O VYHLEDÁVÁNÍ MODELŮ

Tomáš Havránek

SVT ČSAV, Praha

Poměrně často se můžeme ve statistice setkat se situací, kdy máme dánu množinu modelů /hypotéz/ a tuto množinu chceme rozdělit na modely, které můžeme na základě dat zamítnout, a na modely, které zamítnout nemůžeme /říkejme pracovně, že tyto modely akceptujeme - s vědomím jak málo toto akceptování znamená, použijeme-li jako rozhodovací pravidlo běžný test dobré shody/. Množina modelů může být snadno velká, například pro n regresorů v lineární regresi může obsahovat 2^n modelů.

Algoritmus popisovaný v tomto článku vykrystalizoval postupně ze speciálních případů grafových a obecně log-lineárních hierarchických modelů v mnohorozměrných kontingenčních tabulkách /Havránek, 1982a,b, 1984a,b a Edwards a Havránek, 1985a,b/. Pro práci algoritmu je důsledně využívána struktura množiny modelů daná částečným uspořádáním podle jejich vzájemných vztahů / jednodušší-složitější model/; této struktuře je věnována v článku pozornost, včetně zopakování potřebných matematických pojmů. Článek se opírá o práci /Edwards a Havránek, 1985b/ , obsahuje však navíc úvod do základních pojmů teorie svazů, příklady různých svazů, se kterými se ve statistice můžeme setkat, a příklady použití algoritmu v těchto jednotlivých speciálních případech. Neobsahuje důkazy a detailní popis aplikace algoritmu na lineární regresi včetně rozboru použitých testů. V oblasti log-lineárních a grafových modelů navazuje na předcházející články publikované ve sbornících konference ROBUST /Havránek, 1982a, 1984a/.

1. ZÁKLADNÍ POJMY

<u>Částečně uspořádaná množina</u> je neprázdná množina X spolu s relací $\leq$, kde platí pro každé $x,y,z \in X$:

A1/ $x \geq x$ /reflexivnost/,

A2/ je-li $x \geq y$ a $y \geq x$, pak $x = y$ /antisymetrie/,

A3/ je-li $x \geq y$, $y \geq z$, pak $x \geq z$ /transitivita/.

<u>Svaz</u> je částečně uspořádaná množina $(X,\leq)$,pro kterou platí:

B1/ pro každé $x,y \in X$ existuje největší dolní závora /označme si ji $x \wedge y$/,

B2/ pro každé $x,y \in X$ existuje nejmenší dolní závora /označme si ji $x \vee y$/.

Viz obr. 1: $(A,B) \vee (A,C)$ je (A,B,C) , $(A,B) \wedge (A,C)$ je (A) .

Na $\wedge$ a $\vee$ se můžeme dívat jako na binární operace; říkáme jim <u>průsek</u> a <u>spojení</u>. Tyto operace mají následující základní vlastnosti, které někdy bývají používány jako vlastnosti definující svaz:

C1/ $x \wedge x = x$, $x \vee x = x$,

C2/ $x \wedge y = y \wedge x$, $x \vee y = y \vee x$,

C3/ $(x \wedge y) \wedge z = x \wedge (y \wedge z)$, $(x \vee y) \vee z = x \vee (y \vee z)$,

C4/ $x \wedge (x \vee y) = x$, $x \vee (x \wedge y) = x$.

Je-li množina X konečná, nazýváme svaz $(X,\leq)$ <u>konečným</u> <u>svazem</u>.

Svaz se nazývá distributivní, platí-li :

D1/ $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$,

D2/ x y z x y x z .

Distributivní svaz je booleovský , jestliže obsahuje nejmenší prvek O a největší prvek I a pro každé x existuje právě jediné $-x \in X$ /doplněk/ tak, že $x \wedge -x = O$, $x \vee -x = I$. Příklad: systém všech podmnožin konečné množiny spolu s inklusí tvoří booleovský svaz. Průsek odpovídá průniku a spojení odpovídá sjednocení.

Definujeme nyní další pojmy /obecně ve svazu/:
Prvek x je $\wedge$-nerozložitelný, jestliže $z = x \wedge y$ má za důsledek, že $z = x$ nebo $z = y$. Podobně x je $\vee$-nerozložitelný, jestliže $z = x \vee y$ má za důsledek, že $z = x$ nebo $z = y$. Je-li $(X, \leq)$ konečný svaz , pak každý prvek x může být vyjádřen jako spojení $\vee$-nerozložitelných prvků a zároveň jako průsek $\wedge$-nerozložitelných prvků. Můžeme tedy napsat např. $x = e_1 \vee \ldots \vee e_n$, kde $e_1, \ldots, e_n$ jsou $\vee$-nerozložitelné. Důležitá jsou pouze taková vyjádření, kde $e_1, \ldots, e_n$ jsou nesrovnatelná; taková vyjádření jsou neredundantní.

Platí /Birkhoff, 1967, kap.IX/: Je-li konečný svaz distributivní, pak každý prvek x má jednoznačné neredundantní vyjádření $x = e_1 \vee \ldots \vee e_n$, kde $e_1, \ldots, e_n$ jsou $\vee$-nerozložitelné. Protože operace $\vee$ a $\wedge$ jsou duální, platí totéž i pro $\wedge$-nerozložitelné prvky a operaci průseku. Pochopitelně v konečném svazu je konečný počet $\vee$-nerozložitelných prvků $e_1, \ldots, e_n$ a konečný počet $\wedge$-nerozložitelných prvků $e^1, \ldots, e^m$. Obecně $n \neq m$.

Platí /Edwards a Havránek, 1985b/: Je-li konečný svaz distributivní, pak existuje následující jednoznačná korespondence mezi $\vee$-nerozložitelnými a $\wedge$-nerozložitelnými prvky. Označíme-li si e^i i-tý $\wedge$-nerozložitelný prvek, pak mu odpovídá v této korespondenci $\vee$-nerozložitelný prvek e_i takový, že

$$e_i = \min \{ x ; x \nleq e^i \} .$$

Důkaz je cvičení opírající se o skutečnosti z /Birkhoff, 1967, kap.IX/: Množina $I_e = \{x ; x \leq e\}$ je hlavní ideál, $J_f = \{x ; f \leq x\}$ hlavní filtr. Hlavní ideál je prvoideál, jestliže $X - I_e = J_f$ pro nějaké f. Podstatná je věta 8, §7, která říká, že hlavní ideál je prvoideál, právě když e je $\wedge$-nerozložitelné /pro $\vee$ dostáváme vše duálně/. Stačí nyní si uvědomit, že k e^i máme hlavní ideál I_{e^i} ; $X - I_{e^i}$ je pak jednoznačně určený hlavní filtr J_{e_i} .

Duálně platí $e^i = \max \{x ; e_i \nleq x \}$.

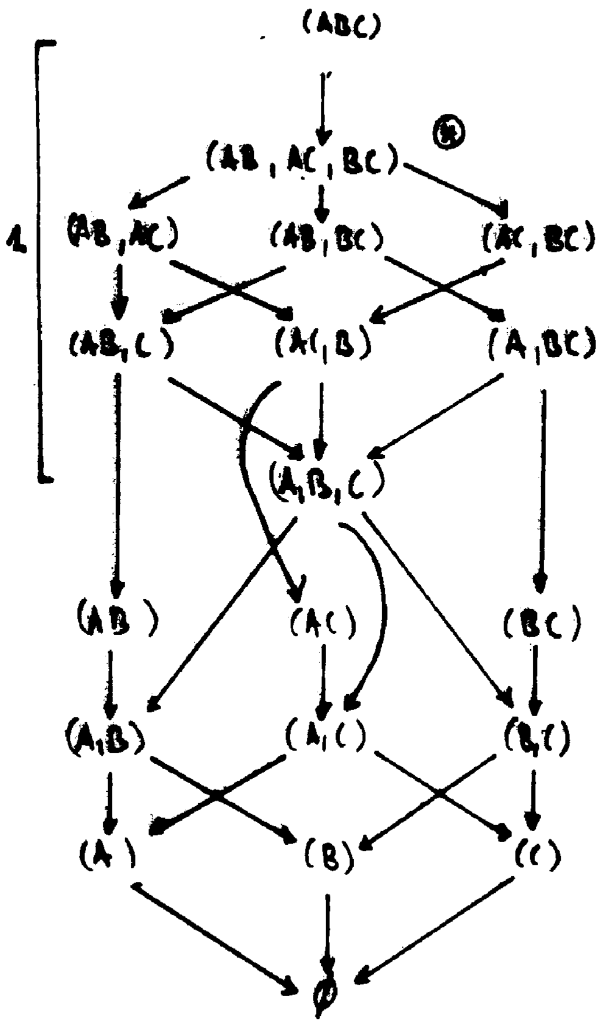
Rozkladový svaz: Uvažujeme /konečnou/ množinu Y. Nechť X je množina všech rozkladů této množiny / t.j. $x \in X$ jestliže $x = \{x_1, \ldots, x_n\}$, $\bigcup x_i = Y$, $x_i \cap x_j = \emptyset$ pro $i \neq j$ /. Definujme uspořádání $\leq$ na X takto: $x \leq y$ pro $x = \{x_1, \ldots, x_n\}$ a $y = \{y_1, \ldots, y_n\}$, je-li každé $x_i \supseteq y_j$ pro nějaké j / y je zjemněním x/ . Tento svaz $(X, \leq)$ se nazývá rozkladový svaz. V definici jsme použili obrácené uspořádání než je obvyklé v učebnicích o teorii svazů; námi použité uspořádání se lépe hodí ve statistické aplikaci.

Platí slavná věta /Pudlák a Tůma, 1980/: Každý konečný svaz může být vnořen do konečného rozkladového svazu /jde o injektivní $\vee, \wedge$ -homomorfismus/.

Protože, jak dále uvidíme, s rozkladovými svazy umíme dobře pracovat, zdálo by se, že by tato věta mohla být užitečná - uměli bychom každý konečný svaz /t.j. každou rozumnou konečnou množinu modelů/ vnořit do rozkladového svazu. Problém je tem, že zmíněná věta a její důkaz nedávají žádný konstruktivní návod, jak v konkrétním případě takové zobrazení hledat. Pro konstrukci konkrétních algoritmů se nám spíše hodí jednoduché skutečnosti až triviality z teorie svazů.

2. PŘÍKLADY SVAZŮ V ANALÝZE DAT

2.1 Hierarchické log-lineární modely v analýze kontingenčních tabulek. Uvažujme např. kontingenční tabulku dimenze $n = 4$, t.j. čtyři kategoriální veličiny A, B, C, D /při nezávislém multinomickém výběru/. Nechť p_{ijkl} je pravděpodobnost, že A nabyde hodnoty i, B hodnoty j atd.; předpokládejme, že $p_{ijkl} > 0$. Pak tyto pravděpodobnosti vyjadřujeme v log-lineárním rozvoji jako $\log p_{ijkl} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{ij}^{AC} + \lambda_{jk}^{BC} + \lambda_{kl}^{CD} + \lambda_{ijk}^{ABC}$. V tomto konkrétním případě jde o model podmíněné nezávislosti AB a D podmíněno C. Modely zapisujeme pomocí maximálních indexů /generujících tříd/ jako (ABC,CD) . Blíže viz /Havránek,1982a,1984a/. Svazové operace lze zde formálně definovat jako operace na generujících třídách, např. $(ABC,CD) \vee (ACD) = (ABC,ACD)$ a $(ABC,CD) \wedge (AB,C,D) = \cancel{(AB,C,D)}$. Pro $n = 3$ máme množinu modelů zobrazenu na obr.1.



Obr.1 log-lineární modely pro n=3. → označuje směr nerovnosti, relace plynoucí z transitivity neznačeny

Můžeme si zkontrolovat, že např. $(AB,C) \wedge (AC,B) = (A,B,C)$ a $(AB,C) \vee (AC,B) = (AB,AC)$. Celý svaz tedy obsahuje i modely kolapsovaných tabulek, t.j. i modely obsahující jen některé hlavní efekty; např. (AB)/ $\log p_{ijkl} = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$ / a $(\emptyset)$/ $\log p_{ijkl} = \theta$ /. Je to konečný distributivní svaz s nejmenším a největším prvkem $(\emptyset)$ a (ABC) . Přehled o $\vee$-nerozložitelných a $\wedge$-nerozložitelných prvcích dává tabulka 1. Tento svaz uvažuje Whittaker /1985b/.

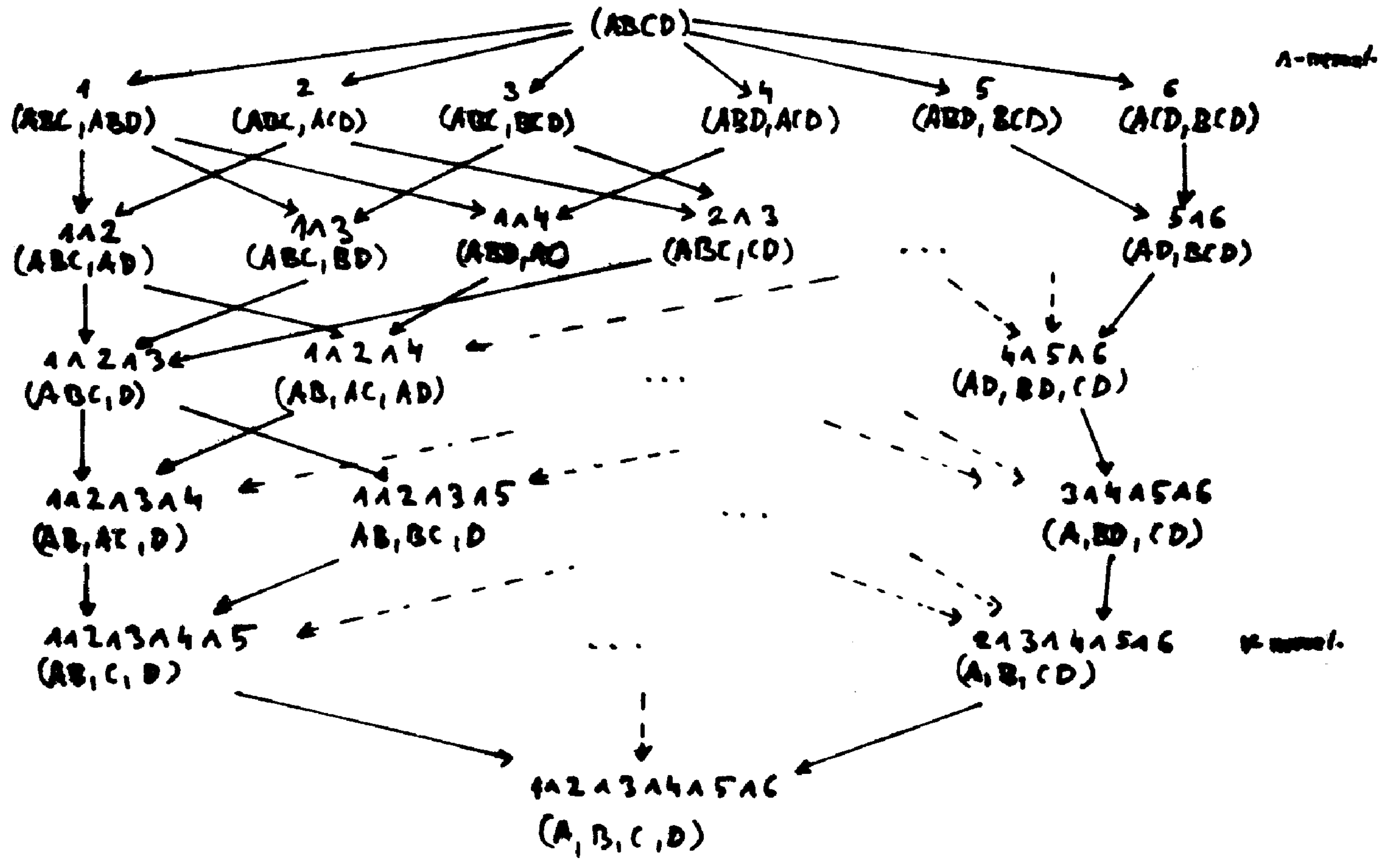
Důležitý je distributivní podsvaz ozančený 1 v obr.1 a obsahující pouze modely se všemi hlavními efekty $/\lambda^A, \lambda^B, \lambda^C/$. Tento svaz byl zkoumán v /Havránek,1982b/ K vyjádření ostatních prvků slouží $\wedge$-nerozložitelné prvky (AC,BC) , (AB,BC) a (AB,AC) a jim odpovídající $\vee$-nerozložitelné (A,BC) , (AC,B) a (AB,C).

Obecně pro tento podsvaz a obecnou dimenzi n jsou $\wedge$-nerozložitelne prvky sloužící k vyjádření ostatních právě prvky tvaru $(a_1,\ldots,a_k)$,kde každé a_i obsahuje právě n-1 písmen a $k \geq 2$. Je-li $e^i = (a_1,\ldots,a_k)$ takový $\wedge$-nerozložitelný prvek, pak odpovídající $\vee$-nerozložitelný prvek je $e_i = (b,A_1,\ldots,A_s)$,kde b je množina těch písmen, které chybí postupně v $a_1,\ldots,a_k$ a $A_1,\ldots,A_s$ jsou ostatní písmena. Např. $e^i = (ABCD,ABCE,ABDE)$ pro $n=5$,pak $e_i = (CDE,A,B)$ / je $e_i = \min\{ x; x \nleq e^i \}$ a musí tedy obsahovat CDE /.

Tabulka 1.

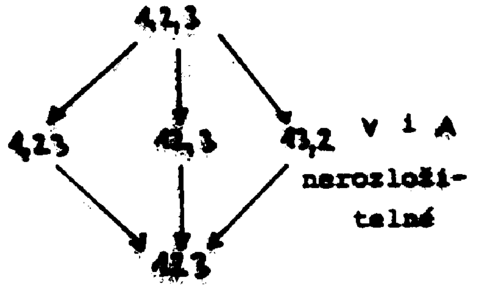
e_i (V-nerozložitelné)		e^i (Λ - nerozložitelné)
∅		ABC
A	pouze tyto prvky se používají k vyjádření ostatních pomocí	BC
B		AC
C		AB
AB		AC,BC
AC		AB,BC
BC	spojení / průseků	AB,AC
ABC		AB,AC,BC

2.2 <u>Grafové modely</u>. Vynechámeli z podsvazu 1 na obr.1 model ⑫ ,dostáváme svaz grafových modelů /obsahujících hlavní efekty/. Tento svaz je <u>booleovský</u> /a tedy distributivní/. Obecně lze uvnitř svazu hierarchických log-lineární modelů s hlavními efekty vymezit svaz grafových modelů jako svaz obsahující /1/ úplný saturovaný model /např. pro n=5 model (ABCDE) / a /2/ obsahující všechny modely vyjádřitelné průseky Λ-nerozložitelných prvků tvaru (a_1,a_2) / kde a_1 i a_2 obsahuje n-1 písmen/. Je-li $e^i=(a_1,a_2)$,kde a_i neobsahuje písmeno A_i,pak $e_i = (A_1A_2,A_3,\dots,A_n)$,kde $A_3,\dots,A_n$ jsou ostatní písmena. Např. při n=4 je $e^1=$(ABCD,ABCE) a $e_1=$(DE,A,B,C) . Na obr.2 je tento svaz,ve kterém je vyznačeno jak jsou jednotlivé modely generovány z Λ-nerozložitelných prvků.

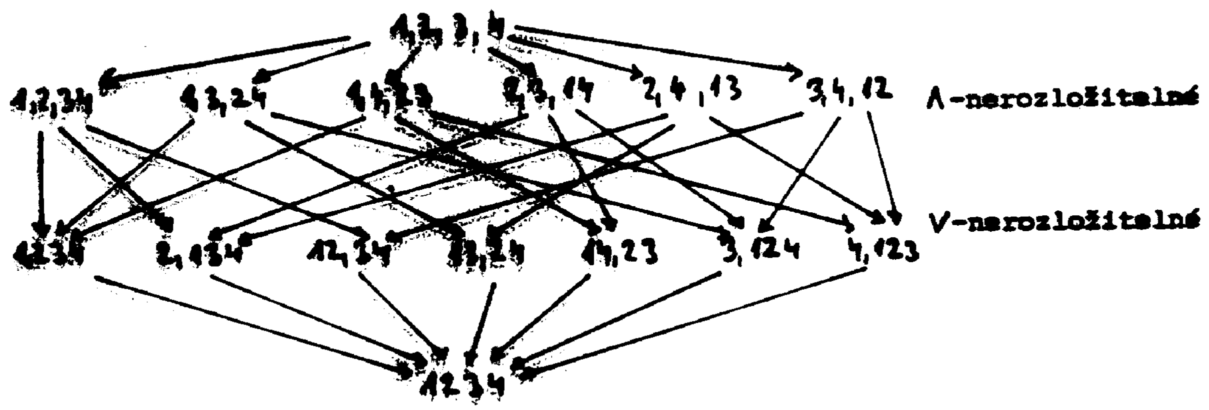


Obr.2 Grafové modely pro n = 4.

2.3 Rozkladové svazy v analýze rozptylu. Uvažujme zobecnění problému řešeného v /Cox a Spjøtvoll,1982/. Mějme t parametrů $\mu_1,\dots,\mu_t$ /např. t středních hodnot v t výběrech/. Hypotézy, které nás zajímají jsou tvaru $\mu_i = \Theta_k$ pro $i \in a_k, k = 1,\dots,q$ kde $a_1,\dots,a_q$ tvoří rozklad množiny $a = \{1,\dots,t\}$. Hypotéza /model/ je tedy dána rozkladem $\{a_1,\dots,a_q\}$ množiny $\{1,\dots,t\}$. Svaz tvoří prvky $x = \{a_1,\dots,a_q\}$ spolu s uspořádáním $\leq$: $x \leq y$, je-li $y = \{b_1,\dots,b_s\}$ zjemněním x. Je to konečný rozkladový svaz. Operace sjednocení a průseku mohou být popsány přímo takto: /1/ $x \vee y$ $\{a_1 \cap b_1, a_1 \cap b_2, \dots, a_q \cap b_s\}$ při vynechání prázdných průniků. /2/ $x \wedge y$ dostaneme ve dvou krocích; nejprve vytvoříme $\{a_1,\dots,a_q,b_1,\dots,b_s\}$ a pak sjednotíme nedisjuktní množiny v tomto výrazu. Modely tvaru $\{a_1,a_2\}$ jsou $\vee$-nerozložitelné, každý ostatní model /kromě $\{a\}$ / může být jednoznačně vyjádřen jako jejich spojení. Modely tvaru $\{a_1,\dots,a_{t-1}\}$ jsou $\wedge$-nerozložitelné / jedno z a_i obsahuje dva elementy, ostatní jsou jednoelementové/.

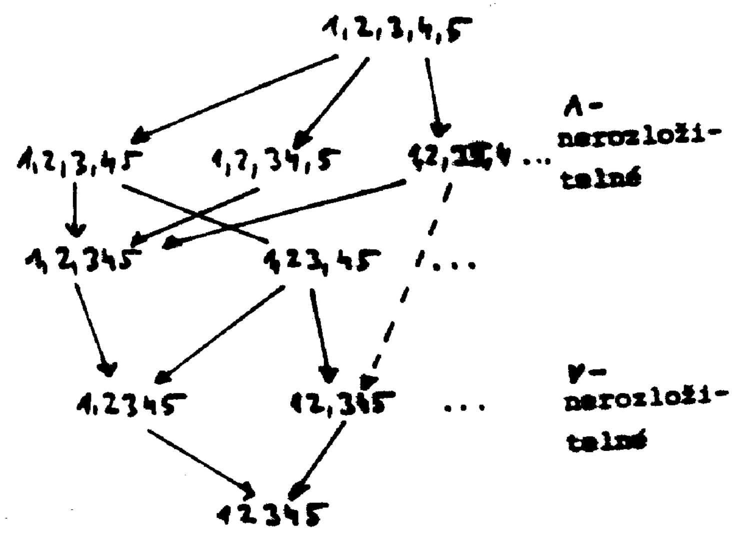


Obr.3 t = 3



Obr.4 t = 4

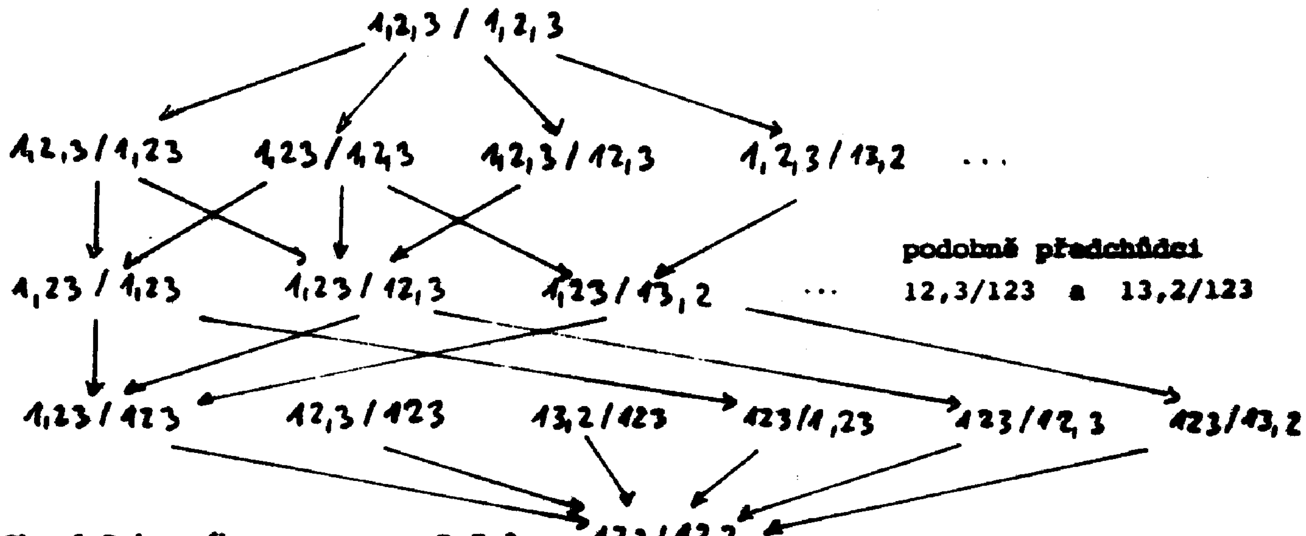
Pro t=5 uvažme, že $(123,4,5) = (12,3,4,5) \wedge (1,23,4,5) = (12,3,4,5) \wedge (13,2,4,5)$, takže zde není jednoznačná neredundantní representace prvků pomocí průseků $\wedge$-nerozložitelných prvků. Svaz tedy není distributivní. Na obr.3 je tento svaz pro t=3, na obr.4 pro t=4. Vidíme, že v této dimenzi je vše ještě jednoduché. Pro t=5 /obr.5/ je už situace složitější máme zde už jiné prvky než nerozložitelné.



Obr. 5 t = 5

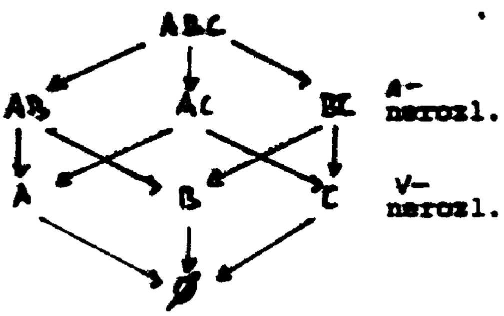
2.4 Pokorného svaz. Uvažujme kontingenční tabulku RXC. Kolapsovaná tabulka je dána rozkladem x na $\{1,\dots,R\}$ a rozkladem y na $\{1,\dots,C\}$ /přitom uvažujeme nejméně dvojčlenné rozklady/. Při kolapsování slučujeme ty řádky/sloupce tabulky, jejichž indexy jsou ve stejných členech rozkladů. Vzniklý svaz je součinem dvou rozkladových svazů:

$(x_1,y_1) \leq (x_2,y_2)$ jestliže $x_1 \leq x_2$ a $y_1 \leq y_2$. Pro R=C=3 máme tento svaz na obr. 6

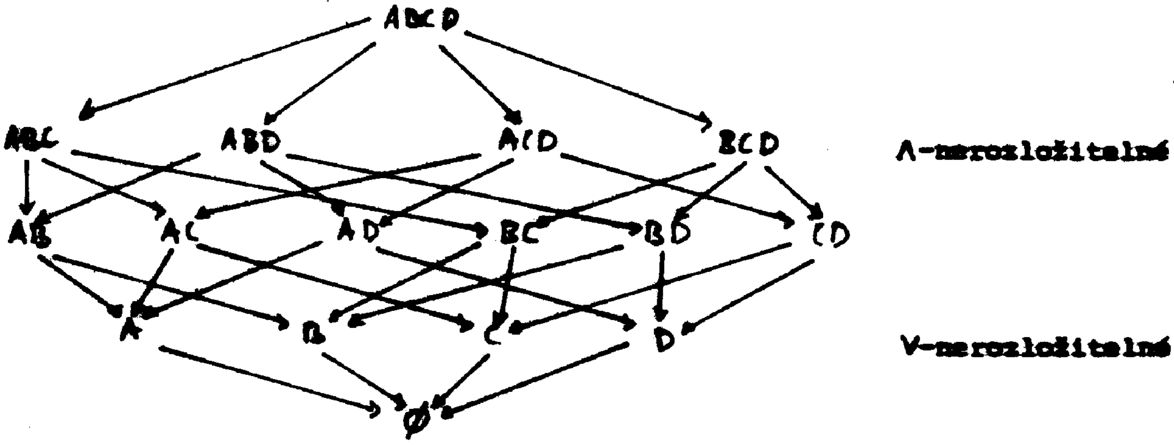


Obr.6 Pokorného svaz pro R=C=3

2.5 <u>Lineární regrese</u>. Uvažujme regresní modely $y=\mu+\beta_1 A+\beta_2 B+\dots\beta_n N+e$, kde A,B,C,...,N jsou "nezávislé" veličiny. Při problému výběru "nejlepší" množiny regresorů uvažujeme modely tvaru AB, BCD atd., t.j. různé množiny regresorů. Uspořádání $\leq$ je dáno inklusí $\subseteq$. Operace průseku a spojení odpovídají průniku a sjednocení. Důležité /t.j. sloužící k vyjádření ostatních/ Λ-nerozložitelné prvky jsou množiny o n-1 veličinách; je-li $e^i=(A_1,\dots,A_{i-1},A_{i+1},\dots,A_n)$, kde je vynecháno A_i, je $e_i=A_i$. Jde o velmi jednoduchý distributivní svaz. Pro n = 3 máme tento svaz na obr. 7 - je ekvivalentní se svazem grafových modelů pro n = 3. Pro n=4 máme tento svaz na obr.8 /zde už ekvivalenci s grafovými modely nemáme/.

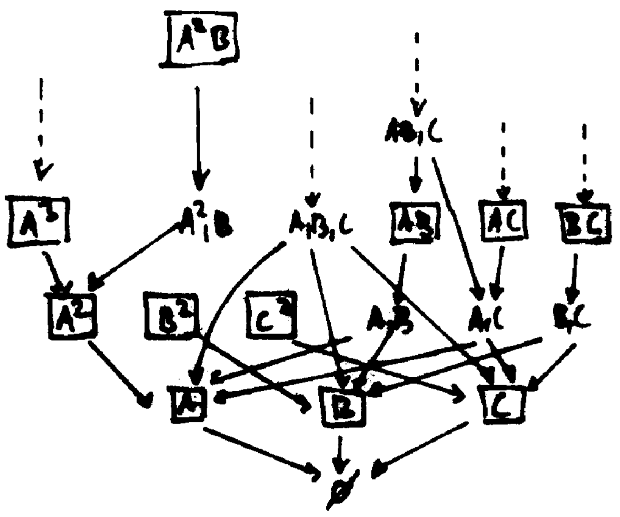


Obr.7 lineární regrese, n=3



Obr.8 lineární regrese , n= 4 .

2.6 <u>Obecná polynomiální regrese</u> /poněkud vyumělkovaný příklad/. Jde o regresi tvaru $y=\mu+\beta_1 A+\beta_2 A^2+\beta_3 B+\beta_4 A\cdot B$ ... + e. Regresory jsou tedy veličiny A,B,..., N ale i A^2, B^2, $A\cdot B$, $A^3\cdot B$ atd. Předpokládejme, že jde o hierarchický případ, t.j je-li v modelu A^2, je tam i A, je-li v modelu AB, je tam A i B, přítomnost

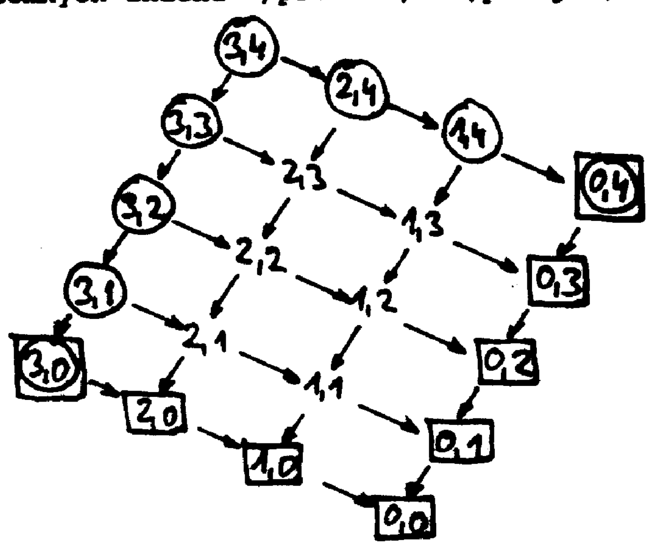


Obr.9 Polynomiální hierarchický svaz

A^2B znamená přítomnost AB i A^2 atd. Modely jsou popsány jako množiny maximálních elementů /regresorů/, obsažených v lineárním vyjádření modelu. Uspořádání $\leq$ odpovídá inklusi vzhledem k celým obsaženým množinám regresorů, ne jen k maximálním prvkům, svazové operace jsou pak sjednocení a průnik. Jde o nekonečný distributivní svaz, splňující podmínku, že každý řetěz má nejmenší prvek - každý prvek má pak jednoznačné vyjádření pomocí -nerozložitelných prvků. Tyto prvky jsou zde jednoelementové modely /ve zkráceném zápisu/, t.j. (A), (A^2), (AC), (BC), ... Viz obr.9 pro n=3 /nezávislé veličiny A,B,C /.

Pro nehierarchický případ bychom měli ω nerozložitelných prvků již v prvním patře : $A, A^2, A^3, \dots B, B^2, \dots$ AB, ... atd. Celkově by bylo 2^{ω} možných modelů.

2.7 ARMA modely jsou uvažovány v kontextu vyhledávání modelů v /Whittaker, 198 Nechť $\{x(t)\}$ je "autoregressive moving average" proces, t.j. $x(t) + a_1 x(t-1) + \dots a_p x(t-p) = e(t) + b_1 e(t-1) + \dots b_q e(t-q)$, kde $\{e(t)\}$ je gaussovský bílý šum. Jde o hledání modelů s nejmenším počtem parametrů. Odpovídající svaz je při použití hierarchického principu velmi jednoduchý. Jeho prvky jsou (i,j), t.j. dvojice nejvyšších přítomných indexů /pro i=0,...,p a j=0,...,q/. Uspořádání je : $(i,j) \leq (k,l)$ je-li $i \leq k$ a $j \leq l$. Spojení a průsek: $(i,j) \vee (k,l) = (\max(i,k), \max(j,l))$, $(i,j) \wedge (k,l) = (\min(i,k), \min(j,l))$, V-nerozložitelné jsou prvky tvaru (i,0),(0,j) , Λ-nerozložitelné jsou (i,q), (p,j). Pro p=3,q=4 viz obr. 10. Každý prvek lze jednoznačně vyjádřit pomocí V-nerozložitelných i Λ-nerozložitelných prvků: $(i,j) = (i,0) \vee (0,j) = (i,q) \wedge (p,j)$. Korespondence: $e_i = \min\{x; x \nleq e^i\}$, kde $e^i = (i,q)$. Např. $\min\{x; x \nleq (2,4)\} = (3,0)$, t.j. k (i,q) koresponduje (i+1,0), k (p,j) koresponduje (0,j+1) . Je to analogie Pokorného svazu; jde o součin dvou lineárních uspořádání $0 \leq 1 \leq \dots \leq p$, $0 \leq 1 \leq \dots \leq q$.



Obr.10 ARMA svaz

V-nerozložitelné Λ-nerozložitelné

3. ALGORITMUS VYHLEDÁVÁNÍ MODELŮ

Jde tedy o rozdělení množiny model na modely zamítnuté a nezamítnuté /akceptované/. Nechť $(X,\leq)$ je částečně uspořádaná množina modelů. Předpokládejme, že máme rozhodovací pravidlo d, které nám pro daná data M řekne zda daný model $x \in X$ máme zamítnout /r/ či akceptovat /a/ / $d(x,M)=a$ nebo r /. Přijmeme dva následující principy:

I/ Jestliže $x \leq y$ a $d(x,M)=a$, pak bychom měli přijmout i y /tj. složitější model/. Řekneme, že takový model y je <u>slabě</u> akceptovaný /aniž bychom počítali $d(y,M)$/.

II/ Jestliže $x \leq y$ a $d(y,M)=r$, pak bychom měli zamítnout i x /jednodušší model/. Říkáme, že x je slabě zamítnut.

Víme-li, že model je slabě akceptován nebo slabě zamítnut, pak již nepočítáme $d(x,M)$, t.j. neptáme se v datech zda ho můžeme zamítnout či akceptovat.

Předpokládejme nyní, že $(X,\leq)$ je <u>konečná</u>. Nechť $S \subseteq X$ je /nějaká/ množina nesrovnatelných modelů. Definujeme <u>a-duál</u> množiny S jako množinu nejjednodušších modelů z X, které nejsou obsaženy v žádném modelu z S, t.j. $D_a(S) = \min\{x;\ x \nleq y$ pro každé $y \in S\}$. Kdyby modely v S byly zamítnuté, pak $D_a(S)$ obsahuje nejjednodušší modely, které by ještě mohly být akceptovány. Podobně definujeme <u>r-duál</u> jako $D_r(S) = \max\{x;\ y \nleq x,\ y \in S\}$. Kdyby modely v S byly akceptovány, pak $D_r(S)$ obsahuje nejsložitější modely, které by ještě mohly být zamítnuty.

Je nutné si uvědomit, že díky I/ a II/ nám přináší nejvíce informace akceptování jednoduchých modelů a zamítnutí složitých modelů.

Příklady duálů: Lineární regrese pro $n=5$. $D_a(\{(ABD)\}) = \{(C),(E)\}$, $D_r(\{(ABD)\}) = \{(BCDE),(ACDE),(ABCE)\}$. Lineární regrese pro $n=4$ /obr.5/: $D_a(\{(AB),(AC)\}) = \{(D),(BC)\}$, $D_r(\{(AB),(AC)\}) = \{(BCD),(AD)\}$, $D_a(\{(BCD)\}) = \{(A)\}$, $D_r(\{(A)\}) = \{(BCD)\}$.

Nechť A je množina akceptovaných modelů /u určitém stadiu práce procedury/, nechť R je množina zamítnutých modelů. Předpokládejme, že nedochází ke sporu, t.j. není $x \leq y$, $x \in A, y \in R$ /navíc je lépe předpokládat, že každá z množin A a R je nesrovnatelná - pro vyloučení redundance/.

Platí: Nechť T je množina modelů, o kterých na základě A a R a principů I/ a II/ ještě neumíme nic říci, t.j. $T = \{x \in X;\ a \nleq x$ pro všechna $a \in A$ a $x \nleq r$ pro všechna $r \in R\}$. Pak $\max(T) = D_r(A) - R$ a $\min(T) = D_a(R) - A$.

Toto lemma odůvodňuje naši snahu zkoumat duály. Jde tedy vždy o "okraje" množiny T.

3.1 <u>Obecný algoritmus</u>.

<u>Vstup</u>: data M, konečná částečně uspořádaná množina modelů $(X,\leq)$, rozhodovací pravidlo d a počáteční nesrovnatelná množina modelů S_0.

1/ Počáteční množina S_0 je testována /t.j. pro každé $x \in S_0$ zjistíme $d(x,M)$ /. Položíme $A = \{x;\ d(x,M)=a\}$, $R = \{x;\ d(x,M)=r\}$.

2/ Je-li $A = \emptyset$ jdeme na 3/, je-li $R = \emptyset$, jdeme na 4/. Jinak <u>zvolíme</u> mezi 3/ a 4/.

3/ Testujeme modely z $D_r(A)-R$. Jsou-li všechny zamítnuty, <u>stop</u>. Jinak upravíme A a R /t.j. přidáme akceptované a zamítnuté modely a vynecháme redundantní vzhledem k $\leq$ / a jdeme na 2/.

4/ Testujeme modely z $D_a(R)-A$. Jsou-li všechny akceptovány, <u>stop</u>. Jinak upravíme A a R a jdeme na 2/.

V bodě 2/ se můžeme rozhodovat podle toho, zda je menší $D_r(A)-R$ či $D_a(R)-A$. Obecně je výsledek procedury /t.j. množiny A a R/ závislý na S_0 a rozhodnutích v bodě 2/. O jednoznačnosti viz 4.1.

Příklad. Lineární regrese pro $n=4$ / viz obr.8/. Položíme $S_0 = \{(A),(B),(C),(D)\}$ /t.j. nejjednodušší modely/.

1.krok: $R=\{(A),(B),(C)\}$, $A=\{(D)\}$. $D_a(R)=\{(D)\}$, $D_r(A)=\{(ABC)\}$. $D_a(R)-A=\emptyset$. Jdi na 3/.

2.krok: Testujeme (ABC), je akceptován, $R=\{(A),(B),(C)\}$, $A=\{(ABC),(D)\}$, $D_a(R)=\{(D)\}$,
D_r A AB AC BC .

3.krok: Testujeme $D_r(A)-R=\{(AB),(AC),(BC)\}$. $R=\{(AB),(C)\}$, $A=\{(AC),(D),(BC)\}$.
$D_a(R)=\{(D)\}$, $D_r(A)=\{(AB),(C)\}$, $D_a(R)-A=\emptyset$, $D_r(A)-R=\emptyset$, <u>stop</u>.

Při použití algoritmu je problémem <u>hledání duálů</u>. Je potřebné použít jinou metodu než přímo podle definice. K tomu lze využít svazovou strukturu.

3.2 <u>Algoritmus pro obecný /konečný/ svaz</u>.

A. Platí: Pro libovolné $S,T\subseteq X$, $S\cup T$ nesrovnatelné, je $D_r(S\cup T)=\max\{s\wedge t;\ s\in D_r(S),\ t\in D_r(T)\}$, $D_a(S\cup T)=\min\{s\vee t;\ s\in D_a(S),\ t\in D_a(T)\}$.

B. Platí: Nechť $P\subseteq X$ jsou $\wedge$-nerozložitelné prvky a $Q\subseteq X$ jsou $\vee$-nerozložitelné prvky. Nechť $x\in X$. Pak $D_r(x)=\max\{z\in P;\ x\nleq z\}$, $D_a(x)=\min\{z\in Q;\ z\nleq x\}$.

Máme tedy návod jak hledat pro $S=\{m_1,\dots,m_k\}$ duály: Hledáme-li např. $D_r(S)$, najdeme nejprve $D_r(m_i)$ pro $i=1,\dots,k$ pomocí $\wedge$-nerozložitelných prvků a pak položíme $D_r(S)=\max\{s_1\wedge\dots\wedge s_k;\ s_1\in D_r(m_1),\dots,s_k\in D_r(m_k)\}$.

V nekonečných svazech mohou být potíže, např díky nekonečnosti P a Q.

C. Platí: Nechť $x=x_1\wedge\dots\wedge x_q=\bigwedge x_i$ /nebo $x=y_1\vee\dots\vee y_h=\bigvee y_i$ /. Pak $D_a(\bigwedge x_i)=\min\{\bigcup D_a(x_i)\}$ / a $D_r(\bigvee y_i)=\max\{\bigcup D_r(y_i)\}$ /.

Nyní, je-li $(X,\leq)$ konečný svaz, kde tedy každé $x\in X$ lze vyjádřit pomocí $\wedge$-nerozložitelných i $\vee$-nerozložitelných prvků, vidíme, že nám stačí znát pouze r-duály pro $\vee$-nerozložitelné prvky a a-duály pro $\wedge$-nerozložitelné prvky. Pak můžeme pro každé m_i v S použít <u>přímo</u> C a pak skonstruovat např. $D_r(S)$ jako $\max\{s_1\wedge\dots\wedge s_k;$ kde s_i jsou přímo prvky duálů $\vee$-nerozložitelných prvků $\}$.

Celý tento postup může být ještě dosti složitý.

3.3 <u>Algoritmus pro konečný rozkladový svaz</u>.

Protože $D_r(x)=\max\{p\in P;\ x\nleq p\}$, vidíme, že $D_r(X)$ se skládá z $\wedge$-nerozložitelných prvků, jejichž dvou elementová složka obsahuje elementy, které v $x=\{a_1,\dots,a_k\}$ byly v různých a_i, a_j.

Podobně $D_a(x)=\min\{q\in Q;\ q\nleq x\}$ a tedy $D_a(x)$ se skládá z $\vee$-nerozložitelných prvků tvaru $\{b_1,b_2\}$ takových, že alespoň pro jedno a_i je $a_i\cap b_1$ i $a_i\cap b_2$ neprázdné.

Příklad. t=5 /Obr.5/. $x=\underset{\{a_1,a_2,a_3\}}{\{1,2,345\}}$ pak $D_r(x)=\{\{12,3,4,5\},\{13,2,4,5\},\{14,2,3,5\},\{15,2,3,4\},\{23,1,4,5\},\{24,1,3,5\},\{25,1,3,4\}\}$ a $D_a(x)=\{\{123,45\},\{124,35\},\{125,34\},\{1234,5\},\{13,245\},\dots\}$.

Zde tedy máme přímý <u>popis</u> $D_a(m_i)$ i $D_r(m_i)$ pro každé m_i z množiny S, pro kterou bychom chtěli konstruovat duály.

3.4 <u>Algoritmus pro konečný distributivní svaz</u> /či svaz s odpovídající 1-1 korespondencí mezi nerozložitelnými prvky/.

Nechť $e_1,\dots,e_n$ jsou $\vee$-nerozložitelné prvky a $e^1,\dots,e^n$ jim odpovídající $\wedge$-nerozložitelné prvky. Je-li nyní $x=\bigvee_{j\in W} e_j$ pro nějaké $W\subseteq\{1,\dots,n\}$, pak $D_r(x)=\{e^j\}_{j\in W}$. Podobně, je-li $x=\bigwedge_{j\in W} e^j$, pak $D_a(x)=\{e_j\}_{j\in W}$.

Příklad /viz obr.2, grafové modely n=4/.

$$m_1 = (ABC,D) = \underset{\substack{\updownarrow \\ e_1}}{(ABC,ABD)^{e^1}} \wedge \underset{\substack{\updownarrow \\ e_2}}{(ABC,ACD)^{e^2}} \wedge \underset{\substack{\updownarrow \\ e_3}}{(ABC,BCD)^{e^3}}$$

$$D_a(m_1) = \{(A,B,CD),\ (A,BD,C),\ (AD,B,C)\}$$

podobně $m_2 = (AB,AC,AD) = e^1 \wedge e^2 \wedge (ABD,ACD)^{e^4}$,

$D_a(m_2) = \{(A,B,CD),(A,BD,C),(A,BC,D)\}$. Je-li $R=\{m_1,m_2\}$, pak

$$D_a(R) = \min\{s\wedge t;\ s\in D_a(m_1),\ t\in D_a(m_2)\} = \min\{\underset{e_1}{(A,B,CD)},\ \underset{e_1\vee e_2}{(A,BD,CD)},\ \underset{e_1\vee e_4}{(A,BC,CD)},$$
$$\underset{e_2}{(A,BD,C)},\ \underset{e_2\vee e_4}{(A,BC,BD)},\ \underset{e_3\vee e_1}{(AD,B,CD)},\ \underset{e_3\vee e_2}{(AD,BD,C)},\ \underset{e_3\vee e_4}{(AD,BC,D)}\} = \{(A,B,CD),(A,BD,C),(AD,BC,D)\}.$$

Příklad. ARMA svaz, aplikujeme první krok algoritmu:
start $S_0=\{(1,0),(0,1)\}$, oba $\vee$-nerozložitelné.
A R

$D_r(A)=\{(0,4)\}$ korespondující $\wedge$-nerozložitelný; pro $D_a(R)$ musíme vyjádřit $(0,1)$ pomocí $\wedge$-nerozložitelných prvků: $(0,1)=(0,4)\wedge(3,1)$, $D_a((0,1)) = \{(1,0),(0,2)\}$, $D_a((0,1))-A=\{(0,2)\}$.

Můžeme si vakutku vybrat. Díky akceptování $(1,0)$ se pak pohybujeme rychle je po jedné větvi /viz obr.10/.

3.5 Algoritmus pro polynomiální regresi / příklad nekonečného svazu/. Zřejmé jsou $\vee$-nerozložitelné prvky, je možné hledat $D_a(x) = \min\{z\in Q;\ z \not\leq x\}$. Např. $D_a((AC,B,C)) = \{(A^2),(B^2),(C^2),(AB),(AC),(BC)\}$, t.j. budeme používat pouze bod 4/.

Příklad /viz obr.9/. Testujeme $S_0=\{(A),(B),(C)\}$. Nechť $A=\{(A)\}$, $R=\{(B),(C)\}$. $D_a((B))=\{(B^2),(C),(A)\}$, $D_a((C))=\{(C^2),(A),(B)\}$, $D_a(R)=\{(C^2),(A),(B,C),(B^2)\}$, $D_a(R)-A = \{(C^2),(B,C),(B^2)\}$ atd.

Zde je otázka, zda tento algoritmus je rozumně zvládnutelný; je zde nutné doplnit zvnějšku pravidlo pro zastavení.

4. STATISTICKÉ PROBLÉMY

4.1 Vraťme se k rozhodovacímu pravidlu. Mějme $(X,\leq)$ konečnou částečně uspořádanou množinu modelů. Rozhodovací pravidlo d je koherentní, jestliže pro žádná data M a žádné $x\leq y$ se nestane, že $d(x,M)=a$ a $d(y,M)=r$, t.j. nenastane případ, že by byl akceptován jednodušší model a zamítnut složitejší model. Je-li d koherentní, je výsledek práce algoritmu jednoznačný. Pro některé případy lze použít rozumně zdůvodněné koherentní pravidlo. Například pro lineární regresi lze použít pravidlo: $d(m,M)=a$, je-li $R^2_m(M) \geq 1-(1-R^2_I(M))(1+F_{1-\alpha}(n,N-n-1)/(N-n-1))$, kde R^2 je čtverec mnohonásobného koeficientu korelace / pro model m a pro model I se všemi regresory/ a N je rozsah dat M /viz Edwards a Havránek,1985b/.

Někdy máme možnost použít jako koherentní, tak nekoherentní pravidlo; to je případ hierarchických log-lineárních modelů, kde bychom mohli použít informační statistiku /chí-kvadrát poměru věrohodnosti/ s pevnou kritickou mezí pro všechny modely /t.j. nezávisloou na jednotlivých stupních volnosti/. Takové pravidlo by bylo koherentní, ale bylo by při běžném smyslu slova slabé. Použijeme-li kritické meze pro danou hladinu významnosti α závislé na stupních volnosti, je výsledné pravidlo nekoherentní /ale doufáme,že jen slabě nekoherentní/.

V takových případech používáme často pravidla,která nejsou koherentní,ale skoro koherentní,kde tímto vágním pojmem míníme,že doufáme,že pravděpodobnost nekoherence je malá. Ve skutečnosti o tom,nakolik jsou některé konkrétní pravidla nekoherentní, víme velmi málo. Určité drobné úvahy v tomto směru jsou v /Havránek a Pokorný,1985/.

4.2 Dalším problémem jsou řídká data; jde zhruba řečeno o to,že pro posuzování složitějších modelů nemáme často dostatek dat. Příkladem problémů,které mohou nastat,jsou numerické problémy v regresi,je-li příliš mnoho regresorů,či numerické problémy při odhadování parametrů složitějších hierarchických log-lineárních modelů. K těmto problémům přistupuje špatná asymptotika testů v mnohorozměrných kontingečních tabulkách právě v oblasti složitějších modelů.Celkově lze v takových situacích doporučit používat jednoduché startovací množiny / např. $S_0=\{(A),(B),(C),(D)\}$ v lineární regresi, či $S_0=\{(A,B,C,D)\}$ v grafových modelech / a v bodě 2/ se rozhodovat pro 4/ . Je možné doufat,že se algoritmus zastaví dříve,než dojde na složité modely.

4.3 Použijeme-li jako rozhodovací pravidlo test dobré shody na hladině α a je-li tento test koherentní,platí,že množina akceptovaných modelů po skončení práce procedury $A_1=\{x;\ y \leq x \text{ pro některé } y \in A\}$ obsahuje platnou hypotézu /model/ s pravděpodobností $1-\alpha$. Je to formulováno nepřesně,ale doufám,že je zřejmé o co jde. Vytváříme tedy jakousi konfidenční oblast pro skutečný model. Jsou-li testy skoro koherentní, platí toto tvrzení pouze "skoro". Problém je jistě velikost množiny A i A_1 . To souvisí se silou použitých testů pro danou nulovou hypotézu vzhledem k blízkým jednodušším alternativám. Otázka podrobnějšího hodnocení procedur tohoto typu je otevřená.Nebylo by vhodné konstruovat pro takové situace jiná rozhodovací pravidla ?

LITERATURA

G.Birkhoff/1967/: Lattice theory, Amer.Math.Soc.Coll.Publ.25,3rd.ed./rusky 1985/

D.R.Cox,E.Spjøtvoll /1982/: On partitioning means into groups, Scand.J.Statist. 9 147-152.

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

D.Edwards,T.Havránek /1985b/: A fast model selection procedure for large families of models /zasláno do tisku/.

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

T.Havránek /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.

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

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

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

P.Pudlák,J.Tůma /1980/: Every finite lattice can be embedded in a finite partition lattice, Algebra Universalis 10,74-95.

J.Whittaker /1985a/: Additive elements of ARMA models, J.Time Series Analysis /v tisku/.

J.Whittaker /1985b/: Factorisation,ireducible components and additive elements of log-linear models /zasláno do tisku/.