POPIS A VÝBĚR MODELŮ VE VÍCEROZMĚRNÝCH KONTINGENČNÍCH TABULKÁCH

Tomáš Havránek, ÚIVT ČSAV, Praha

Tato práce je podmnožinou materiálů, přednesených na semirářích ROBUST 82, 84, 86 a 90.

1. Popis modelů

Podívejme se na následující frekvenční tabulku:

E	D	C	B	A		%
				0	1	
<3.0	<140	0	0	95	1	1.0
			1	201	8	3.8
		1	0	295	5	1.7
			1	38	2	5.0
	≥140	0	0	49	5	9.3
			1	117	15	11.4
		1	0	192	15	7.2
			1	20	3	13.0
≥3.0	<140	0	0	58	7	10.8
			1	158	6	3.9
		1	0	145	10	6.5
			1	22	3	12.0
	≥140	0	0	47	6	11.4
			1	137	17	11.0
		1	0	117	16	12.0
			1	22	9	29.0

Tabulka 1

kde A je výskyt ischemické choroby srdeční v průběhu pěti let od vyšetření (0 ne, 1 ano), B psychická náročnost práce (subjektivně, 0 ne, 1 ano), C fyzická náročnost práce subjektivně, 0 ne, 1 ano), D systolický krevní tlak (pod 140, nad 140 včetně), E index alfa a beta lipoproteinů. Hodnoty B až E jsou zjišťovány při vstupním vyšetření. Hodnoty veličin D a E nejsou ve

zdrojových datech kategorizovány, ale kategorizace se provádí zejména pro tabulační účely. Data pocházejí z výzkumu prováděného v minulých letech Angiologickou laboratoří fakulty všeobecného lékařství UK pod vedením prof.MUDr. Z. Reiniše, DrSc.

Čeho si na tabulce všimneme? Především neobsahuje žádné nulové frekvence. To nám při zpracování ušetří mnohé starosti. Dále je pravděpodobně zřejmé, že A závisí na čtveřici B,C,D,E. Otázkou je však struktura této závislosti i nezávislosti B,C,D a E mezi sebou. Dále je asi vhodné konstatovat, že žádná z veličin není veličinou řízenou.

1.1. Stojíme nejprve před otázkou, **jak vyjadřovat hypotézy o nezávislosti resp. závislosti náhodných veličin**, jejichž pozorováním (nezávislým homogenním náhodným výběrem) vznikne tabulka tohoto typu. Jedna z možností, která se v posledních letech čím dál tím více používá je logaritmicko-lineární vyjádření. O co jde si nejprve ukážeme na případě tří náhodných veličin. Pro jednoduchost budeme předpokládat, že všechny nabývají pouze hodnot 0 a 1. Veličiny si označíme A,B,C. Nechť nyní P^{ABC}_{ijk} je pravděpodobnost, že trojice $\langle A,B,C\rangle$ nabývá hodnoty $\langle i,j,k\rangle \in \{0,1\}^3$. Tyto pravděpodobnosti můžeme rozepsat jako

$$\log P^{ABC}_{ijk} = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{j} + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk} \qquad (1)$$

kde požadujeme, aby $\sum_i \lambda^{A}_{i} = \sum_j \lambda^{B}_{j} = \sum_k \lambda^{C}_{k} = \sum_i \lambda^{AB}_{ij} = \sum_i \lambda^{AC}_{ik} = \sum_j \lambda^{AB}_{ij} = \sum_k \lambda^{AC}_{ik} = \sum_k \lambda^{ABC}_{ijk} = 0$.

Tento pohled je podobný pohledu při obvyklém modelu analýzy rozptylu s více faktory. Čísla $\lambda^{A}_{i}, \lambda^{B}_{j}, \lambda^{C}_{k}$ nazýváme *hlavními efekty* (nebo *efekty prvního řádu*), $\lambda^{AB}_{ij}, \lambda^{AC}_{ik}, \lambda^{BC}_{jk}$ efekty *druhého* řádu a λ^{ABC}_{ijk} *efektem* třetího řádu (často se též říká dvoufaktorový efekt) atd. Zpravidla, mluvíme-li o efektech, pak efektem myslíme vždy např. λ^{AB}_{ij} pro každé $i\in\{0,1\}$ a $j\in\{0,1\}$. Model (1) resp. struktura závislosti jemu odpovídající, "vysvětluje" jakoukoliv trojrozměrnou tabulku. Problém je, zda tabulku nelze "vysvětlit" jednodušším způsobem, jednodušší strukturou závislosti vzniklou vynecháním některých efektů, např. zda nestačí předpokládat, že

$$\log (P^{ABC}_{ijk}) = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{BC}_{jk}, \qquad (2)$$

což odpovídá nezávislosti páru B,C na A.

Různé struktury závislosti dostáváme vynecháváním efektů; budeme předpokládat, že vynecháme-li některý efekt druhého řádu, vynecháme i efekt třetího řádu, t.j. λ^{ABC}_{ijk}. Takto získané modely nazýváme *hierarchické logaritmicko-lineární modely* a můžeme je zapisovat zkráceným zápisem. Zapisujeme vždy jen horní indexy nejvyšších obsažených efektů (efekt druhého řádu je vyšší než efekt prvního řádu atd.). Tedy pravděpodobnostní model (1) má zápis (ABC), model (2) má zápis (A,BC) a model úplné

nezávislosti $\log(P_{ijk}^{ABC})=\theta+\lambda_i^A+\lambda_j^B+\lambda_k^C$ má zápis (A,B,C). Tento způsob zápisu je dostatečně přehledný a mnohdy snadno interpretovatelný: (A,B,C) znamená, že A,B a C jsou navzájem nezávislé atd. Model ABC se nazývá *saturovaný* model,nebo *úplný* model *třetího řádu*, model (AB,AC,BC), který obsahuje všechny efekty prvního a druhého řádu, se nazývá *úplný* model *druhého* řádu, podobně (A,B,C) je úplný model prvního řádu.

Důležité je si uvědomit, že o modelech můžeme mluvit pouze na základě jejich velmi stručného zápisu, který je tvořen vlastně výrazy (slovy), určitého velmi jednoduchého formálního jazyka. S těmito výrazy lze provádět operace, které mohou být popsány syntaktickými pravidly a které mohou mít sémantický význam v prostoru odpovídajících pravděpodobností (pravděpodobnostních modelů).

Podívejme se nyní na seznam logaritmicko-lineárních hierarchických modelů, které připadají rozumně v úvahu u trojrozměrné tabulky (pomíjíme modely nižší dimenze a modely vzniklé vynecháním efektů prvního řádu):

model	vyjádření v pravděpodobnostech (horní indexy vynecháváme)	stupně volnosti
(A,B,C)	$p_{ijk}=p_{i..}p_{.j}p_{..k}$	4
(AB,C)	$p_{ijk}=p_{ij.}p_{..k}$	3
(AC,B)	$p_{ijk}=p_{i.k}p_{.j.}$	3
(A,BC)	$p_{ijk}=p_{i..}p_{.jk}$	3
(AB,AC)	$p_{ijk}=p_{ij.}p_{i.k}/p_{i..}$	2
(AB,BC)	$p_{ijk}=p_{ij.}p_{.jk}/p_{.j.}$	2
(AC,BC)	$p_{ijk}=p_{ij.}p_{.jk}/p_{..k}$	2
(AB,AC,BC)	$(p_{111}p_{001})/(p_{101}p_{011})=(p_{110}p_{000})/(p_{100}p_{010})$	1

Všimněme si, že prvních sedm modelů lze snadno vyjádřit v řeči nezávislosti a podmíněné nezávislosti. Navíc, maximálně věrohodné odhady pravděpodobností p_{ijk} při platnosti každého z těchto modelů dostaneme snadno součinem (či podílem) přímých odhadů marginálních pravděpodobnosti. Tyto modely se nazývají *rozložitelné* nebo multiplikativní. Poslední model (AB,AC,BC) má jiný charakter; jednak jeho interpretace je jiná a jednak odhady metodou maximální věrohodnosti musíme získávat iteračním postupem (např. iterative proportional fitting).

1.2. Logaritmicko lineární modely lze použít i pro tabuky vyšší dimenze, kdy vyjadřování v pravděpodobnostech je velmi složité a nepřehledné. Ve vyšších dimenzích je zřetelně vidět jednoduchost formálního zápisu modelů. Logaritmicko-lineární model odpovídající (A,BCDE) si jistě každý umí sestavit. Uvažujme nyní obecně tabulky dimenze n a nechť veličiny

jsou očíslovány 1,...,n. Pak každý zápis $(a_1,\ldots,a_k)$, kde $a_i \subseteq \{1,\ldots,n\}$ a pro každé $i \neq j$ není ani $a_i \subseteq a_j$, ani $a_i \supseteq a_j$, definuje jednoznačně logaritmicko-lineární hierarchický model. $(a_1,\ldots,a_k)$ se pak nazývá *generující sentencí* (množinou) *modelu*. Množiny $a_1,\ldots,a_k$ nazýváme *generárory*. Obvykle se ovšem prakticky používají místo čísel písmena, neboť konkrétně zatím nikdo pravděpodobně nepracuje s tabulkou dimenze větší než např.n = 22. Navíc místo pedantického zápisu {{A,B},{B,C,D,E}} atp. se používá zjednodušený zápis (AB,BCDE).

Uvažujeme obecně 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 dimenzionální tabulce. Množinu veličin vytvářejících danou tabulku značíme V={A,B,C,...,N}. Jednotlivé frekvence v tabulce značíme m(i). Odpovídající pravděpodobnosti značíme p(i).

Pro generující sentence máme definovány operace průseku a spojení $\wedge,\vee$:

$$(a_1,\ldots,a_k)\wedge(b_1,\ldots,b_l)=(a_1\cap b_1, a_1\cap b_2,\ldots,a_k\cap b_l)$$

při vynechání redundantních množin a 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 $\wedge$ odpovídá průniku či konjunkci modelů, proto běžně píšeme pro dva modely

$$\varphi\&\psi = (a_1,\ldots a_k)\&(b_1,\ldots b_l) = (a_1,\ldots,a_k)_{\varphi}\wedge(b_1,\ldots,b_l)_{\psi}.$$

Maximálně věrohodné odhady frekvencí (resp. pravděpodobností) při daném modelu φ značíme $m^{\varphi}(i)$ (resp. $p^{\varphi}(i)$). Obecně je nutné tyto odhady při HLL modelech stanovovat iteračním postupem (iterative proportional fitting). Ve speciálním případě **rozložitelných** modelů interpretovaných 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 $m(i_a)= \sum_{i \sqcap a^c} m(i)$, t.j. sčítá se přes hodnoty indexů odpovídajících veličinám z $a^c=V_n-a$. Například pro model φ=(AB,BC) potřebujeme marginální tabulky T_{AB} a T_{BC}. Odhad je pak $p_{ijk}=(m_{ij.},m_{.jk})/(m_{.j.}m)$ v běžné notaci $(m_{ij.}=\sum_k m_{ijk})$.

Ve vyšších dimenzích drtivě převažují hierarchické logaritmicko-lineární modely, které nejsou rozložitelné. Viz následující tab. 2.

dimenze: n=	2	3	4	5	výraz
hierachické	5	19	167	7580	2^{2^n-1}
z toho					
grafové	5	18	113	1450	$\sum_{i=0}^{n} \binom{n}{i} 2^{\binom{i}{2}}$
rozložitelné	5	18	110	1230	není znám

Tabulka 2

S třídou grafových (resp.ZPA) logaritmicko-lineárních hierarchických modelů se seznámíme za chvíli.

Jak rozlišíme rozložitelné a nerozložitelné modely? Jednoduchý algoritmus využívající generující sentence lze nalézt v knize (Bishop et al., 1975). Zde ho uvádíme v nepatrně pozměněné formulaci:

* Měj $(a_1,\ldots,a_k)$. Je-li k = 2, zastav.

(1) Vyhoď písmeno, které se vyskytuje ve všech a.

(2) Vyhoď písmeno, které se vyskytuje v jediném a_i.
Nebylo-li možné použít ani (1) ani (2) zastav. Tím vznikne $(a'_1,\ldots,a'_k)$ (některé a'_i může být prázdné).

(3) Vyhoď každé a'_i, které je vlastní částí jiného. Je-li $a'_i=a'_j$ pro i<j, vyhoď a'_j (proveď pro všechna i,j).
Tím vznikne $(a''_1,\ldots,a''_l)$, l≤k. Polož k = l a $(a_1,\ldots,a_k)=(a''_1,\ldots,a''_l)$ a jdi na *.

Model $(a_1,\ldots,a_k)$ vstupující do prvního cyklu je rozložitelný právě když je výpočet zastaven v některém cyklu pro podmínku k = 2.

Uveďme si příklady pro n = 5:

(ABC,ACDE) je triviálně rozložitelný.

(ABE,BCE,ADE,CDE) $\rightarrow^{(1)}$ (AB,BC,AD,CD) a již nelze aplikovat (1) a (2); model není rozložitelný.

(ABE,ADE,BC)$\rightarrow^{(1)}$(ABE,ADE,B)$\rightarrow^{(3)}$(ABE,ADE) - model je rozložitelný.

Pro n=7: (ABD,BCE,ACF,EG,ABC)$\rightarrow^{(2)}$ (ABD,BCE,AC,EG,ABC) $\rightarrow^{(3)}$ (ABD,BCE,EG,ABC)$\rightarrow^{(3)}$ (ABD,BCE,E,ABC) $\rightarrow^{(3)}$ (ABD,BCE,ABC) $\rightarrow^{(1)}$ (AD,CE,AC) $\rightarrow^{(2)}$ (A,CE,AC) $\rightarrow^{(3)}$ (CE,AC) - model je rozložitelný.

Zde je opět na místě zdůraznit, že jde o čistě syntaktický postup, používající pouze formální vyjádření modelů.

Někdy je vhodné použít duální reprezentaci HLL modelů, kdy model je popsán minimálními horními indexy efektů, které jsou nulové. Máme tedy duální generující sentenci $(a_1,\ldots,a_k)^d$. Např. (ABC,BCD) = $(AD)^d$, (ABD,ACD,BCD) = $(ABC)^d$ a (BD,AD,CD) = $(AB,BC,AC)^d$. Jednoduchý algoritmus umožňující přechod mezi oběmi reprezentacemi byl popsán v (Edwards a Havránek, 1985).

1.3. **Z hlediska interpretačního rozumná třída modelů** by měla být uzavřena vzhledem ke konjunkci. Naneštěstí třída rozložitelných modelů uzavřená na konjunkci není. Viz následující příklad: (ABC,BCD)&(ABD,ACD) = (AB,AC,BD,CD) ; viz (Enke, 1980). Je tedy vhodné hledat nejmenší třídu modelů, uzavřenou na konjunkci a obsahující třídu rozložitelných modelů (samozřejmě jako podtřídu hierarchických logaritmicko-lineárních modelů). Cesta k hledání takové třídy vede přes standardní representaci hierarchických logaritmicko-lineárních modelů.

Uvažujme modely dané pevné dimenze n. Generující sentenci sestávající se z množin kardinality právě n-1 nazveme *elementární*. Příklad pro n = 5 : (ABCD,ACDE,ABCE). Počet množin v generující sentenci nazveme velikostí sentence.

Platí, že každou generující sentenci lze jednoznačně vyjádřit konjunkcí $\varphi_1\&\ldots\&\varphi_r$, kde $\varphi_1,\ldots,\varphi_r$ jsou elementární sentence (Havránek, 1982). Tuto konjunkci nazýváme *standardní representací*.

Poznamenejme (Edwards a Havránek, 1985), že pro duální popis modelů máme $(a_1,\ldots,a_j)^d\&(\mathfrak{b}_1,\ldots,\mathfrak{b}_k)^d = (a_1,\ldots,a_j,\mathfrak{b}_1,\ldots,\mathfrak{b}_k)^d$ po vynechání redundantních množin. Pak pro libovolný model je $(\mathfrak{b}_1,\ldots,\mathfrak{b}_k)^d = (\mathfrak{b}_1)^d\& \ldots \& (\mathfrak{b}_k)^d$ a po přechodu k obvyklému vyjádření modelů odpovídají $(\mathfrak{b}_1)^d, \ldots, (\mathfrak{b}_k)^d$ právě elementárním sentencím.

Maximální velikostí sentence v $\varphi_1\&\ldots\&\varphi_r$ můžeme měřit složitost původní sentence, jejímž standardním vyjádřením daná konjunkce je. Modely, jejichž standardní vyjádření obsahuje sentence velikosti maximálně 2, nazýváme *ZPA modely*. Z praktických důvodů se většinou omezujeme na případ, kdy standardní vyjádření obsahuje právě sentence velikosti dvě (jinak by chyběl některý efekt prvního řádu). Tuto třídu modelů označíme $\mathcal{H}_1$.

Příklady standardních vyjádření:
(A,B,C,D) = (ABC,ACD) & (ABC,BCD) & (ABC,ABD) & (ABD,ACD) & (ABD,BCD) & (ACD,BCD) je ZPA i rozložitelný.
(AB,AC,AD,BC) = (ABC,ACD)&(ABD,ACD) je ZPA a není rozložitelný.
(AC,AD,BD) = (ABC,ABD)&(ABD,ACD)&(ACD,BCD) je ZPA i rozložitelný.

Co znamená ZPA? "Zero partial association". Model ABC,ACD odpovídá nulové parciální asociaci mezi B a C, t.j. podmíněné nezávislosti B a C podmíněno AD (jde o parciální asociaci v Birchově smyslu).

Platí následující skutečnosti:

1. Třída $\mathcal{H}_1$ je uzavřena vzhledem ke konjunkci,
2. třída $\mathcal{H}_1$ obsahuje třídu rozložitelných modelů
3. třída $\mathcal{H}_1$ je nejmenší třída hierarchických logaritmicko-lineárních modelů obsahující třídu rozložitelných modelů a uzavřená vůči konjunkci.

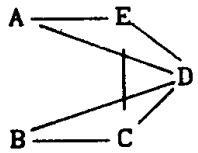
Vidíme, že třída ZPA modelů má vlastnosti naší hledané třídy. ZPA modely, které nejsou rozložitelné, neumožňují ovšem přímé odhady pravděpodobnstí a mají obtížnější interpretaci. Rozdíl v počtu ZPA modelů a rozložitelných modelů je vidět z tabulky 2.

1.5. Generující třídu $(a_1,\ldots,a_k)$ můžeme zobrazit na neorientovaný graf s množinou uzlů V_n následujícím způsobem:

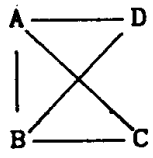
$$i((a_1,\ldots,a_k))=(V_n,\mathcal{E}),$$

kde $\mathcal{E}=\{(x,y);\ (x,y)\in V_n X V_n$ a xy je obsaženo v některém $a_i\}$.

Modelu (ACDE,BCDE) tedy odpovídá graf,

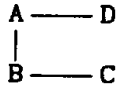
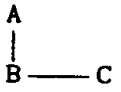
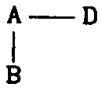


t.j. úplný graf nad V_n bez jedné hrany AB. Modelu (ABC,AD,BD) pro n = 4 odpovídá graf:



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

Pro graf  máme dvě kliky  a 

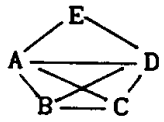
a tedy $\iota(\underline{G})$=(ABC,ABD).

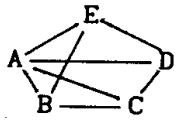
Důležité je nyní toto: $i(\mathcal{H}_1)$ je množina všech grafů (s n uzly) a zároveň $(i\restriction\mathcal{H}_1)^{-1} = \iota$. Při resktrikci i na $\mathcal{H}_1$ dostáváme vzájemně jednoznačné zobrazení $\mathcal{H}_1$ na množinu všech grafů s n uzly. Vzhledem k této vlastnosti je plně oprávněný název, který pro třídu $\mathcal{H}_1$ použil Darroch. Lauritzen a Speed (1980), t.j. třída *grafových modelů*.

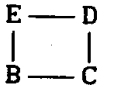
1.6. Charakterizace rozložitelnosti. Pro práci s grafovými modely je vhodné používat techniky teorie grafů. V tomto jazyce také charakterizovali Darroch, Lauritzen a Speed (1980) rozložitelné modely:

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

Příklad (n=5):

(ABCD,ADE)  je rozložitelný,

(ABC,ACD,ABE,ADE)  není rozložitelný

protože podgraf  je kružnicí délky čtyři. Pro rozpoznávání triangulovaného grafu existuje algoritmus (eliminování izolovaných uzlů, Golumbic, 1980), odpovídající algoritmu pro rozpoznávání rozložitelných modelů uvedenému např. v 1.2. To, že rozložitelné modely mají triangulované grafy odpovídá tomu, že jde o jednoduché modely; triangulované grafy jsou v jistém smyslu rovněž jednoduché - jejich klikovost se rovná jejich barevnosti (viz Nešetřil, 1979).

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

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

V jiné formulaci to znamená, že je-li ψ 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); patřil léta k folkloru mezi badateli v dané oblasti. Důvodem ne zcela jasných fomulací i ne všeobecného pochopení různých výsledků je zde patrně stále nejednotný jazyk.

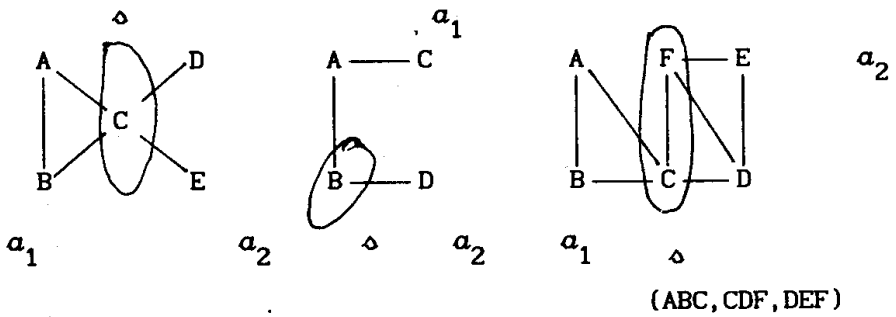
Z 1.3 víme, že grafový model φ lze psát jako konjunkci $\varphi_1\&\ldots\&\varphi_k$ elementárních ZPA modelů. Modely $\varphi_1,\ldots,\varphi_k$ jsou samozřejmě rozložitelné. Platí (Havránek, 1982b):

(B) φ je rozložitelný právě když členy v konjunkci lze uspořádat tak, že $\varphi_1\&\varphi_2\&\ldots\&\varphi_j$ je rozložitelný pro každé j = 1,...,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).

1.7. Řekneme, že množina uzlů Δ *odděluje* množiny uzlů a_1 a a_2, jestliže každá cesta mezi a_1 a a_2 musí vést přes uzly v Δ.

Příklady:



(ABC,CDF,DEF)

Tušíme ihned, že v řeči grafových modelů platí:
(A): a_1 a a_2 jsou podmíněné nezávislé vzhledem k Δ (značíme $a_1 \perp a_2 | \Delta$) právě když je Δ odděluje.

2. Grafové algoritmy

2.1. Máme-li klasickou úlohu testovat danou hypotézou popsanou generující sentencí $\varphi=(a_1,\ldots,a_k)$, "víme" jak postupovat (za určitých podmínek např. na frekvence v tabulce apod.). Můžeme například použít χ^2 test poměrem věrohodnosti. Ukážeme si jej opět pro speciální případ dimenze tabulky, abychom se vyhnuli obtížným formálním zápisům. Nechť tedy $n = 4$. Jsou-li m_{ijkl} napozorované frekvence v tabulce, označíme $m = \sum_{i,j,k,l} m_{ijkl}$ a očekávané frekvence můžeme vyjádřit jako $m_{ijk}(\varphi) = m\, p_{ijk}(\varphi)$, kde $p_{ijkl}(\varphi)$ jsou pravděpodobnosti odhadnuté metodou maximální věrohodnosti za předpokladu platnosti hypotézy popsané pomocí φ. K testování pak použijeme statistiku

$$G^2(\varphi) = 2 \sum_{ijkl} m_{ijkl} \log(m_{ijkl}/m_{ijkl}(\varphi))$$

s příslušným počtem stupňů volnosti (zde 7 - počet odhadovaných parametrů).

Někdy je vhodné zkoumat vztahy mezi hypotézami. Řekneme, že $\varphi \quad \psi$ (φ logicky vyplývá z ψ), plyne-li z pravdivosti modelu definovaného φ pravdivost modelu definovaného ψ (je-li φ splněno, je splněno i ψ). Tuto relaci opět snadno rozpoznáme na syntaktické úrovni:

Platí totiž, že $(a_1,\ldots,a_k)$ $(\&_1,\ldots,\&_l)$ právě když $(a_1,\ldots,a_k) \leq (\&_1,\ldots,\&_l)$, kde relace $\leq$ je definována takto: $(a_1,\ldots,a_k) \leq (\&_1,\ldots,\&_l)$, jestliže pro každé a_i existuje $\&_j$ tak, že $a_i \subseteq \&_j$.

Chceme-li nyní testovat "významnost rozdílů" mezi φ a ψ, $\varphi \quad \psi$, použijeme statistiky

$$G^2(\varphi|\psi) = G^2(\varphi) - G^2(\psi)$$

(stupně volnosti dostaneme rovněž odečtením). Je dobré si všimnout, že φ může být považována za **jednodušší** hypotézu.

2.2. Velmi často se ocitáme v situaci popsané na začátku oddílu 1: nemáme žádnou specifikovanou hypotézu, nebo několik málo specifikovaných hypotéz, ale musíme naopak nějaké hypotézy "podporované daty" nalézt. To pak znamená zkoušet celou řadu hypotéz, které by mohly připadat v úvahu a pozorovat jejich shodu s daty. Úmyslně zde je použito slovo "zkoušet" místo testovat, aby bylo jasné odlišení této situace od klasické situace testování. V našem případě, i když používáme testových statistik, dosažených hladin významnosti atd., chápeme tyto spíše jako míry shody a neshody hypotéz s daty, než jako testy v klasickém smyslu (neuvažujeme zde nyní metody simultánní inference). Metody vyhledávání hypotéz by měly být z teoretického hlediska vlastně hodnoceny jinak, než klasické metody testování. Obrana proti chybě prvního druhu (resp. globální chybě prvního

druhu) zde možná není vůbec tou nejdůležitější věcí. Spíše zde záleží na něčem, co by mohlo být nazváno silou - například na tom, s jakou pravděpodobností metoda odhalí danou (známou) strukturu závislosti (viz Havránek a Soudský, 1989).

Vraťme se nyní k frekvenčním tabulkám. Víme, že hierarchických logaritmicko-lineárních modelů, které by mohly teoreticky připadat v úvahu, je velmi velké množství. Musíme se tedy omezit na nějakou rozumnou podtřídu. Z řady výše uvedených důvodů je prvním kandidátem na takovou třídu právě třída grafových modelů:

(1) je uzavřená vůči konjunkci, (2) obsahuje třídu rozložitelných modelů a je to nejmenší taková třída a (3) všechny grafové modely jsou popsatelné konjukcemi elementárních sentencí, které popisují jednoduché (ZPA) grafové modely vyjadřující nulovou parciální asociaci párů veličin.

Na druhé straně je vidět, že i "grafových" hypotéz může být poměrně mnoho. Je proto nutné využít vztahy mezi hypotézami k tomu, abychom nemuseli zkoumat celou třídu grafových modelů. Zvolíme si pevnou hladinu významnosti;hypotézy s dosaženou hladinou významnosti nižší než zvolená zamítáme, ostatní "akceptujeme".

Zdá se vhodné využít toto pravidlo:

Je-li $\varphi \models \psi$ (a tedy $\varphi \leq \psi$) pak při zamítnutí ψ v daných datech zamítneme i φ (aniž bychom ho testovali).

Mohlo by se zdát, že je vhodné použít i opačného vztahu: při "akceptování" φ "akceptovat" i ψ v logickém smyslu z φ vyplývající. Tedy například, kdybychom akceptovali některé elementární ZPA modely, akceptovali bychom i jejich konjunkci. Víme však, že role zamítnutí a nezamítnutí není v tomto smyslu symetrická a že zamítnutí můžeme považovat za "jistější".

Předběhneme nyní k dokončení našeho příkladu (viz tab.3). Nezamítnutí hypotéz označených 2, 3, 5, 6, 0 při dané hladině významnosti by tedy znamenalo akceptování hypotézy 23560=(AB,BC,ADE), která má však hodnotu χ^2 = 41.56 při 20 stupních volnosti a tedy dosaženou hladinu významnosti 0.00321!).

Samozřejmě, že při použití testu χ^2 poměrem věrohodnosti (i Pearsonova χ^2) se nám může stát, že sice ψ zamítneme, ale φ při testování nikoliv. Při použití výše zmíněního pravidla dojde tedy k chybnému rozhodnutí (jde ovšem o něco jiného než chyba I. a II. druhu) vzhledem k daným datům. Praxe však ukazuje, že tyto případy nejsou příliš časté a takto zamítnuté hypotézy bez testování nemívají dosaženou hladinu významnosti příliš vyšší než je zvolená mez a nepatřily by k "nejlepším" hypotézám podle kritérií naznačených v odst. 2.3. Celá věc by si však zasloužila ještě hlubší zkoumání (viz Havránek a Pokorný, 1985).

Celý postup vyhledávání "akceptovatelných" modelů vypadá s použitím námi navrženého pravidla takto:

Grafové modely generujeme a testujeme v pořadí opačném k uspořádání ≤. To znamená, že v prvním kroku testujeme elementární hypotézy parciálních asociací (ZPA). V dalším kroce testujeme modely odpovídající dvoučlenným konjunkcím elementárních sentencí, v třetím kroce modely odpovídající

	Testováno:	zamítnuto (0.1)	marginální a parciální asociace dle Browna:	
1.	ABCD, ABCE	+	+	+
2.	ABCD, ABDE		+	+
3.	ABCD, ACDE		+	+
4.	ABCD, BCDE	+	+	+
5.	ABCE, ABDE			
6.	ABCE, ACDE			
7.	ABCE, BCDE	+	+	+
8.	ABDE, ACDE	+	+	+
			(nejsilnější vazba)	
9.	ACDE, BCDE	+	+	
10.	ABDE, BCDE			

23 znamená konjunkci elementárních sentencí 2 a 3. Testovány jsou pouze dvoučlenné konjunkce elementárních sentencí nezamítnutých v prvním kroce.

Je tedy testováno pouze 10 modelů místo $\binom{10}{2}$=45 možných.

23.	ABCD, ADE	+	rozložitelný
25	ABC, ABDE		rozložitelný
26	ABC, ACD, ABE, ADE		ne
20	BCD, ABDE		rozložitelný
35	ABC, ABD, ACE, ADE		ne
36	ABC, ACDE		rozložitelný
30	ABD, BCD, ADE, CDE,	+	ne
56	ABCE, ADE		rozložitelný
60	ABE, BCE, ADE, CDE		ne

V třetím kroce, jsou testovány pouze ty konjunkce, které nejsou zamítnuty již na základě zamítnutí konjunkcí v předchozích krocích. Tedy netestujeme např. 125, nebo 350.

256	ABC, ABE, ADE	rozložitelný
250	BC, ABDE	rozložitelný
260	BC, CD, ABE, ADE	ne
356	ABC, ACE, ADE	rozložitelný
560	ABE, BCE, ADE	rozložitelný

Testování 5 hypotéz místo 105 možných.

Ve čtvrtém kroce máme již pouze jedinou konjunkci, které není zamítnutá na základě předchozích výsledků.

2560	ABE, ADE, BC	rozložitelný

Tabulka 3

trojčlenným konjunkcím atd. Tak máme zajištěno, že pro dané φ (= φ_1&...&φ_2) budou všechny modely ψ takové, že $\psi \models \varphi$ (a tedy je už nemusíme generovat a testovat) uvažovány později než φ a můžeme tedy informaci o zamítnutí využít. Výsledkem je množina "akceptovaných" modelů, t.j. takových modelů, které byly testovány a nezamítnuty (hypotézy, které jsou zamítnuty na základě našeho pravidla nejsou testovány).

Takto získaná množina kandidátů může být ještě dosti velká. Je proto vhodné z ní vybírat "nejsilnější". "nejzajímavější", "nejlepší" apod. modely. Výše uvedená slova je nutné nějakým způsobem specifikovat. Jednou z čistě logických možností je hledat minimální prvky množiny akceptovaných modelů vzhledem k uspořádání ≤. Tím dostáváme vlastně i nejjednodušší akceptované modely, pokud měříme složitost počtem obsažených logaritmicko-lineárních parametrů (efektů). To je ve shodě se strategií propagovanou v (Benedetti a Brown, 1978).

2.3. Dokončíme nyní náš příklad: Výsledky jsou přehledně uvedeny v tabulce 3.

Množina "akceptovatelných" hypotéz je tedy tvořena všemi (modely) sentencemi, které byly testovány a nezamítnuty. Zkráceně jde tedy o modely zapsané výše kódy 2, 3, 5, 6, Q, 25, 26, 35, 60, 256, 250, 260, 356, 560, 2560.

Které z uvedených modelů jsou ty "nejlepší" a měly by být reprezentovány jako možné modely pro další zkoumání? Jeden z možných přístupů je hledat minimální prvky množiny "akceptovaných" sentencí vzhledem k uspořádání ≤ . Platí totiž, že na teoretické úrovni jsou všechny ostatní sentence v množině "akceptovaných" sentencí jejich logickými důsledky. Zde jsou dva takové prvky a to 356 a 2560, naštěstí oba rozložitelné.

Pro vyhledávání těchto prvků je vhodné si všimnout vztahu relace ≤ mezi sentencemi a inkluze jejich kódů. Vztahu $\varphi \geq \psi$ je ekvivalentní kód(φ)≤kód(ψ). Např. : (ABE.BCE,ADE) ≥ (ABE,ADE,BC) a 560 ≤ 2560. Hledání minimálních prvků v ≤ se tedy převádí na hledání maximálních kódů vzhledem k inkluzi.

Někdy může být užitečná i celá množina "akceptovaných" modelů. Navržené kriterium pro výběr "nejzajímavějších" modelů není jediné možné. Při výběru mohou hrát roli i apriorní znalosti zadavatele o tom, co s čím může souviset, nebo jiná kriteria výběru "nejlepších" modelů, například podle nejvyšší dosažené hladiny významnosti chápané jako kriterium shody modelu s daty, či podle různých modifikací χ^2 jako míry shody, např. AIC (Akaike information criterion), viz Sakamoto a Akaike , (1978); AIC = χ^2 - 2 d.f.. V našem příkladě jsou dvě hypotézy s nižším AIC (a tedy lepším) než hypotézy vybrané a to 256 a 56 (viz Havránek, 1984a).

Při větších příkladech to znamená uchovávat strukturovanou množinu "akceptovaných" modelů spolu s některými numerickými údaji a pomocí dalších programů z ní pak vybírat podle různých návrhů a kritérií výše zmíněných i podle dalších kriterií, která mohou být z různých důvodů uplatněna.

2.4. Zopakujme, že právě navržený přístup (Havránek, 1983, 1984a,b) 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 $\psi \leq \varphi$.

Připomeňme, že $\varphi \leq \psi$, jestliže graf modelu φ je podgrafem modelu ψ. Postup byl počítačově realizován M. Řebíčkovou - Horákovou (Michálek a kol.,1984) a B.P. Murphym (1984).

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

Proč je zde použit 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 dále oddíl 3.4). K principu (ii): na teoretické úrovni je jisté, že platí-li model φ, musí platit i ψ pro $\varphi \leq \psi$,t.j. nezamítnutí φ by mělo vést k nezamítnutí ψ (pozor, je to jiná věc než příklad s konjunkcí 23560 v 2.2). 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 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, 1985) *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).

Víme, že ani testová statiština poměrem věrohodnosti ani Pearsonův χ^2 nejsou dobré míry shody a pomocí nich nelze modely, které nejsou v relaci $\leq$, srovnávat; v příkladě byly uvažovány dosažení hladiny významnosti a AIC kriterium. Poznamenejme, že zde by bylo vhodné hledat takový test, či obecněji rozhodovací pravidlo, které by mělo 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.

2.5. V (Edwards a Havránek, 1985), 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 $\varphi \leq \psi$.

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 $\leq$. Tyto modely se otestují a tím rozdělí na zamítnuté (množina R) a na nezamítnuté (množina A). Dále se analyzuje $\mathcal{H}_1$-S. Hledají se $\leq$-minimální modely $D_a(R)$, které nejsou slabě zamítnuté na základě R nebo $\leq$-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 shora 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 použít například efektivní postup pro hledání $D_r(A)$ atd. (viz odstavec 2.6).

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

Dvě poznámky: (a) startujeme-li tuto proceduru 2.5 (resp. proceduru 2.4) s počáteční množinou elementárních ZPA modelů, pohybujeme se velmi často v oblasti velmi špatné aplikability asymtotických testů. Navíc zamítnutí elementárních modelů vede k velké redukci množiny dále uvažovaných modelů (možných modelů). Proto je žádoucí zde použít exaktní test.

(b) Ideální test pro tuto situaci by měl být takový, že by slabé zamítnutí a slabé přijetí slývalo se zamítnutím i přijetím. Pak by inference na teoretické úrovni přesně odpovídala chování testů na datech (dedukci na datové úrovni, viz oddíl 4).

Procedura 2.5 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 14 elementárních grafových modelů a 32753 neelementárních grafových modelů. V citované práci byla použita procedura

(2.4) s dalším trikem, který odpovídá jednomu kroku procedury (2.5). Přesto bylo nutné explicitně testovat více než 100 modelů. Procedura (2.5) 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 programem GRAPH napsaný Edwardsem v jazyce Pascal (viz dále odst. 2.6).

2.6 Nyní popíšeme proceduru 2.5 podrobněji. Uvažujme grafové modely ($\mathcal{H}_1$) s hlavními efekty, ti graf popisují závislost veličin z množiny V má jako uzly všechny tyto veličiny.

Model m může být pak specifikován dvěma možnými způsoby: (i) $m = (e_i,e_j,\ldots,e_k)^p$, kde $e_i,e_j,\ldots,e_k$ jsou hrany přítomné v *interakčním grafu* modelu. Zpravidla specifikujeme hrany svými (koncovými) uzly, t.j. $(AB,BC)^p$ odpovídá modelu s právě dvěma hranami AB a BC.

(ii) **Duálně** může být model specifikován jako $m = (e_i',e_j',\ldots,e_k')^d$, kde e_i', ... ,e_k' jsou hrany chybějící v interakčním grafu. Poznamenejme, že $(\mathcal{H}_1,\vee,\wedge)$, kde operace $\wedge$ a $\vee$ jsou definovány pomocí průniku a sjednocení množin hran tvoří booleovský svaz.

Nechť nyní $S = \{m_1,\ldots,m_p\}$, $m_i \in \mathcal{H}_1$ pro $i = 1,\ldots,p$ je množina grafových modelů a předpokládejme, že tyto modely jsou akceptovány. Množina modelů z $\mathcal{H}_1$, které můžeme pokládat za slabě akceptované je

$$S^+ = \{m \in \mathcal{H}_1;\ m_i \leq m \text{ pro některé } m_i \in S\};$$

o modelech z

$${}^cS^+ = \{m \in \mathcal{H}_1; \text{neplatí } m_i \leq m \text{ pro } i=1,\ldots,p\}$$

nemůžeme říci nic. Přitom ${}^cS^+ = \mathcal{H}_1 - S^+$.

Položme nyní

$$D_r(S) = \max {}^cS^+$$

(max (A) je množina maximálních prvků množiny A vzhledem k uspořádání $\leq$). $D_r(S)$ nazýváme r-*duál* množiny S. Jsou to nejsložitější modely, které můžeme ještě zamítnout, jsou-li akceptovány modely v S. Všimněme si, že jestliže by byly modely z $D_r(S)$ zamítnuty, pak všechny modely z ${}^cS^+ - D_r(S)$ jsou slabě zamítnuté a každý model z $\mathcal{H}_1$ je již akceptován či slabě akceptován nebo zamítnut či slabě zamítnut. Množiny S, S^+-S, $D_r(S)$ a ${}^cS^+$ tvoří disjunktní rozklad S.

Podobně můžeme definovat a-*duál* množiny S: nejprve položíme

$$S^- = \{m \in \mathcal{H}_1;\ m \leq m_i \text{ pro některé } m_i \in S\}$$

(množina slabě zamítnutých modelů, jsou-li modely z S **zamítnuty**) a a-duál je pak

$$D_a(S) = \min {}^cS^-$$

kde ${}^cS^- = \mathcal{H}_1 - S^- = \{m \in \mathcal{H}_1; m \leq m_i,\ i=1,\ldots,p\}$.

Platí, že ${}^cS^+ = (D_r(S))^-$ a ${}^cS^- = (D_a(S))^+$. Je-li S nesrovnatelná množina, pak

$$D_a(D_r(S)) = D_r(D_a(S)) = S.$$

Pro realizaci procedury 2.5 je důležité nalezení duálů, jinak než prohledáváním množin ${}^cS^+$ a ${}^cS^-$.

Nechť $S = \{m_1,\ldots,m_p\}$, kde $m_j = (e_{i1},\ldots,e_{in(i)})^p$ pro $i = 1,\ldots,p$. Pak v každém modelu z $D_r(S)$ musí chybět alespoň jedna hrana z každého m_i, $i = 1,\ldots,p$. Platí tedy, že $D_r(S) = \max \{(e_{11},e_{21},\ldots,e_{p1})^d,(e_{12},e_{21},\ldots,e_{p1})^d,\ldots,(e_{1n(1)}, e_{2n(2)},\ldots,e_{pn(p)})^d\}$.

Při hledání $D_a(S)$ je nutné využít duální representace modelů z S, t.j. $m = (e_{11},\ldots,e_{1m(i)})^d$ pro $i = 1,\ldots,p$. Pak $D_a(S) = \min \{(e'_{11},e'_{21},\ldots,e'_{p1})^p, (e'_{12},e'_{21},\ldots,e'_{p1})^p, \ldots, (e'_{1n(1)},e'_{2n(2)},\ldots,e'_{in(i)})^p\}$.

Vytváření $D_r(S)$ i $D_a(S)$ je **asociativní** v následujícím smyslu: $D_r(S_1 \cup S_2) = \max \{s \wedge t;\ s \in D_r(S_1),\ t \in D_r(S)\}$ a $D_a(S_1 \cup S_2) = \max \{s \vee t;\ s \in D_a(S_1),\ t \in D_a(S_2)\}$. Horáková (1989) definuje takto novou operaci: $D_r(S_1 \cup S_2) = D_r(S_1) \hat{\wedge} D_r(S)$ a $D_a(S_1 \cup S_2) = D_a(S_1) \hat{\wedge} D_a(S_2)$.

Popišme nyní formálněji proceduru 2.5:

V každém kroku jsou A a R množiny modelů, které byly **akceptovány** resp. **zamítnuty** ($A \cap R = \emptyset$).

Krok 1: Vstupem je nesrovnatelná množina modelů S_0.
Modely z S_0 jsou klasifikovány do A a R (přitom $A \cup R = S$, $A \cap R = \emptyset$).

Krok 2: Je-li $A = \emptyset$ jdi na krok 4, je-li $R = \emptyset$ jdi na krok 3, jinak zvol mezi krokem 3 a krokem 4..

Krok 3: Klasifikuj modely z $D_r(A)-R$ jako zamítnuté (R_1) resp. akceptované (A_1). Je-li $D_r(A) = R_1$, **stop**, jinak polož $A := A \cup A_1$, $R := R \cup R_1$.

Krok 4: Klasifikuj modely z $D_a(R)-A$ jako zamítnuté (R_1) či akceptované (A_1). Je-li $D_a(R)-A = A_1$, **stop**, jinak $A = A \cup A_1$, $R = R \cup R_1$.

Proceduru jsme popsali jako **symbolickou** manipulaci, bez ohledu na způsob jakým jsou modely klasifikovány jako zamítnuté či akceptované.

Výsledkem procedury jsou množiny A a R takové, že $\mathcal{H}_1 = A^+ \cup R^-$. Z intuitivního hlediska je zřejmé, že by nebylo dobré, aby některý model byl současně (slabě) akceptován či (slabě) zamítnut. Musí být tedy $A^+ \cap R^- = \emptyset$.

Platnost $A^+ \cap R^- = \emptyset$ je zaručena, jestliže pro žádný model $m_1 \in A$ a žádný model $m_2 \in R$ neplatí $m_1 \leq m_2$.

Postačuje, aby tato podmínka byla splněna pro $A \cup R = S_0$ v prvním kroku, kroky 3 a 4 ji již nemohou porušit, např. pro $m \in D_r(A) - R$ je $m \in \max({}^cA^+)$ a tedy nemůže být $m \leq m'$ pro $m' \in R$ ani $m'' \leq m$ pro $m'' \in A$; bez ohledu na klasifikaci m jako přijatého či zamítnutého nemůže být podmínka porušena. Nejjednodušší **racionální** a interpretačně racionální je právě zvolit S_0 jako **nesrovnatelné**.

Konstatujme tři důležitá fakta o diskutované proceduře:

(i) Procedura skončí po konečném počtu kroků (neboť A a R monotonně rostou).

(ii)Není potřebné uchovávat A a R celé, stačí minimální a maximální prvky. V prvním kroku je $A := \min(A)$ a $R := \max(R)$. V kroku 3 a 4 pak klademe $A := \min(A \cup A_1)$ a $R := \max(R \cup R_1)$. Takto definované A a R již neobsahuje všechny akceptované či zamítnuté modely, ale stále je $A^+ \cup R^- = \mathcal{H}_1$.

(iii) Díky asociativitě je při hledání duálů možné duály pouze po kroku 3 a 4 upravovat, $D_r(A \cup A_1) = D_r(A) \wedge D_r(A_1 - A)$, $D_a(R \cup R_1) = D_a(R) \vee D_a(R_1 - R)$.

Volba v kroku 2 mezi krokem 3 a krokem 4 závisí na **ceně**, (ve smyslu např. počítačového času), jakou připisujeme klasifikaci modelu. Je-li klasifikování modelů drahé (nezávisle na modelu) je vhodné volit podle velikosti $D_r(A) - R$ resp. $D_a(R) - A$. (viz odd. 5).

Obecně závisí výsledné A a R na volbě S_1 a na rozhodnutích v kroku 2. Je-li však klasifikace modelů **koherentní** v tom smyslu, že nemůže nastat situace, že $m_1 \leq m_2$, m_1 akceptován a m_2 zamítnut, je výsledek procedury jednoznačně dán (vzhledem k dané klasifikaci); viz dále oddíl 4).

3. Algoritmus IPF a ověřování modelů v datech

Jak je celkem snadno vidět, je možné a vhodné rozdělit proceduru vyhledávání modelů na nejméně dvě úrovně - symbolickou a numericko-datovou. Nyní se budeme pro případ grafových a HLL-modelů věnovat právě numerické a datové úrovni.

Na rozdíl od jiných metod vyhledávání modelů /generování hypotéz/ zde je pevně dána množina veličin, jejichž strukturu závislosti uvažujeme. Proto je zpravidla možné přejít od datové matice ke kontingenční tabulce, která obsahuje všechny potřebné údaje. Pro tři veličiny A,B,C je to tedy tabulka $T_{ABC} = \{m_{ijk}\}$, kde m_{ijk} je četnost (frekvence) výskytu konfigurace v datech. Datová úroveň výpočtů není tedy zpravidla prováděna vícenásobně. Pro tabulky vyšších dimensí, zejména obsahují-li prázdná políčka (nulové frekvence) je ovšem možné uvažovat o co nejvýhodnější representaci. Je si nutné uvědomit, že navrženými metodami lze zpracovávat pouze omezený počet veličin, např. maximálně N=10: při dvouhodnotových veličinách má tabulka 1024 prvků, při trojhodnotových 59 049, při pětihodnotových 9 765 625 atd. Důvody nejsou ovšem pouze výpočetní (v oblasti potřebné paměti), ale zejména statistické - pro ověřování modelů používáme asymptotické testy, jejichž chování při nízkých četnostech není jasné.

3.1 Při posuzování toho **jak model souhlasí s daty** se používá srovnání napozorovaných četností s četnostmi odhadnutými za předpokladu platnosti modelu. Např. při nezávislosti dvou veličin jde o srovnávání m_{ij} s $m_{i.}m_{.j}/m$, kde m je celkový počet případů. Pro porovnání těchto četností je možné používat celé řady statistik (viz Bishop at al, 1975), z různých důvodu je vhodné soustředit pozornost na statistiku

$$G^2 = 2 \sum O \log \left(\frac{O}{E}\right),$$

kde O symbolizuje napozorované četnosti a E odhanuté (teoretické) četnosti při platnosti modelu. Měli bychom psát přesněji $G^2(\varphi)$ a $E(\varphi)$ pro model φ. Tato statistika má za standardních podmínek a za předpokladu platnosti modelu χ^2-rozložení se stupni volnosti odpovídajícím rozdílu počtu "volných" parametrů saturovaného modelu a počtu "volných" parametrů při daném testovaném modelu φ (jde o rozdíl dimenzí odpovídajících parametrických prostorů).

Abychom však mohli hodnotu $G^2(\varphi)$ vypočítat, je nutné **odhadnout parametry modelu** např. v HLL reprezentaci. Při použití principu maximální věrohodnosti vede tato úloha na minimalizování výrazu $G^2 = 2\sum O \log(\frac{O}{E})$, kde jak jsme řekli E symbolizuje odhadnuté četnosti při platnosti modelu (teoretické četnosti) a O napozorované četnosti v tabulce, t.j. např. pro n = 3 model (AB,AC,BC) jde o minimalizaci v parametrech $\theta, \lambda^A_i, \lambda^B_j, \lambda^C_k, \lambda^{AB}_{ij}, \lambda^{AC}_{ik}, \lambda^{BC}_{jk}$ za platnosti obvyklých omezení; E je pak pro danou hodnotu i,j,k rovno $mp_{ijk}(\hat{\theta}, \hat{\underline{\lambda}})$, kde m je celkový počet pozorování.

Ohodnocení modelu φ tedy probíhá dvoustupňově:

(i) nejprve je nutné odhadnout parametry a

(ii) pak posoudit velikost výrazu G^2.

Výpočetní problém spočívá v kroku (i), krok (ii) je spiše problémem statistickým.

Pro výpočet odhadů parametrů, t.j. je nalezení minima (a odpovídajících hodnot parametrů) se v této situaci používá algoritmus IPF (iterative proportional fitting).

Algoritmus IPF pracuje následujícím způsobem: Uvažme např. n = 3 a HLL model (AB,AC,BC), t.j. nejjednodušší případ, kdy je potřebné tento algoritmus používat. Postačující statistiky odpovídají generátorům AB,AC a BC a jsou to tři marginální tabulky $T_{AB} = \{m_{ij.}\}$, $T_{AC} = \{m_{i.k}\}$a $T_{BC} = \{m_{ijk}\}$ vytvořené sčítáním z tabulky $T_{ABC} = \{m_{ijk}\}$. Odhadované pravděpodobnosti budeme značit $\hat{p}_{ijk}$ a odpovídající teoretické četnosti $\hat{m}_{ijk} = m\hat{p}_{ijk}$, kde m je celkový počet objektů ($\sum_{i,j,k} m_{ijk}$). Dále položíme $\hat{m}_{.jk} = \sum_i \hat{m}_{ijk}$ atd.

Jako počáteční odhad položíme $\hat{m}^{(0)}_{ijk} = 1$. Pak upravujeme v prvním kroce postupně pro jednotlivé generátory :

$$\text{pro AB:} \qquad \hat{m}^{(1)}_{ijk} = \hat{m}^{(0)}_{ijk} \frac{m_{ij.}}{\hat{m}^{(0)}_{ij.}} \quad ,$$

$$\text{pro AC:} \qquad \hat{m}^{(2)}_{ijk} = \hat{m}^{(1)} \frac{m_{i.k}}{\hat{m}^{(1)}_{j.k}} \quad ,$$

$$\text{pro BC:} \qquad \hat{m}^{(3)}_{ijk} = \hat{m}^{(2)}_{ijk} \frac{m_{.jk}}{\hat{m}^{(2)}_{.jk}} \quad .$$

Další krok opět probíhá stejným způsobem opět přes generátory AB, AC a BC. Postup se zastavuje, je-li po dokončení úplného kroku

$$|\hat{m}^{(3r)}_{ijk} - \hat{m}^{(3r-1)}_{ijk}| \leq \delta \ ,$$

kde δ je předem zadaná konstanta (např. $\delta = 0.01$ či $\delta = 0.1$). Algoritmus pro obecný HLL či grafový model je z tohoto příkladu zřejmý. Algoritmus byl navržen Demingem a Stephanem (1940) a jeho konvergence dokázána Csiszárem (1975).

3.2. Je nutné si všimnout dvou věcí:

(a) pro práci algoritmu využíváme pouze **symbolické informace** o tom, které jsou generátory modelu (tedy generující sentenci; HLL vyjádření $\log p_{ijk} =$

... není potřebné)

(b) pro některé modely není použití IPF nutné, resp. IPF se zastaví po prvním úplném kroku.

Tato skutečnost platí pro **rozložitelné** modely (viz oddíl 3.1). Např. model AB,BC je rozložitelný. Pro konfiguraci AB počítáme

$$\hat{m}^{(1)}_{ijk} = \hat{m}^{(0)}_{ijk} * \frac{m_{ij}}{\hat{m}^{(0)}_{ijk}} = m_{ij.} \quad ,$$

pro konfiguraci BC počítáme

$$\hat{m}^{(2)}_{ijk} = \hat{m}^{(1)}_{ijk} * \frac{m_{jk}}{\hat{m}^{(1)}_{.jk}} = m_{ij.} * \frac{m_{.jk}}{m_{.j.}}$$

což je skutečně odhad při podmíněné nezávislosti A a C podmíněno B, t.j. za platnosti

$$p_{ijk}/p_{.j.} = (p_{ij.}/p_{.j.}) \cdot (p_{.jk}/p_{.jk})$$

dostáváme $m \hat{p}_{ijk} = m \cdot \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} = \frac{m_{ij.}\, m_{.jk}}{m_{.j.}}$.

V IPF bychom dále počítali

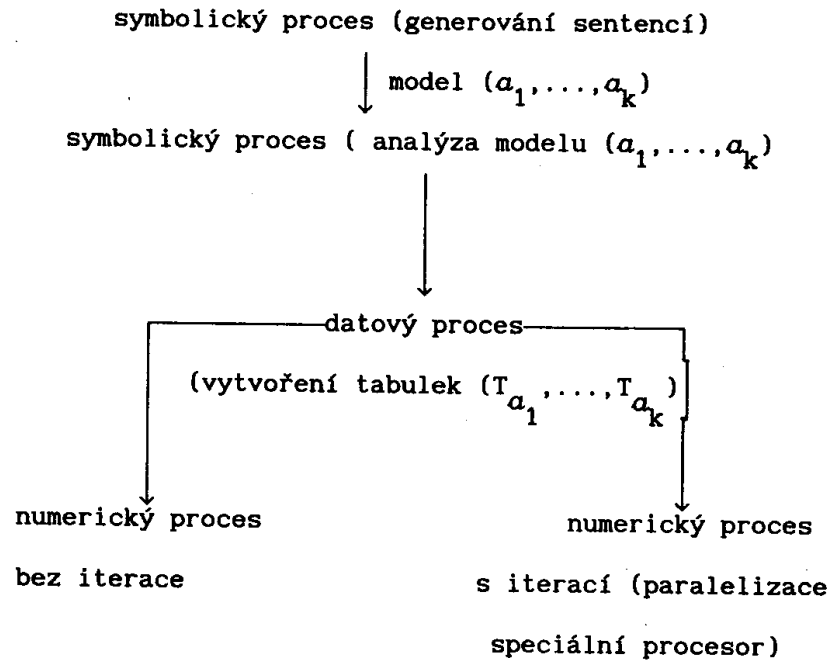
$$\hat{m}^{(3)}_{ijk} = \hat{m}^{(2)}_{ijk} * \frac{m_{ij}}{\hat{m}^{(2)}_{ij}} \quad ,$$

kde ovšem $m_{ij.}/\hat{m}^{(2)}_{ij.} = 1$.

Podíváme-li se na tuto situaci z výpočetního hlediska, je podstatné, že to, že IPF algoritmus se ihned zastaví, **víme předem** na **základě symbolické analýzy** modelu. Tento postup je realizován v programu MIMAS (Horáková, 1988b).

Budeme-li nyní uvažovat o možné paralelizaci algoritmu IPF, či využití speciálního procesoru, je zřejmé, že můžeme vyloučit předem zbytečnou komunikaci (viz obr. 1)

3.3 Vyhodnocování grafových a HLL modelů v datech je tedy v některých případech velmi jednoduché, v jiným může iterovat **velmi pomalu** (jiná věc je, že celý výpočet může být porušen výskytem nulových marginálních četností - to je opět další důvod pro omezení dimenze zpracovávaných tabulek). Pro nerozložitelné modely je tedy vhodné se pokud možno vyhnout ověřování modelu v datech (každé ověřování má ovšem svou prahovou cenu danou potřebnými datovými manipulacemi).



Obr. 1

Situace se stane ještě výraznější, kdybychom opustili používání obvyklých asymptotických kritických mezí (t.j. pro zamítnutí modelu je používáno $G^2(\varphi)>\chi^2(\varphi)$, kde $\chi^2(\varphi)$ je kritická hodnota χ^2-rozložení), a nastoupili cestu exaktních a pseudoexatních testů (Kreiner, 1987), kdy je nutné vytvářet buď všechny možné tabulky, nebo alespoň rozsáhlý náhodný výběr (např. 1 000 až 10 000) tabulek s danými marginálními četnostmi.

Je zajímavé, že i ve složitějších případech (např. smíšených hierarchických modelů, viz oddíl 5) je možné využívat symbolické analýzy modelu k rozhodnutí o vhodné metodě výpočtu, viz např. program MIM (Edwards, 1988, 1989, Fryndeberg a Edwards, 1989).

4. Koherence, kvasikoherence a paralelismus

Zopakujme, že v algoritmu 2.6 resp. 2.5 používáme dvou pravidel,které mají vlastně za účel **nahradit numerické/datové výpočty výpočty symbolickými**. Je definován pojem slabě zamítnutých a slabě akceptovaných modelů. Otázkou je, kdy výše zmíněná náhrada je zcela adekvátní.

4.1 Podmínkou proto je požadavek, aby rozhodovací pravidlo d, klasifikující modely při daných datech **M** jako akceptované (d(m,**M**) = a) či zamítnuté (d(m,**M**) = r) bylo **koherentní**:

pro žádní data **M** a pro žádný pár modelů m_1 a m_2 takových, že $m_1 \leq m_2$ nesmí nastat d(m,**M**) = a a d(m_2,**M**) = r.

Je-li rozhodovací pravidlo koherentní, pak:

(A) Model m je slabě zamítnut právě tehdy, může-li být zamítnut; podobně pro akceptování. Není tedy z interpretačního hlediska nutné rozlišovat mezi slabě zamítnutými a zamítnutými modely.

V důsledku toho platí dále:

(B) Výsledek práce algoritmu 2.6 nezávisí na volbě počáteční nesrovnatelné množiny modelů S_0, ani na rozhodování v kroku 2 mezi krokem 3 a 4.

Výsledek práce algoritmu je tedy při daných datech jednoznačně dán.

Je nutné si uvědomit, že výsledek práce je sice jednoznačně dán, ale počet ověřovaných modelů v datech, kterým můžeme měřit složitost práce algoritmu závisí při daných datech silně na volbě počáteční množiny S_0 i na rozhodnutích v kroce 2.

Zhruba řečeno, záleží na tom, jak je S_0 blízké minimálním akceptovaným modelům (je-li např. $S_0 = A$, práce končí). Proto je vhodné volit počáteční množinu modelů na základě některých předběžných testů (viz Benedetti a Brown, 1978).

Doporučená volba (volit 3 nebo 4) podle srovnání počtu modelů v $D_r(A)$-R resp. $D_a(R)$-A minimalizuje počet ověřovaných modelů v dalším kroku algoritmu, nikoliv však celkový počet ověřovaných modelů.

4.2 Naneštěstí v námi zatím uvažovaném případě grafových či HLL modelů, není obvyklé rozhodovací pravidlo:

$$d(m,M) = r, \text{ je-li } G^2(m,M) \geq \chi^2(m) \qquad (4)$$

koherentní. Pravidlo lze změnit na koherentní, nahradíme-li $\chi^2(m)$ závislé na počtu stupňů volnosti (na modelu) pevnou mezí, řekněme K. Protože platí, že $G^2(m_1,M) \geq G^2(m_2,M)$ pro $m_1 \leq m_2$, pravidlo

$$d'(m,M) = r, \text{ je-li } G^2(m,M) \geq K \qquad (5)$$

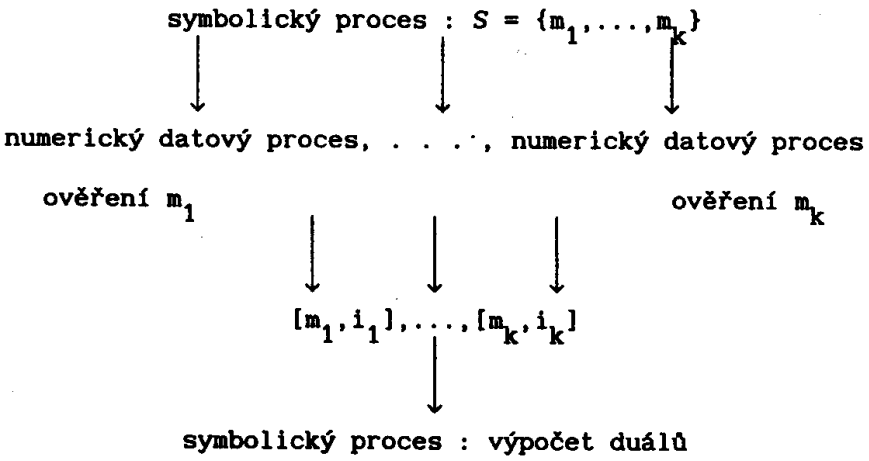
je pak koherentní. Toto pravidlo má velkou nevýhodu, má totiž malou rozlišovací schopnost (sílu), je-li k zvoleno tak, aby pro každý model šlo o (konzervativní) test dobré shody na hladině α (viz Havránek a Soudský, 1989): s velkou pravděpodobností může docházet K nezamítnutí "nesprávného" modelu. Stačí si uvědomit, že pak pro většinu modelů je $\chi^2(m) << K$.

Na druhé straně, na základě zkušenosti je možné soudit, že pravidlo (4) je **kvasikoherentní** (Edwards, Havránek, 1985). To znamená, že porušení koherentnosti není příliš časté. Tato otázka není dosud zdaleka prozkoumána; problém je komplikován tím, že je zde nutné počítat různé pravděpodobnosti zejména pro necentrální χ^2-rozložení. Dílčí výsledky jsou obsaženy v práci (Havránek, Pokorný, 1985).

4.3 Uvažujme nyní pro jednoduchost dále **koherentní případ.** Nejjednodušší forma paralelismu může při práci algoritmu 2.6 spočívat v tom, že v kroce 3 resp. 4 je vždy volána procedura ověření modelu v datech souběžně. Jde o klasickou fork operaci:

$$\text{call (ověření; } m_1,\ldots,m_k),$$

kde $m_1,\ldots,m_k$ jsou modely z S_0, $D_r(A)$-R či $D_a(R)$-A. Každá paralelné probíhající operace není ovšem primitivní; pro daný model m_i je vždy nutné vytvořit ze sdílených (a nemodifikovaných dat) protřebné marginální tabulky a pak aplikovatovat proceduru odhadu parametrů pro daný model a vyhodnotit rozhodovací kriterium (jde tedy o situaci vhodnou pro multiprocesorový systém se sdílenou pamětí a bez synchronizace). Symbolický proces čeká na "návrat" všech modelů, pak provádí výpočet duálů:



(zde i_j je a nebo r).

Jak již jsme řekli výše, numerické/datové procesy si musí vytvořit z celkové tabulky svoje vlastní kopie marginálních tabulek; pak již mohou pracovat (např. iterovat) zcela asynchronně .

Výpočet duálů je možné provádět rekursivním algoritmem (viz Horáková, 1988 a,b) s využitím skutečnosti, že

$$D_a(S_1 \cup S_2) = \min\{m_1 \vee m_2;\ m_1 \in D_a(S_1), m_2 \in D_a(S_2)\}$$

a

$$D_r(S_1 \cup S_2) = \max\{m_1 \wedge m_2;\ m_1 \in D_r(S_1), m_2 \in D_r(S_2)\}$$

(Edwards a Havránek, 1987). Jak jsme již řekli, můžeme formálně psát

$$D_a(S_1 \cup S_2) = D_a(S_1) \vee D_a(S_2)$$

a

$$D_r(S_1 \cup S_2) = D_r(S_1) \wedge D_r(S_2)$$

Výpočet a-duálu pak může být realizován voláním následující schématicky popsané rekursivní procedury (Horáková, 1988a):

procedura a-duál množiny modelů:

(var S : set of models, var h = $|S|$; integer);
begin $D_a(S) := 0$;
$S_1 := \{m_1, m_2, \ldots, m_{h/2}\}$;
$S_2 := \{m_{h/2+1}, \ldots, m_k\}$;
if $|S_1| > |$ then a-duál množiny modelů (S_1, h/2)
else if $|S_1| = 1$ then a-duál modelu (S_1);
if $|S_2| > 1$ then a-duál množiny modelů (S_2, h/2);
else it $|S_2| = 1$ then a-duál modelů (S_2);
end;

Podobně pro r-duál. Jde tedy o proces typu binárního stromu.

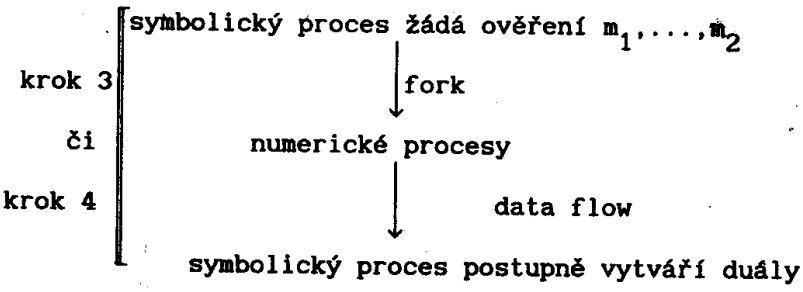
Při důsledné paralelizaci algoritmu 2.6 je však zřejmě výhodnější jiná cesta: k "návratu" ohodnocených modelů z numerických procesů $p_1(m_1), \ldots, p_k(m_k)$ nedochází v jednom čase, ale postupně; je tedy výhodnější, aby symbolický proces nečekal na návrat všech modelů, ale pracoval vždy po návratu libovolného modelu $[m_j, i_j]$ (data flow) s použitím

(a) výpočtu duálu modelu $D_a(m_j)$ nebo $D_r(m_j)$ podle hodnoty i_j,

(b) aplikace operace $D_a(R) \vee D_a(m_j)$ resp. $D_r(a) \wedge D_r(m_j)$.

Bod (a) a (b) probíhá pouze, je-li ověřeno, že není $m_j \leq m$ pro nějaké $m \in R$, resp. $m_j \geq m$ pro nějaké $m \in A$ (přitom probíhá odstraňování redundatních modelů v R i A).

Dojde-li v průběhu (a), (b) k návratu dalšího modelu, je uložen v zásobníku a zpracován až po ukončení práce symbolického procesu pro předchozí model. Zřejmě je možná situace, kdy pak je vhodné zpracovat více modelů zamítnutých (R_1) či akceptovaných (A_1) najednou. Pak je možné, aby symbolický proces (po ověření nerendundance) aplikovat rekursivní proceduru na $D_a(R_1)$ resp. $D_u(A_1)$ v bodě (a) a pak v (b) použil $D_a(R_1)$ místo $D_a(m_j)$. Situace je tedy:



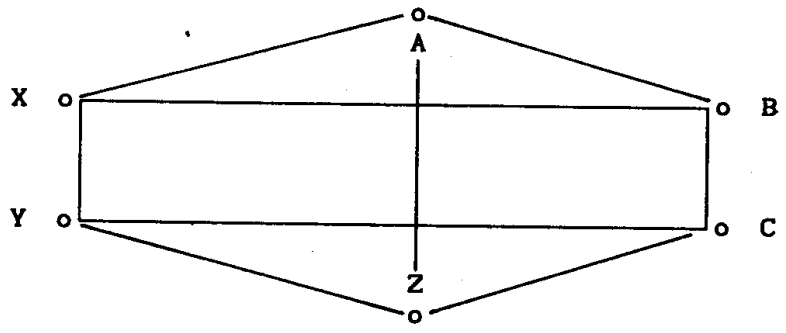
Vytvoření duálů je tedy řízeno návratem informace o zamítnutí či akceptování modelů (data flow); díky "asociativitě" není nutná synchronizace.

5. Zobecnění pro hierarchické a smíšené hierarchické modely

Úvahy předchozích oddílů se týkaly především grafových modelů, ale je možné je velmi snadno zobecnit pro další třídy modelů.

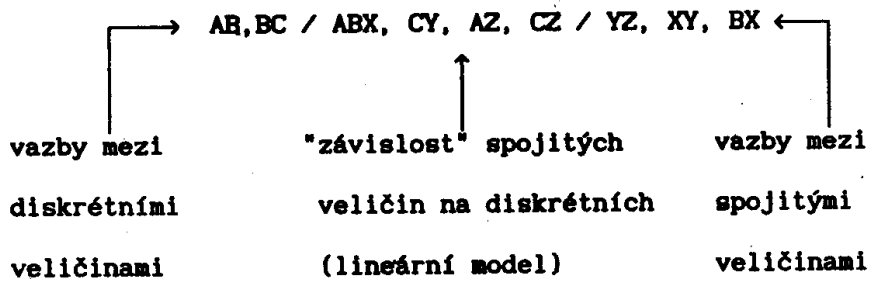
5.1 První z nich je **třída hierarchických logaritmicko-lineárních (HLL) modelů** (viz oddíl 1). Tyto modely na rozdíl od grafových netvoří Booleovskou algebru, ale pouze distributivní svaz. Algoritmus výpočtu duálů je pak modifikován způsobem popsaným v (Edwards a Havránek, 1987), ale byl popsán pro tento speciální případ již v (Edwards a Havránek, 1985).

5.2 Další **třídou je třída smíšených grafových modelů**, kdy zhruba řečeno uvažujeme vrcholy dvou typů - pro spojité veličiny (o - circle - continuous) a pro diskrétní veličiny (o - dot - discrete). Model je pak popsán grafem, např.



Statistické předpoklady vycházejí z toho, že podmíněné rozdělení spojitých veličin v modelu je vícerozměrné normální rozložení (pro každou kombinaci hodnot diskrétních veličin) a z toho, že chybějící hrana v grafu (stejně jako v grafovém) znamená podmíněnou nezávislou odpovídajících uzlů podmíněně ostatními.

Model může být zapsán opět pomocí generující sentence:



Přítomnost YZ znamá, že je nenulová vazba (korelace) mezi Y a Z, není však různá pro různé hodnoty diskrétních veličin; BX znamená, že variabilita X je různá pro různé hodnoty veličiny B.

Velmi podobně jako jsou HLL modely zobecněním grafových modelů pro diskrétní veličiny, lze zde definovat třídu hierarchických smíšených modelů, popsaných generujícími sentencemi výše zmíněného typu (Edwards, 1989).

Ověřování těchto modelů v datech (včetně odhadování parametrů) je poměrně složité; opět je nutné používat iterační procedury (Fryndeberg a Edwards, 1989). Pro ověřování jednotlivých modelů byl napsán program MIM (Mixed Interaction Models, Edwards, 1988). Při práci tohoto programu je využívána symbolická analýza modelu k volbě vhodného ověřovacího algoritmu i k řešení dalších problémů; ověřuje se symbolicky zda model je grafový, rozložitelný, lineární ve střední hodnotě, homogenní, grafový atd.

Program je realizován v jazyce TURBO PASCAL 5.0 na IBM PC (rozsah je 6000 řádek); využívá bohatě rekursivních konstrukcí navrhovaných Murphym (1986) a Horákovou (1988a,b).

Algoritmus výběru modelů pro třídu grafových smíšených modelů byl navžen na základě algoritmu 2.8 Horákovou (1988a,b) a realizován v programu MIMAS, který je podstatným rozšířením programu MIM právě v oblasti symbolických řídících struktur. Realizovaný postup využívá opět toho, že i zde jde o Booleovskou algebru, jejíž každý prvek je vyjadřitelný pomocí dvou typu atomů.

Literatura

Antoch, J.(1986): Algoritmic development in variable selection procedures, COMPSTAT'86, Physica-Verlag, Heidelberg, 83-90.

Antoch, J. (1987): Variable selection in linear model based on trimmed least squares estimator. Y.Dodge (ed.):Statistical Data Analysis Based on L_1-norm and Related Methods, Elsovier, Amsterdam, 231-245.

Antoch, J.(1985): Numerical behaviour of L-estimators in general linear model, Comp.Stat. Quaterly 1, 363-381.

Aitkin, M.A. (1974): Simultaneous inference and the choice of variable subsets in multiple regression, Technometric 16, 221-227.

Aitkin, M.A. (1980): A note on the selection of log-linear models, Biometrics 36, 173-178.

Benedetti, J.K., Brown, M.B. (1978): Strategies for the selection of log-linear models. Biometrics 34, 680-686.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975): Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, Mass.

Cziszar. I. (1975): I - divergence geometry of probability distributions and minimization problems, Ann.Probability 3, 146-158.

Darroch, J.N., Speed, T.P. (1978): Multiplicative and additive models and interactions. Dept.of theoretical statistics, res.rep.49, University of Aarhus.

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

Deming, W.E. (1940): On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, Ann.Math.Statist. 11, 427-444.

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

Edwards, D. (1988): A guide to MIM, Statistical Research Unit, Copenhagen University.

Edwards, D. (1989): Hierachical interactions models, J. Royal Statist.Soc. B (v tisku).

Edwards, D. a Havránek, T.(1985): A fast procedure for model search in multidimensional contingency tables, Biometrika 72, 339-351.

Edwards, D., Havránek, T. (1987): A fast model selection procedure for large families of models. J. Amer.Statist. Assoc.82, 205- 213.

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

Enke, H.(1980): To some reasonable test procedures in multiple contingency tables to investigate certain epidemiological or medicin-sociological relationships. Biometrical J. 22, 779-793.

Frydenberg, M., Edwards, D. (1989): A modified iterations scaling algorithm for estimation im regular exponential families, Computational Statistics and Data Analysis (v tisku).

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

Hájek, P., Havránek, T. (1978a): Mechanizing Hypothesis Formation. Springer-Verlag, Heidelberg (rusky: Nauka, Moskva 1984).

Hájek, P., Havránek, T. (1978b): The GUHA method - its aims and Techniques, Int.J.Man-Machine Studies 10, 193-308. Hájek, P., Havránek, T., Chytil, M. (1983): Metoda GUHA, Academia. Praha.

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

Havránek, T. (1982b): Some complexity considerations concerning hypotheses in multidimensional contingency tables, Trans. IX. Prague Conf.Inf.Theory, Statist. Dec. Func. and Rand. Processes, Academia, Praha, 281-286.

Havránek, T.(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.

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

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

Havránek, T. (1986a): O vyhledávání modelů, ROBUST 86, JČMF, Praha, 35-45.

Havránek, T. (1987a): Model search in large model families, invited lecture, Proc. of the First World Congress of the Bernoulli Society, VNU Press. Utrecht, vol. 2, 327-338.

Havránek, T. (1988): On a general algorithm for model choice in multivariate analysis, Statistics 19, 465-475.

Havránek, T. (1989): Vytváření struktur znalostí, in: J. Mařík, Z. Zdráhal (eds.): Metody umělé inteligence a expertní systémy IV, ČSVTS FEL ČVUT, Praha, 46-60.

Havránek, T. (1991): Statistika pro biologické a lékařské vědy, Academia, Praha (v tisku).

Havránek,T., Edwards, P.(1986): On variable selection and model choice in multivariate analysis, Proceedings of DIANA II, Matematický ústav ČSAV, Praha, 161-174.

Havránek, T., Pokorný, D. (1985): On the GUHA approach to model in connection to generalized linear models, Generalized linera models, R. Gilchrist, B. Francis, J. Whittaker (eds), Lecture Notes in Statistics 32, Springer-Verlag, Heidelberg, 82-92..

Havránek, T., Soudský, O.(1989): Model choice in the context of simultaneous inference, in: Y.Dodge(ed.),International Conference in Honor of C.R. Rao, North Holland, Amsterdam,165-176.

Horáková, M. (1988a): Structura závislosti dat. Práce ke kandidátskému minimu. SVT ČSAV, Praha.

Horáková, M. (1988b): Struktura závislosti dat, kandidátská disertační práce, SVT ČSAV, Praha.

Michálek, J., Němcová, M., Popelinský, L., Řebíčková, M.: The system for multivariate data processing, COMPSTAT'84, Physica-Verlag, 335-340.

Murphy, B.P. (1984): Similarities between linear, logistic and log-linear models for surviral analysis, COMPSTAT 84, Physica-Verlag, Wien, 378-382.

Murphy, B.P. (1989): Updating the sufficient configurations for fitting ZPA models for multidimensional contingency tables, Applied Statistics 38, 412-420.

Murphy, B.P., Rohl, J.S., Cribb, R.L. (1986): Recursive techniques in statistical programming, COMPSTAT 86, Physica-Verlag, Hieldelberg, 338-344

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

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

Sakamoto, Y, Akaike, H. (1978): Analysis of cross-classified data by AIC, Annals of the Inst. Statist. Math. 30, 185-197.

Sundeberg, R. (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.

Wermuth, N. (1976): Model search among multiplicative models, Biometrics 32, 253-263.

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