SMÍŠENÉ INTERAKČNÍ MODELY

Marta Horáková, ÚSEB ČSAV Brno

Předpokládejme, že máme analyzovat data zapsaná ve tvaru tabulky objekt-znak, která jsou realizací náhodného výběru obecně ze smíšeného rozložení pravděpodobností, a předpokládejme, že nás zajímá vzájemná struktura závislosti pozorovaných znaků - náhodných veličin.

Problematiku vhodné definice struktury závislosti takových dat a její ověření s použitím smíšených interakčních modelů budeme sledovat na příkladu souboru dat z literatury, která analyzoval např. Morricon a s použitím námi sledovaných postupů pak znovu Edwards (1987). Jedná se o data, která jsou výsledkem měření váhových ztrát po jednom a dvou týdnech od aplikace tří různých látek, z nichž každá byla podána čtyřem náhodně vybraným samcům a čtyřem samicím krys. Počet objektů je v tomto příkladě 24, sledované náhodné veličiny jsou: A = pohlaví, B = aplikovaná látka, X = váhová ztráta po 1 týdnu, Y = váhová ztráta po 2 týdnech.

Obecně budeme předpokládat, že máme výsledek pozorování p diskrétních a q spojitých náhodných veličin a že objekt je popsán p+q vektorem (i,y), kde $i = (i_1,\dots,i_p)$ jsou hodnoty diskrétních proměnných, $y = (y_1,\dots,y_q)$ jsou hodnoty spojitých proměnných. Množinu diskrétních náhodných veličin označíme D, množinu spojitých S.

Předpokládejme, že smíšené rozložení pravděpodobností, z něhož je prováděn náhodný výběr rozsahu N, je tzv. CG-rozložení (Conditional Gaussian) s hustotou

$$\log f(i,y) = \alpha_i + \beta_i.y - \frac{1}{2}.y^T.\Delta_i.y \quad ,$$

kde skalár α_i, qxl vektor $\beta_i = (\beta_i^1,\dots,\beta_i^q)^T$ a qxq symetrická pozitivně definitní matice koncentrace $\Delta_i = (\delta_i^{jk})_{\substack{j=1,\dots,q\\k=1,\dots,q}}$ jsou po řadě diskrétní, lineární a kvadratické kanonické parametry rozložení CG.

Název CG-rozložení vystihuje skutečnost, že podmíněné rozložení spojitých náhodných veličin při pevných hodnotách diskrétních veličin je q-rozměrné normální $N_q(\mu_i, \Sigma_i)$, kde $\mu_i = \Delta_i^{-1}.\beta_i$, $\Sigma_i = \Delta_i^{-1}$. Marginální rozdělení diskrétních náhodných veličin má pravděpodobnostní funkci

$$p_i = (2.\pi)^{q/2}.\det(\Delta_i)^{-1/2}.\exp\{\alpha_i + \frac{1}{2}.\beta_i^T.\Delta_i^{-1}.\beta_i\} \quad .$$

Parametry p_i, μ_i, Σ_i se nazývají přirozené, $m_i = p_i.N$ je očekávaná četnost třídy i odpovídající kontingenční tabulky z diskrétních náhodných veličin, μ_i je qxl vektor průměrů spojitých náhodných veličin a Σ_i je jejich qxq kovariační matice.

CG-rozložení pravděpodobností pak popisuje smíšený interakční model. Zadat smíšený interakční model tedy po stránce pravděpodobnostní znamená zadat diskrétní, lineární a kvadratické kanonické, resp. přirozené, parametry odpovídajícího CG-rozložení. V praxi se přitom obvykle omezují úvahy na hierarchické a zejména pak grafové smíšené interakční modely (Edwards 1986). Situace je analogická s problematikou analýzy vícerozměrných kontingenčních tabulek pro výlučně diskrétní proměnné, kterou popisuje ve sbornících ROBUST 82, 84, T. Havránek z SVT ČSAV , z jehož podnětu tento příspěvek vznikl.

Popis struktury závislosti je názornější, vyjdeme-li z kanonických parametrů. Zadat hierarchický interakční model znamená zadat generující třídu

hierarchického rozkladu pro diskrétní kanonické parametry α_i pro všechna i; pro každou složku β_i^Y, Y∈S, lineárních kanonických parametrů β_i generující třídu pro všechna i; pro každý prvek δ_i^{XY}, X,Y∈S, matice koncentrace Δ_i generující třídu hierarchického rozkladu pro všechna i. (Hierarchickým rozkladem přitom rozumíme situaci podobnou analýze rozptylu.)

V příkladě s krysami je D = {A,B}, S = {X,Y} a lze specifikovat například následující modely:

M_1:	parametr	hierarchický rozklad	generátor	generující třída
	α_i	$\lambda + \lambda_i^A + \lambda_i^B$	A,B	A,B
	β_i^X	$\xi + \xi_i^A + \xi_i^B$	A,B	AX,BX,AY
	β_i^Y	$\eta + \eta_i^A$	A	
	δ_i^{XX}	φ	1	
	δ_i^{YY}	ψ	1	X,Y
	δ_i^{XY}	0	0	

M_2:	parametr	hierarchický rozklad	generátor	generující třída
	α_i	$\lambda + \lambda_i^A + \lambda_i^B + \lambda_i^{AB}$	AB	AB
	β_i^X	$\xi + \xi_i^A$	A	AX,AY
	β_i^Y	$\eta + \eta_i^A$	A	
	δ_i^{XX}	$\varphi + \varphi_i^A$	A	
	δ_i^{YY}	$\psi + \psi_i^A$	A	AXX,AYY,AXY stručně: AXY
	δ_i^{XY}	$\tau + \tau_i^A$	A	

M_3:	parametr	hierarchický rozklad	generátor	generující třída
	α_i	$\lambda + \lambda_i^A + \lambda_i^B + \lambda_i^{AB}$	AB	AB
	β_i^X	$\xi + \xi_i^B$	B	BX,BY
	β_i^Y	$\eta + \eta_i^B$	B	
	δ_i^{XX}	$\varphi + \varphi_i^B$	B	
	δ_i^{YY}	ψ	1	BXX,YY,XY stručně: BX,XY
	δ_i^{XY}	τ	1	

M_4:	parametr	hierarchický rozklad	generátor	generující třída
	α_i	$\lambda + \lambda_i^A + \lambda_i^B + \lambda_i^{AB}$	AB	AB
	β_i^X	$\xi + \xi_i^B$	B	BX,BY
	β_i^Y	$\eta + \eta_i^B$	B	
	δ_i^{XX}	φ	1	
	δ_i^{YY}	ψ	1	XX,YY,XY stručně: XY
	δ_i^{XY}	τ	1	

Pro stručný zápis modelu Edwards (1986) definuje generující sentenci hierarchického smíšeného modelu jako zápis diskrétních, lineárních a kvadratických generujících tříd oddělených lomítkem. Generující sentence modelů M_1, M_2, M_3 a M_4 tedy jsou : M_1: A,B/AX,BX,AY/X,Y M_3: AB/BX,BY/BX,XY

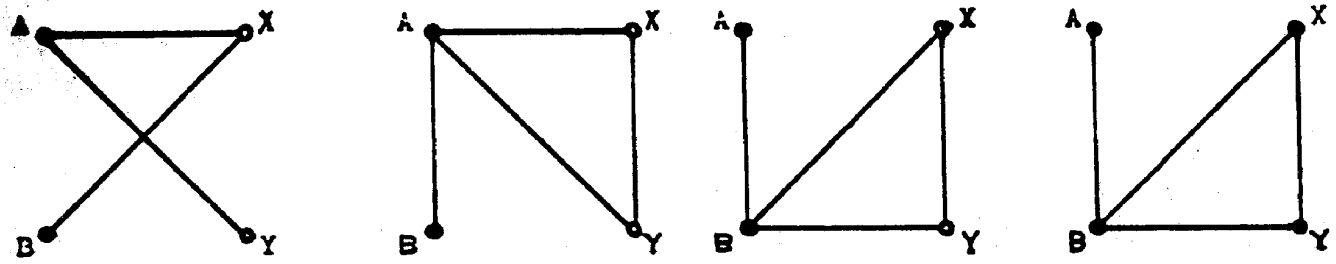
M_2: AB/AX,AY/AXY M_4: AB/BX,BY/XY

Edwards (1986) také uvádí požadavky, které musí parametrizace modelu splňovat:

1. Pouze nediagonální prvky matice Δ_i mohou být nulové a zbývající musí být alespoň konstantní.
2. Každý průnik lineárního generátoru s množinou diskrétních náhodných veličin D musí být obsažen v některém diskrétním generátoru.
3. Každý průnik kvadratického generátoru s množinou diskrétních náhodných veličin D musí být při sjednocení s každou spojitou náhodnou veličinou obsažen v lineárním generátoru.

Generující sentenci smíšeného interakčního modelu lze přiřadit její graf. Grafem generující sentence hierarchického smíšeného interakčního modelu rozumíme obyčejný graf s dvěma typy uzlů. Diskrétním náhodným veličinám odpovídají „diskrétní" uzly (značené obvykle • podle anglického dot - discrete), spojitým náhodným veličinám „spojité" uzly (značené obvykle ○ podle anglického circle - continuous). Dva uzly jsou pak spojeny hranou právě tehdy, když existuje generátor, který obě jim odpovídající náhodné veličiny obsahuje.

Grafy generujících sentencí modelů uvedených dříve jsou například:
M_1: A,B/AX,BX,AY/X,Y M_2: AB/AX,AY/AXY M_3: AB/BX,BY/BX,XY M_4: AB/BX,BY/XY



Nechť G = (V,E) je graf generující sentence hierarchického smíšeného modelu. Grafový (smíšený) model je pak právě takový model, jehož diskrétní generátory jsou právě kliky (= maximální úplné množiny uzlů) grafu G(D) = (V(D),E(D)), kde V(D) je množina všech diskrétních uzlů, množina hran E(D) = {e∈E: e⊆V(D) ; lineární generátory jsou právě kliky grafu G(D∪{Y}) = (V(D∪{Y}),E(D∪{Y})) pro každou spojitou veličinu Y S, kde V(D∪{Y}) = D∪{Y}, E(D∪{Y}) = {e∈E: e⊆V(D∪{Y}) ; kvadratické generátory jsou právě ty kliky grafu G, které obsahují aspoň jeden spojitý uzel.

Z modelů uvedených v příkladu s krysami jsou tedy grafové modely M_1, M_2, modely M_3 a M_4 grafové nejsou.

Z pravděpodobnostního hlediska jsou grafové právě ty smíšené modely, jejichž hustota pravděpodobností je markovská vzhledem ke grafu G; tj. dvě náhodné veličiny odpovídající nesousedním uzlům v grafu generující sentence jsou podmíněně nezávislé při daných hodnotách zbývajících proměnných. Interpretaci grafových modelů lze vyčíst přímo z grafu generující sentence: Jestliže a, b jsou množiny náhodných veličin, které jsou v grafu G odděleny množinou c, pak náhodné veličiny v a jsou podmíněně nezávislé s náhodnými veličinami v b při pevné hodnotě náhodných veličin v c.

Interpretace modelu M_2 například je: a = {B}, b = {X,Y}, c = {A}, tedy při daném pohlaví jsou aplikovaná látka a váhové ztráty po jednom týdnu spolu s váhovými ztrátami po dvou týdnech nezávislé náhodné veličiny.

Ucelený systém programů pro hierarchické smíšené interakční modely MIM (Mixed Interaction Models) nabízí Edwards (1987).

V systému MIM jsou náhodné veličiny označovány jedním písmenem A až Z. Diskrétní náhodné veličiny jsou spolu s počtem úrovní specifikovány příkazem jazy-

ka MIMu MIM -> FACTOR A 2 B 2 C 3 D 2 ,

spojité proměnné jsou specifikovány příkazem

MIM -> CONTINUOUS X Y Z W ,

model je zadáván generující sentencí modelu příkazem

MIM -> MODEL A,B/AX,AY/AXY ,
MIM -> MODEL ABC,BCD// ,
MIM -> MODEL // WXY,XYZ .

Užitím příkazů PLOT, DESCRIBE, PRINT je možno zobrazit graf generující sentence, vypsat určité vlastnosti běžného modelu a jeho generující sentenci.

Ověření adekvátnosti hierarchického smíšeného modelu v systému MIM provádí příkazy FIT, PRINT, TEST, BASE. Příkazem FIT se provede odhad parametrů pro daný model a data, spočte se odpovídající hodnota věrohodnostní funkce L_M pro daný model M a pro saturovaný model I a určí se hodnota testové statistiky

$$-2.\log Q((i,y)^{(1)},\ldots,(i,y)^{(N)}) = -2.(\log L_M - \log L_I).$$

Příkaz PRINT umožní vypsat podle volby odhady parametrů. Příkaz TEST je určen pro test daného modelu proti saturovanému modelu založený na poměru věrohodností a asymptotickým rozložením χ^2 s počtem stupňů volnosti rovnému rozdílu počtu volných parametrů v těchto modelech. Příkazem BASE lze určit jiný model než saturovaný, proti němuž má být model testován. Odhad parametrů je prováděn iteračním procesem, který může být řízen příkazy MAXCYCLES a CONVCRIT jazyka MIMu se zřejmým významem specifikování maximálního počtu cyklů a udáním požadované přesnosti konvergence.

Načtení vstupních dat v systému MIM je možné buď ve tvaru tabulky objekt-znak, nebo ve tvaru četností, průměrů a kovariančních matic (s koeficientem 1/n místo obvyklého 1/(n-1)). V prvním případě jsou po příkazu READ s udáním jmen načítaných proměnných zadávány jejich hodnoty postupně na jednotlivých objektech. Ve druhém případě se při výlučně diskrétních náhodných veličinách zadávají četnosti jednotlivých tříd kontingenční tabulky pro p diskrétních proměnných, při výlučně spojitých náhodných veličinách se zadává celkový počet objektů N, q výběrových průměrů pro q spojitých náhodných proměnných a q.(q+1)/2 odpovídajících kovariancí (s dělitelem n nikoli n-1), ve smíšeném případě se postupně pro každou třídu p-rozměrné kontingenční tabulky zadává četnost třídy a q výběrových průměrů spojitých náhodných veličin pro objekty z dané třídy a q.(q+1)/2 výběrových kovariancí spojitých znaků pro objekty z dané třídy. Užitím příkazu STATREAD se načítají údaje ve standardním uspořádání tříd (1,...,1),(1,...,1,2),....., (rozsah prvního diskrétního znaku,...,rozsah p-tého diskrétního znaku), užitím příkazu CELLREAD je nutno zadat třídu explicitně, a uspořádání se tedy nevyžaduje.

V příkladu s krysami Edwards (1987) postupně ověřoval následující modely s cílem najít model co nejjednodušší:

```
MIM -> MODEL AB/ABX,ABY/AXY
MIM -> FIT
DEVIANCE: 24.4014      DF: 12
CYCLE: 6
MIM -> TEST
TEST OF HO: AB/ABX,ABY/AXY
AGAINST THE FULL MODEL.
LR: 24.4014      DF: 12      P: 0.0178
```

```
MIM -> MODEL AB/ABX,ABY/BXY
MIM -> FIT
DEVIANCE: 15.7173      DF: 9
CYCLE: 6
MIM -> TEST
TEST OF HO: AB/ABX,ABY/BXY
AGAINST THE FULL MODEL.
LR: 15.7173      DF: 9      P: 0.0725
```

```
MIM -> BASE
MIM -> MODEL AB/ABX,ABY/BX,XY
MIM -> FIT
DEVIANCE: 21.4050      DF: 13
CYCLE: 5
MIM -> TEST
TEST OF HO: AB/ABX,ABY/BX,XY
AGAINST H:  AB/ABX,ABY/BXY
LR: 5.6877      DF: 4      P: 0.2225
```

```
MIM ->BASE
MIM -> MODEL AB/BX,BY/BX,XY
MIM -> FIT
DEVIANCE: 26.7960      DF: 19
CYCLE: 5
MIM -> TEST
TEST OF HO: AB/BX,BY/BX,XY
AGAINST H : AB/ABX,ABY/BX,XY
LR: 5.3911      DF: 6      P: 0.5041
```

```
MIM - BASE
MIM - MODEL AB/BX,Y/BX,XY
MIM - FIT
DEVIANCE: 51.4148      DF: 21
CYCLE: 3
MIM - TEST
TEST OF HO:  AB/BX,Y/BX,XY
AGAINST H:   AB/BX,BY/BX,XY
LR: 24.6188      DF: 2      P: 0.0000
```

Výsledkem Edwardsovy analýzy dat o krysách je tedy model s generující sentencí AB/BX,BY/BX,XY (zde označovaný jako model M_3). V Morrisonovi (1967) jsou tato data analyzována s použitím exaktních testů vícerozměrné analýzy s výsledkem, který se zde užívanou symbolikou odpovídá modelu M_4: AB/BX,BY/XY .

Systém MIM je psán v jazyce Pascal Turbo na IBM PC, vyžaduje alespoň 320 KB RAM a numerický koprocesor.

Literatura:

Edwards, D. (1986): Mixed interaction models.
Stat. Research Unit Univ. Copenhagen, Research Report.

Edwards, D. (1987): A guide to MIM.
Stat. Research Unit Univ. Copenhagen, Research Report.

Morrison, D.F. (1967): Multivariate Statistical Methods.
McGraw-Hill.

Havránek, T. (1982): O analýze mnohorozměrných kontingenčních tabulek.
ROBUST 82, JČSMF, Praha, str. 11-18.

Havránek, T. (1984): O logaritmicko-lineárních modelech pro mnohorozměrná kategoriální data. ROBUST 84, JČSMF, Praha, str. 31-41.