Rekonstrukce diskrétního rozdělení psti metodou maximální entropie

Příklad Lze nalézt ‚četnosti' nepozorovaných stavů tak, abychom si „vymýšleli co nejméně"?

Nechť n_i, $i = 1, 2, \ldots, N$ jsou známé (absolutní) četnosti z neznámého diskrétního rozdělení pravděpodobnosti a nechť m_j, $j = 1, 2, \ldots, M$ jsou neznámé četnosti z tohoto rozdělení. Hledáme hodnoty m_j tak, aby entropie celého rozdělení byla maximální.

$$J = -\sum_{i=1}^{N} \frac{n_i}{n} \log \frac{n_i}{n} - \sum_{j=1}^{M} \frac{m_j}{n} \log \frac{m_j}{n} + \lambda (n - \sum_{i=1}^{N} n_i - \sum_{j=1}^{M} m_j),$$

neznámé jsou m_j, $j = 1, 2, \ldots, M$, n a λ.

Výsledek: ⊛1

$$m_j = m = n_0\, e^{-H_0}, \qquad \text{kde} \qquad n_0 = \sum_{i=1}^{N} n_i \qquad \text{a} \qquad H_0 = -\sum_{i=1}^{N} \frac{n_i}{n_0} \log \frac{n_i}{n_0}$$

$$H = H_0 + \log \frac{n}{n_0}, \qquad \text{kde} \qquad n = n_0 + M\, m$$

- m nevyjde jako celé číslo!

Rekonstrukce spojitého rozdělení ze známých momentů

Příklad

Nechť μ je střední hodnota a σ rozptyl neznámého rozdělení pravděpodobnosti s hustotou $f(x)$. Jak vypadá $f(x)\colon \mathbb{R} \mapsto [0,1]$, která má za těchto podmínek maximální diferenciální entropii?

$$Q = -\int_{-\infty}^{\infty} f(x)\log f(x)\mathrm{d}x + \lambda_0\left(1 - \int_{-\infty}^{\infty} f(x)\mathrm{d}x\right) + \lambda_1\left(\mu - \int_{-\infty}^{\infty} x f(x)\mathrm{d}x\right) + \\ + \lambda_2\left(\sigma^2 - \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x\right) \tag{1}$$

Výsledek

$$f(x) = \frac{1}{T}\, e^{a_1 x + a_2 x^2} = \cdots = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

Důkaz

1. pomocí Gibbsovy nerovnosti (viz dále)
2. pomocí variačního počtu z rovnice (1) ⊛1

Zobecnění předchozího výsledku

Nechť $g_j(x)$, $j = 1, 2, \ldots, K$ jsou známé funkce a $c_j = \int_{-\infty}^{\infty} g_j(x) f(x) \mathrm{d}x$ jsou známé hodnoty. Rekonstruujeme z nich neznámou hustotu $f(x)$ tak, aby maximalizovala entropii $H(x)$.

Výsledek

$$f(x) = \frac{1}{Z} e^{-\sum_{j=1}^{K} a_j g_j(x)}$$

$Z = Z(a_1, a_2, \ldots, a_K)$

a

$$H(x) = -\int_{-\infty}^{\infty} f(x) \log f(x) \mathrm{d}x = \log Z + \sum_{j=1}^{K} a_j c_j$$

Důkaz Nechť $\phi(x)$ je hustota rozdělení, které má větší entropii než $f(x)$ a vyhovuje podmínkám (má stejné hodnoty c_i). Platí varianta Gibbsovy nerovnosti

$$H_\phi(x) = -\int_{-\infty}^{\infty} \phi(x) \log \phi(x) \mathrm{d}x \leq -\int_{-\infty}^{\infty} \phi(x) \log f(x) \mathrm{d}x =$$

$$= \int_{-\infty}^{\infty} \phi(x) \left(\log Z + \sum_{j=1}^{K} a_j g_j(x) \right) \mathrm{d}x = \log Z + \sum_{j=1}^{K} a_j c_j = H_f(x), \quad \text{spor.}$$

$a_j \cdot \underbrace{\int \phi(x) g_j(x) \mathrm{d}x}_{c_j}$

pozn: $c_j = \int_{-\infty}^{\infty} g_j(x) \phi(x) \mathrm{d}x = \int_{-\infty}^{\infty} g_j(x) f(x) \mathrm{d}x$

Výpočet konstant a_j

Pro dané funkce $g_j(x)$ napíšeme vztah pro Z jako

$$Z(a_1, \ldots, a_K) = \int_{-\infty}^{\infty} e^{-\sum_{j=1}^{K} a_j g_j(x)} \mathrm{d}x$$

Platí

$$\frac{\partial Z}{\partial a_i} = \int_{-\infty}^{\infty} -g_i(x) e^{-\sum_{j=1}^{K} a_j g_j(x)} \mathrm{d}x = -Z \int_{-\infty}^{\infty} g_i(x) f(x) \mathrm{d}x = -Z\, c_i$$

Takže a_i jsou řešením soustavy nelineárních rovnic

$$-\frac{1}{Z}\frac{\partial Z}{\partial a_i} = c_i, \qquad i = 1, 2, \ldots K, \tag{3}$$

kde neznámé jsou a_i, kde c_i jsou dány a kde $g_i(x)$ jsou známé funkce.

Pozn

- soustava (3) je většinou obtížně řešitelná analyticky

⊛1 použijte pro důkaz výsledku (2) na str 49.

Identifikace struktury

Nejlepší dekompozice systému S na množinu podsystémů $G = \{{}^1S, {}^2S, \ldots, {}^qS\}$

Podproblémy

1. systematické generování rekonstrukčních hypotéz G
   - neredundance $\forall i, j \in \{1, 2, \ldots, q\}: {}^iS \not\sqsubseteq {}^jS$
   - pokrytí $\bigcup_i {}^iS = S$

   (S . . . množina proměnných systému)
2. Rekonstrukce celkového systému S^* z G.
3. Vyhodnocení rekonstrukční chyby $\Delta(G_0, G) = L(p, p^*)$ $G_0 = \{S\}$

Bez podmínky nerendundance a pokrytí by byl počet hypotéz

$$2^{2^n-1} - 1.$$

(n – počet proměnných, $n = 2 \Rightarrow$ 7 hypotéz, $n = 4 \Rightarrow$ 32767 hypotéz)

Zjemnění rekonstrukční hypotézy

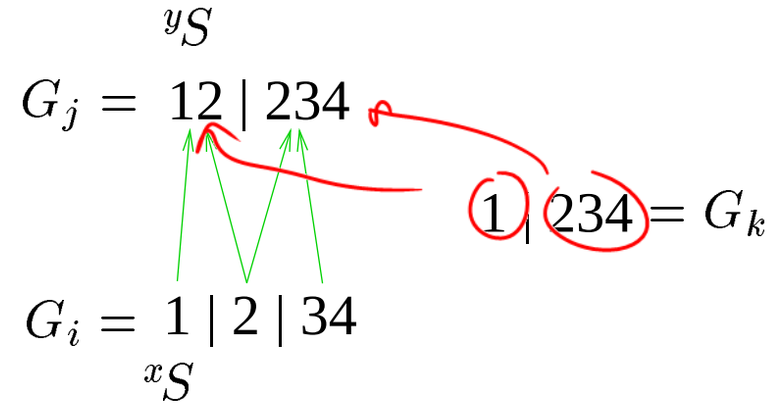
Rekonstrukční hypotéza $G = \{{}^1S, {}^2S, \ldots, {}^qS\}$

př: ${}^1S = \{s_1, s_2, s_3\}$, ${}^2S = \{s_2, s_4\}$, $G = \big\{\{s_1, s_2, s_3\}, \{s_2, s_4\}\big\}$, zkráceně $G = 123 \mid 24$

Zjemnění rekonstrukční hypotézy $\prec$

G_i je zjemněním G_j když pro každé ${}^xS \in G_i$ existuje ${}^yS \in G_j$ takové, že ${}^xS \sqsubseteq {}^yS$. Značíme

$$G_i \prec G_j$$



Bezprostřední zjemnění $\overset{*}{\prec}$

G_i je bezprostředním zjemněním G_j jestliže neexistuje G_k takové, že

$$G_i \prec G_k \quad \text{a} \quad G_k \prec G_j$$

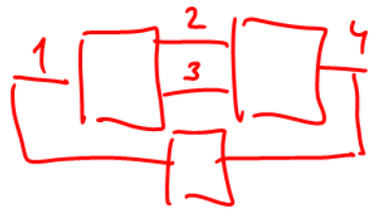
Relace bezprostředního zjemnění tvoří svaz na množině všech rekonstrukčních hypotéz.

Pozn: nejde o zjemňování rozkladu, rekonstrukční hypotézy netvoří rozklad (proměnné se opakují).

Generátor rekonstrukčních hypotéz

Vstup: Rekonstrukční hypotéza $G = \{{}^1S, {}^2S, \ldots, {}^qS\}$

Výstup: Všechna bezprostřední zjemnění G



Procedura (prohledávání svazu $(G, \overset{*}{\prec})$)

1. Pro $i = 1, 2, \ldots, q$ dělej kroky 2,3.
2. Jestliže $|{}^iS| \geq 2$ potom nahraď ${}^iS$ množinou všech podmnožin ${}^iS$ o velikosti $|{}^iS| - 1$.
3. Po odstranění redundantních podmnožin v každém rozkladu dostaneme seznam bezprostředních zjemnění G.

Pozn: může se opakovat na stejné úrovni zjemnění.

$$G = 123 \mid 234 \mid 14 = \{{}^1S, {}^2S, {}^3S\}$$

$$G_1 = 12 \mid 13 \mid 234 \mid 14$$

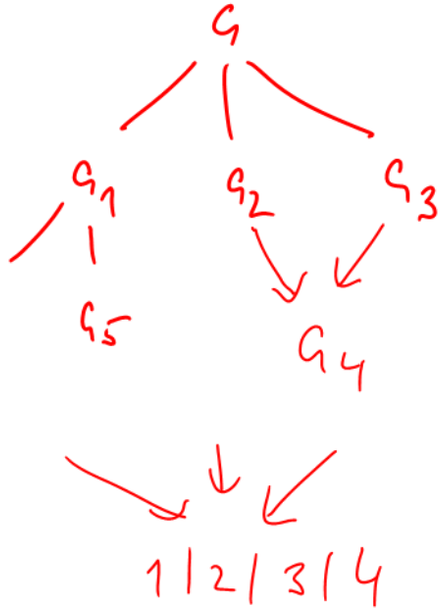
$$G_2 = 123 \mid 24 \mid 34 \mid 14$$

$$G_3 = 123 \mid 234$$

$$G_4 = 12 \mid 13 \mid 23 \mid 34 \mid 24 \mid 14$$

$$G_5 = 12 \mid 13 \mid 234$$

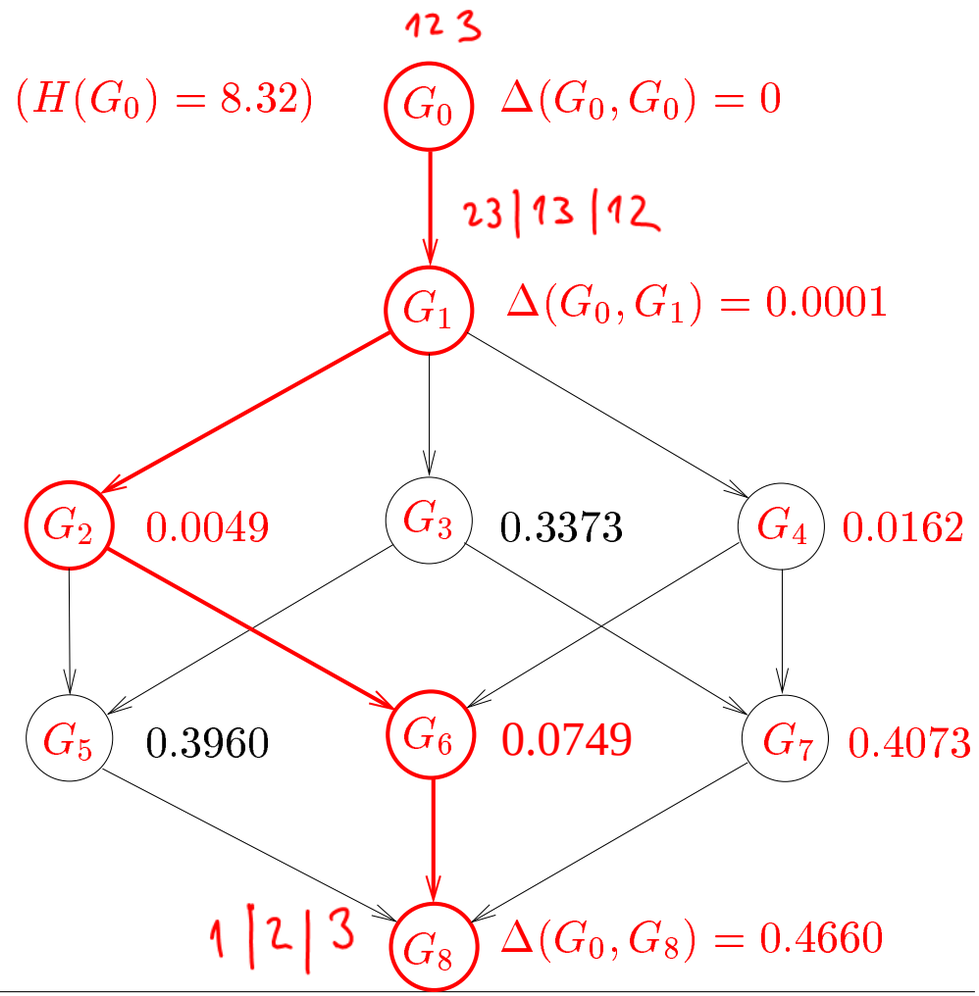
$$G_6 = 123 \mid 24 \mid 34$$



Prohledávání svazu zjemnění

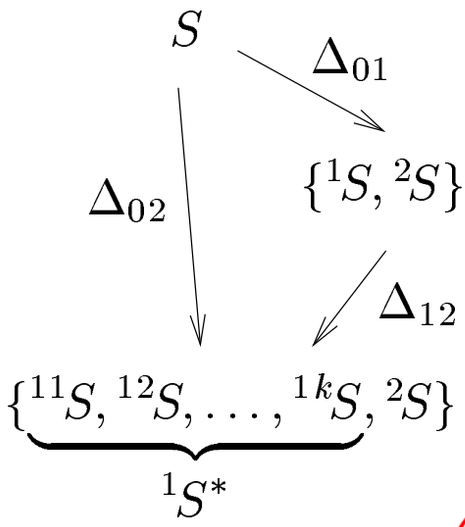
- Rekonstrukční chyba Δ je monotonně neklesající podél každé cesty svazu zjemnění: Je-li $G_i \succ G_k \succ G_j$ potom $\Delta(G_0, G_i) \leq \Delta(G_0, G_k) \leq \Delta(G_0, G_j)$
- Rekonstrukční chyba Δ je aditivní podél každé cesty svazu zjemnění: je-li $G_i \succ G_k \succ G_j$, potom $\Delta(G_i, G_j) = \Delta(G_i, G_k) + \Delta(G_k, G_j)$ [Higashi 1983]

- **Nejjednodušší postup:** Hledáme cestu, podle které Δ roste nejpomaleji
- **Složitější postup:** Na každé úrovni nás zajímá množina řešení s nejmenším Δ
  - Není nutno expandovat všechny větve: postačí metoda větví a mezí [Narenda&Fukunaga 1977]



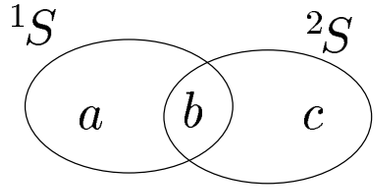
Aditivita podél cesty ve svazu zjemnění rozkladu

- pro jednoduchost předpokládáme, že jednoduchá spojovací procedura stačí k nestrannému spojení



$${}^1S^* = {}^{11}S * {}^{12}S * \cdots * {}^{1k}S$$

$$\Delta_{02} \stackrel{?}{=} \Delta_{01} + \Delta_{12}$$



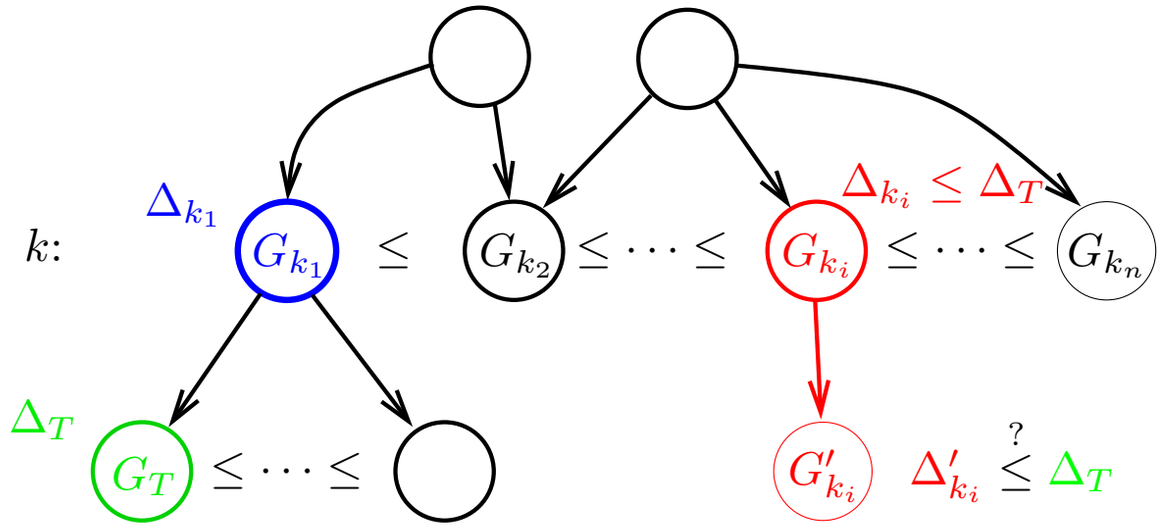
$$\Delta_{01} = L(p, {}^1p * {}^2p) = \sum_{i,j,k} p(a_i, b_j, c_k) \log \frac{p(a_i, b_j, c_k)}{{}^1p(a_i, b_j)\, {}^2p(c_k \mid b_j)}$$

$$\Delta_{12} = L({}^1p * {}^2p, {}^1p^* * {}^2p) \stackrel{!}{=} L({}^1p, {}^1p^*) = \sum_{i,j} {}^1p(a_i, b_j) \log \frac{{}^1p(a_i, b_j)}{{}^1p^*(a_i, b_j)} = \sum_{i,j,k} p(a_i, b_j, c_k) \log \frac{{}^1p(a_i, b_j)}{{}^1p^*(a_i, b_j)}$$

potom

$$\Delta_{01} + \Delta_{12} = \sum_{i,j,k} p(a_i, b_j, c_k) \log \frac{p(a_i, b_j, c_k)\, {}^1p(a_i, b_j)}{{}^1p(a_i, b_j)\, {}^2p(c_k \mid b_j)\, {}^1p^*(a_i, b_j)} = L(p, {}^1p^* * {}^2p) = \Delta_{02}$$

Metoda větví a mezí na svazu zjemnění



1. vygeneruj úroveň k
2. nalezni uzel G_{k_1} s nejmenším Δ a expanduj ho
3. najdi G_T s nejmenším Δ mezi expandovanými uzly
4. zvol práh $T = \Delta(G_T)$
5. expanduj $G_{k_2}, G_{k_3}, \ldots, G_{k_n}$ **jen pokud** $\Delta(G_{k_i}) \leq T$
6. (aktualizuj T, pokud nalezneš menší hodnotu prahu)
7. pokračuj stejně na úrovni $k+1$

- Takto lze jen proto, že platí monotonicita Δ!
- Toto není obecná metoda větví a mezí

Alternativní (neekvivalentní) postup:

1. založ haldu $H := (G_0; \Delta = 0)$
2. dokud $H \neq \emptyset$ opakuj
   a. $G_x :=$ vyjmi kořen z H
   b. vlož do H všechna bezprostřední zjemnění G_x s prioritami Δ (bez duplikací)

Halda on n prvcích:

- konstrukce z n prvků $O(n)$
- vložení prvku $O(\log n)$
- vyjmutí kořene $O(\log n)$

[Sedgewick: Algorithms]

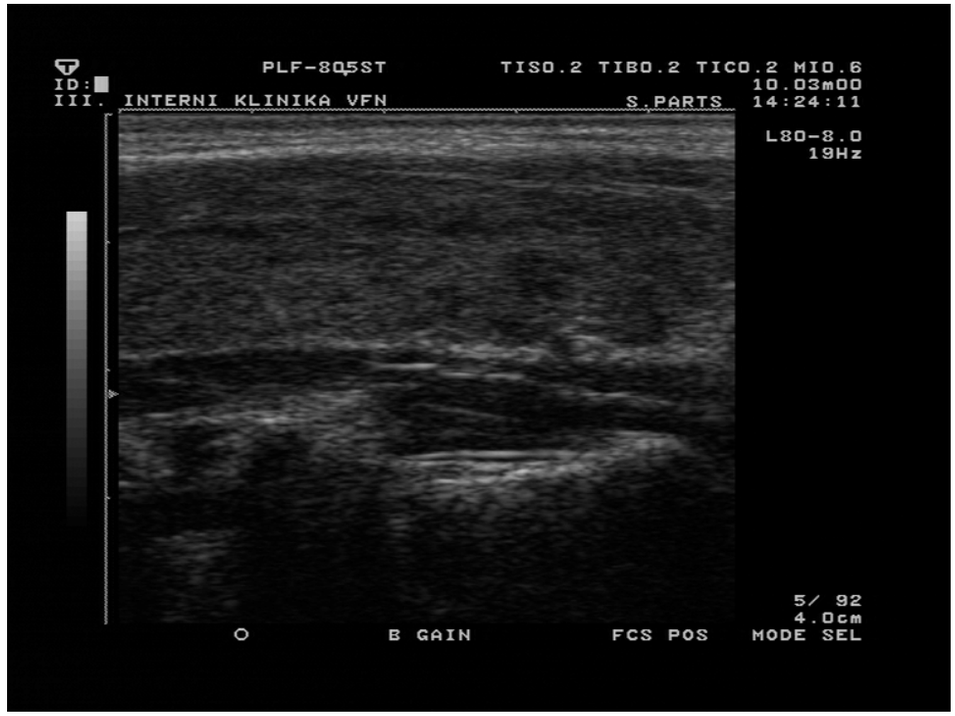
Identifikační procedura

1. Identifikuj zobecněný dynamický systém S z datového systému D.
2. Vytvoř rekonstrukční hypotézu $G_0 = \{S\}$.
3. Prohledej svaz zjemnění hypotézy G_0, na každé úrovni zjemnění zaznamenej hypotézu G_k o nejmenší hodnotě $\Delta(G_0, G_k)$.
4. Výsledkem je soubor nejlepších dekompozičních hypotéz klesajícího stupně strukturní složitosti.

Sonogram štítné žlázy v podélném řezu

zdravá

lymfocitická thyroitida



Příklad: Klasifikace textury

Aplikace

Diagnóza difúzních změn ve štítné žláze.

Pracovní hypotéza

Klasifikace je možná na základě textury.

Volba

Model založený na zobecněném dynamickém systému (kookurenční matici).

Základní problém

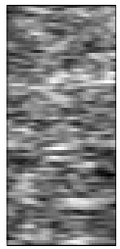
Zvolit co nejmenší model textury založený na funkci přípustnosti.

Důvody:

1. klasifikace je jednodušší
2. klasifikátor je mnohem snadnější naučit

Rekonstrukční analýza

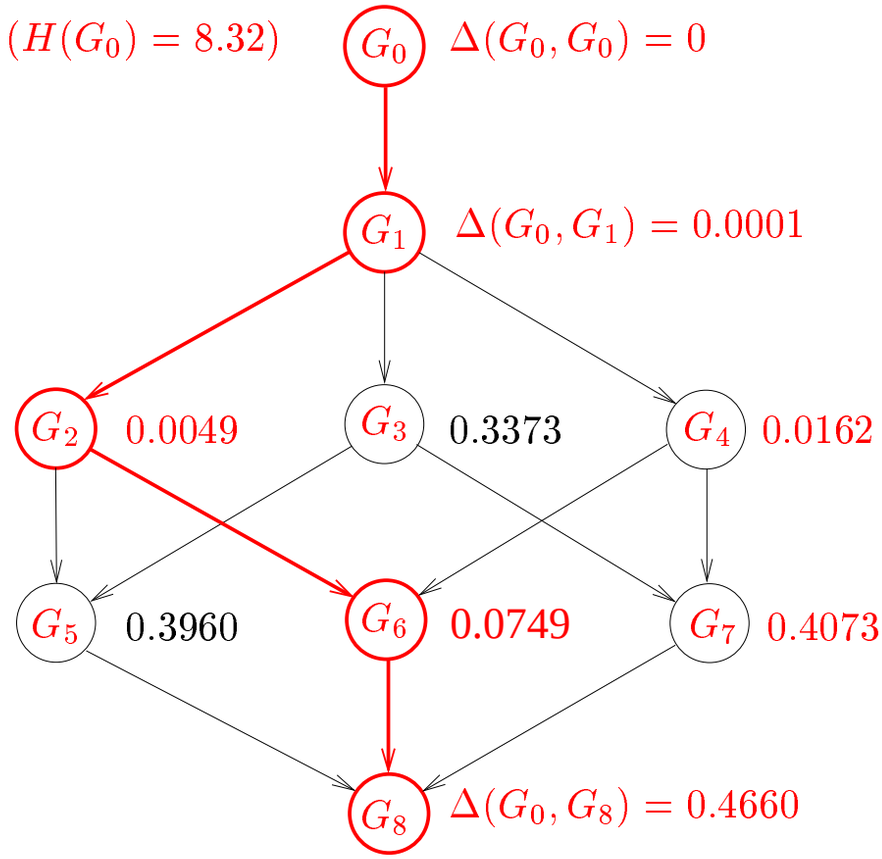
vzorek textury:



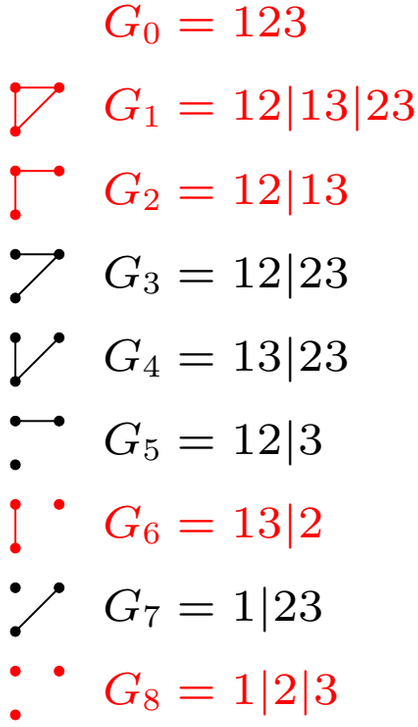
maska:

1	2
3	

svaz zjemnění



rekonstrukční hypotézy



Závěr: kookurenční matici má smysl počítat pouze ve svislém směru