KONSTANTIN MOMIROVIĆ I ŽIVAN KARAMAN
Sveučilišni računski centar, Zagreb

Primljeno 13. 4. 1982.

INDIFF — MODEL, ALGORITAM I PROGRAM ZA ANALIZU PROMJENA STANJA NEKOG OBJEKTA OPISANOG NAD SKUPOM KVANTITATIVNIH VARIJABLI

PRETHODNO PRIOPĆENJE

SAŽETAK

Konstruirani su model i algoritam i napisan je program za komponentnu analizu promjena stanja individualnih objekata. Dekomponiranjem trajektorije promjena na ortogonalne komponente dobiva se uvid u strukturu stohastičkog procesa opisanog kvantitativnim varijablama. Također se analiziraju reziduali kao i relacije među vremenskim točkama.

0. UVOD

Model, algoritam i program, koji su ovdje opisani, namijenjeni su komponentnoj analizi promjena stanja nekog objekta opisanog nad skupom kvantitativnih varijabli registriranih kroz neki vremenski period. Potreba za takvom vrstom analize javlja se prilikom istraživanja u kineziologiji, medicini, ekonomiji i drugim znanostima, onda kada je moguće opisati promjene stanja nekog entiteta nekim skupom izmjerivih varijabli. Tako, npr., u kineziologiji to može biti neki sportaš kojemu se u toku treninga mjere fiziološke, motoričke i druge promjene; u medicini to može biti pacijent kojemu se kroz nekoliko dana ili tjedana svakih nekoliko sati registriraju promjene pulsa, krvnog pritiska, temperature i druge medicinski relevantne promjene; u ekonomiji to može biti prodaja nekih proizvoda u nekoj zemlji praćena mjesčno unatrag nekoliko godina, itd.

Tip modela koji je ovdje predložen namijenjen je analizi nestacionarnih stohastičkih procesa; stacionarne stohastičke procese moguće je analizirati pod modelom koji je implementiran u programima serije TALAMBAS. Promjene stanja objekta koje su originalno opisane nekim skupom od n manifestnih varijabli objašnjavaju se gotovo posve sa $p<n$ latentnih dimenzija — komponenata promjena. Na taj način dobvaimo bitno jednostavniji prikaz promjena stanja proučavanog objekata. Ovakvom analizom dobivamo također uvid u strukturu komponenata promjena, tj. odnose registriranih varijabli i komponenata promjena, kao i jednostavan prikaz relacija među vremenskim točkama. Model je izvorno predložen od L. Tuckera, a opis originalni hmodela i algoritma nalazi se u Momirović (1972). U ovom radu dan je detaljan opis algoritma koji je implementiran u programu INDIFF, napisanom u meta-jeziku SS (Zakrajšek, Štalec, Momirović, 1974) i implementiranom na računalu UNIVAC 1100 Sveučilišnog računskog centra u Zagrebu.

1. MODEL

Neka je nekim eksperimentom dobivena matrica trajektorije promjena B tipa m x n, gdje je m broj vremenskih točaka, a n broj varijabli. Pri tome moraju biti zadovoljeni slijedeći uvjeti:

a) $m > n$

b) rang $(B) = n$ (tj. varijable $v_1, \ldots, v_n$ su linearno nezavisne)

c) vremenske točke su ekvidistantne

d) interval između ma koje dvije susjedne vremenske točke je dovoljno kratak da se promjene stanja u svakoj od n varijabli mogu aproksimirati nekom monotonom funkcijom.

Varijable iz B standardiziramo i matricu standardiziranih varijabli označimo sa Z.

Standardiziranu trajektoriju promjena Z dekomponiramo na $p < n$ aditivnih ortogonalnih komponenata promjena $K_1, K_2, \ldots, K_p$, definiranih na skupu vremenskih točaka, takvih da vrijedi

$$K_1 = \sum_{j=1}^{n} x_{1j} v_j$$

je linearna kombinacija standardiziranih varijabli $v_1, v_2, \ldots, v_n$ čija je varijanca veća od varijance ma koje druge linearne kombinacije čiji koeficijenti zadovoljavaju

uvjet $\sum_{j=1}^{n} x^2_{1j} = 1$

$$K_2 = \sum_{j=1}^{n} x_{2j} v_j$$

je linearna kombinacija standardiziranih varijabli čija je varijanca veća od varijance m koje druge linearne kombinacije varijabli ortogonalne na K_1, a čiji koeficijenti

zadovoljavaju uvjet $\sum_{j=1}^{n} x^2_{2j} = 1.$

$$K_p = \sum_{j=1}^{n} x_{pj} v_j$$

jelinearna kombinacija standardiziranih varijabli čija je varijanca veća od varijance ma koje druge linearne kombinacije varijabli ortogonalne na $K_1, K_2, \ldots, K_{p-1}$, a čiji

koeficijenti zadovoljavaju uvjet $\sum_{j=1}^{n} x^2_{pj} = 1,$

Broj zadržanih komponenata p određen je u skladu sa PB kriterijem (Štalec, Momirović, 1971).

Konkateniramo li vektore K_l, $l=1, 1, \ldots, p$ u matricu K (tipa m x p), dobijemo kondenzirani prikaz trajektorije promjena koji sadrži gotovo sve informacije kao i polazna matrica Z, a mnogo je pregledniji i pogodniji za analizu.

Strukturom komponenata promjena nazivamo korelacije između standardiziranih varijabli (Z) i komponenata (K)

$$H = Z^T K \frac{1}{m}$$

a relacije vremenskih točaka definirane su kao normirani skalarni produkti vektora vremenskih točaka

$$Q = (\mathrm{diag}\ ZZ^T)^{-1/2}\ ZZ^T\ (\mathrm{diag}\ ZZ^T)^{-1/2}$$

2. ALGORITAM

2.1. Parametri varijabli

Razmotrimo sada pobliže eksperimentom dobijenu matricu $B = (b_{ij})$, $i = 1, \ldots, m$; $j = 1, \ldots, n$. Svaki redak matrice B predstavlja jednu vremensku točku $t_i \,\Xi T = \{_i, l = 1, 2, \ldots, m\Xi$ u kojoj su registrirane varijable, dok stupci matrice B odgovaraju pojedinim varijablama $v_j \,\Xi V = \{v_j, j = 1, \ldots, m\}$ praćenim kroz niz vremenskih točaka. Kako ne možemo očekivati da varijable budu u istoj metrici, potrebno ih je na neki način standardizirati da bismo ih mogli međusobno uspoređivati. To napravimo tako da svaka standardizirana varijabla ima očekivanje 0 i varijancu 1.

$v_j \rightarrow v_j$

$E(v_j) = 0 \qquad j = 1, 2, \ldots, n$

$\mathrm{var}\ (v_j) = 1.$

Ako sa $1^T = (1, 1, \ldots, 1)$ označimo m-dimenzionalni sumacioni vektor, tada je matrica kovarijanci varijabli iz V dana sa $C = (B^T B - B^T J B) \frac{1}{m}$, gdje je $J = 1\, 1^T \frac{1}{m}$. Matricu varijanci varijabli iz V označimo sa $S^2 = \mathrm{diag}\ C$, pa je matrica standardiziranih podataka $Z = (B - JB)\, S^{-1}$.

Matrica korelacija varijabli v_j iz V biti će

$$R = S^{-1}\ CS^{-1} = Z^T Z \frac{1}{m}$$

Parcijalne korelacije varijabli iz V dane su sa

$$P = -\ UR^{-1}\ U,$$

gdje je U dijagonalna matrica definirana sa $U^2 = (\mathrm{diag}\ R^{-1})^{-1}$.

U matrici U^2 nalaze se zapravo unikviteti varijabli v_j iz V.

Pogledajmo sada detaljnije uvjete navedene u 1.a) — 1.d). Prva dva uvjeta (1.a) i (1.b) nam omogućuju izračunavanje unikviteta i preko toga primjenu PB kriterija za određivanje broja zadržanih glavnih komponenata. Ukoliko ta dva uvjeta nisu ispunjena nemoguće je analizirati promjene na ovdje predložen način. U tom slučaju moguće je primijeniti algoritam i program TALAMBAS i koji unikvitete varijabli računa primjenom generaliziranog Moore-ePnrosoovog (Moore, 1935; ePnrose, 1955) inverza R^{-} matrice korelacija R. Druga dva uvjeta, (1.c) i (1.d), nisu formalne prirode, te njihovo neispunjavanje neće onemogućiti provođenje analize, ali će tako dobiveni rezultati biti neupotrebljivi budući su uvjeti (1.c) i (1.d) implicitno sadržani u algoritmu. Stoga je važno osigurati da sva četiri uvjeta (1.a) — (1.d) budu zadovoljena.

2.2. Komponente promjena

Standardiziranu trajektoriju promjena Z dekomponirati ćemo na aditivne, ortogonalne komponente. Tražimo takvu linearnu kombinaciju.

$$K^l = x_{l1}\, v_1 + x_{l2}\, v_2 + \ldots + x_{ln}\, v_n$$

standardiziranih varijabli v_j iz Z koja ima maksimalnu varijancu, tj.

$\mathrm{var}\ (K_l) = \max$

uz uvjet da njeni koeficijenti zadovoljavaju dodatni uvjet

$$\sum_{j=1}^{n} x^2_{lj} = 1$$

Označimo li sa $X^T_l = (x_{l1}, x_{l2}, \ldots, x_{ln})$ prethodne relacije, možemo u matričnom obliku pisati

$$K_l = Z X_l \tag{1}$$

$$K_l^T K_l \frac{1}{m} = \max \tag{2}$$

$$X_l^T X_l = 1 \tag{3}$$

Budući je

$$K_l^T K_l \frac{1}{m} = X_l^T Z^T Z X_l \frac{}{} = X^T R X_l \tag{3}$$

lako se pokaže da se problem maksimizacije (2) uz uvjet (3) svodi na rješavanje karakteristične jednandžbe za R

$$(R - \lambda_l\, I)\, X_l = O \tag{5}$$

Da bismo dobili netrivijalno rješenje ($X_l \neq O$) nužno je da bude

$$|\, R - \lambda_l I\, | = 0,$$

a to znači da λ_l mora biti svojstvena vrijednost matrice R, pa je onda i X_l pridruženi svojstveni vektor. Da bismo odredili koju svojstvenu vrijednost treba uzeti, pomnožimo (5) sa X_l^T i, uvaživši (3), dobijemo

$$X_l^T R X_l - \lambda_l X_l^T X_l = X_l^T R X_l - \lambda_l = O$$

tj.

$$X_l^T R X_l = K_l^T K_l \frac{}{} = \mathrm{var}\ (K_l) = \lambda_l \tag{6}$$

pa budući želimo da varijanca od K_I bude maksimalna uzimamo najveću svojstvenu vrijednost matrice korelacija R, i osim toga iz (6) vidimo da je varijanca prve komponente upravo jednaka najvećoj svojstvenoj vrijednosti matrice R.

Za drugu glavnu komponentu

$$K_2 = x_{21}v_1 + x_{22}v_2 + \ldots + x_{2n}x_n \quad (7)$$

tražimo takvu linearnu kombinaciju standardiziranih varijabli v_j iz Z, čija je varijanca veća od varijance ma koje druge linearne kombinacije varijabli, okomite na k_I i čiji koeficijenti zadovoljavaju uvjet

$$\sum_{j=1}^{n} x^2_{2j} = 1.$$

Označimo li sa $X_2^T = (x_{21}, x_2^2, \ldots, x_{2n})$, pišemo

$$\mathrm{var}\,(K_2) = \max$$

$$X_1^T X_2 = 0$$

$$X_2^T X_2 = 1$$

Analognom tehnikom kao i za prvu glavnu komponentu dobijemo da su koeficijenti u (7) jednaki elementima svojstvenog vektora pridruženog drugoj po veličini svojstvenoj vrijednosti matrice R.

Pokazuje se da općenito vrijedi (Morrison, 1967): m-ta glavna komponenta je linearna kombinacija varijabli

$$Kl = xl_1v_1 + xl_2v_2 + \ldots + l_nv_n,$$

čiji su koeficijenti elementi svojstvenog vektora matrice korelacija varijabli R pridruženog m-toj po veličini svojstvenoj vrijednosti λl. Ako je $\lambda_k = \lambda l$, koeficijenti k-te i l-te komponente su nužno ortogonalni; ako je $\eta_k = \eta l$, elementi se mogu izabrati tako da budu ortogonalni, iako takvih vektora ima beskonačno mnogo. Varijaca m-te komponente je λl, pa je ukupna varijanca

$$\lambda_1 + \lambda_2 + \ldots + \lambda_n = \mathrm{trag}\ R = n$$

Važnost m-te komponente u ovakvom jednostavnijem prikazu mjeri se sa $\lambda l/n$, pošto vrijedi $\lambda_1 > \lambda_2 > \ldots > \lambda_n$.

No, mi ne želimo zadržati puni skup komponenata, nego ćemo broj komponenata reducirati na $p<n$. Za određivanje broja značajnih komponenata p postoje mnogi kriteriji, a ovdje se koristi PB kriterij (Štalec, Momirović, 1971).

Označimo li sa $U^2 = (\mathrm{diag} R^{-1})^{-1}$ dijagonalnu matricu unikviteta, tada je

$$\mathrm{trag}\ (I - U^2) = \eta$$

zbroj image varijanci varijabli iz V.

Tada p mora zadovoljavati nejednakosti

$$\sum_{k=1}^{p-1} \lambda_k < \eta < \sum_{k=1}^{p} \lambda_k$$

tj. zadržimo toliko glavnih komponenti da zbroj njihovih varijanci bude veći ili jednak zajedničkom varijabilitetu svih varijabli iz V.

Na taj način dekomponirali smo trajektoriju promjena na p aditivnih, ortogonalnih komponenata promjena. U matričnom obliku to pišemo

$$K = Z\,X \quad (8)$$

gdje je

$K = (K_1, K_2, \ldots, K_p)$ K_i — glavne komponente

$X = (X_1, X_2, \ldots, X_p)$ X_i — svojstveni vektori

pridruženi zadržanim svojstvenim vrijednostima.

Da bismo mogli međusobno uspoređivati p zadržanih glavnih komponenata, standardiziramo ih na varijancu 1.

$$L = ZX\lambda^{-1/2}$$

gdje je Λ dijagonalna matrica p zadržanih svojstvenih vrijednosti matrice R.

Struktura komponenata promjena pomaže nam pri njihovoj interpretaciji, a definirana je korelacijama varijabli glavnih komponenata

$$H = Z^TL\frac{1}{n}\,Z^TZ\frac{1}{n}\,X\Lambda^{-1/2} = RX\Lambda^{-1/2} = X\Lambda^{1/2}$$

2.3. Analiza reziduala

Eckart-Youngova (1936) aproksimacija matrice Z je najbolja aproksimacija (pod kriterijem najmanjih kvadrata) te matrice matricom ranga p

$$Z^* = ZXX^T$$

U toj matrici su sadržane one informacije o promatranim promjenama koje su objašnjive sa p zadržanih glavnih komponenata. Onaj dio informacija koji se ne može objasniti zadržanim komponentama opisan je rezidualnom matricom standardiziranih podataka

$$Z = Z - Z^* = Z\,(I - XX^T)$$

Rezidualne korelacije varijabli iz V definirane su sa

$$R = R - H\,H^T,$$

a to su upravo oni dijelovi korelacija među varijablama koji se ne mogu objasniti sa p zadržanih komponenata promjena.

2.4. Relacije vremenskih točaka

Relacije vremenskih točaka definiraju se kao normirani skalarni produkti vektora vremenskih točaka u prostoru standardiziranih varijabli.

$$Q^* = Z\,Z^T$$

$$Q = (\mathrm{diag}\ Q^*)^{-1/2}\ Q^*\ (\mathrm{diag}\ Q^*)^{-1/2}$$

Teorijske relacije vremenskih točaka definirane su kao skalarni produkti vektora vremenskih točaka u prostoru standardiziranih komponenata promjena, normirani na istu metriku kao i relacije vremenskih točaka Q

$$Q^*_T = K\,K^T$$

$$Q_T = (\mathrm{diag}\ Q^*)^{-1/2}\ Q^*_T\ (\mathrm{diag}\ Q^*)^{-1/2},$$

a to je onaj dio relacija vremenskih točaka koji je objašnjiv sa p zadržanih glavnih komponenata.

Rezidualne relacije vremenskih točaka su razlike relacija vremenskih točaka Q i teorijskih relacija vremenskih točaka Q_T

$$Q_R = Q - Q_T$$

i to je onaj dio relacija vremenskih točaka koji se ne može objasniti sa p zadržanih komponenata promjena.

3. PROGRAM

Program INDIFF napisan je u meta-jeziku SS i implementiran na računalu UNIVAC 1100/U2 Sveučilišnog računskog centra u Zagrebu. U trenutnoj verziji (5.2/M) sistema SS njime je moguće analizirati promjene stanja objekta opisanog sa do 250 varijabli u ne više od 10.000 distinktnih vremenskih točaka. Podacima je potrebno pridružiti naredbe koje opisuju format podataka u skladu s uputama za sistem SS (jednu SEQUENCE naredbu i onoliko VARIABLE naredbi koliko ima varijabli).

INDIFF ne testira hipoteze o značajnosti komponenata, ali, ako su m i n dovoljno veliki brojevi, značajnost neke komponente K_i, $i=1, 2, \ldots, p$ može se testirati ovako:

1) neka je $e_i = m + n + 1 - 2_i$

2) neka je $e = (n - p)\ (m - p)$ gdje je p broj komponenata zadržanih po PB kriteriju.

3) neka je

$$f_i = \frac{\lambda_i e_v}{e_i\ (\sum_{j=p+1}^{n} \lambda_j)}$$

4) tada je f_i distribuiran u skladu sa Snedecorovom raspodjelom sa e, e_i stupnjeva slobode.

4. LITERATURA

1. Eckart, C. T. & Young, G.: The approximation of one matrix by another of lower rank. Psychometrika, 1936, 1, 211—218.
2. Momirović, K.: Metode za transformaciju i kondenzaciju kinezioloških informacija. Institut za kineziologiju Zagreb, 1972, 250—254.
3. Moore, E. H.: General analysis, Part I. Mem. Am. Phil. Soc. 1935, 1, 197.
4. Morrison, D. F.: Multivariate statistical methods. Mc Graw-Hill, New York, 1967.
5. Penrose, R.: A generalized inverse for matrices. Proc. Cambridge Phil, So., 1955, 51, 406—413.
6. Štalec, J., K. Momirović: Ukupna količina valjane varijance kao osnov kriterija za određivanje broja značajnih glavnih komponenata. Kineziologija, 1971, 1, 77—81.
7. Zakrajšek, E., J. Štalec i K. Momirović: SS-Programski sistem za multivarijantnu analizu podataka, I Međunarodni simpozij »Kompjuter na Seučilištu«, Zagreb 1974

INDIFF — THE MODEL, ALGORITHM AND PROGRAM FOR ANALYSIS OF CHANGES IN CONDITION OF AN OBJECT DESCRIBED OVER A GROUP OF QUANTITATIVE VARIABLES

The model and algorithm were construted and the program was written for the component analysis of changes in condition of individual objects. By decomposition of trajectories of changes into orthogonal components we reach an insight into the structure of the stohastic zrocess described by quantitative variables. Also, the residuals and the relation between time points were analysd.

Константин Момирович, Живан Караман

»ИНДИФ« — МОДЕЛЬ, АЛГОРИТМ И ПРОГРАММА ДЛЯ АНАЛИЗА ИЗМЕРЕННИЙ СОСТОЯНИЯ НЕКОТОРОГО ОБЪЕКТА, ОПИСАННОГО НАД МНОЖЕСТВОМ КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

Созданы модель и алгоратм и написана программа для компонентного анализа измерения состояний отдельных объектов. При помощи расчленения траектории изменений на ортогональные компоненты выявляется структура стохастического процесса, описанного при помощи количестванных переменных. Также проводится анализ результатов и взаимоотношений между различными точками во времени.