ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVII, 2 (1982)

ALGORITHM 82

ANNA BARTKOWIAK (Wrocław)

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

1. Procedure declaration. For given matrices C_1 and C_2 stored by lower triangles row-wise in one-dimensional arrays, procedure *disstepw* performs a search for variables with the greatest discriminative power, using the Λ-statistic as the measure of the discriminative power.

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 "within" and "total" corrected cross product matrices, stored row by row, i.e. the elements of the matrix $C = \{c_{ij}\}$, $i, j = 1, 2, \ldots, p$, should be taken in the order: $c_{11}, c_{21}, c_{22}, c_{31}, c_{32}, \ldots, c_{p,p-1}, c_{pp}$;

$l1$ — 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.

```
procedure dirstepw(p,c1,c2,l1,l2,nr,ind,eps,onestep,
 outratio);
 value p,l1,l2,eps;
 real eps;
 integer p,l1,l2;
 array c1,c2;
 integer array nr,ind;
 procedure onestep,outratio;
 begin
  real ratio,x,y,z;
  integer k,l,q,r;
  array d,dt[1:p];
  l:=k:=0;
  ratio:=1.0;
  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
        ratio:=ratio×c1[k]/c2[k];
        l:=l+1;
        onestep(q,1.0,p,c1);
        onestep(q,1.0,p,c2);
        outratio(p,l,ratio,nr,ind)
       end c1[k]>eps ∧ c2[k]>eps
       else ind[q]:=0
```

```
      end ind[q]=1
    end q;
   if l1>p
    then l1:=p;
nextvar:
   if l>l1
    then go to back;
   z:=1.0;
   k:=q:=0;
   for r:=1 step 1 until p do
    begin
     k:=k+r;
     if ind[r]=0
      then
      begin
       x:=c1[k];
       y:=c2[k];
       if x>eps∧y>eps
        then
        begin
         x:=x/y;
         if x<z
          then
          begin
           z:=x;
           q:=r
          end x<z
        end y1>eps ∧ y2>eps
      end ind[r]=0
    end r;
```

```
    if q>0
     then
     begin
      ratio:=ratio×z;
      ind[q]:=1;
      l:=l+1;
      onestep(q,1.0,p,c1);
      onestep(q,1.0,p,c2);
      outratio(p,l,ratio,nr,ind);
      go to nextvar
     end q>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
        x:=c2[k]/c1[k];
        if x>z
         then
         begin
          z:=x;
          q:=r
         end x>z
       end ind[r]=1
```

```
    end r;
   if q>0
    then
    begin
     ratio:=ratio/z;
     ind[q]:=0;
     l:=l-1;
     onestep(q,-1.0,p,c1);
     onestep(q,-1.0,p,c2);
     outratio(p,l,ratio,nr,ind);
     go to back
    end q>0;
fin:
 end disstepw
```

Results:

ind[*1* : *p*] — array indicating the variables contained in $\mathscr{D}$ after finishing the selection procedure: *ind*[*i*] = 1 means that the variable no. *i* belongs to $\mathscr{D}$, *ind*[*i*] = 0 means that the *i*-th variable does not belong to $\mathscr{D}$.

Other results are obtained by the use of procedure *outratio* described in the sequel, which is several times called during a run of *disstepw*.

Other parameters:

onestep — identifier of the procedure performing one step of the modified Gauss-Jordan transformation, given by formulae (3) and (4), including or excluding the variable no. *q*; this procedure should be headed as follows:

procedure *onestep*(*q*, *v*, *p*, *c*);
value *q*, *v*, *p*;
real *v*;
integer *q*, *p*;
array *c*;

The meaning of the formal parameters is as follows:

q — number of the variable for which the transformation is performed;

v — should be set to +1.0 or −1.0: *v* = +1.0 means the forward transformation by the use of formulae (3), *v* = −1.0 means the back transformation by the use of formulae (4);

```
procedure onestep(q,v,p,c);
 value q,v,p;
 real v;
 integer q,p;
 array c;
 begin
  real x;
  integer i,j,k;
  array d,dt[1:p];
  j:=q×(q-1)÷2;
  x:=-v/c[j+q];
  for i:=1 step 1 until q do d[i]:=c[j+i];
  for i:=q+1 step 1 until p do d[i]:=c[i×(i-1)÷2+q];
  for i:=1 step 1 until p do dt[i]:=d[i]×x;
  dt[q]:=v×x;
  for i:=1 step 1 until q do c[j+i]:=dt[i];
  for i:=q+1 step 1 until p do c[i×(i-1)÷2+q]:=dt[i];
  if v=-1.0
   then
   for i:=1 step 1 until p do dt[i]:=-dt[i];
  k:=0;
  for i:=1 step 1 until p do
   begin
    if i≠q then
     for j:=1 step 1 until i do
      if j≠q
       then c[k+j]:=c[k+j]+d[i]×dt[j];
    k:=k+i
   end i;
end onestep
```

p — number of variables under consideration;
c — array containing the lower triangle of the matrix on which the transformation is performed.

An example of realization of procedure *onestep* is given on page 356.

outratio — identifier of the procedure printing intermediate results of the search procedure after a new variable has been introduced in $\mathscr{D}$ or eliminated from it, headed as follows:

procedure *outratio*($p, r, ratio, nr, ind$);
value p, r;
real *ratio*;
integer p, r;
integer array *ind*;

```
procedure outratio(p,r,ratio,nr,ind);
 value p,r;
 real ratio;
 integer p,r;
 integer array nr,ind;
 begin
  integer k;
  format('?␣␣p=123␣␣r=12␣␣x=123456.123456?');
  print(p,r,ratio,'?␣variables␣in␣the␣discriminative␣set:');
  format('123');
  k:=0;
  for r:=1 step 1 until p do
   if ind[r]=1
    then
    begin
     print(nr[r]);
     k:=k+1;
     if k=10
      then k:=line(2)+space(34)
    end ind[r]=1;
  line(3)
 end outratio
```

The meaning of the formal parameters of *outratio* is the following:

p — number of variables under consideration;
r — number of variables actually being in $\mathscr{D}$;
ratio — value of the Λ-statistics (see Section 2) evaluated for the variables actually being in $\mathscr{D}$;
nr[*1* : *p*] — nos. (places) of the *p* variables under consideration in the primary data set;
ind[*1* : *p*] — array indicating for which variables the Λ-statistic has been evaluated.

Procedure *outratio* contains exit procedures specific for the Algol compiler under use. An example of procedure *outratio* prepared for the Algol 1204 compiler is given on page 357.

2. Method used.

2.1. A measure of the discriminative power of a set of variables $X_1, X_2, \ldots, X_p$ is given by the ratio

$$\Lambda = \frac{|W|}{|T|} \tag{1}$$

with the determinant of the within-groups cross product matrix W in the numerator and the determinant of the total cross product matrix T in the denominator [6]. The ratio Λ is called *Wilks Λ-criterion.* If the value of the Λ-statistic is small, then the between-groups variance is large, and this means a great differentiation of the groups under consideration.

2.2. It is known [5] that the value of the determinant of a grammian matrix $C = \{c_{rs}\}$, $r, s = 1, 2, \ldots, p$, can be calculated stepwise by the use of the formula

$$|C| = s_1^2 s_{2.1}^2 s_{3.12}^2 \cdots s_{p.12\ldots(p-1)}^2, \tag{2}$$

where $s_1^2 = c_{11}$, and $s_{i.12\ldots(i-1)}^2$, $i = 2, 3, \ldots, p$, is the residual variance of the i-th variable X_i conditioned on variables $X_1, X_2, \ldots, X_{i-1}$; it is after subtraction from X_i its best linear predictor based on variables $X_1, X_2, \ldots, X_{i-1}$ and evaluated by the least square method.

2.3. Now the residual variance $s_{i.12\ldots(i-1)}^2$ can be calculated by the use of the modified Gauss-Jordan algorithm operating on the sequentially transformed matrix C. This algorithm is realized in a sequence of linear transformations T_q, $q = 1, 2, \ldots, i-1$, applied to the matrix C given by the formulae

$$c'_{qq} = -1/c_{qq}, \quad c'_{rq} = c'_{qr} = c_{rq}c'_{qq}, \quad c'_{rs} = c'_{sr} = c_{rs} + c_{rq}c'_{qs} \quad \text{for } r, s = 1, 2, \ldots, p;\ r, s \neq q, \tag{3}$$

where c_{rs} denotes an element of C before the transformation T_q (but perhaps transformed by previous transformations), and c'_{rs} denotes the corresponding element after the transformation T_q (see [3] and [1]).

The transformation T_q performs the inclusion of the variable X_q into $\mathscr{D}$.

A variable X_q actually being in $\mathscr{D}$ can be excluded from $\mathscr{D}$ by the back transformation $\tilde{T}_q$ executed on the already transformed matrix C by the use of the formulae

$$(4)\qquad c'_{qq} = -1/c_{qq}, \quad c'_{rq} = c'_{qr} = -c_{rq}c'_{qq}, \quad c'_{rs} = c'_{sr} = c_{rs} - c_{rq}c'_{qs}$$
$$\text{for } r, s = 1, 2, \ldots, p;\ r, s \neq q,$$

where c_{rq} stands for an element of the matrix C before the transformation $\tilde{T}_q$, and c'_{rq} denotes the corresponding element after the transformation $\tilde{T}_q$.

By symmetry of the matrix C, it is sufficient to work only on the lower triangle of this matrix. The transformation T_q $(1 \leqslant q \leqslant p)$, given by formulae (3), can be performed in the following manner:

$$(5)\qquad \begin{gathered} c'_{qq} = -1.0/c_{qq}, \\ d_j = \begin{cases} c_{qj} & \text{for } j = 1, 2, \ldots, q-1, \\ c_{jq} & \text{for } j = q+1, q+2, \ldots, p, \end{cases} \\ c'_{qj} = d'_j = d_j c'_{qq} \quad \text{for } j = 1, 2, \ldots, q-1, \\ c'_{jq} = d'_j = d_j c'_{qq} \quad \text{for } j = q+1, q+2, \ldots, p, \\ c'_{rs} = c_{rs} + d_r d'_s \quad \text{for } r = 1, 2, \ldots, q-1, q+1, \ldots, p;\ s = 1, 2, \ldots, r. \end{gathered}$$

An analogous algorithm for the back transformation $\tilde{T}_q$ can be performed on the lower triangle of the matrix C and is described by the following formulae:

$$(6)\qquad \begin{gathered} c'_{qq} = -1.0/c_{qq}, \\ d_j = \begin{cases} c_{qj} & \text{for } j = 1, 2, \ldots, q-1, \\ c_{jq} & \text{for } j = q+1, q+2, \ldots, p, \end{cases} \\ c'_{qj} = d'_j = -d_j c'_{qq} \quad \text{for } j = 1, 2, \ldots, q-1, \\ c'_{jq} = d'_j = -d_j c'_{qq} \quad \text{for } j = q+1, q+2, \ldots, p, \\ c'_{rs} = c_{rs} - d_r d'_s \quad \text{for } r = 1, 2, \ldots, q-1, q+1, \ldots, p;\ s = 1, 2, \ldots, r. \end{gathered}$$

The back transformation $\tilde{T}_q$ given by formulae (6) can be performed only in the case where the variable X_q has been introduced into $\mathscr{D}$ by previous transformations.

The two algorithms given by formulae (5) and (6) may be combined together giving one common formula:

$$\begin{gathered} c'_{qq} = -1.0/c_{qq}, \\ d_j = \begin{cases} c_{qj} & \text{for } j = 1, 2, \ldots, q-1, \\ c_{jq} & \text{for } j = q+1, q+2, \ldots, p, \end{cases} \end{gathered}$$

$$(7) \qquad c'_{qj} = d'_j = v d_j c'_{qq} \quad \text{for } j = 1, 2, \ldots, q-1,$$

$$c'_{jq} = d'_j = v d_j c'_{qq} \quad \text{for } j = q+1, q+2, \ldots, p,$$

$$c'_{rs} = c_{rs} + v d_r d'_s \quad \text{for } r = 1, 2, \ldots, q-1, q+1, \ldots, p;\ s = 1, 2, \ldots, r.$$

Substituting $v = 1$ we can use formulae (7) as the forward transformation T_q, and substituting $v = -1$ we can use them for the back transformation $\tilde{T}_q$ as well.

2.4. Suppose we are seeking a set $\mathscr{D}$ of size $l1 < p$ on the base of the matrices W and T. We proceed as follows:

1° Set initially *ratio* := 1.0, $q := l := 0$, $ind[i] := 0$ $(i = 1, 2, \ldots, p)$.

2° Seek the variable for which the ratio of the diagonal elements w_{qq}/t_{qq} is the smallest, thus finding the single variable with the greatest discriminative power. Set *ratio* := w_{qq}/t_{qq}, obtaining the value of the Λ-statistic for $\mathscr{D}$ of size 1. Set l, the actual number of variables in $\mathscr{D}$, equal to 1. If $l1 > l$, go to point 3°; otherwise, go to point 6°.

3° Seek a new variable q, not being in $\mathscr{D}$, for which the quotient of the corresponding diagonal elements is minimal, but look only on these variables for which the diagonal elements are greater than *eps*, a declared small number. If there is any such variable (say variable no. q), pass to point 4°; otherwise, go to point 6°.

4° Introduce the variable X_q (if any) into $\mathscr{D}$; multiply the actual *ratio* by w_{qq}/t_{qq}, set $ind[q] := 1$, set $l := l+1$; go to point 5°.

5° If the actual size l of $\mathscr{D}$ is less than $l1$, the desired size of $\mathscr{D}$, go to point 3°; otherwise, pass to point 6°.

6° Finish the search procedure.

Points 1°-6° describe an algorithm which enables us to make a stepwise selection of variables with a considerable large discriminative power. We use substantially the following property of the Gauss-Jordan algorithm described by formulae (3):

If the forward transformations $T_{i_1}, T_{i_2}, \ldots, T_{i_r}$ are performed on the whole matrix C comprising all variables, then after performance of these transformations the diagonal elements of the variables not being introduced into $\mathscr{D}$ are virtually the residual sums of squares of these variables conditioned on the variables belonging to $\mathscr{D}$.

2.5. An analogous algorithm can be formulated for the back elimination. Let $R_{(q)}$ denote the ratio $|W_q|/|T_q|$ for q variables. Suppose we eliminate the variable X_q. Then

$$R_{(q-1)} \frac{w'_{qq}}{t'_{qq}} = R_{(q)}$$

or, using formulae (3),

$$R_{(q-1)} \frac{t_{qq}}{w_{qq}} = R_{(q)}.$$

We seek such a variable in $\mathscr{D}$ for which the quotient t_{qq}/w_{qq} is the largest one. After elimination of that variable from $\mathscr{D}$ the remaining quotient $R_{(q-1)}$ is the smallest one.

3. Certification. The results of procedure *disstepw* can be certified in two modes:

1° investigating the values of the Λ-statistic,

2° observing the variables introduced into $\mathscr{D}$.

We compared the values of the Λ-statistic obtained from *disstepw* with those calculated directly from (1) by the use of procedure *det2* [4]. The results were practically the same.

The variables chosen by *disstepw* were compared with those chosen by *dissteptr*, a competitive procedure given in [2]. These variables almost always were the same, but we did notice some small discrepancies in cases where the subset chosen was unstable (there existed other sets with similar discriminative power).

4. Test example. For $p = 5$ we have

$$\begin{aligned} c1[1:15] = [&258.9286 \\ &106.3214 \quad 397.0179 \\ &106.3214 \quad 397.0179 \quad 397.179 \\ &104.0000 \quad 138.7143 \quad 138.7143 \quad 317.7143 \\ &-34.3929 \quad 174.2321 \quad 174.2321 \quad 252.7143 \quad 478.3036], \end{aligned}$$

$$\begin{aligned} c2[1:15] = [&19741.8636 \\ &23411.5909 \quad 28902.7727 \\ &23411.5909 \quad 28902.7727 \quad 28902.7727 \\ &11688.5455 \quad 14066.6364 \quad 14066.6364 \quad 7213.8182 \\ &9666.4545 \quad 11872.3636 \quad 11872.3636 \quad 6005.1818 \quad 5293.8182], \end{aligned}$$

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

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

and using procedures *onestep* and *outratio* we get the following results:

$$ind[1:5] = 1, 1, 0, 0, 0.$$

This means that the variables with the greatest discriminative power are the first and the second variables from the input matrices. After relabelling these variables, as indicated by the array *nr*, the variables no. 2 and no. 3 from the original data set (not given here) have the required property.

During the run of *disstepw* we get the following results by *outratio*, which was called 6 times ($p = 5$):

No. of call	r	x	Variables in $\mathscr{D}$
1	1	0.013736	3
2	2	0.004068	2 3
3	3	0.003642	2 3 6
4	4	0.003528	2 3 5 6
5	3	0.003642	2 3 6
6	2	0.004068	2 3

Checking this result by calculations from the definition of the Λ-statistic for the variables chosen lastly we get: the determinant in the numerator

$$\Delta_1 = 914950489_{10} - 4,$$

the determinant in the denominator

$$\Delta_2 = 224920078_{10} - 1,$$

and

$$\Lambda_{2,3} = ratio = \Delta_1/\Delta_2 = 0.004068.$$

So we get exactly the same result as previously. The competitive procedure *dissteptr* gives the same result (see [2]).

Notice that the second and the third rows and columns in the input matrices *c1* and *c2* are the same, i.e., these matrices are not positive definite. The Gauss-Jordan algorithm avoids this difficulty by omitting the variable corresponding to the third row of these matrices.

The calculations have been carried out on the Odra 1204 computer.

References

[1] A. Bartkowiak, *Algorytmy analizy regresji*, Matematyka Stosowana 7 (1976), p. 101-115.

[2] — *Algorithm 83: Stepwise selection of discriminative variables by the use of the trace criterion*, this fascicle, p. 365-375.

[3] E. M. L. Beale, M. G. Kendall and D. W. Mann, *The discarding of variables in multivariate analysis*, Biometrika 54 (1967), p. 357-366.

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

[5] G. P. McCabe, *Computations for variable selections in discriminant analysis*, Technometrics 17 (1975), p. 103-109.

[6] C. R. Rao, *Linear statistical inference and its applications*, J. Wiley, New York 1965.

INSTITUTE OF COMPUTER SCIENCE
UNIVERSITY OF WROCŁAW
50-384 WROCŁAW

Received on 2. 2. 1978

ALGORYTM 82

ANNA BARTKOWIAK (Wrocław)

KROKOWY WYBÓR ZMIENNYCH DO ZBIORU DYSKRYMINACJI NA PODSTAWIE STATYSTYKI WILKSA

STRESZCZENIE

Procedura *disstepw* wybiera metodą krokową zmienne o największej sile dyskryminacji. Siłę dyskryminacji mierzy się za pomocą statystyki Λ Wilksa, zdefiniowanej jako iloraz wyznaczników macierzy C_1 i C_2, $\Lambda = |C_1|/|C_2|$, gdzie C_1 jest macierzą zmienności wewnątrzgrupowej, a C_2 — macierzą zmienności całkowitej (por. [6]).

Przy obliczaniu wyznaczników stosowany jest wzór (2).

Procedura *disstepw* 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 ll, gdzie ll jest daną liczbą.

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

Wprowadzanie i usuwanie zmiennych odbywa się metodą krokową za pomocą zmodyfikowanego algorytmu Gaussa-Jordana opisanego wzorami (3)-(7). Wykonanie jednego kroku tego algorytmu może odbywać się za pomocą procedury *onestep* załączonej do pracy. W każdym kroku dobiera się lub eliminuje zmienną według zasady, żeby siła dyskryminacji zmiennych znajdujących się w zbiorze $\mathscr{D}$ była możliwie duża.

Dane:

p — liczba rozważanych zmiennych (stopień macierzy C_1 i C_2);

$c1, c2[1 : p \times (p+1) \div 2]$ — tablice zawierające dolne trójkąty macierzy poprawionych iloczynów wewnątrzgrupowych C_1 i całkowitych C_2, zapamiętanych wierszami;

ll — maksymalna liczba zmiennych;

$l2$ — minimalna (i końcowa) liczba zmiennych;

$nr[1 : p]$ — numery rozważanych zmiennych według 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 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 otrzymuje się wyniki częściowe po każdym kroku, zmieniającym zawartość zbioru $\mathscr{D}$. Wyniki te drukowane są za pomocą procedury *outratio*, zawierającej instrukcje wyjścia specyficzne dla maszyny, na której wykonywane są obliczenia.

Podano przykładową realizację tej procedury, przystosowaną do translatora Algolu 1204. Procedura ta drukuje następujące wyniki:

p — liczba rozważanych zmiennych;

r — liczba zmiennych znajdujących się aktualnie w zbiorze $\mathscr{D}$;

ratio — wartość statystyki Λ Wilksa dla zmiennych znajdujących się aktualnie w zbiorze $\mathscr{D}$;

nr[*1* : *p*] — numery zmiennych znajdujących się w zbiorze $\mathscr{D}$ (według numeracji określonej tablicą *nr*); drukowane są tylko numery tych zmiennych, dla których tablica *ind*, będąca parametrem formalnym procedury, wykazuje wartości równe 1.