O LOGARITMICKO-LINEÁRNÍCH MODELECH PRO MNOHOROZMĚRNÁ KATEGORIAÁLNÍ DATA

Tomáš Havránek, Středisko biomatematiky Fyziologického ústavu ČSAV
142 20 Praha 4, Vídeňská 1083

Článek navazuje bezprostředně na příspěvek (Havránek,1982) z předchozí konference ROBUST. Je však jiného charakteru,neboť účelem je nyní dát přehled o některých nových výsledcích v oblasti vyhledávání vhodných modelů pro mnohorozměrná kategoriální data spolu s informací o některých teoretických výsledcích,které mohou mít širší uplatnění. Jde například o charakterizaci kolapsibility či charakterizaci modelů s odpověďmi.Článek obsahuje jisté drobné nové poznatky a návrhy týkající se právě spojení těchto dvou oblastí.Je vhodné upozornit,že se doplňuje s prací J.Michálka (1984) z tohoto sborníku.

1. ÚVOD

Zopakujeme jen velmi stručně některé základní pojmy.

Uvažujeme kategoriální veličiny (t.j. veličiny nabývající jen několika málo hodnot) .Veličiny budeme značit písmeny A,B,C,...,N. Předpokládáme,pokud nebude řečeno jinak,že veličiny,které zkoumáme,mají společné multinomické rozložení a že pozorovaná data vznikají realizací stejně rozložených a nezávislých veličin $(A,B,C,\ldots,N)_i$, i=1,...,m. Frekvenční tabulku,která takto vznikne,budeme značit $T_{ABC\ldots N}$. Je-li veličin n ,mluvíme o n dimensionální tabulce.Množinu veličin vytvářejících danou tabulku značíme $V_n=\{A,B,C,\ldots,N\}$ .Jednotlivé frekvence v tabulce značíme,například pro T_{ABC} jako m_{ijk}, obecně $m(\underline{i})$. Odpovídající pravděpodobnosti značíme $p(\underline{i})$.

Omezme se nyní na tři veličiny A,B a C.Pak při logaritmicko lineárním modelu se všemi efekty ,t.j.modelu <u>saturovanému</u> nebo <u>úplnému</u>,se předpokládá,že

$$\log p^{ABC}_{ijk} = \Theta + \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk} \qquad (1)$$

při obvyklých podmínkách zaručujících jednoznačnost. Hierarchické logaritmicko-lineární modely (<u>HLL</u> modely) vznikají vynecháním některých členů v (1) a to tak,že je-li (pro všechny hodnoty indexů) vynechán některý člen, je vynechán i člen vyššího řádu.T.t. vynechání λ^{AB} znamená vynechání λ^{ABC}. HLL modely se popisují pomocí <u>generujících tříd</u> (<u>výrazů</u>,<u>sentencí</u>). Pro model $\Theta + \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^{AC}_{ik} + \lambda^{BC}_{jk}$ je to (AC,BC). Mějme tabulku určité dimense n.Generují třída obecně má tvar $(a_1,\ldots,a_k)$,kde $a_1,\ldots,a_k$ jsou podmnožiny V_n takové,že pro i≠j není ani $a_i\subseteq a_j$ ani $a_j\subseteq a_i$. Přítomnost a v generující třídě určuje,že v modelu je člen λ^a i všechny členy λ^b pro b⊆a. Množiny a_i , i = 1,...,k nazýváme <u>generátory</u>. Píšeme ovšem (AB,BC) místo korektnějšího zápisu {{A,B},{B,C}} .

Pro generující výrazy máme definovány operace průseku a spojení ∧,∨ : $(a_1,\ldots,a_k)\wedge(b_1,\ldots,b_l) = (a_1\cap b_1, a_1\cap b_2,\ldots,a_k\cap b_l)$ při vynechání redundantních množin. Podobně $(a_1,\ldots,a_k)\vee(b_1,\ldots,b_l) = (a_1,\ldots,a_k,b_1,\ldots,b_l)$ .Jde o distributivní svaz. Operace ∧ odpovídá průniku či konjunkci modelů,proto běžně píšeme pro dva modely $\varphi \& \psi = (\underbrace{a_1,\ldots,a_k}_{\varphi})\&(\underbrace{b_1,\ldots,b_l}_{\psi}) = (a_1,\ldots,a_k)\wedge(b_1,\ldots,b_l)$

Maximálně věrohodné odhady frekvencí (resp.pravděpodobností) při daném modelu φ značíme $\hat{m}^{\varphi}(\underline{i})$ (resp. $\hat{p}^{\varphi}(\underline{i})$).Obecně je nutné tyto odahdy při HLL modelech stanovovat iteračním postupem (iterative proportional fitting ,viz (Prášková,1985)). Ve speciálním případě <u>rozložitelných</u> modelů, interpretovatelných v řeči podmíněných

nezávislostí (a ekvi-pravděpodobností), dostáváme odhady jako součiny a podíly marginálních frekvencí v "uzavřené" formě bez iterací.

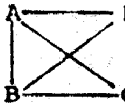
Je-li $a \subseteq V_n$, pak marginální tabulka T_a je tabulka obsahující frekvence $m(i_a) = \sum_{i \wedge a^c} m(i)$, t.j. sčítá se přes hodnoty indexů odpovídajících veličinám z $a^c = v_n - a$. Například pro model (AB,BC) potřebujeme marginální tabulky T_{AB} a T_{BC}. Odhad je pak $\hat{p}_{ijk} = (m_{ij.} . m_{.jk}) / (m_{.j.} \; m)$ v běžné notaci $(m_{ij.} = \sum_k m_{ijk})$.
V (Havránek,1982b) byly rozložitelné modely nazývány multiplikativní; zdá se, že nyní se ustálil termín rozložitelné.

2. GRAFOVÉ MODELY

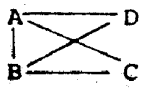
Uvažujme všechny HLL modely dané dimense n. Zajímavou třídou modelů je nejmenší třída modelů obsahující rozložitelné modely a uzavřená vůči konjunkci modelů; označme si ji H_1. Tato třída je charakterizována v (Havránek,1982a) jako třída modelů generovaných průniky (elementárních) modelů nulové parciální asociace (ZPA modelů). Model nulové parciální asociace je model typu (ACDE,BCDE) pro n=4, t.j. model podmíněné nezávislosti A a B vzhledem k ostatním veličinám.

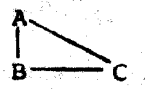
Generující třídu $(a_1,\ldots,a_k)$ můžeme zobrazit na necrientovaný graf s množinou uzlů V_n následujícím způsobem: $i((a_1,\ldots,a_k)) = (V_n,E)$, kde $E = \{(x,y); (x,y) \in V_n \times V_n$ a xy je obsaženo v některém $a_i\}$.

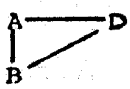
Modelu (ACDE,BCDE) tedy odpovídá graf 

t.j. úplný graf nad V_n bez jedné hrany AB.
Modelu (ABC,AD,BD) pro n=4 odpovídá graf: 

. Tentýž graf však odpovídá i rozložitelnému modelu (ABC,ABD) podmíněné nezávislosti C a D. Vidíme, že zobrazení z množiny generujících sentencí do množiny grafů není vzájemně jednoznačné.

Uvažme nyní zobrazení j zobrazující grafy s n uzly $\{A,\ldots,N\}$ do množiny generujících sentencí. Je-li G nyní graf, pak j(G) je definováno jako $(a_1,\ldots,a_k)$, kde $a_1,\ldots,a_k$ jsou množiny uzlů klik grafu G (klika je maximální souvislý podgraf). Pro 

máme dvě kliky 

a 

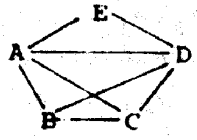
a tedy j(G) = (ABC,ABD) .

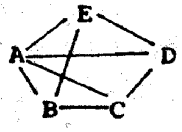
Důležité je nyní toto: i (H_1) je množina všech grafů (s n uzly) a zároveň $(i \wedge H_1)^{-1} = j$. Při restrikci i na H_1 dostáváme vzájemně jednoznačné zobrazení H_1 na množinu všech grafů s n uzly. Vzhledem k této vlastnosti je plně oprávněný název, který pro třídu H_1 použili Darroch, Lauritzen a Speed (1980), t.j. třída grafových modelů.

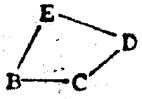
3. CHARAKTERIZACE ROZLOŽITELNOSTI

Pro práci s grafovými modely je výhodné používat techniky teorie grafů. V tomto jazyce také charakterizovali Darroch, Lauritzen a Speed (1980) rozložitelné modely:

grafový model je rozložitelný, neobsahuje-li jeho grafová representace kružnici délky větší než tři jako podgraf (t.j. je to triangulovaný graf; viz Neštřil,1979).

Příklad (n=5):
(ABCD,ADE) 

je rozložitelný, (ABC,ACD,ABE,ADE) 

není rozložitelný, protože podgraf 

je kružnicí délky čtyři. Pro rozpoznání triangulovaného grafu existuje algoritmus (eliminování izolovaných uzlů, Golumbik,1980), odpovídající algoritmu pro rozpoznávání rozložitelných modelů uvedenému např. v (Havránek,1982b). To, že rozložitelné modely mají triangulované grafy odpovídá to-

mu,že jde o jednoduché modely;triangulované grafy jsou v jistém smyslu rovněž jednoduché - jejich klikovost se rovná jejich barevnosti (viz Neštřil,1979).

Platí důležitý výsledek poprvé explicitně publikovaný a grafově dokázaný Edwardsem (1984) :

Máme-li nesaturovaný rozložitelný model φ ,pak existuje rozložitelný model ψ lišící se od φ pouze přidáním jediné hrany. (A)

V jiné formulaci to znamená,že je-li φ daný rozložitelný model,pak postupným přidáváním hran (a to takovým,že výsledkem je opět rozložitelný model) se dostaneme k (elementárnímu) ZPA modelu . Tento výsledek je v poněkud zašifrované podobě obsažen v práci Sundbergově (1975) ; pak patřil léta k folkloru mezi badateli v dané oblasti.Důvodem ne zcela jasných formulací i ne všeobecného pochopení různých výsledků je zde patrně stále nejednotný jazyk.

Z (Havránek,1982a,b) víme,že grafový model φ lze psát jako konjunkci $\varphi_1 \& \ldots \& \varphi_k$ elementárních ZPA modelů. Modely $\varphi_1, \ldots, \varphi_k$ jsou samozřejmě rozložitelné. Platí : φ je rozložitelný právě když lze členy v konjunkci uspořádat tak, že $\varphi_1 \& \varphi_2 \& \ldots \& \varphi_j$ je rozložitelný pro každé $j = 1, \ldots, k-1$. Přeloženo zpět do grafové řeči: rozložitelný model φ dostaneme tak,že postupně po jedné ubíráme hrany. Každý model takto postupně získaný musí být rozložitelný.Navíc: vyjdeme-li z elementárních ZPA modelů a zkoumáme postupně pouze všechny rozložitelné modely vzniklé postupným odebíráním hran,nemůžeme žádný rozložitelný model pominout.Na této skutečnosti je založena procedura pro vyhledávání rozložitelných modelů navržená Edwardsem (1984) .

4. KOLAPSIBILITA

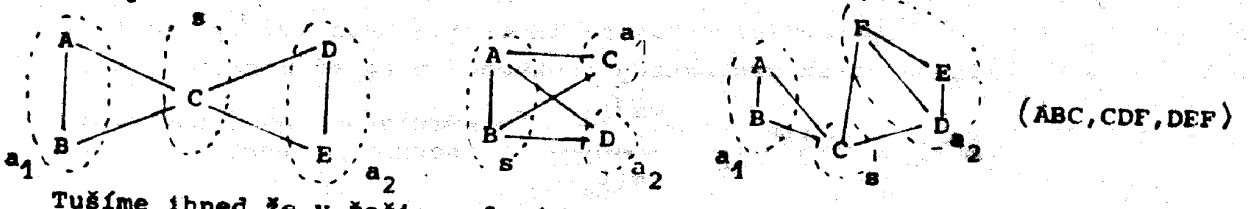
Velmi důležitou věcí pro analýzu mnohorozměrných dat je kolapsibilita.Jde o to, že za jistých okolností můžeme zkoumat vztahy některých veličin bez ohledu na ostatní veličiny,t.j. v marginálním rozložení příslušných veličin či v marginální tabulce.

Platí-li model (ABCD,EF) při $n=6$, můžeme vztahy veličin A,B,C a D zkoumat v marginální tabulce vzniklé sečtením přes indexy odpovídající E a F ,t.j. $m_{ijkl..}$. Problém je zde ovšem zpravidla ten,že my nevíme zda (ABCD,EF)platí a přijetí takového modelu nezávislosti na základě testování jde silně proti stávající statistické filosofii (a konstrukci testů). Pomiňme ale nyní tento problém.Položme si jinou otázku: kdy můžeme zkoumat vztahy některých veličin bez ohledu na jiné veličiny,resp. kdy se můžeme omezovat na marginální tabulky.Pro grafové i obecné HLL modely řeší tento problém dosti obecně Assmusen a Edwards (1983).

Nechť nyní $a \subseteq V_n$, a je tedy množina veličin (resp. vrcholů) . Pro daný grafový model φ definujeme jeho restrikci na a značenou φ_a jako model odpovídající grafu vzniklému vynecháním všech uzlů,které nejsou v a (a všech hran vedoucích do těchto uzlů). Pro obecné HLL modely lze definovat restrikci zcela obdobně. Nechť nyní $\underline{i}_a$ je restrikce indexu do množiny a (t.j. výše ijkl) a $\hat{p}$ odhad při platnosti φ, $\hat{p}_a$ při předpokladu platnosti φ_a. Řekneme,že φ je kolapsibilní na φ_a, je-li pro každé $\underline{i}_a$, $\hat{p}(\underline{i}_a) = \hat{p}_a(\underline{i}_a)$.

Vraťme se na chvíli ke grafům. Řekneme,že množina uzlů s odděluje množiny uzlů a_1 a a_2 ,jestliže každá cesta mezi a_1 a a_2 musí vést přes uzly v s.

Příklady:



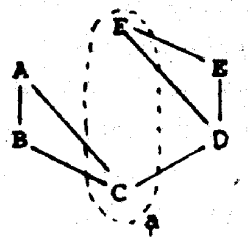
Tušíme ihned,že v řeči grafových modelů:

a_1 a a_2 jsou podmíněně nezávislé vzhledem k s (značíme $a_1 \perp a_2/s$) právě když je s odděluje.

A nyní důležitý výsledek:

Grafový model φ je kolapsibilní na φ_a právě když pro každé a_1, a_2 a s takové, že $a_1, a_2 \subseteq a$, podmíněná nezávislost a_1 a a_2 vzhledem k s implikuje podmíněnou nezávislost a_1 a a_2 vzhledem k $a \cap s$.

Například $\varphi_1 = (ABC, CDF, DEF)$ je kolapsibilní pro $a = \{C,F\}$, neboť C a F nejsou podmíněně nezávislé vzhledem k žádné množině s. Uvažme $\varphi_2 = (ABC, CD, DEF)$ pro $a = \{C,F\}$. C a F jsou podmíněně nezávislé (separovány D), ale nejsou nezávislé, t.j. podmíněně nezávislé v $a \cap s = \{C,F\} \cap \{D\} = \emptyset$ a tedy nelze kolapsovat na a. Uvažme $a = \{C,D,F\}$ - pak $a \cap s = \{D\}$ a je vše v pořádku; model je kolapsibilní na $\{C,D,F\}$.



Co to znamená pro testování ? Chceme-li v rámci modelu φ testovat podmíněnou nezávislost a_1 a a_2 vzhledem k s a φ i výsledný jednodušší model $\varphi \& (a_1 \perp a_2 / s)$ jsou kolapsibilní pro a obsahující a_1 i a_2, je rozumné testovat podmíněnou nezávislost a_1 a a_2 vzhledem k $a \cap s$ bez ohledu na použitý test. Vzhledem kolapsibilitě, zamítnutí $a_1 \perp a_2 / a \cap s$ znamená zamítnutí i $a_1 \cap a_2 / s$. Zřejmě nezamítnutí $a_1 \perp a_2 / a \cap s$ znamená nezamítnutí $a_1 \perp a_2 / s$. Otázka je ale, kdy můžeme $a_1 \perp a_2 / a \cap s$ testovat v marginální tabulce T_a. Pro obvyklé testové statistitiky (poměr věrohodnosti, chi-kvadrát, exaktní test) je to možné (viz pro první dva případy Assmusen a Edwards, 1983, pro poslední Sundberg, 1975) díky faktorizaci pravděpodobnosti na členy týkající se a a na ostatní členy, které jsou pro oba modely shodné, a které se v testových statistikách vykrátí či sečtou; stupně volnosti jsou rozdílem stupňů volnosti obou modelů a nedělají tedy potíže. Bylo by možné obecně definovat třídu statistik, které mají tuto vlastnost, v teoretické rovině jistě intuitivně oprávněnou. Z různých důvodů se zdá, že by bylo vhodné ještě zkoumat nové testové statisky či rozhodovací pravidla pro analýzu mnohorozměrných kategoriálních data (viz zde §6,7). Testování v marginálních tabulkách má velké výhody, zejména díky vyšším teoretickým i pozorovaným frekvencím a tím lepší asymptotice. Výpočetní otázky zde též nejsou zanedbatelné.

Kolapsibilitu lze definovat i obecně pro HLL modely v řeči generujících tříd. Formulace použité v práci Assmusena a Edwardse (1983) nejsou příliš průhledné a bylo by potřebné ještě vhodný jazyk hlouběji promyslet. Poznamenejme, že podmínka kolapsibility uvedená zde pro grafové modely je obecně podmínkou nutnou, ale ne postačující.

5. ODPOVĚDI A VYSVĚTLUJÍCÍ VELIČINY

Podívejme se nyní na situaci, kdy některé veličiny, řekněme z množiny $a \subseteq V_n$ jsou vysvětlující (nezávislé) a jiné z $b = V_n - a$ jsou odpověďmi (závislé veličiny).

Vhodný obecný model pro tuto situaci je faktorizovaná pravděpodobnost

$$p^{\chi}(\underline{i}) = p^{\varphi}(\underline{i}_a)\; p^{\psi}(\underline{i}_b / \underline{i}_a) \qquad (2)$$

kde $p^{\varphi}(\underline{i}_a)$ je marginální pravděpodobnost pro veličiny z a při nějakém HLL modelu φ a p^{ψ} je podmíněná pravděpodobnost b při daných hodnotách a daná HLL modelem pro $V_n = a \cup b$ obsahujícím všechny interakce mezi veličinami v a. Roli modelů φ a ψ ukazuje vztah pro odhady frekvencí v celém modelu, který je zde

$$\hat{m}^{\chi}(\underline{i}) = \hat{m}^{\varphi}(\underline{i}_a)\; \frac{\hat{m}^{\psi}(\underline{i})}{m(\underline{i}_a)}$$

$\hat{m}^{\psi}(\underline{i})$ ← sdružená pravděpodobnost v modelu, všechny interakce z a

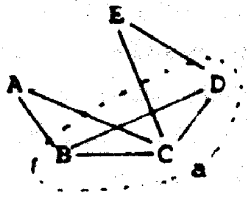
Otázkou je, kdy $p^{\chi}(\underline{i})$ lze vyjádřit přímo HLL modelem a obráceně, kdy HLL model pro celé V_n má tvar odpovídající (2).

Odpověď na obě otázky je spojená s kolapsibilitou:

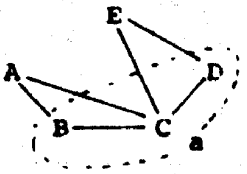
HLL model $\mathcal{X}$ je vyjádřitelný ve tvaru (2) právě když je kolapsibilní vzhledem k a. (B)

Pak φ je dáno $\mathcal{X}_a$ a ψ je dáno jako sjednocení saturovaného modelu pro a a $\mathcal{X}_{cl(b)}$, kde cl(b) je uzávěr b v $\mathcal{X}$, t.j. cl(b) obsahuje b a všechny písmena (uzly) z a, které se vyskytují v generátorech obsahujících veličiny z b.

Pro grafové modely je zde interpretace zcela jasná a snadná. Příklad:



t.j. (ABC,BCD,CDE), kde b={A,E} je odpovědi, a a={B,C,D} vysvětlující veličiny. Model je triválně kolapsibilní na a. Jde o model tvaru (2), kde φ je (BCD) a ψ je (BCD) $\cup \mathcal{X}_{cl(b)}$ = (BCD) $\cup$ (ABCD,BCD,CDE) = (ABC,BCD,CDE), neboť cl(b) = V_n. Podobně model (ABC,CDE) je typu (2) s φ=(BC,CD) a ψ = (BCD) $\cup$ (ABC,CDE) = (ABC,BCD,CDE).

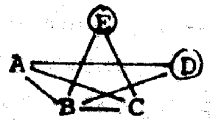


Opačnou otázkou je, kdy je model tvaru (2) HLL modelem. Tvrzení vyslovíme pro speciální případ, tj. pro grafové modely. Nechť V_n = a $\cup$ b. Model $\mathcal{X}$ je dán dvěma modely φ a ψ. Uvažujeme nyní případ, kdy φ i ψ jsou grafové a jsou tedy popsány grafy nad a resp. a $\cup$ b. Množina uzlů b definuje podgraf ψ_b grafu ψ. Tento podgraf se může sestávat s několika komponent (maximálních souvislých částí grafu, viz 2.4.6, Nešetřil, 1979). Hranice komponenty grafu ψ_b v celém grafu ψ je pak definována jako množina těch uzlů grafu ψ, které nejsou v komponentě, ale jsou spojeny s některým uzlem komponenty hranou. Nyní již můžeme vyslovit tvrzení:

Model $\mathcal{X}$ = (φ, ψ) tvaru (2) kde φ a ψ jsou grafové, je grafový právě tehdy, jestliže hranice každé komponenty grafu ψ_b v grafu ψ (C) je obsažena v klice grafu φ.

Výsledný grafový model je pak dán spojením φ a <u>všech</u> $\psi_{cl(b_i)}$, kde b_i je množina uzlů i-té komponenty grafu ψ_b (pozor, zde je v článku Assmusena a Edwardse, 1983 nepřesnost). Ve zmíněném článku je tvrzení (3.2) vysloveno pro HLL modely. To, že z grafovosti φ a ψ plyne grafovost $\mathcal{X}$ je dokázáno v dodatku.

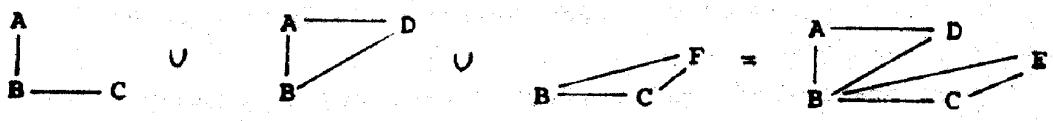
Příklad: Nechť a = {A,B,C} a b = {D,E}. Dílčí modely jsou φ = (AB,BC) a ψ = (ABC,ABD,BCE). ψ má graf a ψ_b je podgraf



Ⓔ Ⓓ

mající dvě komponenty E a D. Hranice komponenty E v ψ je {B,C} a je obsažena v klice BC. Hranice komponenty D je {A,B} a je obsažena v klice AB. Výsledný model je grafový. Pro jeho konstrukci použijeme cl(E) = {B,C,E} a cl(D) = {A,B,D}. Dostáváme spojení tří grafů

(AB,BC) $\cup$ (ABD) $\cup$ (BCE) = (ABD,BCE)



Můžeme ověřit zpětně, že (ABD,BCE) generuje podle (B) z tohoto odstavce ψ = (ABC,ABD,BCE) a φ = (AB,BC) : (ABD,BCE) je kolapsibilní na a = {A,B,C}. Omezení (ABD,BCE) na a dává φ = AB,BC a cl(b) = {A,B,C,D,E}, pro ψ tedy spojíme (ABD,BCE) a (ABC).

6. HLEDÁNÍ GRAFOVÝCH MODELŮ I

Máme-li nyní tabulku zjištěných frekvencí (pro multiresponsní případ) a nemáme předem představu o vhodném modelu struktury závislostí mezi veličinami, je nutné na

základě dat vhodný model navrhnout. Víme, že modelů může být mnoho, například pro $n = 4$ máme 113 grafových (a 110 rozložitelných) modelů, pro $n=5$ je jich již 1450 (resp. 1233). Není tedy vhodné všechny modely testovat a zkoumat jejich shodu s daty, ale je nutné navrhnout způsob, jak rychle nalézt třídu grafových modelů přípustných vzhledem k daným datům. Z této třídy pak můžeme navrhovat další modely pro zkoumání jak z hlediska věcné oblasti aplikací, tak z hlediska shody s dalšími daty.

Pro určení takové množiny modelů můžeme vycházet z toho, že přípustné jsou modely nezamítnuté některou testovou procedurou (zpravidla pro vyšší α). Protože však nechceme z časových důvodů testovat explicitně všechny modely, musíme hledat nějaký postup, který nám umožní třídu přípustných modelů jistým způsobem aproximovat.

Jsou zde možné různé přístupy. Jeden z nich je uplatněn v pracích (Havránek, 1982b a 1984). Nejprve zopakujme, že modely jsou uspořádány relací "být podmodelem": $\varphi \leq \psi$ jestliže graf modelu φ je podgrafem modelu ψ. Můžeme říkat, že model φ je jednodušší než model ψ.

Diskutovaný přístup vychází z toho, že:

(i) postup vyhledávání modelu by měl být nezávislý na použité testové statistice či míře shody modelu s daty,

(ii) zamítneme-li model φ, musíme zamítnout všechny modely jednodušší, t.j. všechny modely ψ takové, že $\psi \leq \varphi$.

Postup je navíc postupem shora dolů; začíná u nejsložitějších modelů, t.j. elementárních ZPA modelů. Ostatní grafové modely jsou pak vytvářeny jejich konjunkcemi.

Proč je zde použit (i když v publikacích implicitně) požadavek (i)? Není zcela zřejmé, který z testů je nejvhodnější; jde o testy asymptotické, různé síly, různého chování v řídkých tabulkách atd. Postup by se neměl měnit, přejdeme-li například k nějakému novému testu o výhodnějších vlastnostech (viz ještě dále). K principu (ii): na teoretické úrovni je jisté, že platí-li model φ musí platit i ψ pro $\varphi \leq \psi$, t.j. nezamítnoutí φ by mělo vést k nezamítnutí ψ. Při daném postupu ale uplatňujeme jen "negativní" důsledky, t.j. zamítnutí má za následek zamítnutí. To souvisí s klasickou koncepcí testování, kdy zamítnutí je považováno vlastně za jediný výsledek testu, na kterém lze dále stavět.

Je zde možné odlišovat modely zamítnuté pomocí testu v datech a modely zamítnuté na základě výše zmíněné dedukce (ii) bez přímého testování v datech. Těmto druhým lze říkat v souladu s (Edwards a Havránek, 1984) slabě zamítnuté.

Při analýze konkrétních dat a při použití daného testu je pak třída přijatelných modelů definována jako třída nezamítnutých ani slabě nezamítnutých modelů. Může být charakterizována svými minimálními (nejjednodušími) prvky vzhledem k uspořádání $\leq$. Podotkněme, že v pozadí celého postupu stojí idea mít celou třídu přijatelných modelů k dispozici, například pro další počítačovou analýzu pomocí jiné míry shody než je použitý test (uspořádání podle shody). Jde ostatně o aplikaci přístupu raženého v knize (Hájek a Havránek, 1978).

Víme, že ani testová statistika poměru věrohodnosti ani Pearsonův chí-kvadrát nejsou dobré míry shody a pomocí nich nelze modely, které nejsou v relaci $\leq$ srovnávat; v (Havránek, 1982b, 1984) jsou uvažovány dosažené hladiny významnosti a AIC kriterium. Poznamenejme, že zde by bylo vhodné hledat takový test, či obecněji rozhodovací pravidlo, který by měl tu vlastnost, že by zamítnutí i slabé zamítnutí splývalo. V příkladech zpracovávaných navrženou procedurou pomocí testu poměrem věrohodnosti se vyskytly (pro $\alpha = 0.1$) případy, že model byl slabě zamítnut, ale pro kontrolu vypočtená dosažená hladina významnosti byla větší než 0.1, ale ne příliš. Například konkrétně bylo pět nejvyšších dosažených hladin významnosti 0.1430, 0.1420, 0.1286, 0.1058, 0.1043.

7. HLEDÁNÍ GRAFOVÝCH MODELŮ II

V (Edwards a Havránek,1984) je použit kromě principů (i) a (ii) ještě princip:

(iii) Je-li některý model φ přijatelný,jsou přijatelné i všechny modely složitější,t.j. takové modely ψ , že $\varphi \leqq \psi$.

Používá se název slabě přijatelný model pro model,který nebyl explicitně testován v datech,ale byl uznán za přijatelný na základě principu (iii) uplatněného na nějaký model φ explicitně v datech testovaný a nezamítnutý.Procedura je navíc zobecněna proti předchozí tak,že je možné začít s libovolnou množinou modelů S,které jsou vzájemně neporovnatelné vzhledem k $\leqq$. Tyto modely se otestují a tím rozdělí na zamítnuté (množina R) a na nezamítnuté (množina A).Dále se analyzuje H_1-S.Hledají se $\leqq$ - minimální modely $D_a(R)$,které nejsou slabě zamítnuté na základě R nebo $\leqq$ - maximální modely ,které nejsou slabě přijatelné na základě A (množina $D_r(A)$).Tyto modely se nazývají minimální možné modely resp. maximální zamítnutelné modely vzhledem k R resp. A . Modely z $D_a(R)$nebo modely z $D_r(A)$ jsou pak explicitně testovány a výsledek pak rozšíří A a R. Celý postup je pak iterován.

Je možné postupovat zhora dolů - používat $D_r(A)$- nebo zdola nahoru - používat $D_a(R)$, případně střídavě. Výsledkem jsou v každém případě třídy modelů přijatelných A a zamítnutých R takové, že každý další model je pak buď slabě přijatelný nebo slabě zamítnutý. Stačí ovšem použít ještě více kondenzovanou informaci,t.j. ukládat pouze minimální modely z A a maximální modely z R.

Praktická realizace je ovšem složitější, je třeba mít například postup na hledání $D_r(A)$ atd.

Jako výsledek dostáváme tedy dvě množiny $\leqq$ - nesrovnatelných modelů (explicitně testovaných v datech) ,které klasifikují všechny ostatní modely do dvou tříd.

Dvě poznámky: (a) startujeme-li tuto proceduru II (resp. proceduru I)s počáteční množinou elementárních ZPA modelů, pohybujeme se velmi často v oblasti velmi špatné aplikability asymptotických testů.Navíc zamítnutí elementárních modelů vede k velké redukci množiny dále uvažovaných modelů (možných modelů) . Proto je žádoucí zde použít exaktní test (blíže viz zde §9) .

(b) Ideální test pro tuto situaci by měl být takový,že by slabé zamítnutí a slabé přijetí splývalo se zamítnutím i přijetím.Pak by inference na teoretické úrovni přesně odpovídala chování testů na datech (dedukci na datové úrovni) .

Procedura II se střídáním kroků dolů a nahoru a při použití počáteční množiny elementárních ZPA modelů byla aplikována na příklad použitý v(Havránek,1983) pro analýzu šesti (n=6) rizikových faktorů ischemické choroby srdeční.Je zde 15 elementárních grafových modelů a 32753 neelementárních grafových modelů.V citované práci byla použita procedura I s dalším trikem,který odpovídá jednomu kroku procedury II. Přesto bylo nutné explicitně testovat více než 100 modelů.Procedura II analyzovala tato data při 27 explicitně testovaných modelech (z toho 15 elementárních,které by bylo možné testovat exaktně) .Celkem 32000 je zamítnutých (15) a slabě zamítnutých, a 768 modelů je přijatelných (12) a slabě přijatelných . Množina minimálních modelů z A obsahuje jen dva modely a to modely stejné jako modely nalezené jako minimální procedurou I. Celý příklad viz (Edwards a Havránek,1984) .Pro výpočet byl použit program GRAPH napsaný Edwardsem v jazyce Pascal.

8. ODPOVĚDI A VYSVĚTLUJÍCÍ VELIČINY II

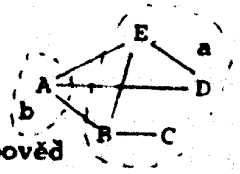
Vraťme se nyní k situaci,kdy některé veličiny jsou odpověďmi a jiné veličinami vysvětlujícími.Výše uvedené procedury je možné adaptovat pro tento případ dvojím způsobem:

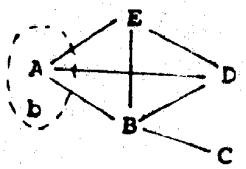
(a) Hledat modely (grafové) bez omezení procedurou I nebo II . Po nalezení minimálních přijatelných modelů aplikovat na tyto modely tvrzení (B) z §5 a rozhodnout te-

dy,zda mají vhodnou strukturu.Modely,které nemají tuto strukturu,vyloučit a hledat další minimální modely atd.

Podívejme se opět na příklad z (Havránek,1982b) . Zde jako výsledek procedury I máme jeden ze dvou grafových minimálních modelů 2560:(ABE,ADE,BC) .Tento model není kolapsibilní na a neboť BC⊥D/AE ale ne BC⊥D/E. Model tedy nepoužijeme jako model pro situaci odpověď-podnět.Vyřadíme-li tento model,je jeden z minimálních modelů model 250: (BC,ABDE) ,který je kolapsibilní na $a=\{B,C,D,E\}$. Model φ (pro nezávislé veličiny)je zde χ_a^{-}(BC,BDE), model ψ(podmíněný pro odpověď)je dán (BCDE)$\cup\chi_{cl(b)}$. Uzavěr b je zde$\{A,B,D,E\}$ a tedy ψ je rovno elementárnímu modelu 10:(ABDE,BCDE) (srovnej dále) .

odpověď

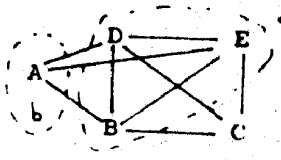




(b) krok 1: Hledat modely pro marginální tabulku definovanou vysvětlujícími veličinami T_a bez omezení. krok 2: Hledat modely pro celou tabulku s omezením,že tyto modely musí obsahovat úplný podgraf G_a se všemi uzly z množiny a. Výsledné modely pak kombinovat do tvaru (2) z §5 (a případně posuzovat podle tvrzení (C) z §5).

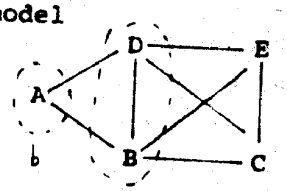
Pro proceduru I je možné realizovat krok 2 velmi snadno.Víme,že zde jsou modely postupně generovány jako konjunkce elementárních grafových (ZPA) modelů.Každý z těchto modelů odpovídá nepřítomnosti jedné hrany.Jde tedy o to generovat modely $\varphi_1\&\ldots\&\varphi_k$ takové,že žádný elementární model φ_i neodpovídá vynechání hrany z G_a.

Pro příklad z (Havránek,1982b),kde $a=\{B,C,D,E\}$, to znamená netestovat elementární modely 1.(ABCD,ABCE), 2.(ABCD,ABDE), 3.(ABCD,ACDE), 5.(ABCE,ABDE), 6.(ABCE,ACDE), 8.(ABDE,ACDE). K testování zbývají čtyři elementární modely 4.(ABCD,BCDE), 7.(ABCE, BCDE), 9.(ACDE,BCDE) (pozor,zde v Havránek,1982b chyba) a 10.(ABDE,BCDE). Modely 4,7 a 9 jsou zamítnuty; pro další postup nám už žádná konjunkce nezbývá.Zde tedy procedura končí triviálně.Dostáváme přijatelný model (ABDE,BCDE).



← separuje A a C; pouze zprostředkované působení C na A

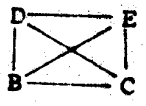
Kdybychom použili jiné α ,např.$\alpha=0.05$, mohou být zamítnuty např. pouze modely 7 a 9. Pro druhý krok nám zbývá jediná konjunkce 4 & 10, tedy model (ABCD,BCDE)&(ABDE,BCDE)=(ABD,BCDE), který dejme tomu na hladině $\alpha=0.05$ opět nezamítáme a máme tak minimální přijatelný model



separuje A od E,C

zprostředkovaného působení E a C na A. Můžeme pak realizovat krok 1. V marginální tabulce dané $a=\{B,C,D,E\}$bychom mohli hledat přijatelné modely (viz (2)v §5) , t.j. vynechávat hrany z úplného grafu

Dimense tabulky je zde menší a celý postup lépe realizovatelný (můžeme použít proceduru I i II) .



Pro krok 2 je asi procedura I s výše uvedenou modifikací vhodná,neboť požadavek inkludování úplného podgrafu G_a drasticky zmenšuje prostor možných modelů.Procedura II by mohla být rovněž odpovídajícím způsobem modifikována.

Celý postup (b) umožňuje nalézt všechny přijatelné či slabě přijatelné modely tvaru (2)v §5 ,jejichž složky jsou grafové. Celkový model ovšem nemusí být vždy ani HLL.Postup (b) je z počítačového a pravděpodobně i věcného hlediska výhodnější než postup (a). Postup (a) má své oprávnění,kdy teprve po analýze zjistíme,že jde o strukturu podnět-odpověď.

9. HLEDÁNÍ GRAFOVÝCH MODELŮ III

Často se setkáváme se situací,kdy tabulka je velmi řídká - mnoho nulových či ma-

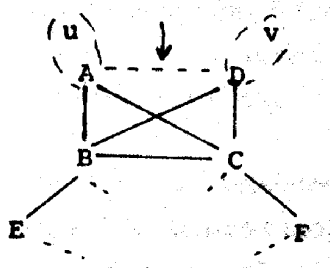
lých i marginálních frekvencí.Postup pro vyhledávání rozložitelných modelů v této situaci navrhuje Edwards (1984) .

Máme-li tři veličiny A,B,C a chceme-li testovat podmíněnou nezávislost A⊥B/C můžeme použít exaktní test založený na generování tabulek.Na tabulku T_{ABC} se můžeme dívat jako na tabulku skládající se z vrstev odpovídajících každé hodnotě veličiny C. V každé vrstvě máme tedy (podmíněnou) tabulku pro A a B. Exaktní test hypotézy A⊥B/C používá podmíněné rozložení při daných marginálních tabulkách T_{AC} a T_{BC} . To odpovídá podmiňování v každé vrstvě sloupcovými a řádkovými frekvencemi.Máme-li n dimensionální tabulku $T_{ABC...N}$,pak A⊥B/C...N můžeme testovat tak,že C...N chápeme jako jedinou veličinu ovšem s patřičně velkým počtem hodnot . Můžeme takto testovat podmíněnou nezávislost dvou veličin ve vícerozměrné tabulce. Algoritmické provedení včetně Monte-Carlo aproximace je zmíněno v Kreinerově (1984) práci a realizováne jeho programem EXABIRCH.

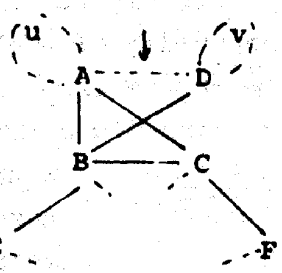
Pro Edwardsovu navrhovanou proceduru je ještě relevantní tento výsledek:

Nechť u a v jsou dvě veličiny.V rámci grafového modelu φ je možné testovat rozdíl k modelu $\varphi' = \varphi \& (u \perp v / V_n - u,v)$ t.j. odstranění hrany u-v testem pro nulovou parciální asociaci (podmíněnou nezavislost) v marginální tabulce T_a pro některé a obsahující u i v tehdy a jen tehdy, je-li φ_a uplný a hrana u-v neleží v hranici komponenty doplňku φ_a ve φ .

Jde o to,že v tomto případě je φ a φ' kolapsibilní na a . Úplnost je potřebná k tomu,aby šlo o ZPA test :



testujeme (ABC,BCD) v T_{ABCD} (proti saturovanému modelu),



zde bychom testovali (ABC,BD) proti(ABC,AD, BD), sice opět v T_{ABCD}.

Procedura II je modifikací postupu zhora dolů uvedeného v práci Edwardse a Kreinera (1983) :

krok 1: Začneme od saturovaného modelu,otestujeme všechny elementární ZPA modely a nejméně signifikatní z nich přijmeme - odstraníme patřičnou hranu.Vznikne model φ_1.

krok 2: Vezmeme model φ_1 a zkoumáme,které hrany můžeme testovat ZPA testem v marginální tabulce (viz výše) ,nejméně signifikatní hranu odstraníme a dostaneme model $\varphi_1 \& \varphi_2$.

krok 3: Vezmeme model $\varphi_1 \& \varphi_2 \ldots$ atd.

Proceduru končíme,nejde-li již žádná hrana testovat nebo jsou-li všechny hrany signifikantní. Všechny použité testy mhou být exaktní.

Vzhledem k tomu,že je-li φ rozložitelný model (a každý φ_1 je jistě rozložitelný) pak možnost testovat odstranění hrany u-v ZPA testem je ekvivalentní s tím,že $\varphi \& (u \perp v / V..)$ je rozložitelný. Výše uvedená procedura tedy dává rozložitelné modely.

Tvrzení (A) z §3 náme říká,že všechny rozložitelné modely jsou touto procedurou dosažitelné.

Výhody procedury III jsou zřejmé.Nevýhody jsou dvě: Jde o krokovou proceduru a i když by tato námitka šla odstranit přebudováním na větvící se proceduru,jsou exaktní testy i jejich rozumné Monte-Carlo aproximace přece jenom počítačově dosti drahé.

10. HLEDÁNÍ GRAFOVÝCH MODELŮ IV

Již předchozí procedura byla omezena použitím určitého typu testů (splňujících určité podmínky vzhledem ke kolapsibilitě). Následující postup navržený Whittakerem (1984) je opřen o použití testu poměrem věrohodnosti, respektive jeho aditivní vlastnosti. Postup má dávat rychlou a výpočetně nenáročnou orientaci v prostoru grafových modelů.

Postup je založen na rozkladu <u>deviance</u> t.j. věrohodnostní statistiky (pro testování proti saturovanému modelu) pro <u>rozložitelné</u> modely na aditivní složky odpovídající třídě ekvi-pravděpodobnostních modelů. Pro n=3 (veličiny A,B,C) máme následující ekvi-pravděpodobnostní modely (ABC), (AB, AC, BC), (A, B, C) a (1), t.j. $\log p_{ijk} =$ 0 pro model (1), $\log p_{ijk} = 0 + \lambda_i^A$ pro model (A) atd. Pro rozložitelné modely dostáváme například dev(AB,AC) = dev(AB) + dev(AC) - dev(A), dev(A,B,C) = dev(A) + dev(B) + dev(C) - 2 dev(1). Whittaker definuje aditivní složky jako:

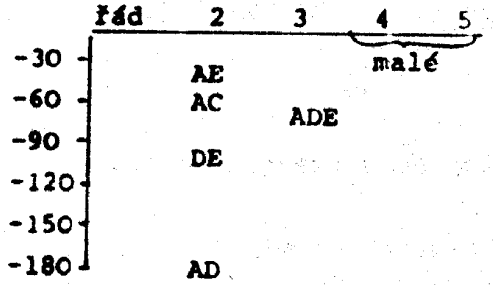
řád		
	G(:ABC) =	dev(ABC)
1	G(A:BC) =	dev(BC) - dev(ABC)
2	G(AB:C) =	dev(C) - dev(AC) - dev(BC) + dev(ABC)
3	G(ABC:) =	dev(1) - dev(A) - dev(B) + dev(AB) - dev(C) + dev(AC) + dev(BC) - dev(ABC),

a podobně pro permutace A,B a C.

Testování podmíněné nezávislosti odpovídají (až na znaménko) složky druhého řádu.

Při navrhovaném postupu vypočteme aditivní složky (bez použití iterací) a z jejich velikosti usuzujeme na vhodné modely. Vysoké absolutní hodnoty složek ukazují, že je nutné odpovídající efekty (kliky) zahrnout do modelu.

V příkladu uváděném Whittakerem (1984) jsou pro případ n=5 graficky znázorněny velké aditivní složky (druhého a vyššího řádu):



Z grafu je vidět, že modely neobsahující ADE, AC, B nemohou "vysvětlit" data. V konkrétním případě podrobnější analýza menších aditivních složek vedla k modelu (ADE, BDE, AC, BC).

Uvedený postup dává pro tabulky nižší dimenze rychlý přehled o celkové situaci a je výpočetně nenáročný. Využití tohoto přehledu je pak věcí statistika analyzujícího daná data.

11. HLEDÁNÍ OBECNÝCH HLL MODELŮ

Procedury I a II mohou být zobecněny pro vyhledávání obecných HLL modelů, připustíme-li, že je takové hledání užitečné (víme, že i pro grafové modely je zde dostatek interpretačních i teoreticko-statistických problémů). V teoretické úrovni jde o přechod od grafů k hypergrafům.

Skladební kameny pro konjunkce v případě procedury I jsou uvedeny v (Havránek, 1982a), i když je zde ještě dosti věcí, které by bylo nutné promyslet.

Zobecnění procedury II je uvedeno v (Edwards a HAvránek, 1984). Výhodou je zde možnost startu od libovolné množiny modelů t.j. např. od třídy minimálních modelů v A (a minimálních v R) vyhledaných procedurou pro grafové modely. Pak zobecněná procedura může skončit překvapivě rychle, jak je vidět z příkladu uvedeného v (Edwards a Havránek, 1984), kdy byly nalezeny místo dvou grafových modelů velice příbuzné minimální HLL modely: šlo o přechod od (AC,F,BE,BC,ADE) a (F,ACE,BC,ADE) k (AC,AD,AE,BC,DE,BE,F) a (AC,AD,AE,BC,DE,CE,F) a tedy o odstranění interakcí druhého řádu. V daném případě byl tento výsledek konsistentní s výsledkem simultánního testu nulovosti efektů třetího a vyššího řádu.

12. DODATEK

Budeme dokazovat, že graf vzniklý spojením grafů $\psi_{cl(b)}$ a φ nemůže mít žádnou kliku, která by nebyla klikou grafů φ nebo $\psi_{cl(b)}$ a tedy by nebyla generátorem společného modelu. Budeme uvažovat ψ_b s jedinou komponentou. Je-li ψ grafový, je i $\psi_{cl(b)}$ grafový (můžeme se na něj dívat jako na průnik ψ a saturovaného modelu nad cl(b). Nechť c je klika (generátor) grafu $\psi_{cl(b)}$ .Připojením φ se nemůže zvětšit (žádný uzel grafu φ ,který by nebyl v c se nemůže přidat,neboť by musel být v cl(b)) . Obráceně,nechť c je klika (generátor) grafu φ mající nějaký uzel v hranici b. Je-li c v nějaké klice grafu $\psi_{cl(b)}$, je vše vpořádku.Není-li tomu tak,obsahuje uzel u,který není v hranici b. Protože hranice leží v klice grafu φ,nemůže spojením grafů přibýt žádná hrana,která by kliku c zvětšila (doplnila např u-x a u-y hranou x-y pro x a y z hranice). Nemá-li c žádný uzel v hranici, nemění se.

LITERATURA

S.Asmussen,D.Edwards (1983) : Collapsibility and response variables in contingency tables, Biometrika 70,567-578.

J.N.Darroch,S.L.Lauritzen,T.P.Speed (1980) : Markov fields and log-linear interaction models for contingency tables, Ann.Statist. 8,522-539.

D.Edwards (1984): A computer intensive approach to the analysis of sparse contingency tables, COMPSTAT 84,Physica-Verlag,Wien, 355-359.

D.Edwards,T.Havránek (1984) : A fast procedure for model search in multidimensional contingency tables, Papport 84/1, RECKU,Kobenhavn

D.Edwards,S.Kreiner (1983) : The analysis of contingency tables by graphical models, Biometrika 70,553-556.

M.C.Columbik (1980) : Algorithmic graph theory and perfect graphs, Academic Press, New York.

P.Hájek,T.Havránek (1978) : Mechanizing hypothesis formation,Springer-Verlag, Heidelberg.

T.Havránek (1982a) : Some complexity considerations concerning hypotheses in multidimensional contingency tables, Transactions of the 9th Prague conference, vol.A, Academia,Praha 281-286.

T.Havránek (1982b) : O analýze mnohorozměrných kontingenčních tabulek,ROBUST 82, JČSMF, Praha, 12-19.

T.Havránek (1983) : Model search in multidimensional contingency tables with epidemiological applications, Biometrie und Biostatistik in der Medizin und verwandten Gebieten, Martin-Luther-Universität, Halle ,92-99.

T.Havránek (1984) : A procedure for model search in multidimensional contingency tables, Biometrics 40,95-100.

J.Nešetřil (1979) : Teorie grafů, SNTL, Praha.

Z.Prášková (1985) : Kontingenční tabulky,skripta MFF UK,Praha

R.Sundberg (1975) : Some results about decomposable (or Markov-type) models for multidimensional contingency tables - distribution of marginals and partitioning of tests, Scandinavian Journal of Statistics 2,71-79.

J.Whittaker (1984) : Fitting all possible decomposable models to multiway contingency tables, COMPSTAT 84,401-406

S.Kreiner (1984) : The analysis of multiple contingency tables by exact tests for zero partial association,Scand.J.of Statistics (v tisku) .