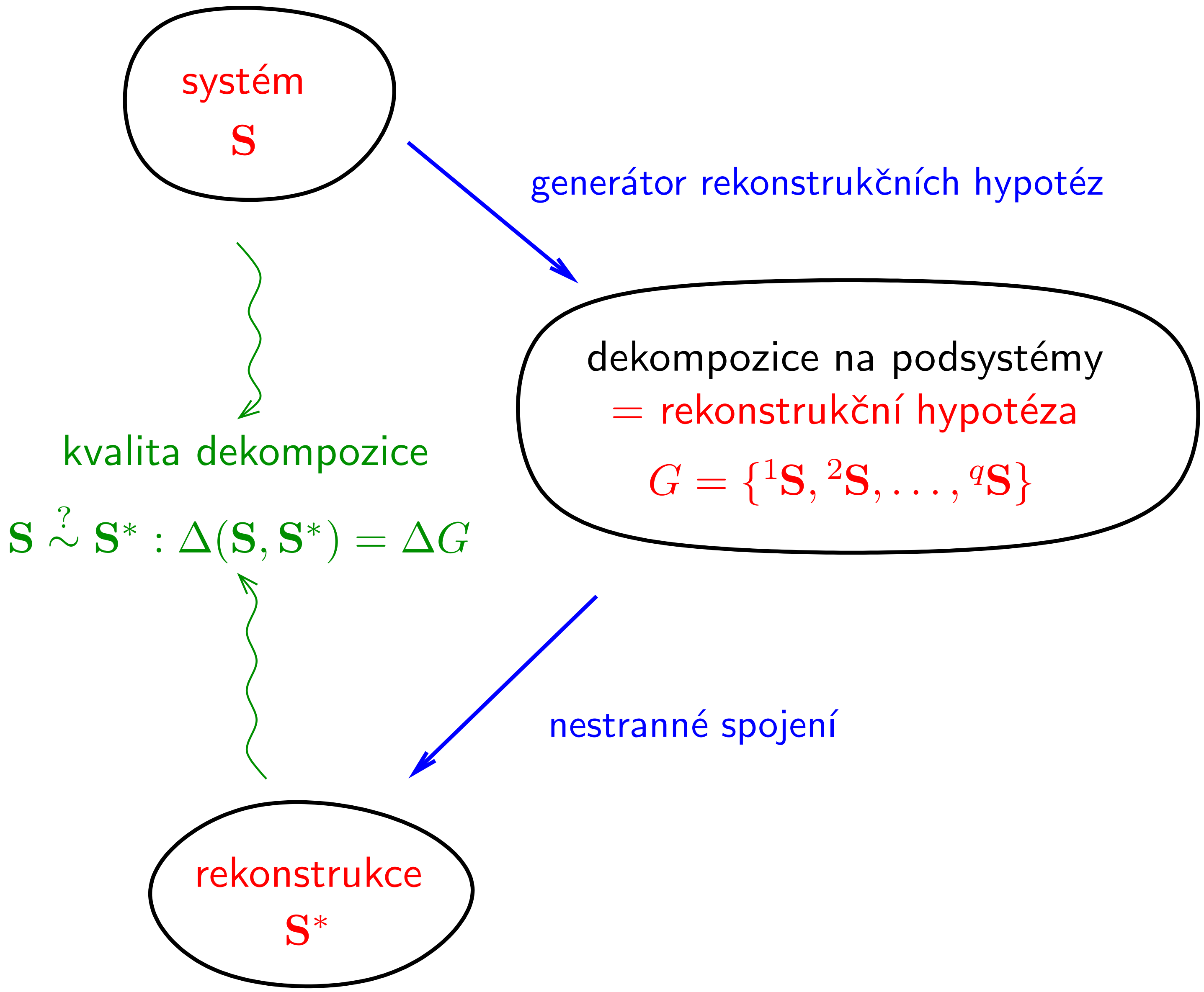
Schéma identifikační procedury



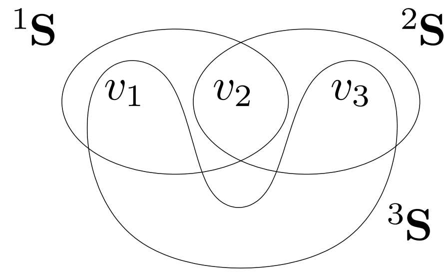
- G a $\mathbf{S}$ nejsou porovnatelné

  nelze srovnat kvalitu G a $\mathbf{S}$
- $\mathbf{S}^*$ a $\mathbf{S}$ jsou porovnatelné

Globální konzistence chování

Sdruženou funkci přípustnosti p_B musí být možno zkonstruovat z marginálních ${}^{i}p_B$, $i \in N_q$.

Př: Globálně nekonzistentní systémy:



Lokálně konzistentní:

v_1	v_2	${}^{1}p_B$
0	0	0 $(= a + x)$
0	1	0.7
1	0	0.3

v_2	v_3	${}^{2}p_B$
0	1	0.3
1	0	0.7
1	1	0 $(= b + y)$

v_1	v_3	${}^{3}p_B$
0	0	0.4
0	1	0.3 $(= a + b)$
1	0	0.3

v_1	v_2	v_3	p_B
0	0	1	a = 0
0	1	1	b = 0

$$a + x = 0 \Rightarrow a = x = 0 \qquad (a, x \geq 0)$$

$$b + y = 0 \Rightarrow b = y = 0 \qquad (y \geq 0)$$

$$a + b = 0.3$$

spor → nekonzistentní!

To je ve sporu.

Rekonstrukce systému

Dáno: Dekompozice systému $G = \{{}^1\mathbf{S}, {}^2\mathbf{S}, \ldots, {}^q\mathbf{S}\}$. ${}^i\mathbf{S}$ je podmnožina proměnných, ${}^i\mathbf{S} \subset \mathbf{S}$

Cíl: Nejlepší hypotéza o celkovém systému $\mathbf{S}$. ($s \in \mathcal{S}$ je stav a $|\mathcal{S}|$ je velikost stavového prostoru)

Postup:

1. **Určit množinu možných rekonstrukcí.**

$$
{}^ip_B({}^i\mathbf{S}) = \sum_{\mathbf{S} \setminus {}^i\mathbf{S}} p_B(\mathbf{S}) \qquad \sum_{i=1}^{q} |{}^i\mathcal{S}| \quad \text{rovnic}
$$

$$
p_B(\mathbf{S}) \geq 0 \qquad |\mathcal{S}| \quad \text{nerovnic}
$$

2. **Vybrat nejlepší z nich.**

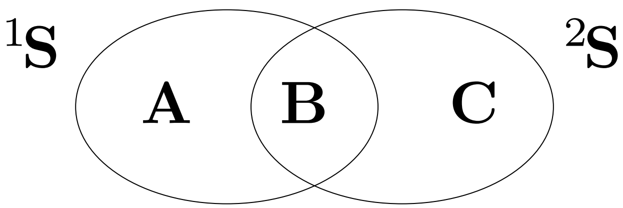
Volba (nestranná rekonstrukce):
$\mathbf{S}$ neobsahuje jinou informaci než tu obsaženou v množinách $\{{}^i\mathbf{S},\ i = 1, 2, \ldots, q\}$.

Implementace: Spojovací procedura.

Spojení dvou marginálních funkcí připustnosti

Marginální funkce připustnosti:

rozklad množiny proměnných:



$$^{1}p_B \colon R(\mathbf{A}) \times R(\mathbf{B}) \longrightarrow \langle 0, 1 \rangle$$

$$^{2}p_B \colon R(\mathbf{B}) \times R(\mathbf{C}) \longrightarrow \langle 0, 1 \rangle$$

zde budu používat malé a, b, c místo A, B, C i když význam zůstane stejný (množina proměnných). Pro jednoduchost si lze představit, že a, b, c jsou vždy jen jedna proměnná

Spojení: $\quad {}^{1}p_B * {}^{2}p_B \colon R(\mathbf{A}) \times R(\mathbf{B}) \times R(\mathbf{C}) \longrightarrow \langle 0, 1 \rangle$

Nestranné spojení (o maximální entropii):

$$p^{*}_B(a, b, c) = ({}^{1}p_B * {}^{2}p_B)(a, b, c) \stackrel{\text{def}}{=} {}^{1}p_B(a, b) \cdot {}^{2}p_B(c \mid b)$$

Pozn: ${}^{1}p_B(b) = {}^{2}p_B(b)$ (kompatibilita), ${}^{1}p_B * {}^{2}p_B = {}^{2}p_B * {}^{1}p_B$

Speciální případy:

$$\mathbf{A} = \emptyset: \quad ({}^{1}p_B * {}^{2}p_B)(b, c) = {}^{1}p_B(b) \cdot {}^{2}p_B(c \mid b)$$

$$\mathbf{B} = \emptyset: \quad ({}^{1}p_B * {}^{2}p_B)(a, c) = {}^{1}p_B(a) \cdot {}^{2}p_B(c)$$

Příklad

podsystémy: lokální kompatibilita:

a	b	${}^1p_B(a,b)$
0	0	0.4
0	1	0.3
1	0	0.2
1	1	0.1

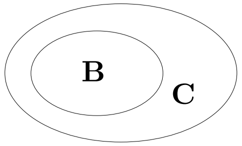
b	c	${}^2p_B(b,c)$
0	0	0.4
0	1	0.2
1	0	0.1
1	1	0.3

b	${}^1p_B(b)$	${}^2p_B(b)$
0	0.6	0.6
1	0.4	0.4

- vyplníme tabulku přípustnosti stavu pro $p^*_B(a,b,c) = {}^1p_B(a,b) * {}^2p_B(c \mid b)$

a	b	c	$p^*_B(a,b,c)$
0	0	0	$0.4 \cdot 0.4/0.6 = 0.266$
0	0	1	$0.4 \cdot 0.2/0.6 = 0.133$
0	1	0	0.075
0	1	1	0.225
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Poznámky



1. Připojením podsystému ‚se nestane nic'

$$({}^1p * {}^2p)(b,c) = {}^1p(b) \cdot {}^2p(c \mid b) = {}^2p(b) \cdot {}^2p(c \mid b) = {}^2p(b,c)$$

2. Je-li $p^*(a,b,c) = p(a,b) \cdot p(c \mid b)$, potom marginální funkce $p(a,b)$ a $p(b,c)$ jsou zachovány:

$$p^*(a,b) = \sum_{c_k} p^*(a,b,c_k) = \sum_{c_k} p(a,b) \cdot p(c_k \mid b) = p(a,b)$$

$$p^*(b,c) = \sum_{a_i} p^*(a_i,b,c) = \sum_{a_i} p(a_i,b) \cdot p(c \mid b) = p(b,c)$$

3. Marginální funkce $p(a,c)$ obecně zachována není.

a	b	c	p	p^*
0	0	0	0.0370	0.0106
0	0	1	0.1111	0.1376
0	1	0	0.0741	0.0684
0	1	1	0.3704	0.3761
1	0	0	0.0000	0.0265
1	0	1	0.3704	0.3439
1	1	0	0.0000	0.0057
1	1	1	0.0370	0.0313

a	c	p^*	p
0	0	0.0790	0.1111
0	1	0.5136	0.4815
1	0	0.0322	0.0000
1	1	0.3753	0.4074

Věta o maximalitě entropie spojení

Nechť $\bar{\mathbf{S}}$ s funkcí přípustnosti stavu $\bar{p}(a, b, c)$ je libovolná rekonstrukce systému $\mathbf{S}$ s funkcí přípustnosti stavu $p(a, b, c)$ a nechť $\mathbf{S}^*$ je rekonstrukce spojením $p^*(a, b, c) = p(a, b)\, p(c \mid b)$. Potom $H(\bar{\mathbf{S}}) \le H(\mathbf{S}^*)$.

Pozn: Platí i pro případ, kdy $\bar{\mathbf{S}}$ je původní systém $\mathbf{S}$.

Důkaz

Předpokládejme, že existuje $\bar{p}$ tak, že $H(\bar{\mathbf{S}}) > H(\mathbf{S}^*)$. Použijeme Gibbsovu nerovnost

$$H(\bar{\mathbf{S}}) = -\sum_{i,j,k} \bar{p}(a_i, b_j, c_k) \log \bar{p}(a_i, b_j, c_k) \le \boxed{-\sum_{i,j,k} \bar{p}(a_i, b_j, c_k) \log p^*(a_i, b_j, c_k)}$$

a to, že $p^*(a, b, c) = p(a, b) \cdot p(c \mid b)$, $\bar{p}(a, b) = p(a, b)$ a $\bar{p}(b, c) = p(b, c)$. Dostaneme $H(\bar{\mathbf{S}}) \le H(\mathbf{S}^*)$, což je spor.

budeme upravovat,
viz následující strana

$$-\sum_{i,j,k} \bar{p}(a_i,b_j,c_k)\log p(a_i,b_j)\,p(c_k|b_j)$$

$$= -\underbrace{\sum_{i,j,k} \bar{p}(a_i,b_j,c_k)\log p(a_i,b_j)}_{-\sum_{i,j}\underbrace{\bar{p}(a_i,b_j)}_{p(a_i,b_j)}\log p(a_i,b_j)} - \underbrace{\sum_{i,j,k} \bar{p}(a_i,b_j,c_k)\log p(c_k|b_j)}_{-\sum_{j,k}\underbrace{\bar{p}(b_j,c_k)}_{p(b_j,c_k)}\log p(c_k|b_j)}$$

$$= -\sum_{i,j} p(a_i,b_j)\log p(a_i,b_j) - \sum_{j,k} p(b_j,c_k)\log p(c_k|b_j)$$

$$1 = \sum_{k} p(c_k|b_j)$$

$$p(b,c) = p(c|b)\cdot p(b_j)$$

$$\sum_{i} p(a_i,b_j)$$

$$= -\sum_{i,j,k} p(a_i,b_j)\,p(c_k|b_j)\log p(a_i,b_j) - \sum_{j,k} p(c_k|b_j)\left(\sum_{i} p(a_i,b_j)\right)\log p(c_k|b_j)$$

$$= -\sum_{i,j,k} \underbrace{p(a_i,b_j)\,p(c_k|b_j)}_{p^*(a_i,b_j,c_k)}\log p(a_i,b_j)\cdot p(c_k|b_j) = -\sum_{i,j,k} p^*\cdot\log p^* = H(S^*)$$

QED

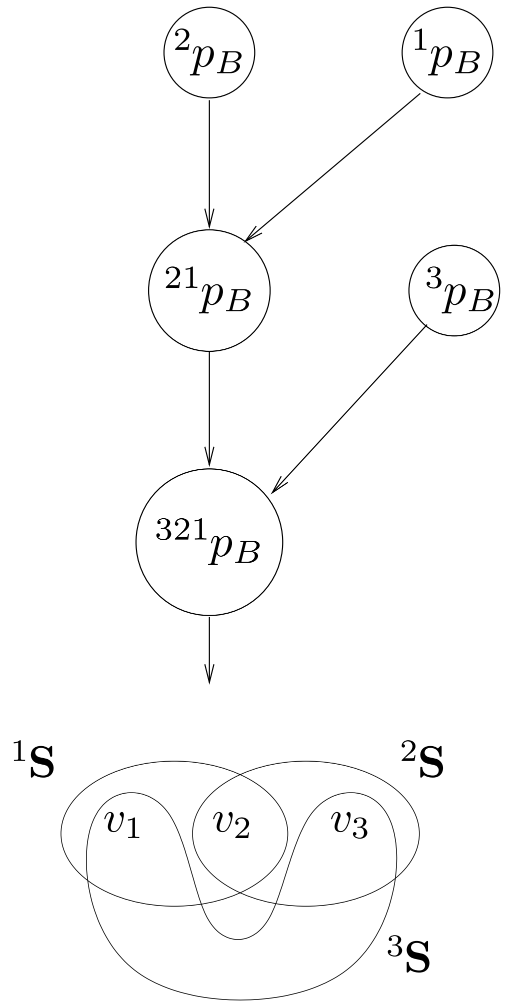
Spojovací procedura

Jednoduchá procedura pro $\{{}^1\mathbf{S}, {}^2\mathbf{S}, \ldots, {}^q\mathbf{S}\}$:

$$p_B^* = {}^q p_B * \left({}^{q-1}p_B * \left(\cdots * \left({}^2 p_B * {}^1 p_B\right)\right)\right)$$

libovolné pořadí spojování

Problém: cyklické vazby mezi proměnnými



Je-li splněna podmínka $[p_B^* \downarrow {}^i\mathbf{S}] = {}^i p_B$, potom je spojení nestranné. Jinak nutno použít iterativní proceduru (viz dále).

Iterativní spojovací procedura

Odstraní vliv cyklických vazeb.

Iterativní spojovací procedura $\mathbf{S}^*({}^1\mathbf{S}, {}^2\mathbf{S}, \dots, {}^q\mathbf{S}; \varepsilon)$

1. $k := 0, \qquad p_0^* := {}^q p_B * {}^{q-1} p_B * \cdots * {}^1 p_B$
2. $k := k + 1,\ p_k^* := {}^q p_B * {}^{q-1} p_B * \cdots * \left({}^1 p_B * p_{k-1}^*\right)$
3. jestliže $\max_{s \in \mathcal{S}} |p_k(s) - p_{k-1}(s)| < \varepsilon$ stop, jinak opakuj krok 2.

Jestliže $\sum_{s \in \mathcal{S}} p_k(s) = 1$, pak řešení nalezeno s chybou $\pm\varepsilon$, jinak $\{{}^1\mathbf{S}, {}^2\mathbf{S}, \dots, {}^q\mathbf{S}\}$ není globálně konzistentní.

Výsledkem je nestranná rekonstrukce s chybou ε

- nutná jen, pokud pro nějaké $i = 1, 2, \dots, q$ platí $\left[p_B \downarrow {}^i\mathbf{S}\right] \neq {}^i p_B$
- připojování nekonzistentních podsystémů ‚přepisováním marginálních funkcí'

$$ {}^2 p_B(b) \cdot {}^1 p_B(c \mid b) = \frac{{}^2 p_B(b)}{{}^1 p_B(b)} \cdot {}^1 p_B(b, c) $$

Důkaz konvergence: [Brown, *Information and Control* 4(2):386-393, 1959]

Příklad 1

Máme tři podsystémy systému s funkcí přípustnosti stavu p_{abc}

a	b	p_{ab}	b	c	p_{bc}	a	c	p_{ac}
0	0	0.1481	0	0	0.0370	0	0	0.1111
0	1	0.4444	0	1	0.4815	0	1	0.4815
1	0	0.3704	1	0	0.0741	1	0	0.0000
1	1	0.0370	1	1	0.4074	1	1	0.4074

Marginální funkce přípustnosti stavu po jednoduchém spojení $p_{ac} * (p_{ab} * p_{bc})$ p_{ac} se připojil poslední

a	b	$\hat{p}_{ab}$	b	c	$\hat{p}_{bc}$	a	c	$\hat{p}_{ac}$
0	0	0.1438	0	0	0.0149	0	0	0.1111
0	1	0.4487	0	1	0.5023	0	1	0.4815
1	0	0.3734	1	0	0.0962	0	0	0.0000
1	1	0.0340	1	1	0.3866	1	1	0.4074

Srovnání výsledku jednoduchého a iterativního spojení

a	b	c	p_{abc}	$p_{ac} * (p_{ab} * p_{bc})$	p^*_{abc}
0	0	0	0.0370	0.0149	0.0370
0	0	1	0.1111	0.0962	0.1111
0	1	0	0.0741	0.1290	0.0741
0	1	1	0.3704	0.3525	0.3704
1	0	1	0.3704	0.3734	0.3704
1	1	1	0.0370	0.0340	0.0370

Příklad 2

Máme tři podsystémy, které nejsou globálně konzistentní (viz str. 13)

v_1	v_2	${}^1p_B$	v_2	v_3	${}^2p_B$	v_1	v_3	${}^3p_B$
0	0	0.0000	0	0	0.0000	0	0	0.4000
0	1	0.7000	0	1	0.3000	0	1	0.3000
1	0	0.3000	1	0	0.7000	1	0	0.3000
1	1	0.0000	1	1	0.0000	1	1	0.0000

Výsledek jednoduchého i iterativního spojení

v_1	v_3	v_2	p_k
0	0	0	0.0000
0	0	1	0.4000
0	1	0	0.0000
0	1	1	0.0000
1	0	0	0.0000
1	0	1	0.0000
1	1	0	0.0000
1	1	1	0.0000

- podmínka $\sum_{s \in \mathcal{S}} p_k(s) = 1$ není splněna

Rekonstrukční chyba

Dvě rozdělení $p^*(s)$, $p(s)$, $s \in \mathcal{S}$ nad stavovým prostorem $\mathcal{S}$.

Kullbackova-Leiblerova vzdálenost

(informační vzdálenost, poměrná entropie)

$$L(p, p^*) = \sum_{s \in \mathcal{S}} p(s) \log \frac{p(s)}{p^*(s)}$$

Vlastnosti:

- $L(p, p^*) \geq 0$ (⇐ Gibbsova nerovnost)
- $L(p, p^*) \neq L(p^*, p)$ (není metrika)
- $\boxed{L(p, p^*) = 0 \quad \Leftrightarrow \quad p \equiv p^*}$ hlavní vlastnost
- je-li p^* nestranné spojení, pak
  1. $p^*(s) = 0 \;\Rightarrow\; p(s) = 0$
  2. $L(p, p^*) = H(\mathbf{S}^*) - H(\mathbf{S})$
  3. ale $L(p^*, p) \neq H(\mathbf{S}) - H(\mathbf{S}^*)$

$p^* = p(a,b) \cdot p(c|b)$

$\sum_k p(a,b,c_k) = p(a,b) = 0$

$\cancel{H(S^* | S)}$ nemá význam podmíněné pravděpodobnosti

Pro generativní systém

$$L(p, p^*) = \sum_{\overline{g} \in \overline{\mathbf{G}}} p(\overline{g}) \sum_{g \in \mathbf{G}} p(g \mid \overline{g}) \log \frac{p(g \mid \overline{g})}{p^*(g \mid \overline{g})}$$

$$= \cdots = H(\mathbf{G}^* \mid \overline{\mathbf{G}}^*) - H(\mathbf{G} \mid \overline{\mathbf{G}})$$

Rekonstrukční chyba: $\boxed{\Delta G = \Delta(\mathbf{S}, \mathbf{S}^*) = L(p, p^*)}$

Příklad

a	b	$p_B(a,b)$	$p^*_B(a,b) = {}^1p_B(a)\,{}^2p_B(b)$
0	0	0	0.21
0	1	0.3	0.09
1	0	0.7	0.49
1	1	0	0.21

a	${}^1p_B(a)$
0	0.3
1	0.7

b	${}^2p_B(b)$
0	0.7
1	0.3

$$L(\mathbf{S},\mathbf{S}^*) = 0.3 \log \frac{0.3}{0.09} + 0.7 \log \frac{0.7}{0.49} = 0.6109$$

$L(S, S^*)$

$$\cancel{H(\mathbf{S}^* \mid \mathbf{S})} = H(\mathbf{S}^*) - H(\mathbf{S}) = 1.2217 - 0.6109 = 0.6109$$

$$\begin{array}{cc|cc} & & \multicolumn{2}{c}{b} \\ & & 0 & 1 \\ \hline a & 0 & 0 & 0.3 \\ & 1 & 0.7 & 0 \end{array} \qquad \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} \qquad \begin{bmatrix} 0.21 & 0.09 \\ 0.49 & 0.21 \end{bmatrix}$$

$$\begin{bmatrix} 0.7 & 0.3 \end{bmatrix}$$

Test statistické významnosti struktury systému

$$H_0 : p(a_i, b_j, c_k) = ({}^1p * {}^2p)(a_i, b_j, c_k)$$

skutečná četnost $\quad n(a_i, b_j, c_k)$

teoretická četnost $\quad n^*(a_i, b_j, c_k) = n \cdot ({}^1p * {}^2p)(a_i, b_j, c_k)$

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n(a_i, b_j, c_k) - n^*(a_i, b_j, c_k)\right)^2}{n^*(a_i, b_j, c_k)}$$

- $I \cdot J \cdot K - J(I + K) + 1$ stupňů volnosti
- Lze jen u identifikace struktury systému (máme n i n^*)

poznámky

Počet stupňů volnosti

1. celý stavový prostor má velikost $I \cdot J \cdot K$
2. 1 podmínka
$$\sum_{i,j,k} p(a_i, b_j, c_k) = 1$$
3. $p(a, b)$ musí být marginální funkce: $I \cdot J - 1$ nezávislých vedlejších podmínek
   protože $\sum_{i,j} p(a_i, b_j) = 1$
4. $p(b, c)$ musí být marginální funkce: $J \cdot K - 1$ nezávislých vedlejších podmínek

$$I \cdot J \cdot K - 1 - (I \cdot J - 1) - (J \cdot K - 1) = I \cdot J \cdot K - J(I + K) + 1$$

[1] Mood, AM – Graybill, FA – Boes, DC. *Introduction to the Theory of Statistics.* McGraw-Hill 1974, str. 454ff.