } \underset{\sim}{e}'\underset{\sim}{N} = 1 \text{ and } \underset{\sim}{N} \geqslant 0, \end{array} \right.$

where $\underset{\sim}{e}' = (1\ 1 \ldots\ldots 1)$ and where $\underset{\sim}{N}$ and $\underset{\sim}{C}$ are defined by :

$$\underset{\sim}{N} = (N_j) \text{ where } N_j = U_j\, M_j = U_j m_j^2 ,$$

$$\underset{\sim}{C} = (C_{ij}) \text{ where } C_{ij} = \frac{A_{ij}}{U_i U_j} = \frac{v_{ij}^2}{U_i U_j} .$$

The authors have solved Problem P2 for many sets of statistical data. Numerical evidence on these test cases will not be given here but reported on elsewhere [CAMBON, 1974] .

To be noted is the fact that the constraint $\underset{\sim}{e}'\underset{\sim}{N} = 1$ will force most of the currently proposed quadratic programming algorithms into the difficulty of a degenerated case. For this reason the authors have developped a specific algorithm which will now be justified and summarized.

$\underset{\sim}{C}$ being (nxn), denote by I a subset of $\{1,2,\ldots, n\}$ and by J the complement of I. Denote by $\underset{\sim}{C}_I$ the submatrix of $\underset{\sim}{C}$ formed by the C_{ij}'s with $i \in I$, $j \in I$; by $\underset{\sim}{C}_J$ the submatrix of $\underset{\sim}{C}$ formed by the C_{ij}'s with $i \in J$, $j \in I$. From the Kuhn-Tucker conditions for Problem P2, it can be proved that :

<u>Proposition 3</u>. Let $\underset{\sim}{N}_I = \underset{\sim}{C}_I^{-1}\ \underset{\sim}{e}$. If

$$\underset{\sim}{N}_I \geqslant 0 \quad \text{and} \quad \underset{\sim}{C}_J \underset{\sim}{N}_I \geqslant e \quad , \tag{1}$$

then the vector $\underset{\sim}{N} = (N_i)$:

$$N_i = \begin{cases} \dfrac{1}{\underset{\sim}{e}'\underset{\sim}{N}_I}\ N_{Ii} & \text{if } i \in I, \\ 0 & \text{if } i \in J, \end{cases} \tag{2}$$

is an optimal solution of Problem P2, and $\underset{\sim}{N}'\underset{\sim}{C}\underset{\sim}{N} = (\underset{\sim}{e}'\underset{\sim}{N}_I)^{-1}$.

A technique to find an optimal solution would be to start computing $\underset{\sim}{N}_I = \underset{\sim}{C}_I^{-1} \underset{\sim}{e}$ for all subsets of indices I and to retain the first found to satisfy (1). The number of sets I to be considered can be greatly restricted in accordance with the following propositions.

Proposition 4. Let $\underset{\sim}{P} = (P_i) = \frac{\underset{\sim}{C}^{-1}\underset{\sim}{e}}{\underset{\sim}{e}'\underset{\sim}{C}^{-1}\underset{\sim}{e}}$, $I^* = \left\{ i : P_i \geqslant 0 \right\}$, $J^* = \left\{ i : P_i < 0 \right\}$. There is an optimal solution $\underset{\sim}{N} = (N_i)$ of Problem P2 such that $N_j = 0$ for at least one index $j \in J^*$.

Proof. It is known that $\underset{\sim}{P}$ is an optimal solution of P2 without the positivity constraint. Consideranу feasible vector $\underset{\sim}{Q} = (Q_i)$ for P2 such taht $Q_j > 0$ for all $j \in J^*$. Let $r \in J^*$ be such taht $Q_r/P_r \geqslant Q_j/P_j$ for all $j \in J^*$, $k = (Q_r/P_r)/(1-Q_r/P_r)^{-1}$, and $\underset{\sim}{Q}^k = (1-k)\underset{\sim}{P} + k\underset{\sim}{Q}$. Since the function F to be minimized is convex and $0 \leqslant k \leqslant 1$, $F(\underset{\sim}{Q}^k) \leqslant F(\underset{\sim}{Q})$ But $\underset{\sim}{Q}^k$ is feasible for P2 and $Q_r^k = 0$. The conclusion follows. ∎

Proposition 5. An optimal solution of Problem P2 can be found as the best of the optimal solutions of the $|J^*|$ (n-1)-dimensional problems on $\underset{\sim}{N}$:

$$\left\{ \begin{array}{l} \text{minimize} \quad \underset{\sim}{N}' \underset{\sim}{C}_{(\mathcal{N} - j)} \underset{\sim}{N}, \\ \text{subject to } \underset{\sim}{e}'\underset{\sim}{N}'=1 \text{ , } \underset{\sim}{N} \geqslant 0, \end{array} \right. \tag{3}$$

where $\mathcal{N} = \left\{ 1,2,\ldots,n \right\}$ and where j spans J^*.

Proof. From Proposition 4, it is seen that one can find an optimal solution of P2 by looking at the solutions of each of the $|J^*|$ problems :

$$\left\{ \begin{array}{l} \text{minimize } \underset{\sim}{N}'\underset{\sim}{C}\underset{\sim}{N} \\ \text{subject to } \underset{\sim}{e}'\underset{\sim}{N} = 1 \text{ , } \underset{\sim}{N} \geqslant 0 \\ \text{and the additional constaint } N_j = 0, \end{array} \right.$$

for $j \in J^*$. For each j, this last problem is equivalent to (3). ∎

Clearly (3) is a problem of type P2 with dimensions reduced by 1. We may therefore repeatedly invoke the above proposition to justify the following algorithm. A queue $\mathcal{Q}$ is built which elements are sets of indices I for which $\underset{\sim}{N}_I = \underset{\sim}{C}_I^{-1} \underset{\sim}{e}$ has to be calculated in accordance with Proposition 3. (We recall that in a "queue", elements are added "at the bottom" and removed "from the to the second element moving up at the top).

1 - Set the queue Q to the single element $\{1,2,\ldots,n\}$ and $\alpha = 0$;

2 - If Q is empty, print error diagnostic and stop ; otherwise, select and remove the first element I. Compute $N_I = C_I^{-1} e$;

3 - Let $K = \{k : N_{Ik} < 0\}$. If K is empty, go to Step 5 ; otherwise :

4 - For each $k \in K$, if $(I-\{k\})$ is not a subset of an element of Q, then add $(I-\{k\})$ to Q. Go to Step 2 ;

5 - If $\alpha \geq e'N_I$, go to Step 2 ; otherwise, set $\alpha = e'N_I$;

6 - If $C_j N_I \geq e$ (see (1)) , set N_i as per (2) and stop ; otherwise go to Step 2.

The numerical efficiency of the algorithm will depend greatly on the particular code used to compute $C_I^{-1} e$. It will be seen in Section IV that, by using the Choleski decomposition of C_I^{-1}, one can make the test $C_j N_I \geq e$ of step 6 an immediate byproduct of step 2.

IV. <u>THE SECOND PROBLEM</u> - Let Y and X be two vectors, possibly equal. Assume X to be n-dimensional and for $k \leq n$ denote by $\mathcal{V}_k$ the set of all k-dimensional subvectors of X. If I is a set of k distinct indices, we shall denote by X_I the subvector of X retaining those elements of X with indices in I and by $\hat{Y}(X_I)$ the vector (ΔX_I) solution of the first problem studied in Section III. Thus this second problem can be stated as :

- Determine the vector $X_I \in \mathcal{V}_k$ which maximizes $RV[Y, \hat{Y}(X_I)]$ over $\mathcal{V}_k$.

When n and k are large, it becomes unpractical to try all k-dimensional subvectors of X and for this reason we propose a sequential algorithm analogous to the one used in step-wise regression.

From the analysis carried in Section III, it is easily seen that for a given set I the vector X_I is feasible for the present problems if and only if the vector $N_I = C_I^{-1} e$ is non negative and that the coefficient $RV[Y, \hat{Y}(X_I)]^2$ is then proportional to $\alpha_I = e'C_I^{-1} e$. Thus our problem becomes that of maximizing

$$\alpha_I = e'C_I^{-1} e \quad \text{subject to} \quad C_I^{-1} e \geq 0 ,$$

over all subsets I of k distinct indices.

Suppose now that I is a subset of only (k-1) indices which is known to be feasible (i.e. $\underset{\sim}{N}_I = \underset{\sim}{C}_I^{-1}\underset{\sim}{e} \geq 0$). The step-wise technique consists in the determination of that index $j \notin I$ such that, with $(I,j) = I \cup \{j\}$, $\underset{\sim}{N}_{(I,j)}$ will be feasible and $\alpha_{(I,j)}$ will be maximized (over all choices of j).

The technicalities will now be described, assuming for simplicity that the set I contains the first (k-1) indices : $I = \{1,2,\ldots,k-1\}$.

Consider the matrix

$$\underset{\sim}{C}^+ = \left[\begin{array}{c|c} \underset{\sim}{C} & \begin{matrix}1\\1\\\vdots\\1\end{matrix} \\ \hline 11 1\cdots 1 & \end{array}\right]$$

where $\underset{\sim}{C}$ is given by (3), and let $\underset{\sim}{C}^+ = \underset{\sim}{R}\,\underset{\sim}{R}'$ be the Choleski decomposition [FORSYTHE, MOLER, 1967] of $\underset{\sim}{C}^+$ (which is positive semi-definite), i.e $\underset{\sim}{R}$ is uppertriangular. We assume that the choice of the elements of the set I has been made sequentially by this same algorithm and that the first (k-1) steps of the Choleski decomposition of $\underset{\sim}{C}^+$ have been carried, column-wise, leading to the (n+1)x(k-1) matrix

$$\left[\begin{array}{c} \underset{\sim}{R}_I \\ \hline \underset{\sim}{r}'_{k\bullet} \\ \vdots \\ \hline \underset{\sim}{r}'_{(n+1)\bullet} \end{array}\right] \tag{4}$$

where $r'_{j\bullet} = (r_{j1}\ r_{j2}\ \ldots\ r_{j(k-1)})$ for $j \geq k$ and

$$\underset{\sim}{C}_I = \underset{\sim}{R}_I\,\underset{\sim}{R}'_I\,,$$

$$\underset{\sim}{R}_I\,\underset{\sim}{r}_{(n+1)\bullet} = \underset{\sim}{e}\,.$$

For $j \notin I$ $(j \neq n+1)$, define

$$w_j = c_{jj} - \sum_{i=1}^{k-1} r_{ji}^2$$

$$s_j = 1 - \sum_{i=1}^{k-1} r_{ji}\, r_{(n+1)i}\,. \tag{5}$$

$(s_j = 1 - \underset{\sim}{r}'_{j\bullet}\,\underset{\sim}{R}_I^{-1}\,\underset{\sim}{e})$. If the index j is to be added to the set I with the feasible vector $\underset{\sim}{X}_{(I,j)}$, we must have the corresponding matrix $\underset{\sim}{C}_{(I,j)}$ non-singular, hence $w_j > 0$. The Choleski decomposition $\underset{\sim}{C}_{(I,j)} = \underset{\sim}{R}_{(I,j)} \underset{\sim}{R}'_{(I,j)}$ would then give :

$$\underset{\sim}{R}^{-1}_{(I,j)} = \begin{bmatrix} \underset{\sim}{R}^{-1}_I & 0 \\ \dfrac{-1}{\sqrt{w_j}} \underset{\sim}{r}'_{j\bullet} \underset{\sim}{R}^{-1}_I & \dfrac{1}{\sqrt{w_j}} \end{bmatrix}, \quad \underset{\sim}{R}^{-1}_{(I,j)} \underset{\sim}{e} = \begin{bmatrix} \underset{\sim}{R}^{-1}_I \underset{\sim}{e} \\ s_j / \sqrt{w_j} \end{bmatrix}$$

The coefficient RV $[\underset{\sim}{Y}, \widehat{\underset{\sim}{Y}}(\underset{\sim}{X}_{(I,j)})]^2$ will be proportional to

$$\alpha_{(I,j)} = \underset{\sim}{e}' \underset{\sim}{R}'^{-1}_I \underset{\sim}{R}^{-1}_I \underset{\sim}{e} + (s_j^2 / w_j) = \alpha_I + (s_j^2/w_j) \qquad (6)$$

The feasibility condition of non-negativity of $\underset{\sim}{N}_{(I,j)} = \underset{\sim}{C}^{-1}_{(I,j)} \underset{\sim}{e}$ must be respected. It can be verified by substitution of the above relations that

$$\underset{\sim}{N}_{(I,j)} = \underset{\sim}{R}'^{-1}_{(I,j)} \underset{\sim}{R}^{-1}_{(I,j)} \underset{\sim}{e} = \begin{bmatrix} \underset{\sim}{N}_I - \dfrac{s_j}{w_j} \underset{\sim}{Z}_j \\ \dfrac{s_j}{w_j} \end{bmatrix}$$

where $\underset{\sim}{Z}_j$ is the solution of the triangular system $\underset{\sim}{R}'_I \underset{\sim}{Z}_j = \underset{\sim}{r}_{j\bullet}$.

We can now summarize one basic cycle of the algorithm assuming that k is given, that I is a set of indices with less than k elements and that the corresponding matrix (4) is available, as s_j and w_j for all $j \in I$:

1 - Define $J = \{ j : j \notin I, w_j > 0 \text{ and } s_j > 0 \}$;

2 - If J is empty, stop. Otherwise select $j_o \in J$ which maximizes (s_j^2/w_j) (see (6)) ;

3 - Solve for $\underset{\sim}{Z}_{j_o}$ the triangular system $\underset{\sim}{R}'_I \underset{\sim}{Z}_{j_o} = \underset{\sim}{r}_{j_o\bullet}$;

4 - If $\underset{\sim}{N}_{(I,j_o)} = \underset{\sim}{N}_I - (s_{j_o}/w_{j_o}) \underset{\sim}{Z}_{j_o}$ is non-negative, then accept j set $I := I \cup \{j_o\}$, $\underset{\sim}{N}_I := \underset{\sim}{N}(I,j_o)$ and go to Step 5. Otherwise set $J := J - \{j_o\}$ and repeat Step 2 ;

5 - Carry one more step of the Choleski decomposition of $\underset{\sim}{C}^+$, computing the column corresponding to column "j_o" of $\underset{\sim}{C}^+$ (to obtain $\underset{\sim}{R}_{(I,j_o)}$) and, for $j \in J$, set ((5)) :

$$s_j := s_j - r_{(n+1)j_o} r_{jj_o}, \quad w_j := w_j - r^2_{jj_o} ;$$

6 - If $k_o < k$, set $k_o := k_o + 1$ and repeat a complete cycle from Step 1. Otherwise, stop.

Note that a "stop" a Step 2 implies that the set of indices already chosen connot be enlarged by addition of a single element so as to increase the RV coefficient.

To initiate the algorithm one may choose any one element X_i of $\underset{\sim}{X}$ and set $I = \{i\}$. The square of the RV coefficient is then proportional to $C^{-1}_{i\,i}$; this suggests initiation with that index for which the diagonal element of $\underset{\sim}{C}$ is minimum.

It must be made clear that this algorithm does not necessarily produce the optimal solution. As in the step-wise regression technique, it produces the best solution of order k knowing which solution has been selected to the order (k-1).

Subject to this conditionality restriction upon convergence to the optimal solution, a count of the number of operations for a k-th cycle shows the algorithm to be quite efficient.

We conclude the study of this second problem by pointing out that the first problem (Section III) can be interpreted as a special case of the second : the case where k = n . Thus the algorithm of this section could be used in an attempt to solve Problem P2. The following proposition, deduced from the Kuhn-Tucker conditions for Problem P2 and the above analysis, provides a criterion to detect the optimal solution :

<u>Proposition 6</u>. A solution produced by the algorithm of this Section is optimal for Problem P2 of Section III if and only if, at stop,

$$s_j < 0 \quad \text{for all} \quad j \in J$$

or all n variables have been accepted (The proof is omitted).

<u>CONCLUSION</u>. - The theory summarized in Section II has shed new lights on two practical problems of interest to the applied statistician. Arguments on convexity and a factorization of the matrices involved have led to algorithms to solve those problems. A deeper numerical analysis of the problems, particularly the first one, should lead to more efficient algorithms and computer programs.

REFERENCES

BEALE E.M.L., KENDALL M.G. and MANN D.W. (1967) - The Discarding of Variables in Multivariate Analysis - Biometrika, 54, 3 and 4, p. 357-365.

CAMBON J. (1974) - Vecteurs equivalents à un autre au sens des composantes principales : Applications hydroliques - Note du laboratoire d'Hydrologie et d'Aménagement des eaux , USTL-Montpellier n°13-7

ESCOUFIER Y. (1970) - Echantillonage dans une population de variables aléatoires réelles - Publ. Inst. Stat. Univ. Paris - XIX-4 p. 1 à 4

ESCOUFIER Y. (1973-a) - Le traitement des variables vectorielles - Biometrics XXIX, p. 751-760.

ESCOUFIER Y. (1973-b) - Vecteurs aléatoires équivalents du point de vue de l'analyse en composantes principales. Rap. Techn. 7301 , Cen de Recherche en Informatique et Gestion, Av. d'Occitanie , 34000-MONTPELLIER.

FORSYTHE G., MOLER C.B. (1967) - Computer Solution of Linear Algebr Systems - Prentice Hall.

GOWER J.C. (1971) - A General Coefficient of Similarity and Some of its Properties - Biometrics, XXVII, p. 857-874.

JOLIFFE I.T. (1973) - Discarding Variables in a principal Component Analysis I : Artificial Data. Applied Statistics, Vol. 21 n° 2 , p. 160-173.