PUC

Series: Monographs in Computer Science
and Computer Applications

Nº 9/69

IDENTIFICATION OF MINOR SUBSETS PERTAINING TO A MAJOR SET THROUGH SIMULTANEOUS ANALYSIS OF RANDOM SAMPLES

by

Fernando Olavo Franciss

Abelardo L. Puccini

Computer Science Department - Rio Datacenter

CENTRO TÉCNICO CIENTÍFICO
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 209 — ZC-20
Rio de Janeiro — Brasil

IDENTIFICATION OF MINOR SUBSETS PERTAINING TO A MAJOR SET THROUGH SIMULTANEOUS ANALYSIS OF RANDOM SAMPLES

CONTENTS

Fernando Olavo Franciss
Associate Professor
Civil Engineering Dept.

Abelardo L. Puccini
Associate Professor
Computer Science Dept.

1. Introduction

Knowing $\underline{m}$ random samples X_i of a set R, Being each sample refered to a non-Euclidian space by a function of $\underline{n}$ parameters X_{ij}, corresponding to its various characteristics, the recognition of subsets S_I, S_{II}, S_{III},... of the set R, is a difficult problem not so far solved statistical analysis (1), because the complexity grows as the number of parameters x_{ij} increases.

2. Recomended Solution

Assuming that the parameters x_{ij} have properties of equivalence, order, interval and value relationships, and are normally distributed, the recomen - ded solution has the following steps:

Step I: Normalization of the Parameters x_{ij}

The normalization of the parameters x_{ij} is necessary to cancel the effects of the parameters being measured by different system of units, and/or when they have different aplitudes of variation.

It is done as following:

a) Define the matrix of $\underline{m}$ x $\underline{n}$ elements x_{ij}

$$\begin{matrix} x_{11} & x_{12} \cdots\cdots\cdots\cdots x_{1n} \\ x_{21} & x_{22} \cdots\cdots\cdots\cdots x_{2n} \\ x_{31} & x_{32} \cdots\cdots\cdots\cdots x_{3n} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ x_{m1} & x_{m2} \cdots\cdots\cdots\cdots x_{mm} \end{matrix}$$

Which corresponds to the column of $\underline{m}$ random samples x_i:

$$x_1\ x_2\ x_3 \cdots x_m$$

Now determine for each column j, the values $\bar{x}_j$ and s_j, using the formulas:

$$\bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}$$

$$s_j = \{ m \sum_{i=1}^{m} x_{ij}^2 - (\sum_{i=1}^{m} x_{ij})^2 \ / m(m-1) \}^{1/2}$$

b) Replace the former values of the elements x_{ij}, by the normalized elements x'_{ij}, as defined below:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Step II: Equalization of the Normalized Parameters x_{ij}

The equalization of the normalized parameters x_{ij} is necessary to cancel the effects of the parameters corresponding usually to characteristics not equally relevants.

It is done as following:

a) Fix for each characteristics j, based on its meaning, an interval Δx_j, whose value allows us to consider the parameters x_{ij} and $x_{ij} \pm \Delta x_j$ as equivalents.

b) Normalize the n elements Δx_j using the following formula:

$$\Delta x'_j = \Delta x_j / s_j$$

c) Obtain the mean interval $\overline{\Delta x'}$ using the following formula:

$$\overline{\Delta x'} = \frac{1}{n} \sum_{j=1}^{n} \Delta x'_j$$

d) Replace the m x n normalized elements x'_{ij} by the normalized and equalized elements x''_{ij} as defined below:

$$x''_{ij} = x'_{ij} \cdot \overline{\Delta x'} / \Delta x'_j$$

Notice that if Δx_j in equal to s_j, then x''_{ij} will be equal to x'_{ij}.

Step III - Obtaining the Distances Between Samples

The determination of the distances between samples is necessary to measure the affinity between the samples:

It is done as following:

a) Obtain $\frac{m(m-1)}{2}$ distances a_{fk} between the random samples X_f and X_k, using the expression below:

$$a_{fk} = \sum_{j=1}^{n} (x''_{fj} - x''_{kj})^{1/2}$$

b) Define the symmetric matrix of m^2 elements a_{fk}, which indicates the distances a_{fk} between the samples X_f and X_k.

$$\begin{array}{c|cccc} & X_1 & X_2 & \cdots\cdots & X_m \\ \hline X_1 & 0 & a_{12} & \cdots\cdots & a_{1m} \\ X_2 & a_{21} & 0 & \cdots\cdots & a_{2m} \\ \vdots & \vdots & & & \\ X_m & a_{m1} & a_{m2} & \cdots\cdots & 0 \end{array}$$

Step IV - Searching Particular Subsets

The searching of aprticular subsets is done as follwing:

a) Choose an integer number N (the bigger the value of N the best is the search). Define a sequence of values A_r, using the expression:

$$A_r = \frac{r}{N} (a_{max} - a_{min}) + a_{min} \quad r = N, N-1; N-2, \ldots 1$$

Where a_{max} and a_{min} correspond to the maximum and minimum values of the distances between samples.

b) For each value of A_r, define subsets S^r_I, S^r_{II}, S^r_{III},...of order $\underline{r}$ using the following procedure:

b-1) Make zero all the elements of the matrix of distances a_{fk} that are greater than A_r.

b-2) Find for each sample X_k (therefore for each column k) the value of the reduced mean distance α_k^r, using the expression:

$$\alpha_k^r = \frac{1}{(m-u_k^r)} \sum_{f=1}^{m} a_{fk}$$

where u_k^r is the number of distances that were made zero in (b-1) (because they were greater than A_r) at column k.

b-3) Call α_I^r the minimum value of α_k^r

b-4) Define the first subset s_I^r by grouping at column I (associated to α_I^r) the samples corresponding to the distances that were not made zero in (b-1).

b-5) Exclude in the values of α_k^r, the ones whose columns correspond to samples already included in the first subset S_I^r.

b-6) Call α_{II}^r the minimum value of the remaining α_k^r, and then define the second subset S_{II}^r as was done before.

b-7) Define new subsets S_{III}^r,... using the values of α_k^r not included in previous subsets.

c) Stop the searching of subsets when, with the decrease of r one of the following situations happens:

c-1) The number of subsets remains unchanged for various values of r without showing any increase of common elements, which confirms their existence as independent entities.

c-2) The number of subsets remains unchanged for various values of r, showing a progressive reduction of common elements, which confirms their existence as interdependents entities.

c-3) The number of subsets increases up to the value m, without verifying either situation (c-1) or (c-2), which does not confirm their existence as real entities.

Samples that are not included on the identified subsets, will be considered as marginals.

The repetition of situations (c-1) during the process of searching identifies

secondary subsets (2^{nd} order subset, 3^{rd} order subset, etc.)

3. Conclusion

After the recognition of the existence of subsets, each subset may be submitted to the conventional statistical treatment.

BIBLIOGRAPHY

1. Miller,R.L.; Kahn,J.S.; Statistical analysis in Geological Sciences, Appendix A; John Wiley and Sons, Inc. New York, 1962.

```
      REAL N,K,L
      COMMON Q(51),P(51),X(51,51),Y(51,51),FI(51),PSI(51)
  111 FORMAT(4F6.2)
      KK=5
      LL=6
      READ(KK,111)PINIC,PFIM,QINIC,QFIM
  112 FORMAT(4F6.2,E9.2)
      READ(KK,112)H,L,N,DELTAH,K
      PI=3.1415
      IPINIC=IFIX(PINIC)+1
      IPFIM=IFIX(PFIM)+1
      IQINIC=IFIX(QINIC)+1
      IQFIM=IFIX(QFIM)+1
      DO 1000 IP=IPINIC,IPFIM
 1000 P(IP)=FLOAT(IP-1)
      DO 1010 IQ=IQINIC,IQFIM
 1010 Q(IQ)=FLOAT(IQ-1)
      QMAX=Q(IQFIM)
      WRITE(LL,11)  QMAX
   11 FORMAT(//////6H QMAX=,E6.2)
      WRITE(LL,12)H,L,N,K
   12 FORMAT(1X,2HH=,E9.4,2X,2HL=,E9.4,2X,2HN=,E9.4,2X,2HK=,E16.8)
      DO 1100 IQ=IQINIC,IQFIM
      DO 1100 IP=IPINIC,IPFIM
      SI1=P(IP)*DELTAH
      SR1=Q(IQ)*DELTAH
      SR2=(PI*P(IP)*DELTAH)/(2.*H)
      SI2=(PI*Q(IQ)*DELTAH)/(2.*H)
      SR3=SIN(SR2)*COSH(SI2)
      SI3=COS(SR2)*SINH(SI2)
      IF(SR3.EQ.0.)GO TO 15
      SR4=ATAN(SI3/SR3)
      GO TO 16
   15 SR4=PI/2.
   16 SR5=(2./N)*SR4
      SR6=(SR3**2+SI3**2)**(1./N)
      SR7=COS(SR5)
      SI7=SIN(SR5)
      SR8=SR6*SR7
      SI8=SR6*SI7
      SR9=L*SR8
      SI9=L*SI8
      X(IQ,IP)=SR9-SR1
      Y(IQ,IP)=SI9+SI1
      FI(IP)=-K*P(IP)*DELTAH
      PSI(IQ)=K*Q(IQ)*DELTAH
 1100 CONTINUE
 1110 CALL SAIDA(IPINIC,IQINIC,IQFIM)
      IPINIC=IPINIC+11
      IF(IPINIC.LT.IPFIM)GO TO 1110
      CALL EXIT
      END
```

```
      SUBROUTINE SAIDA(IPI,IQI,IQF)
      COMMON Q(51),P(51),X(51,51),Y(51,51),EI(51),PSI(51)
    1 FORMAT(1H=,/////,37X,30HORDEM DA TABELA  X= ****.****  )
    2 FORMAT (54X,13HY= ****.**** /)
   10 FORMAT(1X,12H*           *)
    3 FORMAT(4X,9H* P E  *,11(1X,2HP=,E6.2,1X))
    4 FORMAT(13H Q     * FI *,11(E9.2,1X))
    5 FORMAT(13H E PSI  *  *)
    6 FORMAT(1X,131(1H*) )
    7 FORMAT(2X,2HQ=,E6.2,2X,1H*,11(F9.4,1X))
    8 FORMAT (1X,E9.2,2X,1H*,11(F9.4,1X))
   11 FORMAT(13H           *)
      KK=5
      LL=6
      WRITE(LL,1)
      WRITE(LL,2)
      WRITE(LL,10)
      IP2=IPI+10
      WRITE(LL,3)(P(IP),IP=IPI,IP2)
      WRITE(LL,4)(FI(IP),IP=IPI,IP2)
      WRITE(LL,5)
      WRITE(LL,6)
      DO 1565 IQ=IQI,IQF
      WRITE(LL,11)
      WRITE(LL,7)Q(IQ),(X(IQ,IP),IP=IPI,IP2)
 1565 WRITE(LL,8)PSI(IQ),(Y(IQ,IP),IP=IPI,IP2)
      RETURN
      END
```

```
      DIMENSION X(120,20),XBAR(20),DELX(20),S(20)
      COMMON AFAST(120,120),M
    1 FORMAT(3I4)
    2 FORMAT(8F10.4)
    3 FORMAT(1H1,16HNUM DE AMOSTRAS=,I4,20H  NUM DE PARAMETROS=,I4,2X
     116HN ARBITRADO COMO,I4,/5X,9HDELTA(J)=,/,(10X,E10.4))
    4 FORMAT(1H1,10HMATRIZ X(I,J)
    5 FORMAT(1H ,7HAMOSTRA,I4/(10X,8E10.4))
    6 FORMAT(1H1,18HMATRIZ XLINHA(I,J))
    7 FORMAT(1H1,19HMATRIZ X2LINHA(I,J))
   10 FORMAT(1HO,10HAFAST MAX=,F10.2,3X,10HAFAST MIN=,F10.2)
C INICIO DAS ETAPAS I E II
      READ(5,1)M,N,NN
      READ(5,2)((DELX(J),J=1,N)
      DO 4444 I=1,M
 4444 READ(5,2)(X(I,J),J=1,N)
      WRITE(6,3)M,N,NN,(DELX(J),J=1,N)
      WRITE(6,4)
      DO 333 I=1,M
  333 WRITE(6,5)I,(X(I,J),J=1,N)
C CALCULO DE XBAR(J) S(J) E DA MATRIZ XLINHA(I,J)
      DO 100 J=1,N
      SOMAX=0.0
      SOMAQX=0.0
      DO 200 I=1,M
      SOMAX=SOMAX+X(I,J)
  200 SOMAQX=SOMAQX+X(I,J)**2
      XBAR(J)=SOMAX/FLOAT(M)
      S(J)=SQRT((FLOAT(M)*SOMAQX-SOMAX**2)/(FLOAT(M)*FLOAT(M-1)))
      DO 210 I=1,M
  210 X(I,J)=(X(I,J)-XBAR(J))/S(J)
      IF(DELX(J).EQ.0.) DELX(J)=S(J)
      DELX(J)=DELX(J)/S(J)
  100 CONTINUE
      WRITE(6,6)
      DO 127 I=1,M
  127 WRITE(6,5)I,(X(I,J),J=1,N)
C CALCULO DA MATRIZ X2LINHA(I,J)
      SOMAD= 0.0
      DO 220 J=1,N
  220 SOMAD=SOMAD + DELX(J)
      DBARL= SOMAD/FLOAT(N)
      DO 230 J=1,N
      DO 230 I=1,M
  230 X(I,J)=X(I,J)*(DBARL/DELX(J)      )
      WRITE(6,7)
      DO 128 I=1,M
  128 WRITE(6,5)I,(X(I,J),J=1,N)
C FIM DAS  ETAPAS I E II
C INICIO DA ETAPA III DETERM AFASTAMEN, E AFAST MAX EMIN
      AFMIN =32000.
      AFMAX=0.0
      DO 240 LF =1,M
      DO 250 LK =1,M
      SOMAQ=0.0
      DO 260 J=1,N
  260 SOMAQ=SOMAQ + (X(LF,J)-X(LK,J))**2
```

```
      AFAST(LF,LK)=SQRT(SOMAQ)
      IF(LF-LK) 888,250,888
  888 IF(AFAST(LF,LK)-AFMIN)88,88,89
   88 AFMIN=AFAST(LF,LK)
      GO TO 250
   89 IF(AFAST(LF,LK)-AFMAX)250,188,188
  188 AFMAX=AFAST(LF,LK)
  250 CONTINUE
  240 CONTINUE
      WRITE(6,10)AFMAX,AFMIN
C FIM  DA ETAPA III INCLUINDO OS AFASTAMENTOS MAXIMO E MINIMO
C INICIO DA ETAPA IV
      DELTA=AFMAX-AFMIN
      DO 55 IR=1,NN
      AR =AFMIN+(FLOAT(NN-IR+1)/FLOAT(NN))*DELTA
      WRITE(6,4422)IR
 4422 FORMAT(1H1,31HREALIZACAO DA ETAPA IV PARA R =,I3,13H OU SEJA PARA,
     1/)
      CALL ETAP4(AR)
   55 CONTINUE
      CALL EXIT
      END
```

```
      SUBROUTINE ETAP4(AR)
      DIMENSION ALFKR(120),ISIR(120)
      COMMON AFAST(120,120),M
   11 FORMAT(1HO,20HVALORES DE ALFKR(LK),/(5X,20F5.2) )
   12 FORMAT(1H ,4HAR =,F10.4)
   13 FORMAT(1HO,5X,12HALFAMINIMO =,F10.4 )
   14 FORMAT(1HO,5X,56HAMOSTRA PRINCIPAL DO SUBCONJUNTO QUE SE SEGUE E A
     1 NUMERO,I4/)
      DO 2000 LK=1,M
      IUKR=0
      DO 1000 LF=1,M
      IF(AFAST(LF,LK))24,24,301
   24 IUKR=IUKR+1
      GO TO 1000
  301 IF  (AFAST(LF,LK)-AR )1000,1000,400
  400 AFAST(LF,LK)=0.0
      IUKR=IUKR+1
 1000 CONTINUE
C SUBTRAIR 1 DE UKR PARA EXCLUIR DISTANCIA ENTRE AMOSTRAS IDENTICAS
      IUKR=IUKR-1
C CALCULO DOS ALFA(I,R)
      SOMAF=0.0
      DO 1010 LF=1,M
 1010 SOMAF=SOMAF+ AFAST(LF,LK)
 2000 ALFKR(LK)=SOMAF/FLOAT(M-IUKR)
      WRITE(6,12)AR
      WRITE(6,11)(ALFKR(LK),LK=1,M)
```

```
C PESQUISA DE ALFAKR(LK) MINIMO
 4000 ALFMI= 32000.
      DO 1050 LK=1,M
      IF(ALFKR(LK))544,1050,544
  544 IF(ALFKR(LK)-ALFMI)44,44,1050
   44 ALFMI=ALFKR(LK)
      ICOL=LK
 1050 CONTINUE
      WRITE(6,13)ALFMI
C DETERMINACAO DOS SUBCONJUNTOS S(I,R)
      WRITE(6,14)ICOL
      II=0
      DO 1099 LF=1,M
      IF(LF-ICOL)99,2099,99
   99 IF(AFAST(LF,ICOL))2099,1099,2099
 2099 ALFKR(LF)=0.0
      II=II+1
      ISIR(II)=LF
 1099 CONTINUE
      WRITE(6,15)(ISIR(I),I=1,II)
   15 FORMAT(1H ,5X,16HAMOSTRAS NUMEROS,/(23X,20I4))
      K=0
      DO 2355 LK=1,M
      IF(ALFKR(LK))2354,2355,2354
 2354 K=1
 2355 CONTINUE
      WRITE(6,11)(ALFKR(LK),LK=1,M)
      IF(K)5000,5000,4000
 5000 RETURN
      END
```