BURT – ALGORITAM I PROGRAM ZA ODREĐIVANJE LATENTNIH DIMENZIJA SKUPA NOMINALNIH VARIJABLI

Konstantin Momirović, Vesna Dobrić, Marijan Gredelj i Lajos Szirovicza
Sveučilišni računski centr, Zagreb

Prispjelo: 28. 04. 1980.

UDK: 159.3
Originalan znanstveni rad

SAŽETAK

Predložen je algoritam i napisan program za određivanje latentnih dimenzija jednog skupa nominalnih varijabli transformiranih u binarni oblik. Algoritam određuje značajne glavne osovine matrica vjerojatnosti, izračunava vrijednosti entiteta na glavnim komponentama, transformira koordinatni sustav vektora varijabli u orthoblique poziciju i određuje vrijednosti entiteta na latentnim dimenzijama koje su definirane ovim postupkom.

0. UVOD

Gotovo u isto vrijeme, ali nezavisno jedan od drugoga, i polazeći od različitih teorijskih osniova, predložili su Burt i Guttman (Burt, 1950; Guttman, 1950) procedure za kvantifikciju nominalnih varijabli. Formalno, Burtova procedura pripadala je onoj klasi postupaka za analizu podataka koja se, obično, naziva faktorskom analizom; u stvari, Burt je predložio, kao postupak za kvantifikaciju nenumeričkih podataka, određivanje glavnih komponenata izvedenih iz glavnih osovina matrice vjerojatnosti pripadanja entiteta podskupovima definiranim intersekcijama kategorija analiziranih nominalnih varijabli, nakon što je iz te matrice parcijaliziran efekt faktora definiranog vjerojatnostima pripadanja entiteta kategorijama tih varijabli. Guttmanova procedura pripdala je, formalno, metodama multidimenzionalnog skaliranja; no lako se moglo utvrditi da se taj postupak može svesti na postupak određivanja latentnih struktura i da je zato ekvivalentan Burtovu postupku.

I postupak McDonalda (McDonald, 1969) može biti tretiran i pod vidom analize latentnih struktura (Momirović, 1972) i pod vidom generaliziranoga faktorskog modela. Taj je drugi pristup, čini se, općenito prihvaćen; većina algoritama za analizu latentnih dimenzija nominalnih varijabli razvijenih kasnije (Hayashi, 1974; Escoufier, Caillez i Pages, 1978, Benzécri, 1976) temelji se, formalno, na faktorskom modelu.

Postupak koji je predložen u ovom radu pripada klasi transformacija na osnovi kojih je latentne dimenzije nominalnih varijabli određivao i Burt. Naime, slično kao i u Burtovu originalnom algoritmu latentne dimenzije definirane su na osnovi komponentnog modela reprodukcije vjerojatnosti pripadanja nekog entiteta

skupovima koji su definirani intersekcijama skupova formiranih u skladu s vrijednostima entiteta opisanih nad skupom analiziranih nominalnih varijabli. Različitost predloženog od originalnog Burtova algoritma sastoji se u transformaciji svih značajnih glavnih osovina, i to ne samo u glavne komponente nego i u parsimonijsku poziciju definiranu orthoblique rotacijom.

1. GLAVNE KOMPONENTE NOMINALNIH VARIJABLI

Neka je $N = \{N_j;\ j = 1, \ldots, m\}$ skup nominalnih varijabli kojima je opisan neki skup entiteta $U = \{e_i;\ i = 1, \ldots, n\}$ koji je definiran kao reprezentativni uzorak iz neke populacije P.

Neka je svaka varijabla N_j, $j = 1, \ldots, m$ definirana kao skup $N_j = \{K_{jk};\ k = 1, \ldots, m_j\}$, gdje su podskupovi K_{jk}, $j = 1, \ldots, m$; $k = 1, \ldots, m_j$ definirani kategorijama nominalne varijable N_j

Definirajmo $q = \sum_{j=1}^{m} m_j$ i pretpostavimo, za sada, da je $q \leq n$.

Uvedimo, za svaki $e_i \in U$. koji je opisan varijablom $N_j \subset N$ reprezentaciju definiranu vektorom $B_{ij}^T = (b_{i1}, \ldots, b_{ik}, \ldots, b_{im_j})_j$ $j = 1, \ldots, m$ koji je konstruiran tako da, ako je $e_i \in K_{jk}$, $b_{ik} = 1$ $e_i \notin K_{jk}$, $b_{ik} = 0$ $j = 1, \ldots, m$ i definirajmo, za svaki entitet $e_i \in U$. vektor $B_i^T = \Gamma_{j=1}^{m} B_{ij}^T$ $i = 1, \ldots, n$ gdje je sa Γ označena operacija konkatenacije.

Organizirajmo, operacijom $\Gamma_{i=1}^{n} B_i^T = B$, vektore B_i^T matricu $B = (B_{ij}^T)$ $i = 1, \ldots, n$ $j = 1, \ldots, m$ i uočimo da matrica B potpuno definira skup U opisan nad skupom N

Matrica $C = B^T B = (C_{zk})$ $j = 1, \ldots, m$ $k = 1, \ldots, m$ sadrži, očito, kao elemente kontingencijske matrice C_{jk} između varijabli N_j i N_k i $j = k$, opisanih nad skupom U Naravno, ako je $j = k$, elementi c_{kjkj} matrica C_{jj} jednaki su num $(e_i \in K_{jk})$ a vandijagonalni elementi nuli, tako je tr $C_{jj} = n\ \forall j$, i otuda tr $C = n * m$. Ako $j \neq k$, elementi c_{kjkk} matrica C_{jk} jednaki su num $(e_i \in K_{jk} \cap K_{kk})$.

Prema tome, ako je U zaista reprezentativan i dovoljno velik uzorak iz P, u matrici $P = C\frac{1}{n} = (P_{jk})$ $j = 1, \ldots, m$ $k = 1, \ldots, m$ procjene su vjerojatnosti $p(e_i \in K_{jk} \cap K_{kk}$

Neka je $\Lambda = (\lambda_r)$, $r = 1, \ldots, s \leq (q - m + 1)$ matrica nenultih svojstvenih vrijednosti matrice P i neka je $X = (X_r)$ matrica njima pridruženih svojstvenih vektora skaliranih tako da je $X^T X = I$. Očito. $P = X \Lambda X^T \sum_{r=1}^{s} \lambda_r X_r X_r^T$

Neka su svojstvene vrijednosti λ_r uređene tako da je $\lambda_r \lambda_{r+1}$, $r = 1, \ldots, s - 1$ i neka su tako uređeni i svojstveni vektori X_r u X. Vektore $K_r = BX_r$ $r = 1, \ldots, s$ nazvat ćemo glavnim komponentama nominalnih varijabli iz N, opisanih na

$U \subset P$, a matricu $K = (K_r)$ matricom glavnih komponenata.

Uočimo da je $K^T K \frac{1}{n} = X^T P X = \Lambda$ i da sukcesivne ortrogonalne komponente K_r maksimalno diferenciraju entitete iz U i, potencijalno, iz P, u prostoru koji je definiran nominalnim varijablama iz N. Jer, ako je $X^* = (X_r^*)$ matrica vektora odabranih tako da je, uz uvjet funkcija

$$X_r^{*T} X_r^* = 1, \qquad r = 1, \ldots, s$$

$$f(X_r^*) = X_r^{*T} B^T B x_r^* \frac{I}{n} = \max \qquad r = I, \ldots, s$$

deriviranjem

$$\partial f(X_r^*)/\partial(X_r^*) = 2Px_r^* - 2\lambda_r^* X_r^* \qquad r = 1, \ldots, s$$

pa, izjednačivanjem parcijalnih izvoda s nulom i dijeljenjem s 2

$$Px_r^* = \lambda_r^* X_r^* \qquad r = 1, \ldots, s$$

odnosno

$$(P - \lambda_r^* I)\, x_r^* = 0 \qquad r = 1, \ldots, s$$

iz čega slijedi da je $\lambda_r^* = \lambda_r$ i $x_r^* = x_r$.

Razmotrimo sada relacije vektora iz B i vektora iz K Matrica skalarnih produkata tih vektora pomnožena konstantom $\frac{1}{n}$ je

$$H^* = B^T K \frac{I}{n} = PX = X\lambda$$

i očito je faktorska matrica od P^2, jer

$$H^* H^{*T} = X\lambda^2 X^T = P^2;$$

prema tome, faktorska matrica od P je matrica glavnih osovina

$$H = H^* \lambda^{-1/2} = X\lambda^{1/2}$$

zato jer je

$$HH^T = X\lambda X^T = P$$

2. REDUKCIJA DIMENCIONALNOSTI KOMPONENTNOG PROSTORA

Glavne komponente međusobno se razlikuju s obzirom na važnost za reproduciranje informacija sadržanih u matrici B, a time, očito, i s obzirom na njihovu interpretativnu vrijednost. Stoga uvodimo t kao cjelobrojnu vrijednost

s t l

koja je upravo dovoljna da reproducira sve značajne i važne informacije iz Najpodesniji način za definiranje t jest postojanje potvrđene hipoteze o dimenzionalnosti prostora reprezentiranog linearizirani varijablama iz B(1). Ipak, mnogo je češća situacija u kojoj nema valjane hipoteze a t. Program u toj situaciji primjenjuje algoritam koji je formalno identičan DMEAN kriteriju Momirovića i Štaleca (1973).

$$t = \text{num}(\lambda_r \geqslant \alpha)$$

gdje je

$$\alpha = \sum_{r-1}^{s} \lambda_r / q \qquad (2)$$

ili, kao stroži kriterij,

$$t = \text{num}(\lambda_r\, \alpha^*)$$

gdje je

$$\alpha^* = \sum_{r=1}^{s} \lambda_r / s \qquad (3)^*$$

Sada se matrica P može definirati kao zbroj matrica ranga 1, čije su svojstvene vrijednosti jednake vrijednostima spektra matrice P, tj.

$$P = \sum_{r=1}^{t} \lambda_r X_r X_r^T + \sum_{r=t+1}^{s} \lambda_r X_r X_r^T$$

* Program standardno primjenjuje strategiju (3).

pa ako su sve svojstvene vrijednosti uređene tako da vrijedi

$$\lambda_r < \lambda_{r+1},$$

matrica

$$Q = \sum_{r=1}^{t} \lambda_r X_r X_r^T = P - E$$

pri čemu je, očito,

$$E = \sum_{r=t+1}^{s} \lambda_r X_r X_r^T$$

bit će, uz rang t, najbolja aproksimacija matrice P

Ako uvedemo

$$X^* = (X_r) \qquad r = 1, \ldots, t$$

$$\Lambda^* = (\lambda_r) \qquad r = 1, \ldots, t$$

zadržane glavne komponente bit će u matrici $L^* = B\Lambda^*$,
koja, uz rang t, najbolje reproducira informacije sadržane u matrici B, jer je

$$B^* = BXX^T$$

najbolja aproksimacija matrice B uz reducirani rang t.

$$H^* = B^T L \Lambda^{*-1/2} \frac{1}{n} PX^* \Lambda^{*-1/2} = \Lambda^{*1/2}$$

nazvat ćemo matricom značajnih glavnih osovina matrice P. Međutim, kako vrijedi

$$Q = H^* H^{*T}$$

vrijedi i

$$P = H^* H^{*T} + E$$

pa slijedi da je H^* faktorska matrica matrice vjerojatnosti P, a E rezidualna matrica vjerojatnosti, koje se ne mogu pripisati generatorima reprezentiranim zadržanim glavnim komponentama.

3. LATENTNE DIMENZIJE

Latentne dimenzije analiziranog skupa nominalnih varijabli određuje algoritam BURT orthoblique transformacijama svojstvenih vektora, pridruženih zadržanim glavnim komponentama.

Neka je $j = 1, \ldots, q$
$X = (X_{jp})$ $p = 1, \ldots, t$

matrica svojstvenih vektora pridruženih svojstvenim vrijednostima

H_p, $p = 1, \ldots, t$, i neka je
$\Lambda = (\Lambda_p)$ $p = 1, \ldots, t$

dijagonalna matrica tih svojstvenih vrijednosti.

Nađimo, generaliziranim orthomax postupkom (Mulaik, 1972; Fulgosi, 1979), orthonormalnu matricu T reda t koja, uz uvjete $T^T T = TT^T = I$, maksimizira Kaiserovu (Kaiser, 1957) varimax funkciju

$$v = q \sum_{j=1}^{q} \sum_{p=1}^{t} a_{jp}^{*4} - \left(\sum_{j=1}^{q} a_{jp}^{*2}\right)^2$$

nad elementima matrice

$$A^* = XT = (a_{jp}^*)$$

Uočimo da je A^{*T} lijevi pseudoinverz matrice A^* i da, na osnovi generalnog faktorskog modela,

$$B = \Phi^* A^{*-T} + N$$

gdje je Φ^* matrica latentnih dimenzija, N matrica residuala, $\Phi^* = BXT$ s matricom necentriranih kovarijanci

$$M^* = \Phi^{*T} \Phi^* \frac{1}{n} = PXT = X\Lambda T$$

Necentrirane kovarijance binarnih varijabli iz B i latentnih dimenzija iz Φ^* bit će elementi matrice

$$F^* = B^T \Phi \frac{1}{n} = PXT = X\Lambda T$$

Očito, $F^* = A^*M^*$
pa je A^* matrica sklopa, F^* matrica strukture binarnih varijabli kojima su reprezentirane nominalne varijable iz N u prostoru koji je definiran latentnim dimenzijama iz Φ^*

Naravno, $A^*F^{*T} = P - E$
pa su A^* i F^* faktorske matrice od P

Standardizirajmo sada latentne dimenzije tako da su necentrirane varijance tih varijabli jednake 1. Ako je

$$D^2 = \text{diaf } M^*$$

matrica necentriranih varijanci varijabli iz Φ^*, u matrici $\Phi = BXTD^{-1}$
bit će standardizirana (ali, naravno, i dalje necentrirana) latentne dimenzije.

Relacije tako standardiziranih latentnih dimenzija su elementi matrice

$$M = \Phi^T \Phi \frac{1}{n} = D^{-1}T^T \Lambda TD^{-1}$$

a relacije binarnih varijabli iz B i standardiziranih latentnih dimenzija elementi matrice

$$F = B^T \Phi \frac{1}{n} = PXTD^{-1} = X\Lambda TD^{-1}$$

Matrica F je, naravno, također matrica strukture binarnih varijabli, ovaj put definirana u prostoru što ga omeđuju varijable iz Φ. Matrica sklopa binarnih varijabli bit će, u ovoj metrici,

$$A = FM^{-1} = XTD,$$

A i F su, naravno, također faktorske matrice od P, jer

$$AF^T = AMA^T = P - E$$

gdje je, i dalje,

$$E = N^T N \frac{1}{n} = P - X\Lambda X^T$$

matrica residualnih vjerojatnosti u t dimenzionalnom komponentnom modelu.

Algoritam BURT određuje matrice sklopa, strukture i relacija za nestandardizirane i standardizirane latentne dimenzije, i za te varijable određuje raspodjele, izračunava parametre tih raspodjela i testira, metodom Kolmogorova i Smirnova, hipotezu da su raspodjele latentnih dimenzija Gauss-Bernoullieva tipa.

4. PROGRAM BURT

Program BURT napisan je u verziji 4.7/M programskog sistema SS (Zakrajšek, Štalec i Momirović, 1974). Program pretpostavlja (1) da je skup nominalnih varijabli prethodno transformiran u binarne varijable,

(2) da je pripremljena SEQUENCE naredba i onoliko VARIABLE naredbi koliko je ukupno kategorija u analiziranom skupu varijabli,

(3) da broj eniteta nije veći od 10000

(4) da broj kategorija nije veći od 250.

```
OUTPUT (DEVICE = PR 1)
*
*            ***BURT***
*THIS PROGRAM REALIZE AN ALGORITHM FOR THE ANALYSIS OF PRIN-
CIPAL
```

```
*COMPONENTS, FOLLOWED BY ORTHOBLIQUE TRANSFORMATION OF
THE IMPORTANT
*LATENT DIMENSIONS, OF A SET OF QUALITATIVE DATA TRANSFORMED
TO THE
*BINARY FORM. ALGORITHM IS DESCRIBED IN MOMIROVIC, DOBRIC,
GREDELJ
*AND SZIROVICZA, 1980.
*
HEADING (TEXT=LATENT STRUCTURE OF QUALITATIVE DATA)
TEXT (TEXT=                    BURT)
INPUT (DATA=VESNA, SCORE=BOOL, ROWNAME=ENT, VARNAME=
VAR)
MULT (A=BOOL, TA, B=BOOL, M=C)
PRINT (MATRIX=C, TEXT=CONTINGENCY MATRIX)
*
*IN LINEAR, CA=1/N, N=NUMBER OF SUBJECTS.
LINEAR (A=C, CA=                , M=P)
PRINT (MATRIX=P, TEXT=BROBABILITY MATRIX)
DIAGONALISATION (R=P, LAMBDA=L, X=XT, NZ)
HOTELLING (LAMBDA=L, X-XT, F=HT, NAMEF=NMF)
TRANSPOSE (OLD=HT, NEW=H)
PRINT (MATRIX=H, TEXT=PRINCIPAL AXES)
MULT (A=HT, B=H, M=LL)
DIAG (A=LL, C=-0.5, D=LM)
MULT (A=H, B=LM, M=X)
MULT (A=BOOL, B=X, M=K)
PRINT (MATRIX=K, TEXT=PRINCIPAL COMPONENTS)
STATISTICS (SCORE=k, S, CLASS=9, Z=ZK)
DELETE (MATRIX=L)
DELETE (MATRIX=XT)
TRANSPOSE (OLD=X, NEW=XT)
VARIMAX (F=XT, FN=AT, TAU=TT)
TRANSPOSE (OLD=AT, NEW=AZ)
TRANSPOSE (OLD=TT, NEW=T)
MULT (A=X, B=LL, M=COW)
MULT (A=COW, B=T, M=FZ)
MULT (A=TT, B=LL, M=HORSE)
MULT (A=HORSE, B=T, M=MZ)
DELETE (MATRIX=COW)
DELETE (MATRIX=HORSE)
DIAG (A=MZ, C=0.5, D=D)
DIAG (A=D, C=-1.0, D=DM)
MULT (A=AZ, B=D, M=A)
SCALE (C=MZ, R=M)
MULT (A=FZ, B=DM, M=F)
MULT (A=K, B=T, M=S)
RESIDUAL (R=P, F=HT, RES=Q)
```

```
PRINT (MATRIX=T, TEXT=TRANSFORMATION MATRIX)
PRINT (MATRIX=AZ, TEXT=RAW PATTERN MATRIX)
PRINT (MATRIX=MZ, TEXT=RAW RELATIONSHIPS MATRIX)
PRINT (MATRIX=FZ, TEXT=RAW STRUCTURE MATRIX)
PRINT (MATRIX=A, TEXT=PATTERN MATRIX)
PRINT (MATRIX=M, TEXT=RELATIONSHIPS OF LATENT DIMENSIONS)
PRINT (MATRIX=F, TEXT=STRUCTURE MATRIX)
PRINT (MATRIX=Q, TEXT=RESIDUAL PROBABILITY MATRIX)
STATISTICS (SCORE=S, S, CLASS=9, Z=ZS)
DIAG (A=P, C=-0.5, M=DEL)
MULT (A=DEL, B=F, M=NF)
MULT (A=DEL, B=A, M=NA)
PRINT (MATRIX=NA, TEXT=NORMALISED PATTERN MATRIX)
PRINT (MATRIX-NF, TEXT=NORMALISED STRUCTURE MATRIX)
HEADING (TEXT=END)
```

LITERATURA

1. Benzecri, J. P.: *L' analyse des données. L' analyse des correspondances.* 2. ed. Dunod, Paris, 1976.

2. Burt, C.: *The factorial analysis of qualitative data.* Biritsh Journal of Psychology (Statistical Section), 3, 166-185 (1950).

3. Cazes, P., A. Baumerder, S. Bonnefous et J. P. Pages: *Codage et analyse des tableaux logiques: Introduction a la partique des variables qualitatives.* Cahiers du buro, 27, 1977 (Report u *Analyse de données et informatique,* 1. Fontainebleau, 1979).

4. Cazes, P., S. Bonnefous, A. Baumerder et J. P. Pages: *Description coherente des variables qualitative prises globalement et de leurs modalites. Statistique et Analyse des Données, 1976, (Reprint u Analyse des données et informatique,* 1, Fontainebleau, 1979).

5. Cazes, P. et J. P. Lecontre: *Etude de quelques problemes de codage en analyse des correspondences.* Cahiers du buro, 27, 1977 (Reprint u *Analyse de données et informatique,* 1, Fontainebleau, 1979).

6. de Leew: J. *Cannonical analysis of categorical data,* University of Leiden, 1973.

7. Escoufier, Y.: *Echantillonnage dans une population de variables aléatoires réelles.* Institut de Statistique des Universités de Paris, 1970.

8. Escoufier, Y., F. Caillez et J. P. Pagres: *Géometrie et techniques particulieres en analyse factorielle.* Congres Europén de Psychometrie. Upsala, 1978 (Reprint u *Analyse de données et informatique,* 1, Fontainebleau, 1979).

9. Escoufier, Y. and P. Robert: *Choosing variables and metrics by optimizing the RV-coefficient,* Analyse de données et informatique, 1, Fontainebleau, 1979.

10. Fulgosi, A.: *Faktorska analiza,* Školska knjiga, Zagreb, 1979.

11. Guttman, L.: *The principal components of scale analysis.* In S. S. Stouffer, Ed. Measurement and Prediction, Princeton University Press, Princeton, 1950.

12. Hayashi, C.: *Minimum dimension analysis,* Behaviormetrika, 1, 1-24 (1974).

13. Kobilinsky, A.: *Ordre entre formes quadratiques: Application a l'optimalite de sous-espaces en analyse des données.* Analyse de données et informatique, 1, Fontainebleau, 1979.

14. McDonald, R. P.: *The common factor analysis of multicategory data.* The British Journal of Mathematical and Statistical Psychology, 22, 2, 165-175 (1969).

15. Momirović, K.: *Metode za kondenzaciju i transformaciju kinezioloških informacija,* Institut za kineziologiju, Zagreb, 1972.

16. Momirović, K. i J. Štalec: *DMEAN i DMAX kriteriji za određivanje broja značajnih image faktora pri analizi zadataka u psihologijskim testovima,* Stručni skupovi psihologa »Dani Ramira Bujasa«, 1970, 1972, Zagreb, 1973 (95-104).

17. Robert, P. and Y. Escoufier: *A unifying tool for linear multivariate statistical methods: the RV-coefficient.* Applicational Statistics, 25, 3, 257-265 (1976).

18. Zakrajšek, E., J. Štalec i K. Momirović: *SS – programski sistem za multivarijatnu analizu podataka,* Zbornik simpozija »Kompjutor na sveučilištu«, C8.1-C8.16, 1974.

Summary

An algorithm for determination of latent dimensions of a set of nominal variables transformed into a binary ones is proposed, and a program for it described. Significant main axes of the probability matrix, and principal components scores of the entities can be determined by this algorithm, as well as the transformation of the set of variable vectors into orthoblique position and the factor scores of the entities.