ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVII, 2 (1982)

ALGORITHM 83

ANNA BARTKOWIAK (Wrocław)

STEPWISE SELECTION OF DISCRIMINATIVE VARIABLES BY THE USE OF THE TRACE CRITERION

1. Procedure declaration. For given matrices C_1 and C_2 stored by lower triangles row by row in one-dimensional arrays, procedure *dissteptr* performs a search for variables for which the trace criterion attains its maximum.

The procedure allows us for obligatory introduction of variables into the discriminative set (denoted by $\mathscr{D}$).

Data:

p — number of variables under consideration;

$c1, c2[1:p\times(p+1)\div 2]$ — lower triangles of the "between" and "within" corrected cross product matrices, stored row by row;

$l1$ — the largest number of variables to be introduced into $\mathscr{D}$ while selecting upwards;

$l2$ — the smallest number of variables to be retained in $\mathscr{D}$ while selecting downwards;

$nr[1:p]$ — array of nos. (places) of the variables under consideration in the primary (original) data set;

$ind[1:p]$ — array indicating the variables which should be obligatorily introduced into $\mathscr{D}$ before starting the selection procedure: $ind[i] = 1$ means that the variable no. i should be introduced into $\mathscr{D}$, otherwise $ind[i]$ should be put equal to 0;

eps — small number indicating the machine accuracy.

Results:

$ind[1:p]$ — array indicating the variables contained in $\mathscr{D}$ after finishing the selection procedure: *ind*$[i] = 1$ means that the i-th variable belongs to $\mathscr{D}$, *ind*$[i] = 0$ means that the i-th variable does not belong to it.

Other results are obtained by the use of procedure ***result*** which is called after each step including a new variable in $\mathscr{D}$ or excluding it from $\mathscr{D}$.

Other parameters:

trace — identifier of the function evaluating the trace of the product of two matrices A' and B' contained in matrices A and B of larger size, stored by lower triangles row by row. The matrices A' and B', for which the trace criterion is calculated, are identified by an integer array *ind* whose elements equal 1 indicate which rows and columns of the larger matrices should be taken. This function should be headed as follows:

```
real procedure trace(p, c1, c2, ind);
    value p;
    integer p;
    array c2, c2;
    integer array ind;
```

An example of realization of this function is given on page 371.

onestep — identifier of the procedure described in [2];

result — identifier of the procedure printing intermediate results of the search procedure after a new variable has been introduced in $\mathscr{D}$ or eliminated from $\mathscr{D}$; this procedure was described in [2] under the heading *outratio*.

2. Method used.

2.1. The trace criterion is defined as the product of two matrices B and W^{-1}, where B stands for the between-groups cross product matrix, and W for the within-groups or error cross product matrix. Tests of significance between vectors of means of several groups can be formulated on the base of this statistic (see [1] and [4]).

Consider variables labelled $1, 2, \ldots, p$, and let r be a given integer. Our aim is to choose such a subset $i_1, i_2, \ldots, i_r$ for which

$$\operatorname{tr}(B_{i_1 i_2 \ldots i_r} W^{-1}_{i_1 i_2 \ldots i_r}) = \max,$$

where $B_{i_1 i_2 \ldots i_r}$ and $W_{i_1 i_2 \ldots i_r}$ are submatrices of B and W determined by the intersection of rows and columns with indices $i_1, i_2, \ldots, i_r$.

2.2. For stepwise inversion of a grammian matrix W the Gauss-Jordan algorithm may be applied. Having this in mind it is easy to programme an algorithm for stepwise selection of discriminative variables by the use of the trace criterion. Applying sequentially the modified Gauss-Jordan

```
procedure dissteptr(p,c1,c2,l1,l2,nr,ind,eps,trace,onestep,
 result);
 value p,l1,l2,eps;
 real eps;
 integer p,l1,l2;
 array c1,c2;
 integer array nr,ind;
 real procedure trace;
 procedure onestep,result;
 begin
  real x,z;
  integer k,l,q,r,s;
  array c3[1:p×(p+1)÷2];
  l:=k:=0;
  for q:=1 step 1 until p do
   begin
    k:=k+q;
    if ind[q]=1
     then
     begin
      if c1[k]>eps∧c2[k]>eps
       then
       begin
        onestep(q,1.0,p,c2);
        l:=l+1;
       end c1[k]>eps ∧ c2[k]>eps
       else ind[q]:=0
     end ind[q]=1
   end q;
  if l>0
```

```
   then
   begin
    z:=-trace(p,c1,c2,ind);
    result(p,l,z,nr,ind)
   end l>0;
  if l1>p
   then l1:=p;
nextvar:
  if l≥l1
   then go to back;
  z:=.0;
  k:=r:=0;
  for q:=1 step 1 until p do
   begin
    for s:=p×(p+1)÷2 step -1 until 1 do
     c3[s]:=c2[s];
    k:=k+q;
    if ind[q]=0∧c3[k]>eps
     then
     begin
      onestep(q,1.0,p,c3);
      ind[q]:=1;
      x:=-trace(p,c1,c3,ind);
      ind[q]:=0;
      if x>z
       then
       begin
        z:=x;
        r:=q
       end x>z
     end ind[q]=0 ∧ c3[k]>eps
   end q;
  if r>0
   then
   begin
    onestep(r,1.0,p,c2);
    ind[r]:=1;
    l:=l+1;
    result(p,l,z,nr,ind);
    go to nextvar
   end r>0;
back:
  if l2≥l
   then go to fin;
  z:=.0;
  k:=q:=0;
  for r:=1 step 1 until p do
   begin
    k:=k+r;
    if ind[r]=1
     then
     begin
      for s:=p×(p+1)÷2 step -1 until 1 do
       c3[s]:=c2[s];
      onestep(r,-1.0,p,c3);
      ind[r]:=0;
      x:=-trace(p,c1,c3,ind);
      ind[r]:=1;
      if x>z
       then
```

```
        begin
         z:=x;
         q:=r
        end x>z
      end ind[r]=1
   end r;
  if q>0
   then
   begin
    onestep(q,-1.0,p,c2);
    ind[q]:=0;
    l:=l-1;
    result(p,l,z,nr,ind);
    go to back;
   end q>0;
fin:
 end dissteptr
```

transformations to a given matrix W of size $p\times p$ we obtain finally the inverse of the transformed matrix W,

$$W^{-1} = T_p T_{p-1} \dots T_1 W,$$

where $T_p, T_{p-1}, \dots, T_1$ denote the transformations described by formulae (3) in [2].

The transformations T_r ($r = 1, 2, \dots, p$) may be applied for prescribed values of r, say $i_1, i_2, \dots, i_r$. The transformed matrix determined by the intersection of rows and columns with indices $i_1, i_2, \dots, i_r$ is the inverse of the input matrix $W_{i_1 i_2 \dots i_r}$.

Using the forward transformation T_r described by formulae (3) in [2], we include the variable no. r into $\mathscr{D}$. Using the back transformation $\tilde{T}_r$ described by formulae (4) in [2] for a variable actually being in $\mathscr{D}$, we exclude this variable from $\mathscr{D}$. Whether a given variable belongs to $\mathscr{D}$ is indicated by the auxiliary array *ind* whose elements equal 1 signify the presence, and elements equal 0 signify the absence of the variable in $\mathscr{D}$.

```
real procedure trace(p,c1,c2,ind);
 value p;
 integer p;
 array c1,c2;
 integer array ind;
 begin
  real x;
  integer i,j,k;
  x:=.0;
  k:=0;
  for i:=1 step 1 until p do
   begin
    if ind[i]=1
     then
     begin
      for j:=i-1 step -1 until 1 do
       if ind[j]=1
        then x:=x+c1[k+j]×c2[k+j];
      x:=x+.5×c1[k+i]×c2[k+i]
     end ind[i]=1;
    k:=k+i
   end i;
  trace:=x+x
 end trace
```

2.3. The algorithm we use proceeds in the following steps:

1° Perform the Gauss-Jordan transformations on the matrix C for a given set of declared variables (obligatory introduction, if any).

2° Calculate the trace criterion for the submatrices of B and for the transformed matrix W identified by the auxiliary array *ind.*

3° If the actual size of $\mathscr{D}$ is greater than or equal to a given number *l1*, go to 5°; otherwise, go to 4°.

4° If the actual size l is less than the wanted size *l1*, seek additional variables not belonging to $\mathscr{D}$ which give the greatest rise of the trace criterion. Include this variable (if any) into $\mathscr{D}$, after inclusion augment l, and pass to point 3°.

5° If the actual size of $\mathscr{D}$ is greater than a given number *l2*, go to point 6°; otherwise, go to point 7°.

6° Seek variables for which after elimination from $\mathscr{D}$ the trace criterion remains the greatest. If there is any such variable, remove it from $\mathscr{D}$, diminish the actual size l and go to point 5°; otherwise, pass tc point 7°.

7° Finish. End of the selection procedure.

The modified Gauss-Jordan transformations can be performed by the use of procedure *onestep* given in [2]. The trace criterion for variables identified by the array *ind* can be evaluated by the use of procedure *trace* given in this paper.

After each call of procedure *onestep* changing the status quo of $\mathscr{D}$, procedure *result* is called and we obtain information about the actual value of the trace criterion and the nos. of variables actually being in $\mathscr{D}$.

Notice that the Gauss-Jordan transformation T_r is executed only in that case where the diagonal element w_{rr} is greater than *eps*, a given small number. The inequality $w_{rr} < eps$ means that the variable no. q is linearly dependent on variables which have been included previously into $\mathscr{D}$.

3. Certification. The results of *dissteptr* were checked in two modes:

1° by calculating the value of the trace criterion strictly from the definition, using procedure *cholinversion2* [3] for matrix inversion,

2° by comparison of the chosen variables with the numbers of variables chosen by *disstepw* [2].

The results almost always were the same (see the remarks in the certification of procedure *disstepw* in [2]).

4. Test example. Entering *dissteptr* with the data

$$p = 5,$$

$$
\begin{array}{lllll}
\mathit{c1}[\mathit{1}:\mathit{15}] = [19482.9350 & & & & \\
23305.2695 & 28505.7548 & & & \\
23305.2695 & 28505.7548 & 28505.7548 & & \\
11584.5455 & 13927.9221 & 13927.9221 & 6899.1039 & \\
9575.0616 & 11968.1315 & 11968.1315 & 5752.4675 & 4815.5146],
\end{array}
$$

$$
\begin{array}{lllll}
\mathit{c2}[\mathit{1}:\mathit{15}] = [258.9286 & & & & \\
106.3214 & 397.0179 & & & \\
106.3214 & 397.0179 & 397.0179 & & \\
104.0000 & 138.7143 & 138.7143 & 317.7143 & \\
-34.3929 & 174.2321 & 174.2321 & 252.7143 & 478.3036],
\end{array}
$$

$$\mathit{l1} = 5, \quad \mathit{l2} = 2, \quad \mathit{nr}[\mathit{1}:\mathit{5}] = 2, 3, 4, 5, 6,$$

$$\mathit{ind}[\mathit{1}:\mathit{5}] = 0, 1, 0, 0, 0, \quad \mathit{eps} = {}_{10}{-6},$$

and using the function *trace* of this paper, and also procedures *onestep* and *outratio* from [2] we get the following results:

$$\mathit{ind}[\mathit{1}:\mathit{5}] = 1, 1, 0, 0, 0.$$

This means that the variables chosen are labelled by the array *nr* as variables no. 2 and no. 3.

We get also additional results by procedure *result*, which was called 6 times during the run of *disstepw* ($p = 5$):

No. of call	r	x	Variables in $\mathscr{D}$
1	1	71.799671	3
2	2	111.047882	2 3
3	3	113.456186	2 3 6
4	4	116.543725	2 3 5 6
5	3	113.456186	2 3 6
6	2	111.047882	2 3

These are the same results as those obtained by procedure *disstepw* given in [2].

The calculations were done on the Odra 1204 computer.

References

[1] H. Ahrens and J. Läuter, *Mehrdimensionale Varianzanalyse*, Akademie-Verlag, Berlin 1974.

[2] A. Bartkowiak, *Algorithm 82: Stepwise selection of discriminative variables by the use of the Wilks criterion*, this fascicle, p. 351-364.

[3] R. S. Martin, G. Peters and J. H. Wilkinson, *Symmetric decomposition of a positive definite matrix*, Numer. Math. 7 (1965), p. 362-383.

[4] S. J. Press, *Applied multivariate analysis*, New York 1972.

INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW
50-384 WROCŁAW

Received on 2. 2. 1978

ALGORYTM 83

ANNA BARTKOWIAK (Wrocław)

KROKOWY WYBÓR ZMIENNYCH DO ZBIORU DYSKRYMINACJI METODĄ ŚLADU MACIERZY

STRESZCZENIE

Procedura *dissteptr* wybiera metodą krokową zmienne o największej sile dyskryminacji. Siła dyskryminacji zmiennych jest mierzona za pomocą śladu macierzy BW^{-1}, gdzie B oznacza macierz poprawionych iloczynów odchyleń grupowych od średniej generalnej, W zaś macierz poprawionych iloczynów odchyleń wewnątrzgrupowych (macierz błędów przy wielozmiennej analizie wariancji z jednym kierunkiem klasyfikacji). Bliższe omówienie tego kryterium znajduje się w [1] i [4].

Krokowe odwracanie macierzy W odbywa się za pomocą zmodyfikowanego algorytmu Gaussa-Jordana w sposób przedstawiony w [2].

Procedura działa w trzech etapach:

1° Obowiązkowe wprowadzenie l zadeklarowanych zmiennych do zbioru dyskryminacji $\mathscr{D}$ (dopuszcza się możliwość $l = 0$).

2° Dobranie metodą krokową dalszych zmiennych tak, żeby liczebność zbioru $\mathscr{D}$ osiągnęła wielkość *l1*, gdzie *l1* jest daną liczbą.

3° Usunięcie ze zbioru $\mathscr{D}$ odpowiedniej liczby zmiennej tak, aby końcowa liczebność tego zbioru wynosiła *l2*, gdzie *l2* jest daną liczbą.

Po każdym kroku oblicza się wielkość śladu macierzy BW^{-1} dla zmiennych znajdujących się aktualnie w zbiorze $\mathscr{D}$. Zmienne te można identyfikować za pomocą tablicy *ind*, której wartości równe 1 oznaczają przynależność danej zmiennej do zbioru $\mathscr{D}$, a wartości równe 0 — brak przynależności. Do obliczenia śladu macierzy identyfikowanych w opisany sposób służy funkcja rzeczywista *trace*. Szukamy takich zmiennych, dla których wielkość obliczonego śladu byłaby możliwie duża.

Dane:

p — liczba rozważanych zmiennych (stopień obliczanych macierzy);

c1, c2 $[1 : p \times (p+1) \div 2]$ — tablice zawierające dolne trójkąty macierzy poprawionych iloczynów odchyleń międzygrupowych i wewnątrzgrupowych, zapamiętanych wierszami;

l1 — maksymalna liczba zmiennych w zbiorze $\mathscr{D}$;

l2 — minimalna liczba zmiennych w zbiorze $\mathscr{D}$;

nr [*1* : *p*] — numery rozważanych zmiennych według ich pierwotnej numeracji w zbiorze danych;

ind [*1* : *p*] — tablica wskazująca na numery zmiennych (według numeracji w macierzach C_1 i C_2), które mają być obowiązkowo wprowadzone do zbioru $\mathscr{D}$ przed rozpoczęciem postępowania krokowego: *ind* [*i*] = 1 oznacza, że zmienna o numerze *i* powinna być obowiązkowo wprowadzona do zbioru $\mathscr{D}$;

eps — mała liczba oznaczająca dokładność maszynową.

Wyniki:

ind [*1* : *p*] — tablica określająca numery zmiennych (według numeracji w macierzach C_1 i C_2), znajdujących się w zbiorze $\mathscr{D}$:

$$ind[i] = \begin{cases} 1, & \text{gdy } i \in \mathscr{D}, \\ 0, & \text{gdy } i \notin \mathscr{D}. \end{cases}$$

Poza tym za pomocą procedury ***result*** można otrzymać wyniki częściowe, określające po każdym kroku liczbę i numery zmiennych znajdujących się w zbiorze $\mathscr{D}$.