SOUČINOVÉ DISTRIBUČNÍ SMĚSI

II. část: Příklady použití a strukturní model

Jiří Grim

Ústav teorie informace a automatizace AV ČR

Oddělení rozpoznávání obrazů

Květen 2008

Přednáška je volně k dispozici na adrese http://www.utia.cas.cz/RO

ÚTIA

Outline

Statistické řešení problému rozpoznávání

$\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}$: N-rozměrné datové vektory

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\}$: konečný počet tříd, pravděpodobnosti tříd: $p(\omega)$

$P(\mathbf{x}|\omega), \quad \omega \in \Omega$: odhadnuté podmíněné distribuce

BAYESŮV VZOREC: aposteriorní pravděpodobnosti tříd

$$p(\omega|\mathbf{x}) = \frac{P(\mathbf{x}|\omega)p(\omega)}{P(\mathbf{x})}, \qquad P(\mathbf{x}) = \sum_{\omega \in \Omega} P(\mathbf{x}|\omega)p(\omega), \quad \mathbf{x} \in \mathcal{X}$$

BAYESOVA ROZHODOVACÍ FUNKCE: minimalizuje pravděp. chyby

$$d(\mathbf{x}) = \omega_0 = \arg\max_{\omega \in \Omega}\{p(\omega|\mathbf{x})\} = \arg\max_{\omega \in \Omega}\{P(\mathbf{x}|\omega)p(\omega)\}$$

ŘEŠENÍ: odhad neznámých distribucí $P(\mathbf{x}|\omega)$ na základě trénovacích datových souborů $\mathcal{S}_\omega = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K_\omega)}\}, \omega \in \Omega$

POZN. Součinové komponenty umožňují sekvenční rozhodování (postupné doplňování proměnných) výběr informativních příznaků (globálně i lokálně)

▸ Sekvenční rozhodování ▸ Informativní podprostor

Příklad 1: Rozpoznávání obrazců na šachovnici

Problém:

rozpoznávání dvou tříd obrazců vzniklých na šachovnici náhodnými tahy věže (třída ω_1) resp. jezdce (třída ω_2)

dimenze vektorů: N = 8 x 8 = 64
počet náhodných tahů: do obsazení 20 různých políček

$$\mathbf{x} = (x_1, \ldots, x_{64}) \in \{0,1\}^{64}, \quad x_n \in \{0,1\}, \quad \sum_{n=1}^{64} x_n = 20, \quad \Omega = \{\omega_1, \omega_2\}$$

Vlastnosti problému:

netriviální statistický charakter problému, neprázdný průnik tříd, neexistují jednoduché příznaky, možnost generování libovolně velkých trénovacích souborů dat

Příklad 1: Rozpoznávání obrazců na šachovnici

Řešení: (Grim et al., 2003)

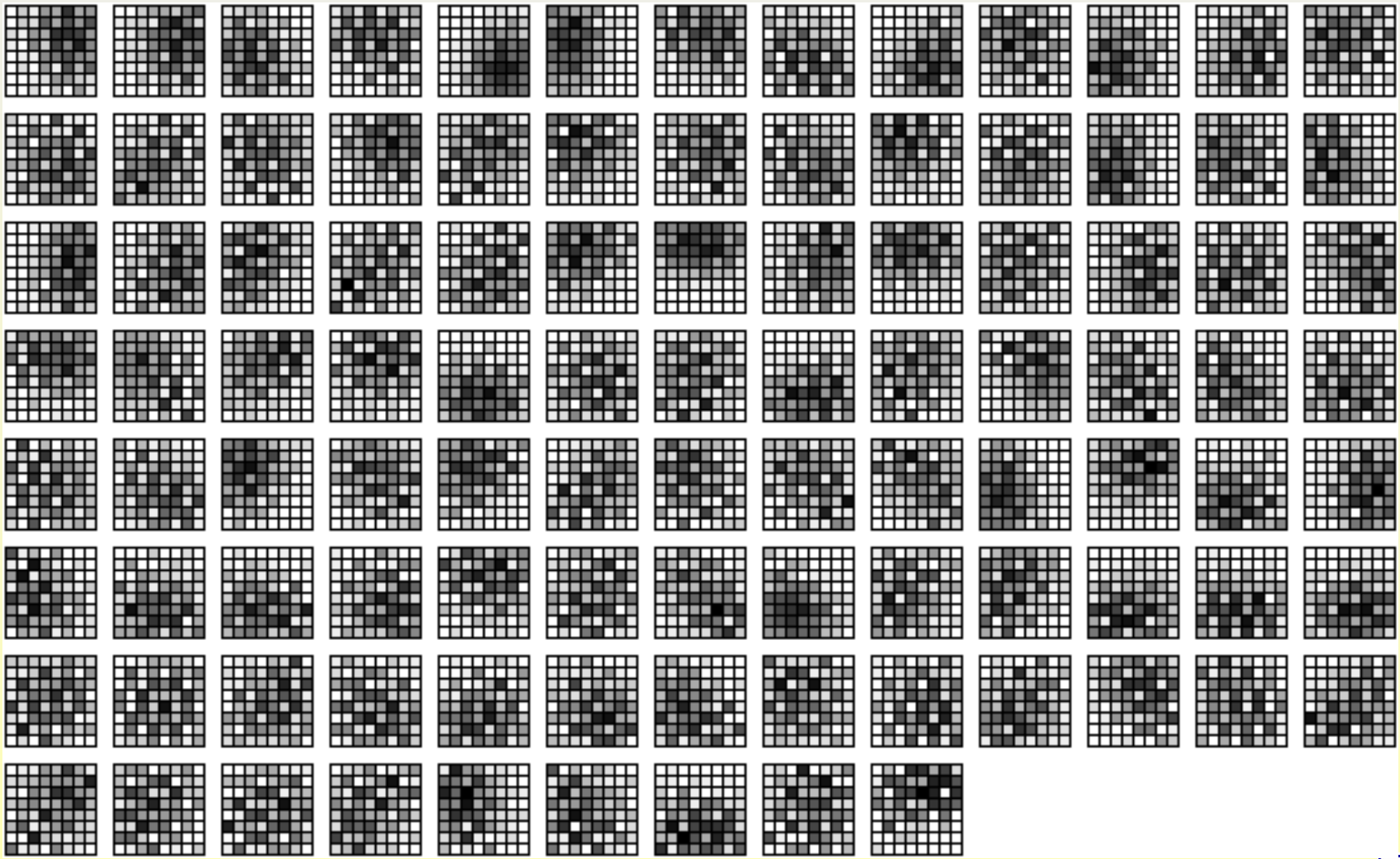
aproximace podmíněných distribucí $P(\mathbf{x}|\omega)$ pomocí směsí Bernoulliho rozložení s dimenzí $N = 64$

$$P(\mathbf{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} f(m) \prod_{n=1}^{64} \theta_{mn}^{x_n} (1 - \theta_{mn})^{1-x_n}, \quad x_n \in \{0, 1\}, \ \omega \in \Omega$$

- počet komponent směsi: $|\mathcal{M}_\omega| = 5$; (20; 100;)
- identické počáteční váhy komponent: $f(m) = 1/|\mathcal{M}_\omega|$
- velikost trénovacího souboru: $|\mathcal{S}_\omega| = 2 \times 10^4$; $(2 \times 10^5;\ 5 \times 10^6;)$
- počáteční hodnoty θ_{mn} generovány náhodně z intervalu $\langle 0.1, 0.9 \rangle$
- počet iterací EM algoritmu omezen podmínkou $(L' - L)/L < 5 \times 10^{-3}$

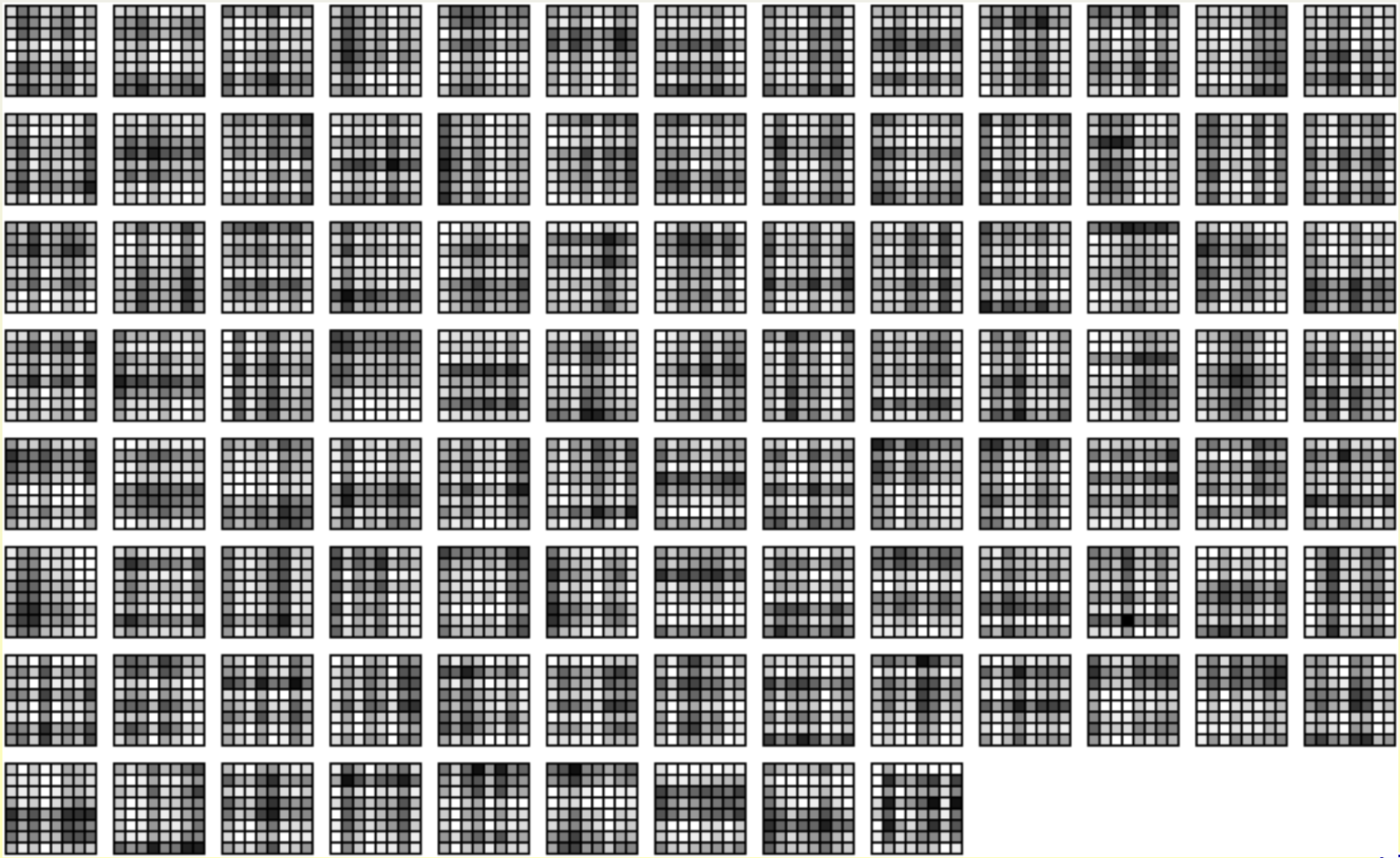
Příklad 1: Rozpoznávání obrazců na šachovnici

Parametry směsi θ_{mn} pro třídu "jezdec"



Příklad 1: Rozpoznávání obrazců na šachovnici

Parametry směsi θ_{mn} pro třídu "věž"



Příklad 1: Rozpoznávání obrazců na šachovnici

PŘESNOST ROZPOZNÁVÁNÍ OBRAZCŮ NA ŠACHOVNICI

	param.	$\lvert\mathcal{S}_\omega\rvert = 2x10^4$	$\lvert\mathcal{S}_\omega\rvert = 2x10^5$	$\lvert\mathcal{S}_\omega\rvert = 5x10^6$
$\lvert\mathcal{M}_\omega\rvert = 5$	324	**80.19**	**83.79**	**83.96**
$\lvert\mathcal{M}_\omega\rvert = 20$	1299	**74.11**	**91.40**	**91.83**
$\lvert\mathcal{M}_\omega\rvert = 100$	6499	**54.97**	**90.36**	**96.35**

- přesnost testována na nezávislém datovém souboru: $|\mathcal{S}_\omega^{test}| = 5 \times 10^5$
- směs 5 komponent je příliš jednoduchá, přesnost rozpoznávání se nezvýší ani po zvětšení trénovacího souboru
- směs 20 komponent dosahuje značnou přesnost již při $|\mathcal{S}_\omega| = 2 \times 10^5$, další zvětšení trénovacího souboru nemá vliv
- směs 100 komponent vyžaduje velký počet dat, ale dosahuje nejpřesnější výsledky rozpoznávání

Příklad 2: Rozpoznávání číslic na binárním rastru

Problém:

rozpoznávání rukou psaných číslic na binárním rastru z databáze Concordia University, Montreal (směrovací čísla z nedoručených dopisů)
4000 trénovacích číslic, tj. 400 pro každou třídu $\omega \in \Omega$
2000 číslic pro nezávislé testování, tj. 200 pro každou třídu

Předzpracování dat:

normalizace číslic na velikost rastru 32x32 (tj. dimenze dat $N = 1024$)
nebyly konstruovány žádné speciální příznaky, databáze byla rozšířena čtyřmi nezávislými posuvy normalizovaných číslic vertikálně i horizontálně $(\pm 1, \pm 2)$, tj. pro každou číslici 5×5 nezávislých pozic

popis číslic: $x_n \in \{0, 1\}, \quad \mathbf{x} = (x_1, x_2, \ldots, x_{1024}) \in \mathcal{X}, \quad \mathcal{X} = \{0, 1\}^{1024}$

počet tříd: $|\Omega| = 10, \quad \Omega = \{\omega_0, \omega_1, \ldots, \omega_9\}$

výsledný soubor: 100000 vzorků číslic $(|\mathcal{S}_\omega| = 5 \times 5 \times 400 = 10000)$

POZN. Pořadí políček rastru ve vektoru **x** může být zvoleno libovolně.

Příklad 2: Rozpoznávání číslic na binárním rastru

Řešení: (Grim et al., 2000a, 2000b, 2002)

spočívá v aproximaci podmíněných distribucí $P(\mathbf{x}|\omega)$ v původním 1024-dimensionálním prostoru pomocí strukturní Bernoulliovské směsi

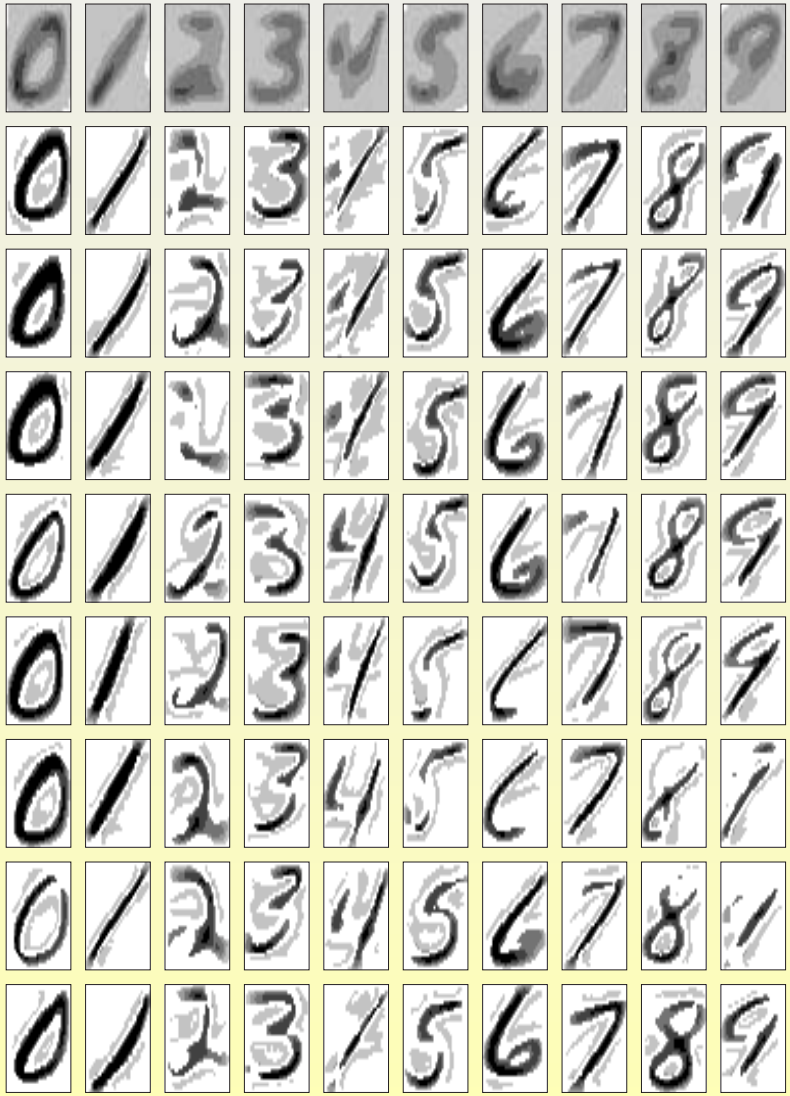
$$P(\mathbf{x}|\omega) = F(\mathbf{x}|0) \sum_{m \in \mathcal{M}_\omega} f(m) \prod_{n \in \mathcal{N}} \left[\left(\frac{\theta_{mn}}{\theta_{0n}} \right)^{x_n} \left(\frac{1-\theta_{mn}}{1-\theta_{0n}} \right)^{1-x_n} \right]^{\phi_{mn}}, \quad \omega \in \Omega$$

$$F(\mathbf{x}|0) = \prod_{n \in \mathcal{N}} \theta_{0n}^{x_n} (1-\theta_{0n})^{1-x_n}, \qquad \theta_{0n} = P\{x_n = 1\} = \sum_{\omega \in \Omega} P_n(1|\omega) p(\omega)$$

- $F(\mathbf{x}|0)$: konstantní distribuce pozadí ($\theta_{0n} \approx$ "průměrná" číslice)
- počáteční počet komponent: $|\mathcal{M}_\omega| = 60$ (váhy $f^{(0)}(m) = \frac{1}{60}$)
- počet nenulových parametrů $\phi_{mn} = 1$: $r_\omega = 15000 - 20000$
- náhodné počáteční hodnoty $\theta_{mn} \in \langle 0.1, 0.9 \rangle$ a $\phi_{mn} \in \{0, 1\}$
- počet iterací EM algoritmu: $20 - 40$

Příklad 2: Rozpoznávání číslic na binárním rastru

Grafické zobrazení parametrů strukturních směsí



Příklad 2: Rozpoznávání číslic na binárním rastru

PŘESNOST ROZPOZNÁVÁNÍ ČÍSLIC

(8 nezávislých řešení problému - náhodné počáteční parametry směsi)

	M	r	r/M	%(1024)	přesnost
Řešení 1	400	180007	450.0	43.9	**0.9330**
Řešení 2	370	175000	473.0	46.2	**0.9325**
Řešení 3	410	200000	487.8	47.6	**0.9325**
Řešení 4	428	159997	373.8	36.5	**0.9315**
Řešení 5	381	160000	419.9	41.0	**0.9310**
Řešení 6	476	240000	504.2	49.2	**0.9300**
Řešení 7	374	175000	467.9	45.7	**0.9295**
Řešení 8	420	179994	428.6	41.8	**0.9270**

r - celkový počet parametrů, M - celkový počet komponent

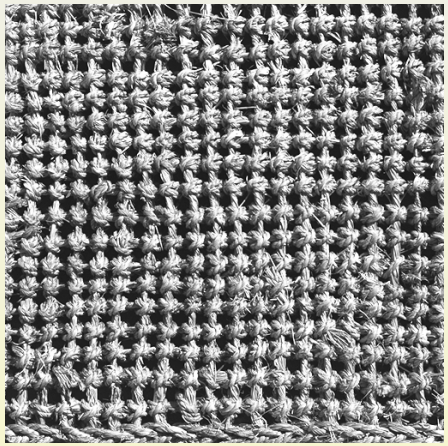
$$r = \sum_{\omega \in \Omega} \sum_{m \in \mathcal{M}_\omega} \sum_{n \in \mathcal{N}} \phi_{mn}$$

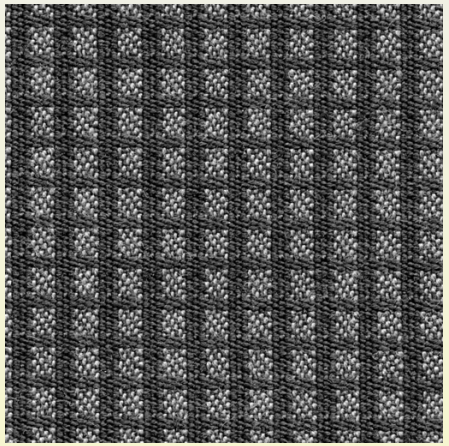
▸ Publikované výsledky klasifikace

Příklad 3: Modelování textur pomocí normální směsi

černobílé textury: $Y = [y_{ij}]_{i=1}^{I}{}_{j=1}^{J}, \quad y_{ij} \in \{0, \ldots, 255\} \approx$ úrovně šedi

příklady textur: rozměry 512×512 pixelů, tj. $I = J = 512$





Předpoklad statistické "homogenity":

předpokládáme, že texturu lze popsat lokálně na základě statistických vlastností vnitřních pixelů $x_1, \ldots, x_N$ nějakého pohyblivého okna

$\mathbf{x} = (x_1, x_2, \ldots, x_N) \approx$ vnitřní pixely pohyblivého okna $(N = 400 \div 900)$

$\mathcal{S} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\} \approx$ data získaná posuvem okna v obrázku textury

POZN. Vektory $\mathbf{x} \in \mathcal{S}$ nejsou nezávislé v důsledku překryvu oken.

Příklad 3: Modelování textur pomocí normální směsi

Princip modelování (Grim et al. 2001, Haindl et al., 2004):

- odhad lokálních statistických vlastností textury uvnitř posuvného okna pomocí normální součinové směsi $P(\mathbf{x})$
- postupná predikce (syntéza) textury (libovolné velikosti) na základě podmíněných distribucí odvozených z $P(\mathbf{x})$
- nahrazení výsledku predikce v každém kroku "nejpodobnější" částí původní reálné textury

$$P(\mathbf{x}) = \sum_{m\in\mathcal{M}} f(m)F(\mathbf{x}|\boldsymbol{\mu}_m,\boldsymbol{\sigma}_m) = \sum_{m\in\mathcal{M}} f(m)\prod_{n\in\mathcal{N}} f_n(x_n|\mu_{mn},\sigma_{mn})$$

$$\begin{aligned}\mathcal{D} = \{j_1,\ldots,j_l\} \subset \mathcal{N} &\approx \text{definovaná část okna}\\ \mathcal{C} = \{i_1,\ldots,i_k\} = \mathcal{N}\setminus\mathcal{D} &\approx \text{nedefinovaná část okna}\end{aligned}$$

$$\begin{aligned}\mathbf{x}_D = (x_{j_1},\ldots,x_{j_l}) \in \mathcal{X}_D, &\quad F(\mathbf{x}_D|\boldsymbol{\mu}_m,\boldsymbol{\sigma}_m) = \Pi_{n\in D}\ f_n(x_n|\mu_{mn},\sigma_{mn})\\ \mathbf{x}_C = (x_{i_1},\ldots,x_{i_k}) \in \mathcal{X}_C, &\quad F(\mathbf{x}_C|\boldsymbol{\mu}_m,\boldsymbol{\sigma}_m) = \Pi_{n\in C}\ f_n(x_n|\mu_{mn},\sigma_{mn})\end{aligned}$$

ÚTIA

Příklad 3: Modelování textur pomocí normální směsi

podmíněné distribuce:

$$P_{C|D}(\mathbf{x}_C|\mathbf{x}_D) = \frac{P_{CD}(\mathbf{x}_C, \mathbf{x}_D)}{P_D(\mathbf{x}_D)} = \sum_{m\in\mathcal{M}} W_m(\mathbf{x}_D) F(\mathbf{x}_C|\boldsymbol{\mu}_{mC}, \boldsymbol{\sigma}_{mC})$$

$$W_m(\mathbf{x}_D) = \frac{f(m)F(\mathbf{x}_D|\boldsymbol{\mu}_{mD}, \boldsymbol{\sigma}_{mD})}{\sum_{j\in\mathcal{M}} f(j)F(\mathbf{x}_D|\boldsymbol{\mu}_{jD}, \boldsymbol{\sigma}_{jD})}$$

očekávané hodnoty $\bar{\mathbf{x}}_C$ při definované části $\mathbf{x}_D$:

$$\bar{\mathbf{x}}_C = \mathrm{E}_{C|D}\{\mathbf{x}_C|\mathbf{x}_D\} = \int \mathbf{x}_C P_{C|D}(\mathbf{x}_C|\mathbf{x}_D) d\mathbf{x}_C = \sum_{m\in\mathcal{M}} W_m(\mathbf{x}_D)\boldsymbol{\mu}_{mC}$$

nahrazení $\bar{\mathbf{x}}_C$ "nejpodobnější částí" původní reálné textury:

$$\mathbf{x}_C^* = \arg\min_{y_C\in\mathcal{S}} \{\|\bar{\mathbf{x}}_C - \mathbf{y}_C\|^2\}$$

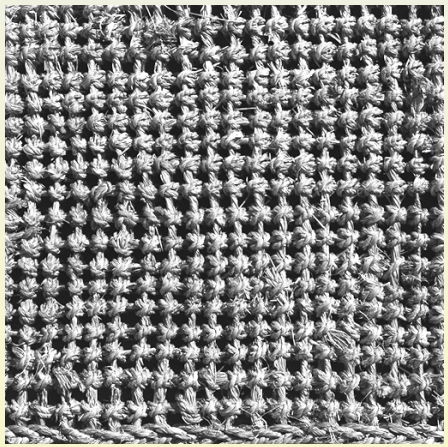
$\mathbf{y}_C$ $\approx$ části původní textury získané posuvem okna v původním obrázku

ÚTIA

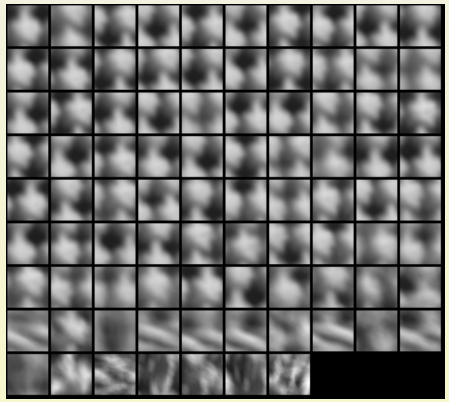
Příklad 3: Modelování textur pomocí normální směsi

model textury "ratan":

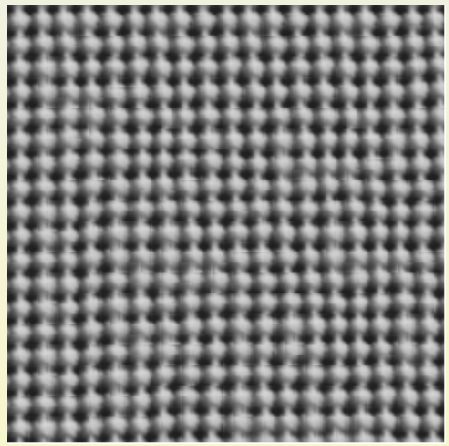
původní textura



průměry komp. μ_m



syntéza vzorkováním

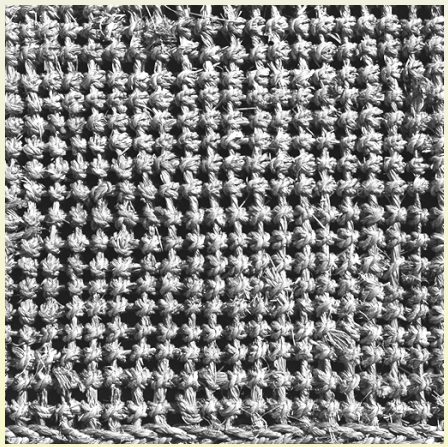


- velikost obrázku: 512x512 pixels $\Rightarrow$ $|\mathcal{S}| \doteq 233000$
- velikost posuvného okna: 30x30 pixelů
- počet komponent: $|\mathcal{M}| = 80$
- počet iterací EM algoritmu: $t = 15$

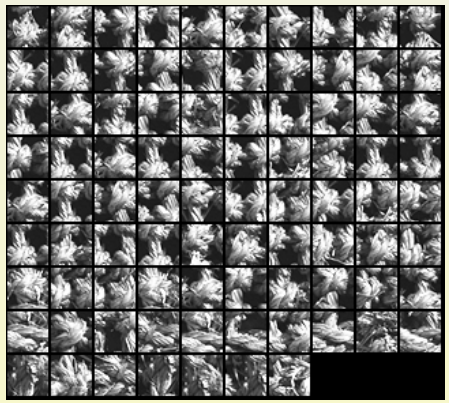
Příklad 3: Modelování textur pomocí normální směsi

model textury "ratan":

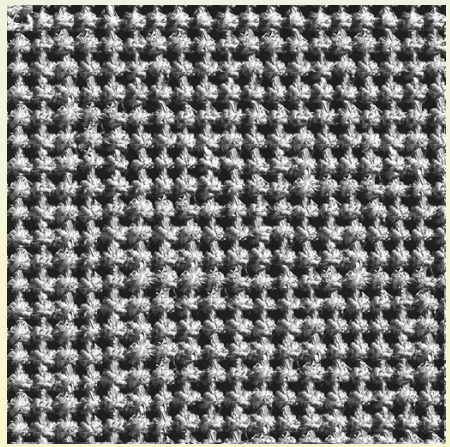
původní textura



optimální "záplaty"



syntéza vzorkováním

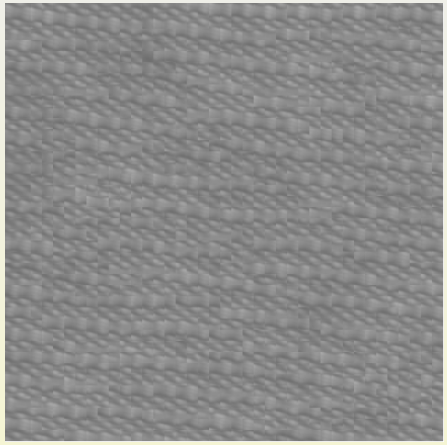
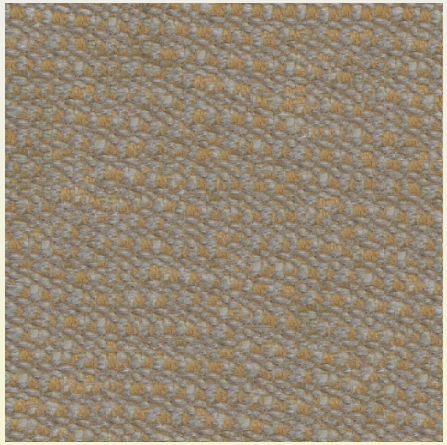


"realistická" syntéza: průměry komponent $\boldsymbol{\mu}_m$ nahrazeny podobnými částmi původní textury $\boldsymbol{\mu}_m^*$ optimálně vyhledanými podle kriteria:

$$\boldsymbol{\mu}_m^* = \arg\min_{\mathbf{x}\in\mathcal{S}}\{\|\mathbf{x}-\boldsymbol{\mu}_m\|^2\}$$

Příklad 3: Modelování textur pomocí normální směsi

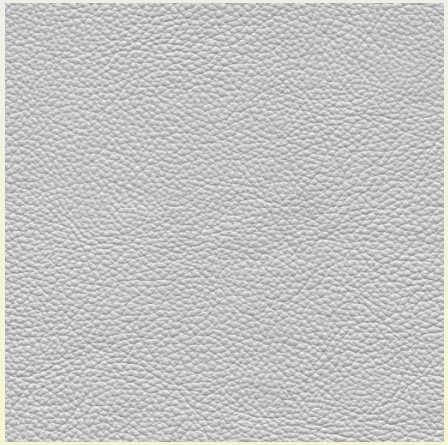
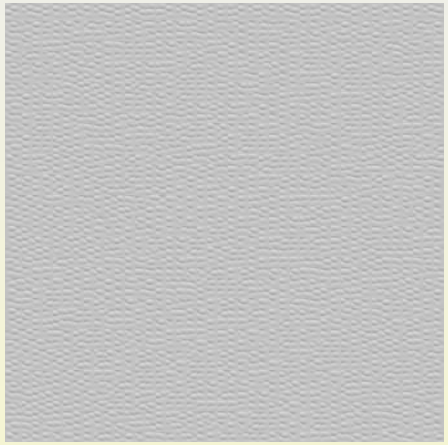
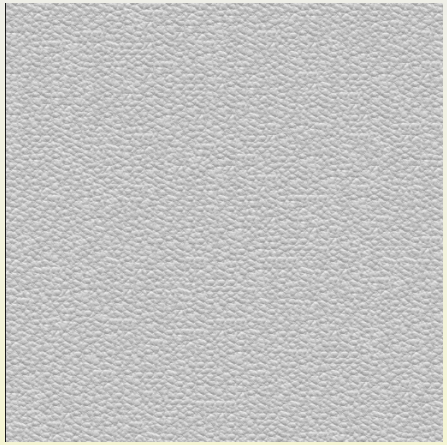
model textury "hrubá látka":

- velikost posuvného okna: 30x30 pixelů
- dimenze směsi: $N = 30x30 = 900$, počet komponent: $|\mathcal{M}| = 128$
- počet vzorků textury získaných pohybem okna: $|\mathcal{S}| \doteq 232000$
- míra vzdálenosti komponent směsi: $\bar{q}_{max} = 0.993$
- posuv okna při syntéze: 13 pixelů
- při nahrazování predikované textury reálnou částí obrázku byly použity části původní barevné textury

Příklad 3: Modelování textur pomocí normální směsi

model textury "světlá kůže":

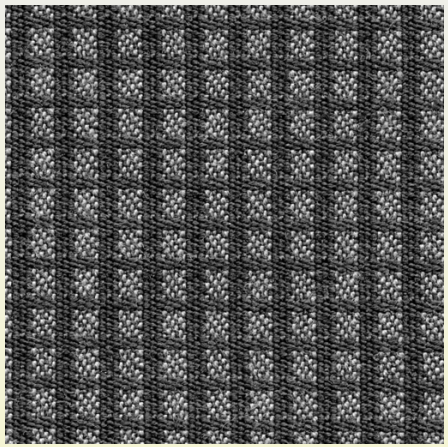
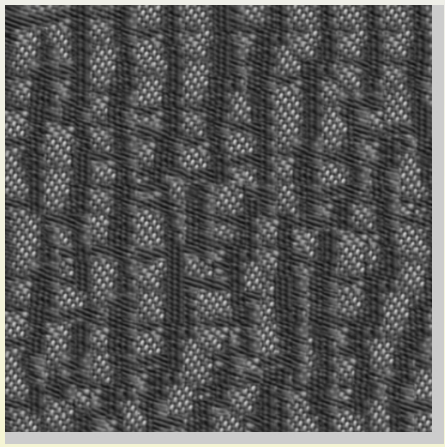
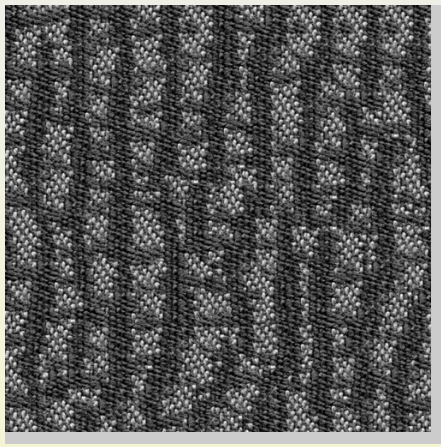




- velikost posuvného okna: 20x20 pixelů
- dimenze směsi: $N = 20x20 = 400$, počet komponent: $|\mathcal{M}| = 50$
- počet vzorků textury získaných posuvem okna: $|\mathcal{S}| \doteq 242000$
- míra vzdálenosti komponent směsi: $\bar{q}_{max} = 0.959$
- posuv okna při syntéze: 12 pixelů

POZN. Paradoxně: nejmenší krok při syntéze není nejlepší, optimální krok odpovídá přibližně polovině strany okna (!?).

Příklad 3: Modelování textur pomocí normální směsi

model textury "koberec":

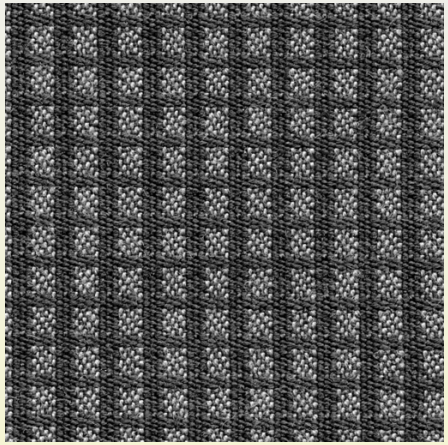
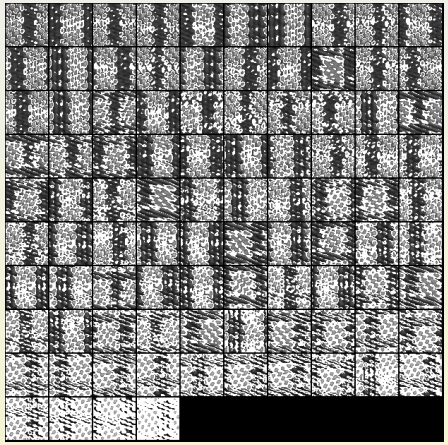
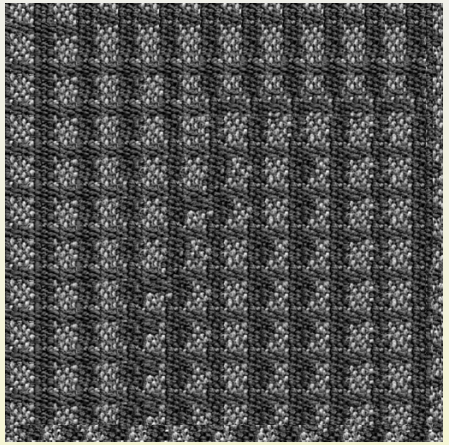




- dimenze směsi: $N = 30x30 = 900$, počet komponent: $|\mathcal{M}| = 90$
- počet vzorků textury získaných posuvem okna: $|\mathcal{S}| \doteq 232000$
- počet iterací EM algoritmu: $t = 20$
- míra vzdálenosti komponent směsi: $\bar{q}_{max} = 0.997$
- posuv okna při syntéze: 18 pixelů

POZN. Při velikosti okna 30x30 pixelů je dobře popsána jemná struktura, ale selhává popis hrubé struktury koberce.

Příklad 3: Modelování textury pomocí strukturní směsi

strukturní model textury "koberec":

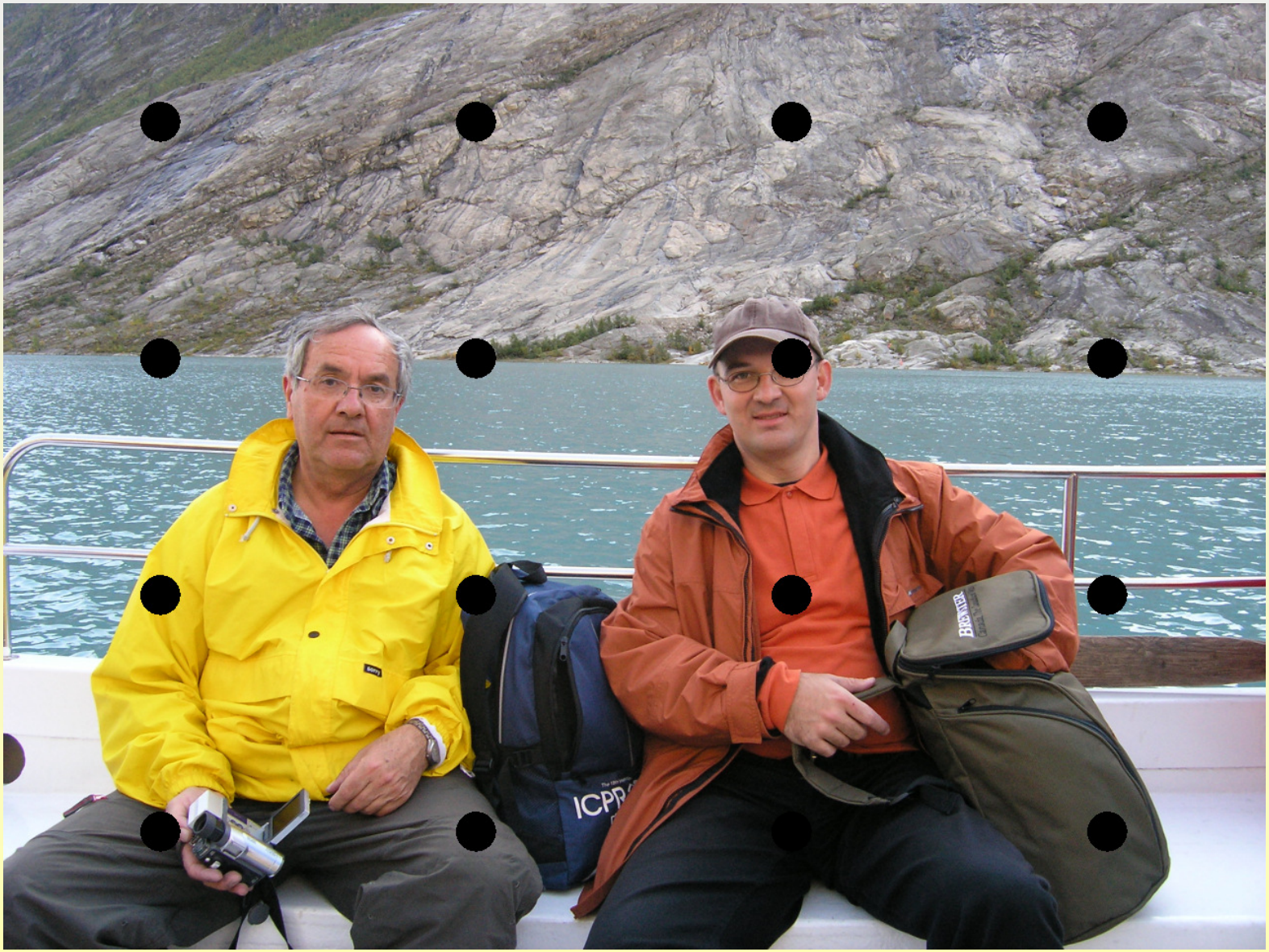




- dimenze směsi: $N = 60x60 = 3600$, počet komponent: $|\mathcal{M}| = 94$
- počet vzorků textury získaných posuvem okna: $|\mathcal{S}| \doteq 205000$
- počet iterací EM algoritmu: $t = 18$
- míra vzdálenosti komponent směsi: $\bar{q}_{\max} = 0.999$
- posuv okna při syntéze: 24 pixelů

POZN. Popis hrubé struktury koberce je zřetelně lepší než při velikosti okna 30x30 pixelů.

Příklad 4: Predikce chybějících částí obrázku

původní poškozený obrázek

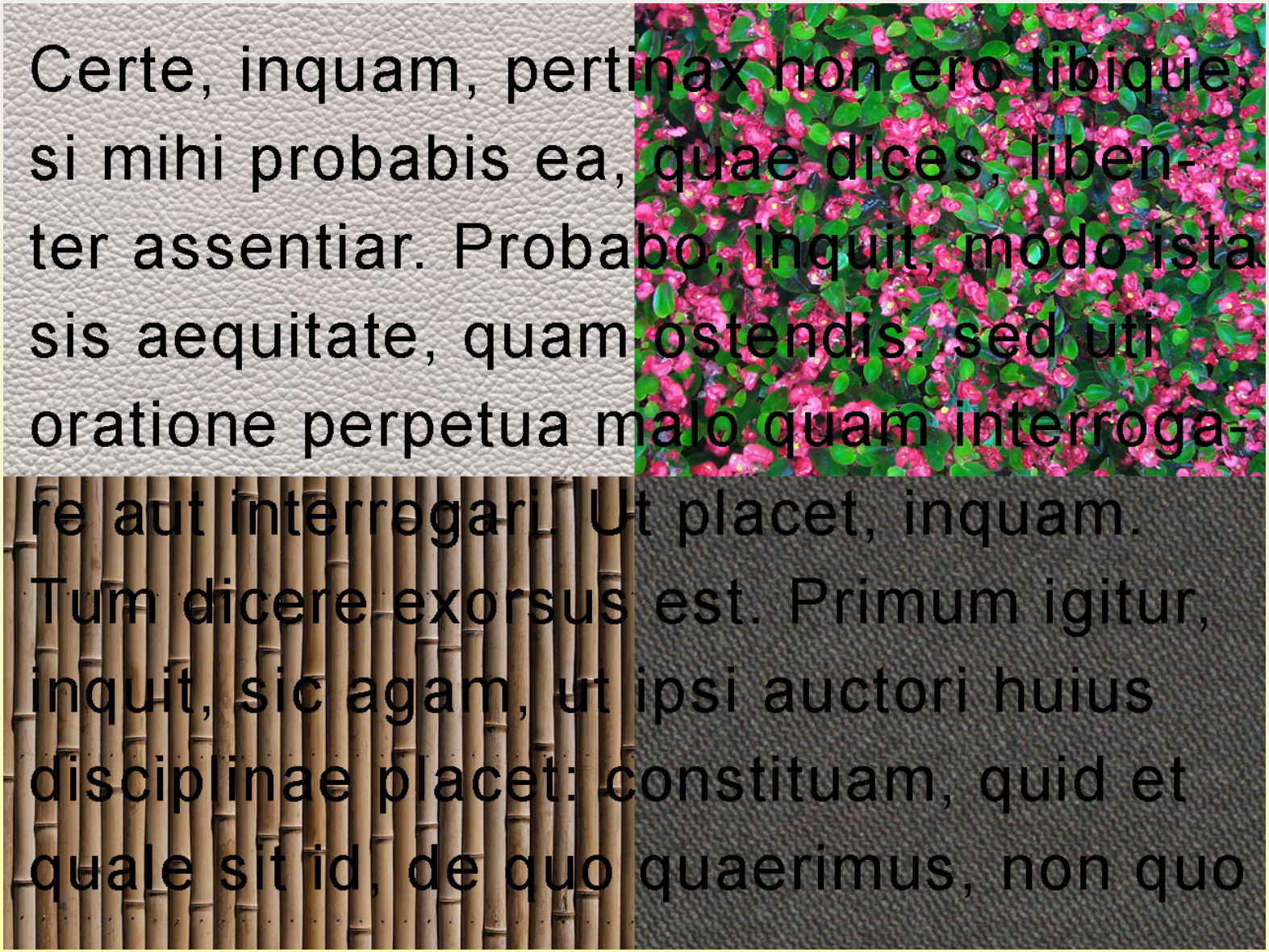


Příklad 4: Predikce chybějících částí obrázku

opravený obrázek

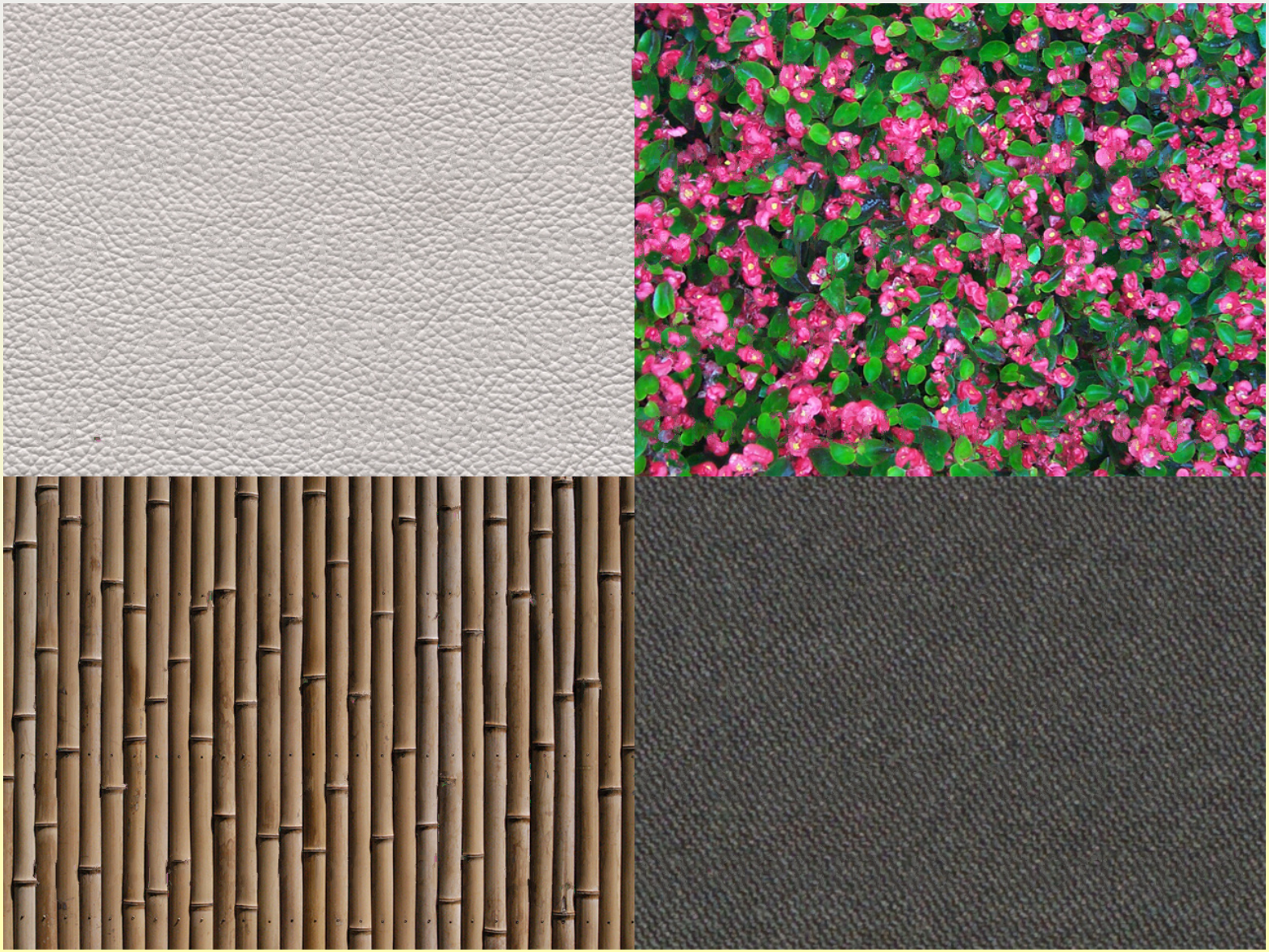
Příklad 4: Predikce chybějících částí obrázku

původní poškozený obrázek



Příklad 4: Predikce chybějících částí obrázku

opravený obrázek



Příklad 4: Predikce chybějících částí obrázku

původní poškozený obrázek

Příklad 4: Predikce chybějících částí obrázku

opravený obrázek

Příklad 5: Interaktivní statistický model dat ze sčítání lidu

náklady sčítání lidu: cca 2.5 mld Kč

dostupnost výsledků sčítání lidu je značně omezena nutností ochrany anonymity dat

Současné metody publikace výsledků

- **aggregovaná data** (územně, např. na úrovni sčítacích okrsků)
  nevýhoda: zcela se znehodnotí informace o subpopulacích
- **publikace tabulek** (tiskem nebo na paměťových médiích)
  nevýhody: až do 10 proměnných, nutné ověřování anonymity dat
- **komerční služby statistických úřadů** (písemný dotaz)
  nevýhody: těžkopádný a zdlouhavý postup, poplatky
- **anonymizované uživatelské soubory mikrodat** ($|\mathcal{S}| \approx 10^6$)
  výhoda: neomezené možnosti formulace otázek
  nevýhody: nutné ověřování anonymity dat, omezená distribuce, omezená přesnost údajů (anonymizační procedury, velikost souboru)

Příklad 5: Interaktivní statistický model dat ze sčítání lidu

Interaktivní odvozování statistických vlastností dat pomocí modelu:

Diskrétní součinová směs může být přímo použita jako báze znalostí pravděpodobnostního expertního systému PES. Umožňuje rychlé odvozování podmíněných histogramů pro libovolně zvolenou subpopulaci interaktivním způsobem při dokonale zabezpečené ochraně dat.

$$\mathbf{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_C, \qquad C = \{i_1, \ldots, i_k\} \subset \mathcal{N}$$

subpopulace: $\mathcal{S}(\mathbf{x}_C) = \{\mathbf{y} \in \mathcal{S} : \mathbf{y}_C = \mathbf{x}_C\} \subset \mathcal{S}$

(Např. dvojicí otázek ($|C| = 2,\ N = 25$) lze určit cca 10^4 subpopulací.)

jednoduchý výpočet podmíněných distribucí $P_{n|C}(x_n|\mathbf{x}_C), n \notin C$**:**

$$P_{n|C}(x_n|\mathbf{x}_C) = \sum_{m \in \mathcal{M}} W_m(\mathbf{x}_C) f_n(x_n|m), \quad W_m(\mathbf{x}_C) = \frac{w_m F_C(\mathbf{x}_C|m)}{\sum_{j=1}^{M} w_j F_C(\mathbf{x}_C|j)}$$

Příklad 5: Interaktivní statistický model dat ze sčítání lidu

PRŮMĚRNÁ CHYBA REPRODUKCE RELATIVNÍCH ČETNOSTÍ

(Sčítání lidu 1991 - Pražské domácnosti, 23 otázek, $|\mathcal{S}| = 535000$)

Počet komponent	Hodnota kriteria L	Trvání výpočtu	Abs. chyba $\bar{\Delta}$	Rel. chyba $\bar{\delta}$
1	-29,2069	2 s	0,00891	54,4 %
5	-22,8797	25 s	0,00274	22,3 %
10	-20,5518	3 m	0,00211	17,9 %
50	-18,4564	39 m	0,00108	10,8 %
100	-17,7872	2,5 h	0,00083	8,1 %
500	-16,6389	12 h	0,00040	4,2 %
1 000	-16,3837	26 h	0,00035	3,3 %
10 000	-15,8367	72 h	0,00018	1,9 %

ϵ-relevantní subpopulace: $|\mathcal{R}_\epsilon| \doteq 87000$ (kombinace 3 otázek, práh $\epsilon = 0.003$)

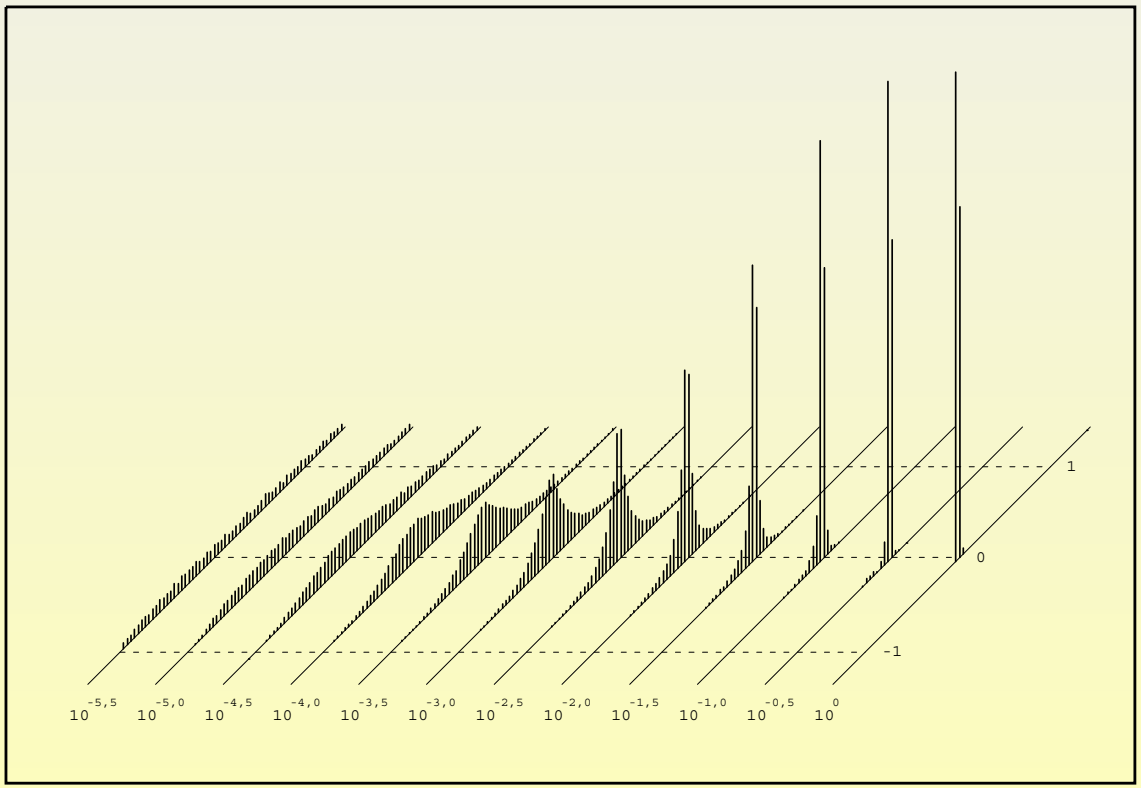
$$\bar{\Delta} = \frac{1}{|\mathcal{R}_\epsilon|} \sum_{A \in \mathcal{R}_\epsilon} |P(A) - \hat{P}(A)|, \qquad \bar{\delta} = \frac{1}{|\mathcal{R}_\epsilon|} \sum_{A \in \mathcal{R}_\epsilon} \frac{|P(A) - \hat{P}(A)|}{\hat{P}(A)}$$

ÚTIA

Příklad 5: Interaktivní statistický model dat ze sčítání lidu

ZÁVISLOST ROZLOŽENÍ RELATIVNÍCH CHYB MODELU NA VELIKOSTI SUBPOPULACE

(Sčítání lidu 1991 - Pražské domácnosti, 23 otázek, $|\mathcal{S}| = 535000, |\mathcal{M}| = 100$)



osa x: horní meze příslušných intervalů velikosti subpopulace

Příklad 6: Vyhledávání poruch a odchylek v textuře

černobílá textura: $\mathcal{Y} = [y_{ij}]_{i=1}^{I}{}_{j=1}^{J}, \quad y_{ij} \approx$ úrovně šedi $(\approx x_n)$

Předpoklad: homogenní textura

lokální statistické závislosti mezi pixely uvnitř zvoleného okna jsou invariantní vůči libovolnému posuvu okna

pixely okna v libovolném pevném pořadí: $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in R^N$

Metoda:

aproximace hustoty pravděpodobnosti $P(\mathbf{x})$ pomocí normální směsi součinových komponent (diagonální kovariační matice)

$$P(\mathbf{x}) = \sum_{m \in \mathcal{M}} w_m F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) = \sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}} f_n(x_n|\mu_{mn}, \sigma_{mn})$$

$$f_n(x_n|\mu_{mn}, \sigma_{mn}) = \frac{1}{\sqrt{2\pi}\sigma_{mn}} \exp\{-\frac{(x_n - \mu_{mn})^2}{2\sigma_{mn}^2}\}$$

Příklad 6: Vyhledávání poruch a odchylek v textuře

Idea:

úspěšná syntéza textury dokazuje, že lokální statistický model ve tvaru distribuční směsi $P(\mathbf{x})$ popisuje původní texturu dostatečně přesně
$\Rightarrow$ lze jej využít pro analýzu výchozího obrázku textury

LOG-LIKELIHOOD: $\log P(\mathbf{x}) \approx$ míra typičnosti (obvyklosti) $\mathbf{x}$

POZN.: Hodnota $\log P(\mathbf{x})$ je citlivá vzhledem k odchylkám úrovní šedi.

$$P_0(\mathbf{x}) = \prod_{n \in \mathcal{N}} f_n(x_n | \mu_{0n}, \sigma_{0n}), \quad \mu_{0n} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} x_n, \quad \sigma_{0n}^2 = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} x_n^2 - \mu_{0n}^2.$$

LOG-LIKELIHOOD RATIO: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})} \approx$ míra strukturní typičnosti části textury $\mathbf{x}$ (jmenovatel potlačuje vliv úrovní šedi)

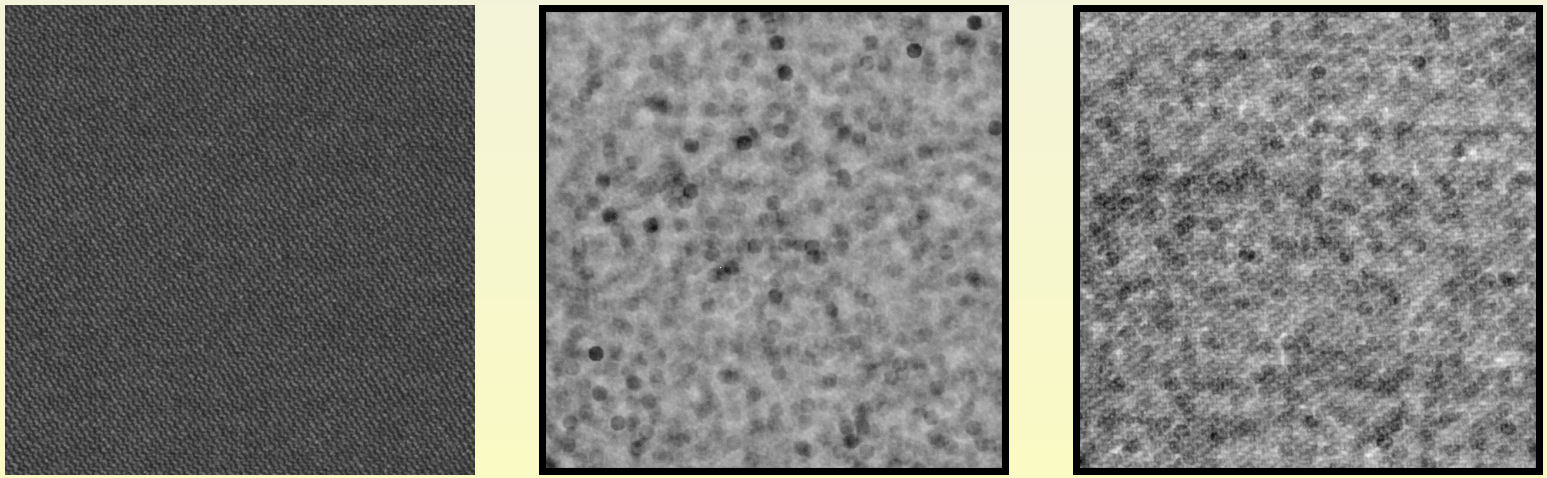
POZN.: Hodnoty průměru a rozptylu μ_{0n}, σ_{0n} jsou téměř identické pro všechna $n \in \mathcal{N} \Rightarrow$ hodnota $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ méně závisí na změnách úrovní šedi a je více ovlivněna odchylkami struktury.

Příklad 6: Vyhledávání poruch a odchylek v textuře

Lokální analýza textury "obklad":

hodnoty $\log P(\mathbf{x})$ resp. $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ jsou zobrazeny jako úrovně šedi centrálního pixelu posuvného okna

původní obrázek | L-věrohodnost | LR-věrohodnost



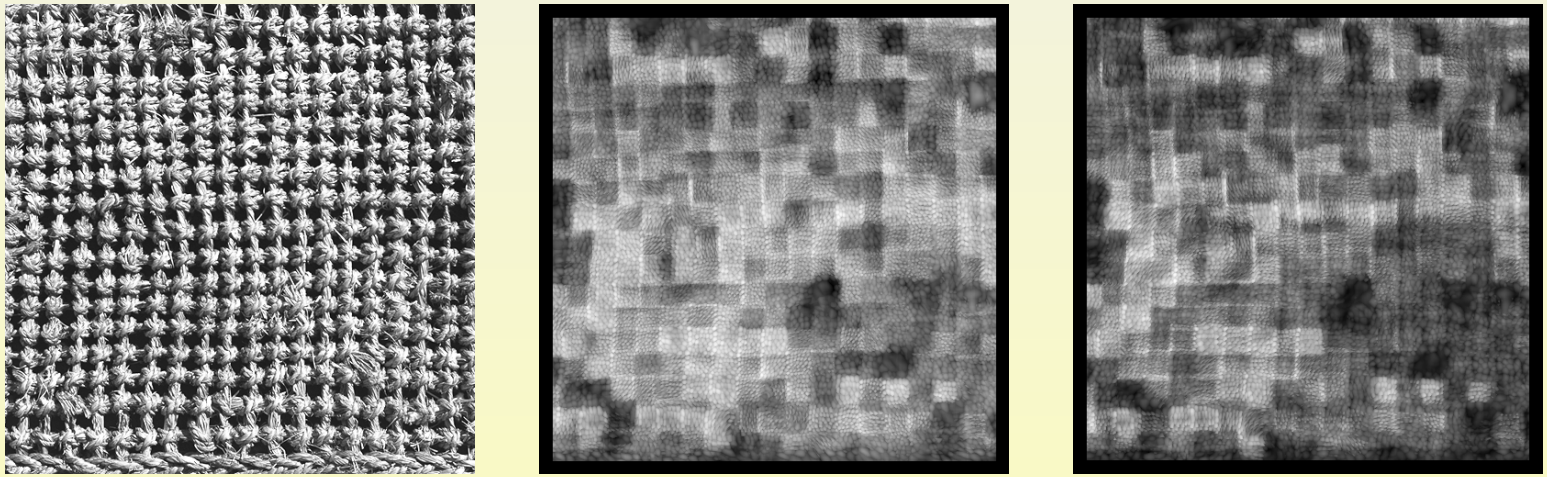
Remark: Hodnoty $\log P(\mathbf{x})$ jsou velmi citlivé na odchylky úrovní šedi. Tak např. stěží viditelné světlejší pixely v textuře "obklad" (levý obrázek) se projeví jako výrazné tmavé skvrny o velikosti okna (střední obrázek).

Příklad 6: Vyhledávání poruch a odchylek v textuře

Lokální analýza textury "ratan":

hodnoty $\log P(\mathbf{x})$ resp. $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ jsou zobrazeny jako úrovně šedi centrálního pixelu posuvného okna

původní obrázek | L-věrohodnost | LR-věrohodnost



POZN. Hodnoty $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ jsou citlivé na strukturní odchylky a méně závisí na úrovních šedi. Nepravidelnosti ve struktuře "ratanu" (levý obrázek) jsou proto zřetelnější na pravém obrázku, který využívá logaritmus věrohodnostního poměru $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$.

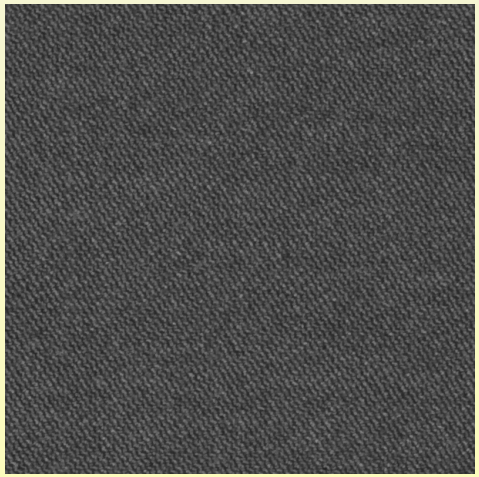
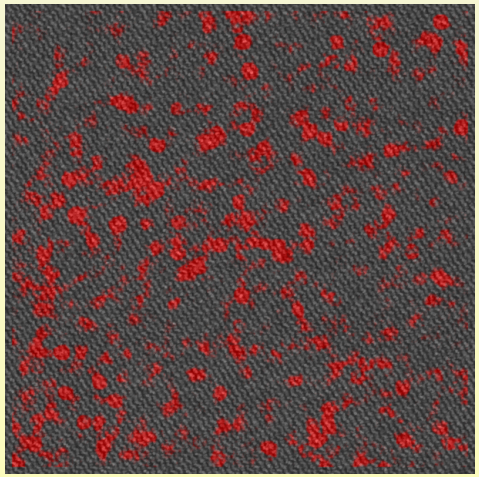
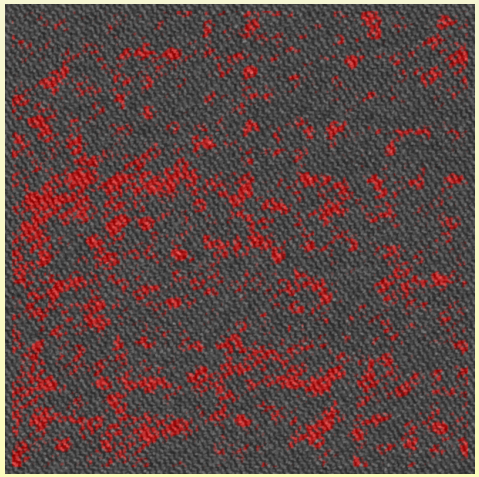
ÚTIA

Příklad 6: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "obklad":

zobrazované vysoké hodnoty $-\log P(\mathbf{x})$ (střední obrázek) resp. $-\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ (pravý obrázek) jsou zvýrazněny červeným zabarvením centrálního pixelu posuvného okna

původní obrázek | L-nevěrohodnost | LR-nevěrohodnost

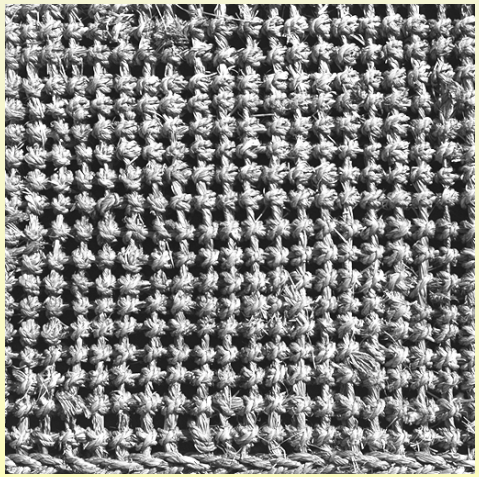
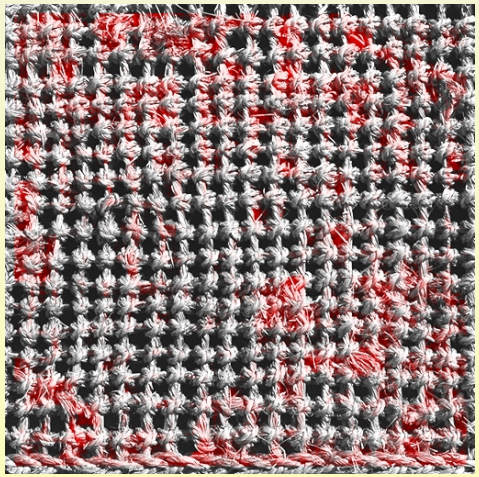
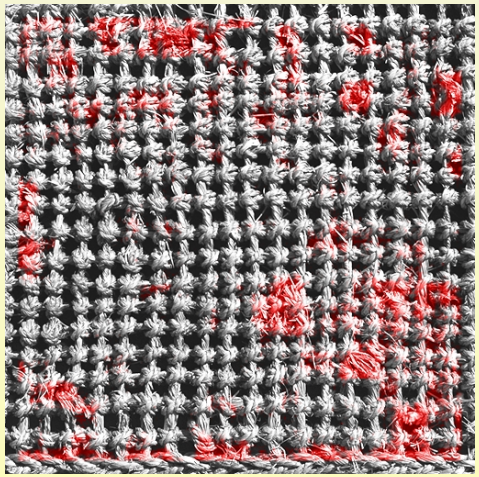




Příklad 6: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "ratan":

vysoké hodnoty $-\log P(\mathbf{x})$ (střední obrázek) resp. $-\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ (pravý obrázek) jsou zvýrazněny červeným zabarvením centrálního pixelu posuvného okna

původní obrázek | L-nevěrohodnost | LR-nevěrohodnost

▸ Další příklady

Příklad 6: Teoretické aspekty lokální analýzy textury

- (!) datové vektory $\mathbf{x}$ generované posuvem okna se překrývají a proto nejsou nezávislé
- datový soubor $\mathcal{S}$ odpovídá pouze "trajektorii" v prostoru $\mathcal{X}$ vzniklé posuvem okna ($\Rightarrow$ není reprezentativní)
- na rozdíl od jiných aplikací (syntéza textury, predikce, rozpoznávání) se odhadnutá směs $P(\mathbf{x})$ aplikuje na původní datový soubor $\mathcal{S}$
- věrohodnostní kritérium optimálně "přizpůsobuje" odhadovanou směs $P(\mathbf{x})$ na výchozí datový soubor $\mathcal{S}$
- $\Rightarrow$ aplikace směsi $P(\mathbf{x})$ na původní data $\mathbf{x} \in \mathcal{S}$ je dobře zdůvodněná metodou odhadu
- $\Rightarrow$ hodnota $\log P(\mathbf{x})$ je vhodnou mírou "typičnosti" vektorů $\mathbf{x} \in \mathcal{S}$
- zhoršená reprezentativnost souboru $\mathcal{S}$ není příliš závažná protože směs $P(\mathbf{x})$ se neaplikuje na data mimo soubor $\mathcal{S}$

Příklad 7: Vyhodnocování screeningových mamogramů

Mamografický screening:

včasná detekce zhoubného nádoru v rámci mamografického screeningu představuje v současnosti jedinou možnost snižování vysoké úmrtnosti

Statistické údaje z mamografického screeningu:

- asi 8 až 10% žen je během života ohroženo rakovinou prsu
- v rámci mamografického screeningu se zhoubný nádor potvrdí jen u 1 až 3 mamogramů z 1000
- 5 až 10% podezřelých nálezů se ověřuje chirurgicky pomocí biopsie (jednoduché vyšetření nicméně fyziky i psychicky traumatizující)
- výsledkem biopsie je v 60 až 80% případů nezhoubný nález
- následné prověřování zhoubných nálezů ukazuje, že výsledky mamografického screeningu jsou v 10 až 20% falešně negativní (tzn. 10 až 20% zhoubných nálezů zůstane nerozpoznáno)
- celkový počet screeningových mamogramů každoročně vyhodnocovaných ve světě se řádově udává v milionech

Příklad 7: Vyhodnocování screeningových mamogramů

Cíl věrohodnostní analýzy:

usnadnit diagnostické vyhodnocování screeningových mamogramů zvýrazněním atypických resp. podezřelých míst

LOKÁLNÍ ZOBRAZENÍ VĚROHODNOSTI:

$\log P(\mathbf{x}) \approx$ míra typičnosti vnitřku okna $\mathbf{x}$

Idea: nízké hodnoty $\log P(\mathbf{x})$ zobrazované jako tmavé pixely by měly odpovídat "neobvyklým" resp. "podezřelým" místům mamogramu

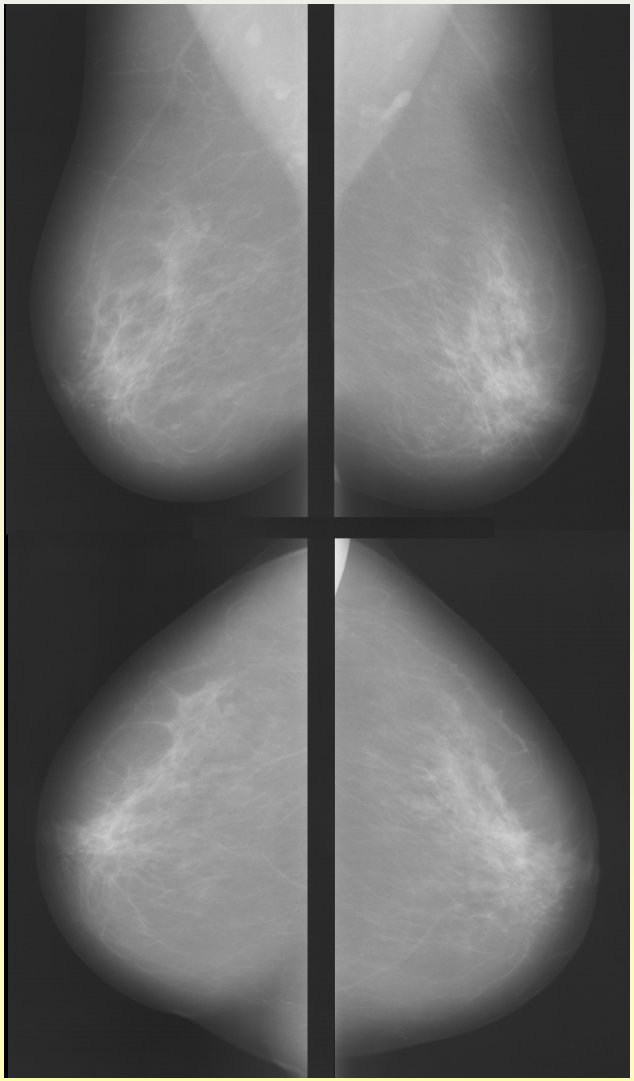
LOKÁLNÍ ZOBRAZENÍ VĚROHODNOSTNÍHO POMĚRU:

$\log P(\mathbf{x})/P_0(\mathbf{x}) \approx$ nebylo použito

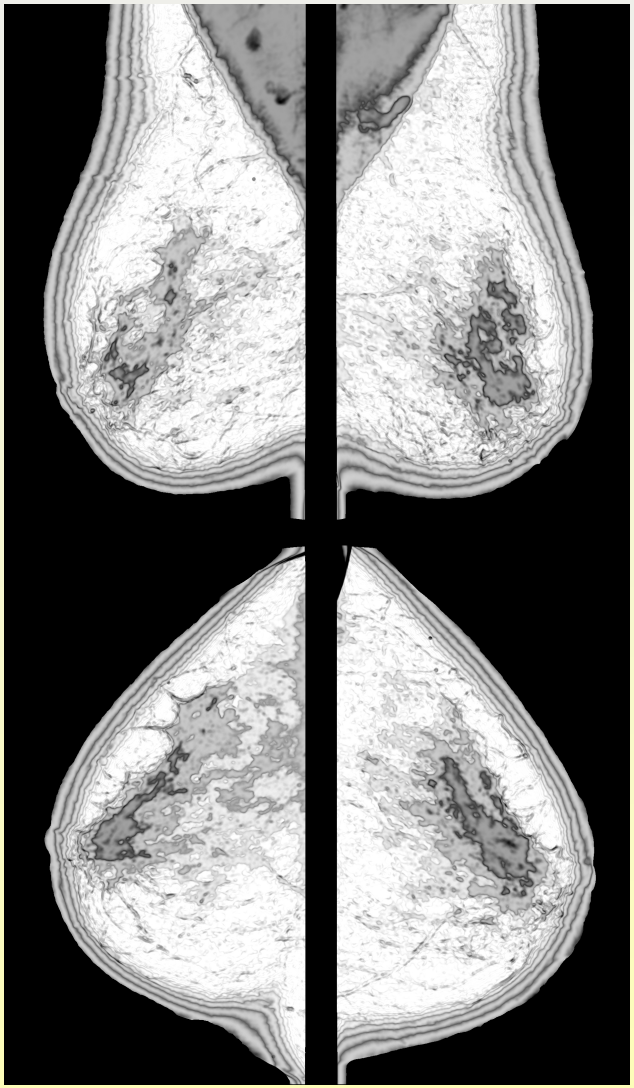
Pozn. Analýza pomocí věrohodnostního poměru potlačuje vliv úrovní šedi, které mají v případě mamogramu diagnostický význam.

Příklad 7: Vyhodnocování screeningových mamogramů

původní mamogram



věrohodnostní analýza

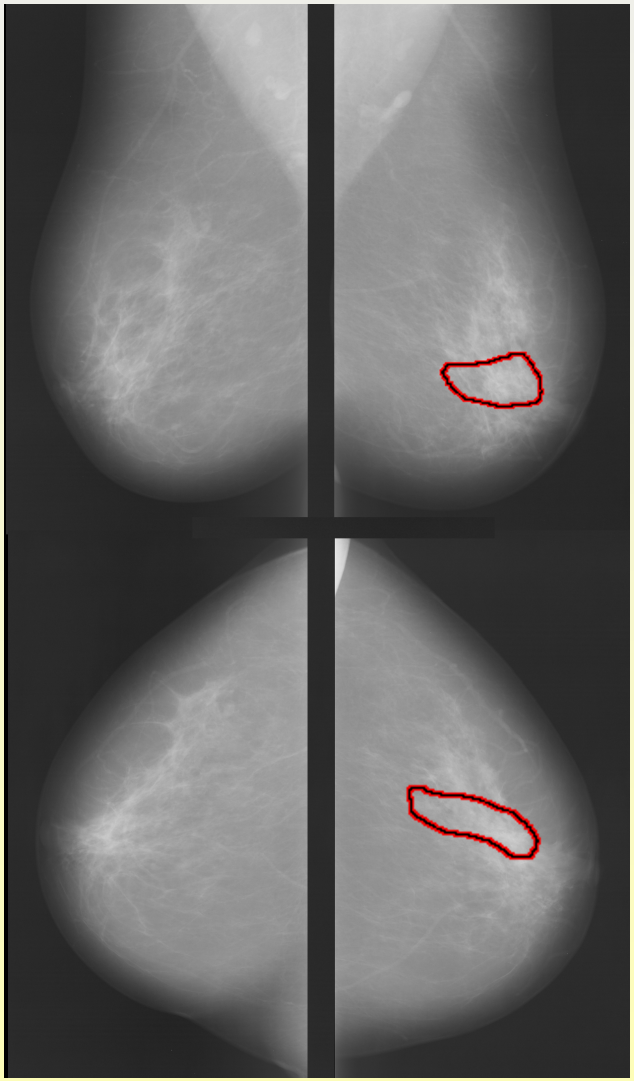


Příklad 7: Výpočetní aspekty vyhodnocování mamogramů

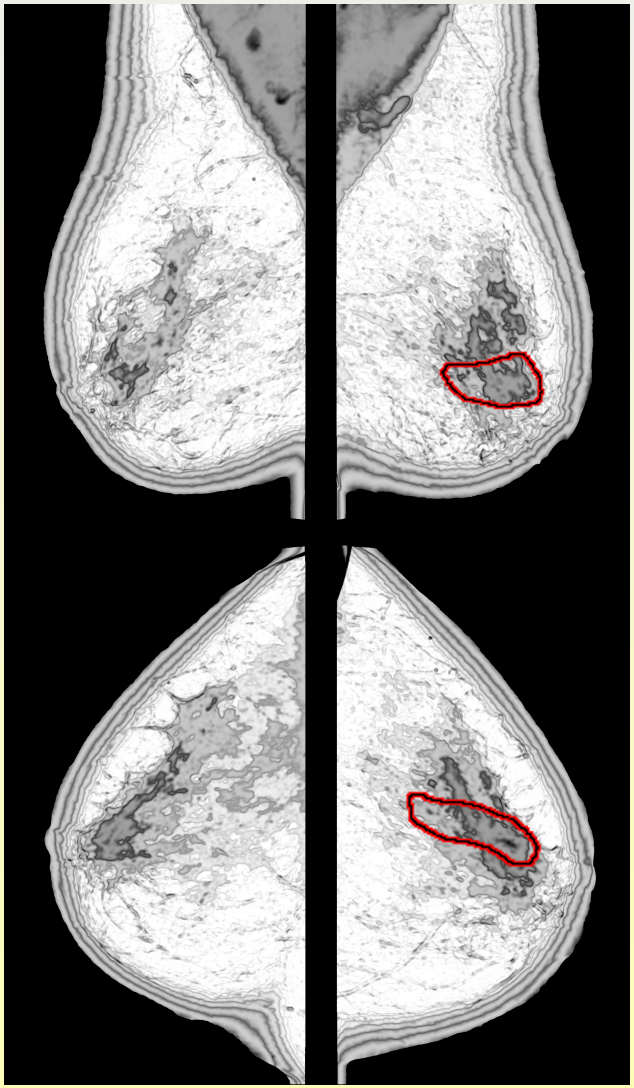
- zdrojová databáze: 2600 tzv. úplných mammogramů, University of South Florida
  http://marathon.csee.usf.edu/Mammography/Database.html
- úplný mamogram zahrnuje 4 snímky: dva medio-laterální pohledy a dva cranio-caudální pohledy a je vyhodnocován jako celek
- pravá část mamogramu je před vyhodnocením zrcadlově transformována aby byla využita pravo-levá symetrie snímků
- lokální analýza využívá čtvercové okno o rozměrech 13 x 13 pixelů s uříznutými rohy, dimenze vnitřku okna $\mathbf{x}$ je $N = 145$ $(= 169 - 4 \times 6)$
- počet komponent odhadované směsi je $M = 36$, parametry jsou inicializovány náhodně
- pro odhad parametrů směsi je k dispozici velký počet dat $|\mathcal{S}| \approx 10^5 - 10^6$ získaných posouváním okna v mamogramu
- lokální statistický model je odhadován individuálně z každého úplného mamogramu, tzn. metoda nevyžaduje trénovací data
- $\Rightarrow$ výsledek věrohodnostní analýzy není ovlivněn vysokou přirozenou variabilitou mamogramů

Příklad 7: Vyhodnocování screeningových mamogramů

ověřený nález

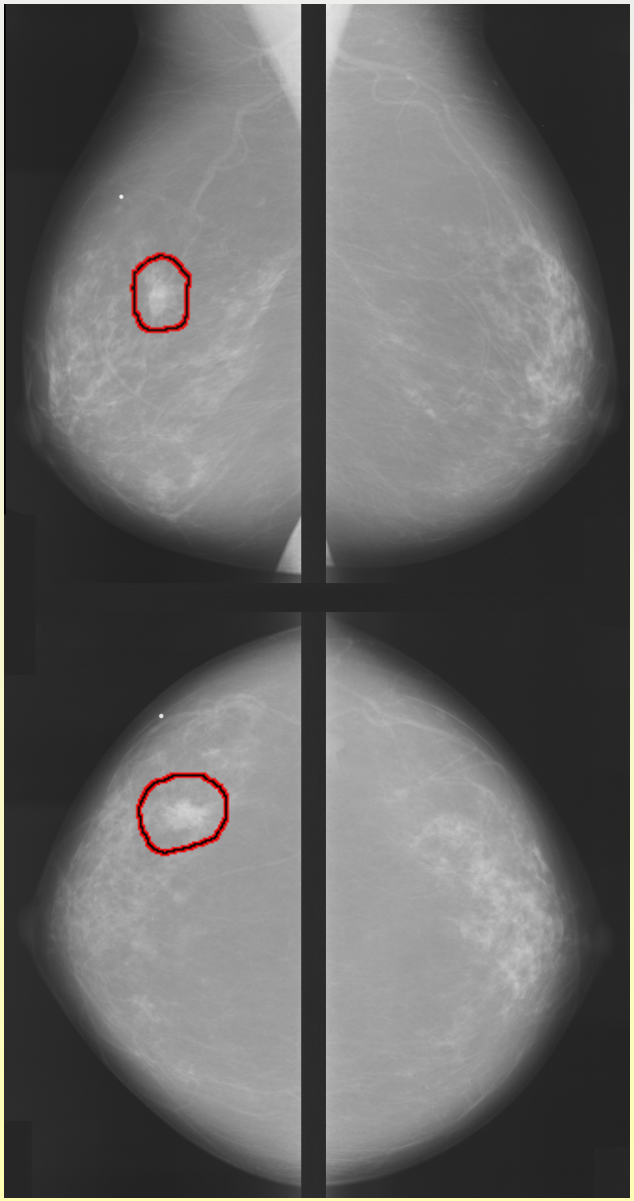


porovnání analýzy a nálezu

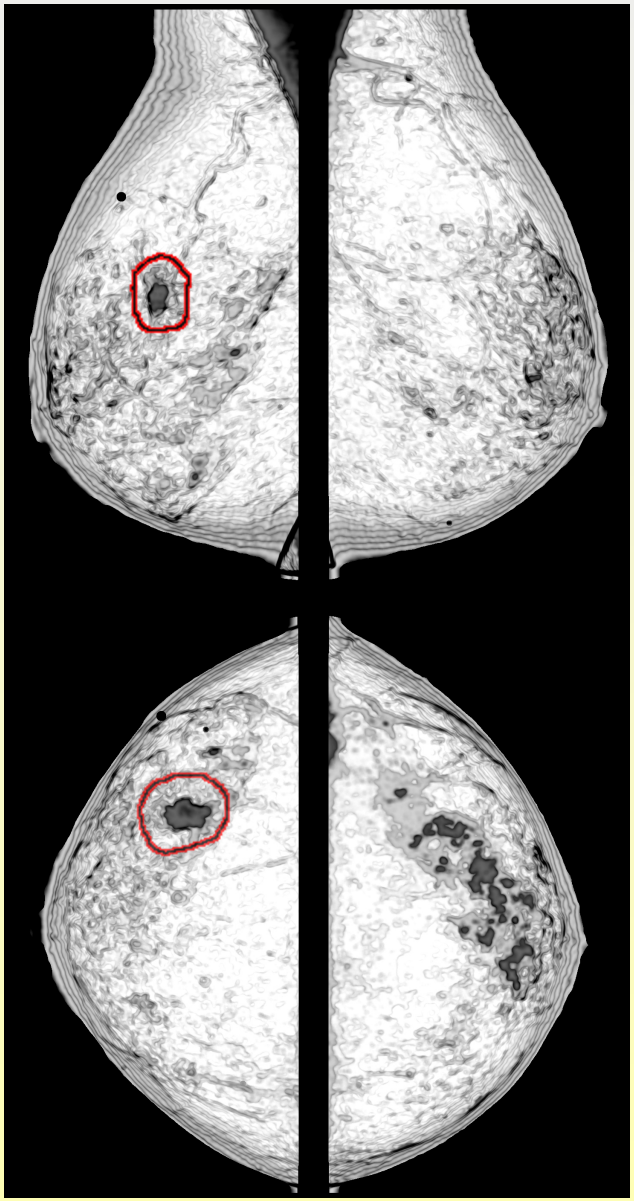


Příklad 7: Vyhodnocování screeningových mamogramů

ověřený nález

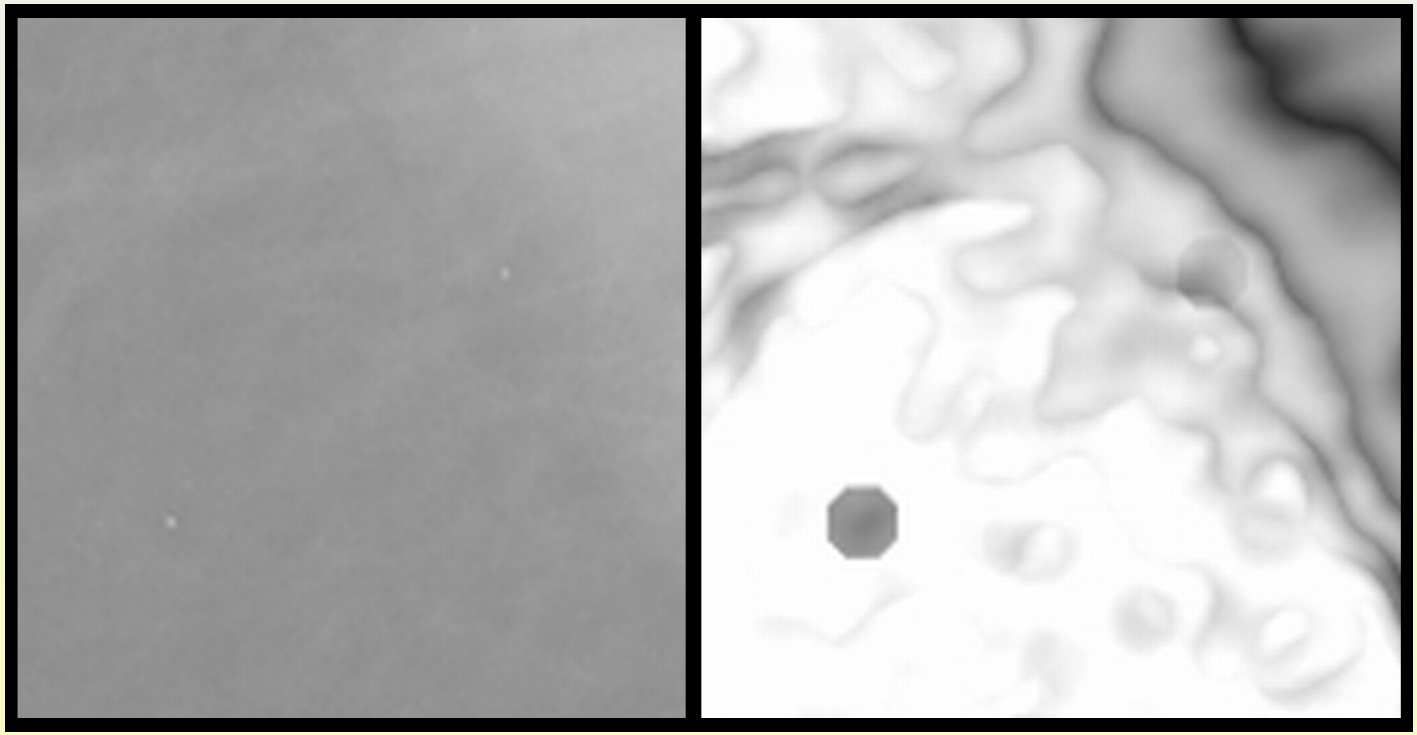


porovnání analýzy a nálezu



Příklad 7: Věrohodnostní analýza mikrokalcifikací

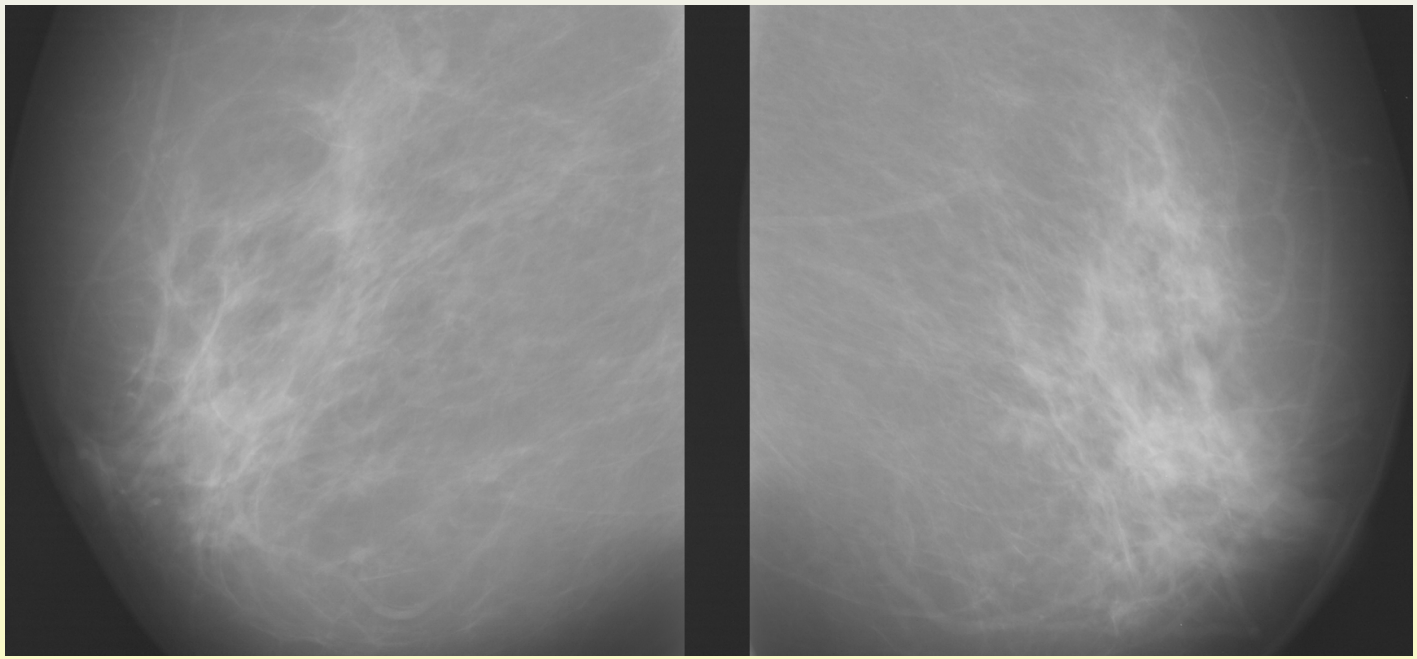
mikrokalcifikace zvýrazněné mikrokalcifikace



Remark: Každá pozice okna obsahující isolovaný světlý pixel implikuje sníženou hodnotu $\log P(\mathbf{x})$. $\Rightarrow$ Světlý pixel se zobrazí jako tmavší skvrna o velikosti okna.

Příklad 7: Identifikace "hmot" pomocí "vrstevnic"

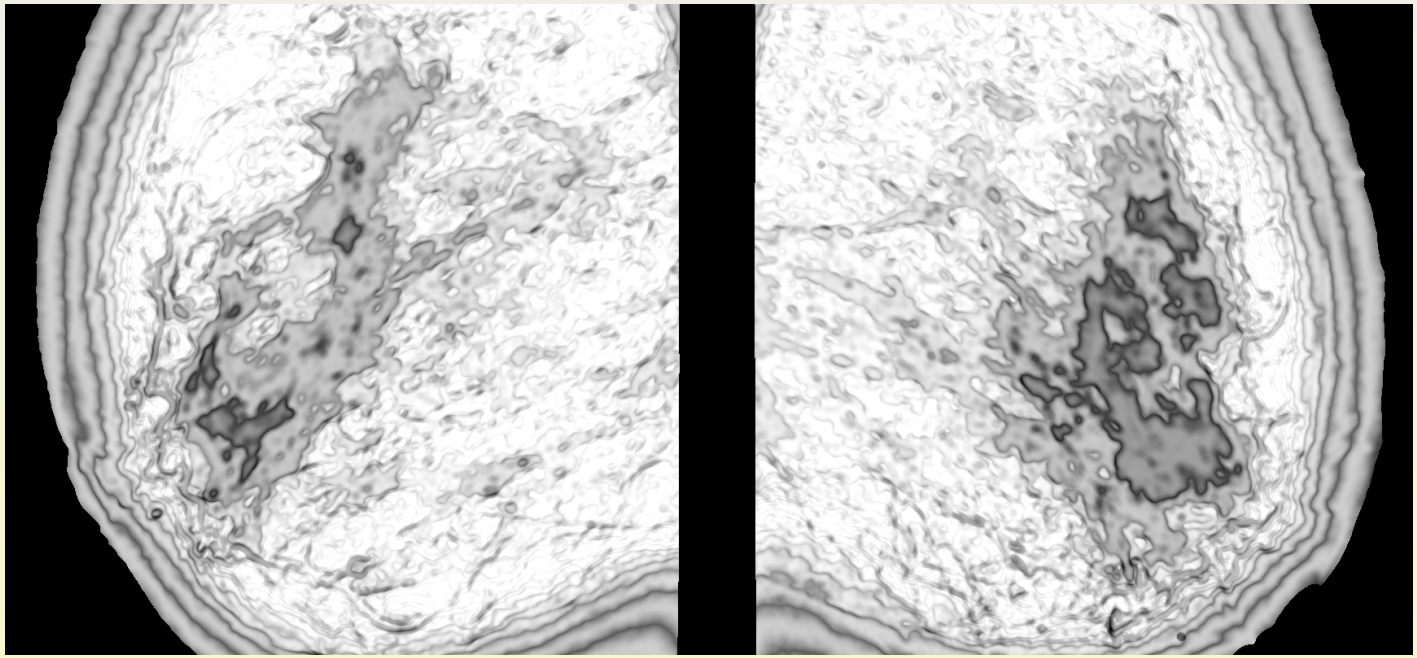
část screeningového mamogramu obsahující podezřelé "hmoty"



Remark: "Zhmotnění" může být velmi malé, může mít nezřetelné hranice a různé tvary. Detekce a klasifikace "hmot" se považuje za obtížnější než detekce mikrokalcifikací.

Příklad 7: Identifikace "hmot" pomocí "vrstevnic"

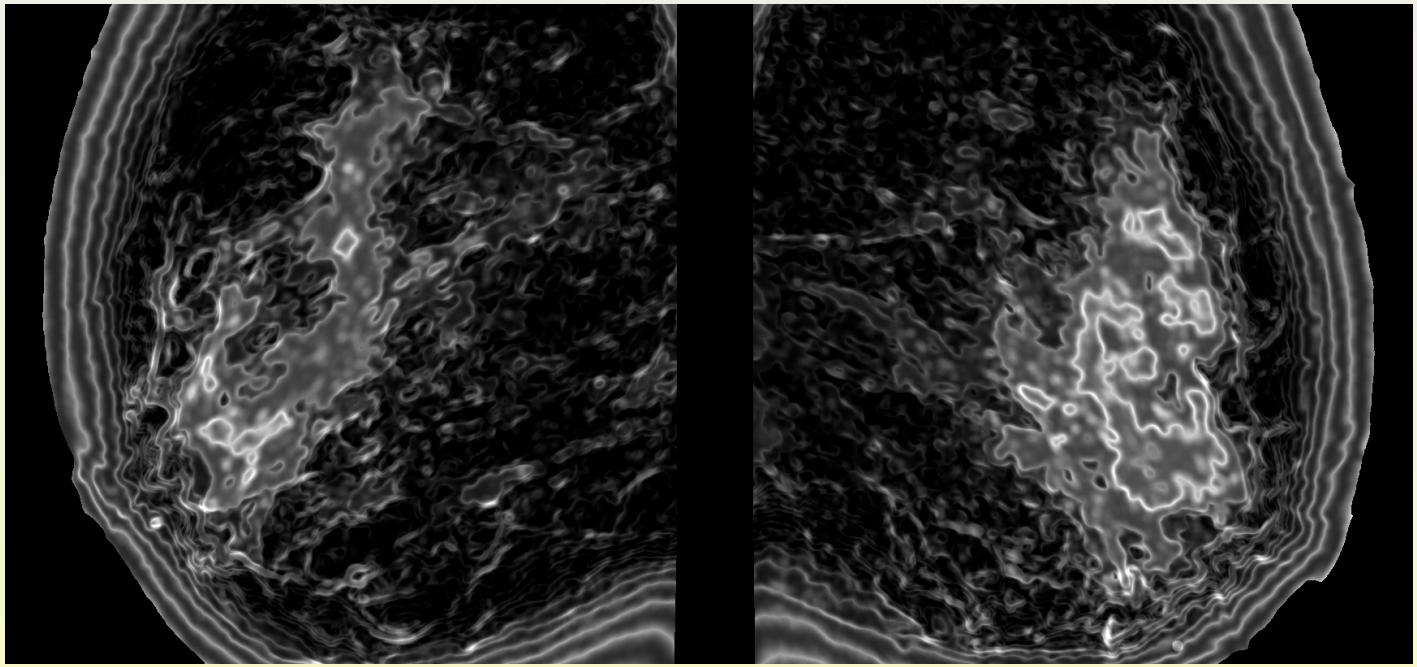
"vrstevnice" zvýrazňující hranice "hmot" a okraje mamogramu



Remark: Jednotlivé věrohodnostní hodnoty $\log P(\mathbf{x})$ jsou typicky určeny jedinou komponentou směsi, která nejlépe odpovídá dané pozici okna. Na okraji různých oblastí mamogramu dochází k záměně komponent, která je spojena s poklesem věrohodnosti $\log P(\mathbf{x})$. Záměna komponent je příčinou vzniku tmavších "vrstevnic" na hranici různých oblastí.

Příklad 7: Identifikace "hmot" pomocí "vrstevnic"

inverzní (negativní) zobrazení "vrstevnic"



Remark: Vrstevnice jsou nejzřetelněji zobrazeny na okraji mamogramu v místě spojitého poklesu úrovní šedi. Vrstevnice mohou usnadnit identifikaci častých kontralaterálních (symetricky lokalizovaných) nálezů a multifokálních nálezů protože oblasti s podobnými vlastnostmi jsou pomocí věrohodnostní analýzy snadno visuálně identifikovatelné.

Celkové shrnutí

Vlastnosti součinových distribučních směsí:

- efektivní výpočet parametrů směsi pomocí EM algoritmu z velkých datových souborů v mnohorozměrném prostoru (!)
- snadný výpočet marginálních rozložení pravděpodobnosti (!)
- vhodné pro aproximaci obecných rozložení pravděpodobnosti
- při velkém počtu komponent se vlastnosti součinové směsi blíží neparametrickému jádrovému odhadu
- EM algoritmus automaticky řeší problém optimalizace vyhlazení, která je nezbytná v případě neparametrického jádrového odhadu
- existuje strukturní modifikace součinové směsi, která umožňuje rozhodování nezávislé na dimenzi prostoru
- možnost odhadu parametrů směsi z neúplných datových vektorů
- EM algoritmus lze použít na vážená data (možnost zrychlení výpočtu agregací dat a jiné aplikace)
- součinové směsi lze interpretovat jako neuronovou síť (až na úrovni funkčních vlastností neuronů)

Příklad 2: Rozpoznávání číslic na binárním rastru

Publikované výsledky přesnosti klasifikace

Autor	rok	přesnost
Lam & Suen	(1988)	0.9310
Legault & Suen	(1989)	0.9390
Krzyzak et al.	(1990)	0.9485
Mai & Suen	(1990)	0.9295
Nadal & Suen	(1990)	0.8605
Suen et al.	(1990)	0.9305
Kim & Lee	(1994)	0.9585
Lee	(1995)	0.9780
Hwang & Bang	(1996)	0.9785
Cho	(1997)	0.9605

Standardní předzpracování dat:

normalizace velikosti číslic + výpočet speciálních informativních příznaků (např. "Kirsch masks", "peripheral directional contributivity" apod.)

Zpět

Příklad 5: Interaktivní statistický model dat ze sčítání lidu

Informační analýza dat (datamining) pomocí statistického modelu:

Výběr subpopulací $\mathcal{S}(\mathbf{x}_C) \in \mathcal{R}_\epsilon$ s nejvyšší/nejnižší hodnotou zvoleného informačního kritéria.

$\mathbf{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_C, \qquad C = \{i_1, \ldots, i_k\} \subset \mathcal{N}$

subpopulace: $\mathcal{S}(\mathbf{x}_C) = \{\mathbf{y} \in \mathcal{S} : \mathbf{y}_C = \mathbf{x}_C\} \subset \mathcal{S}$

(Např. dvojicí otázek ($|C| = 2,\ N = 25$) lze určit cca 10^4 subpopulací.)

ϵ-relevantní subpopulace (větší než $\epsilon|\mathcal{S}|$):

$$\mathcal{R}_\epsilon = \{\mathcal{S}(\mathbf{x}_C) \subset \mathcal{S} :\ P_C(\mathbf{x}_C) > \epsilon,\ \mathbf{x}_C \in \mathcal{X}_C,\ C \subset \mathcal{N}\}$$

jednoduchý výpočet podmíněných distribucí $P_{n|C}(x_n|\mathbf{x}_C), n \notin C$:

$$P_{n|C}(x_n|\mathbf{x}_C) = \sum_{m \in \mathcal{M}} W_m(\mathbf{x}_C) f_n(x_n|m), \quad W_m(\mathbf{x}_C) = \frac{w_m F_C(\mathbf{x}_C|m)}{\sum_{j=1}^{M} w_j F_C(\mathbf{x}_C|j)}$$

Příklad 5: Interaktivní statistický model dat ze sčítání lidu

PŘÍKLADY VOLBY INFORMAČNÍHO KRITÉRIA:

nejvyšší podmíněná pravděpodobnost $P_{n|C}(x_n|\mathbf{x}_C)$:
(např. subpopulace s nejvyšší nezaměstnaností)

$$P_{n|C}(x_n|\mathbf{x}_C) = \sum_{m=1}^{M} W_m(\mathbf{x}_C) f_n(x_n|m), \quad x_n \in \mathcal{X}_n, \ n \notin C$$

kritérium minimální entropie:
(např. možnost vyhledání typické vlastnosti)

$$H_{x_C}(\mathcal{X}_n) = \sum_{x_n \in \mathcal{X}_n} -P_{n|C}(x_n|\mathbf{x}_C) \log P_{n|C}(x_n|\mathbf{x}_C)$$

maximální množství informace mezi dvěma proměnnými:
(korelační koeficient není definován pro nominální proměnné)

$$I_{x_C}(\mathcal{X}_n, \mathcal{X}_r) = H_{x_C}(\mathcal{X}_n) + H_{x_C}(\mathcal{X}_r) - H_{x_C}(\mathcal{X}_n, \mathcal{X}_r)$$

POZN. Problém přesnosti reprodukce relativních četností v souboru $\mathcal{S}$ pomocí statistického modelu.

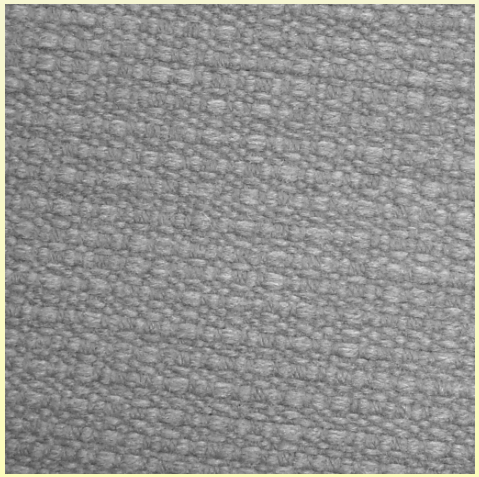
◀ Zpět

Příklad 6: Vyhledávání poruch a odchylek v textuře

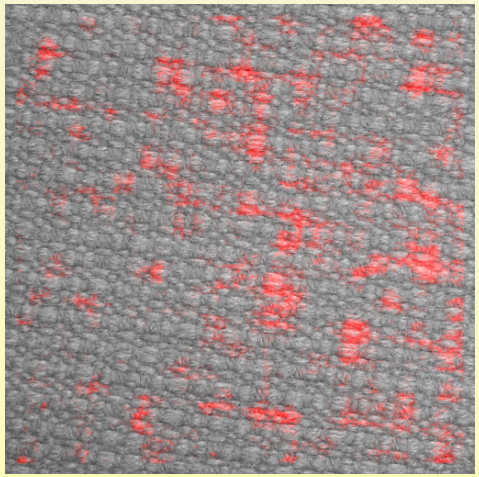
Analýza nepravidelnosti textury "látka":

krajní hodnoty $-\log P(\mathbf{x})$ (střední obrázek) resp. $-\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ (pravý obrázek) jsou zvýrazněny červeným zabarvením centrálního pixelu posuvného okna

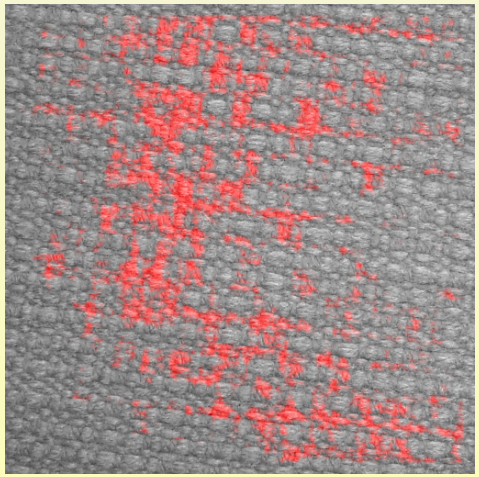
původní obrázek



L-nevěrohodnost



LR-nevěrohodnost

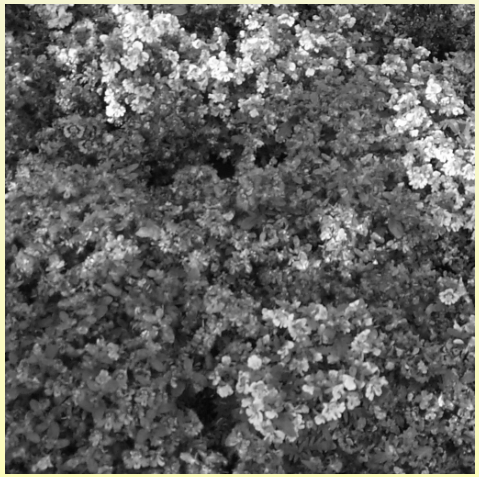


Příklad 6: Vyhledávání poruch a odchylek v textuře

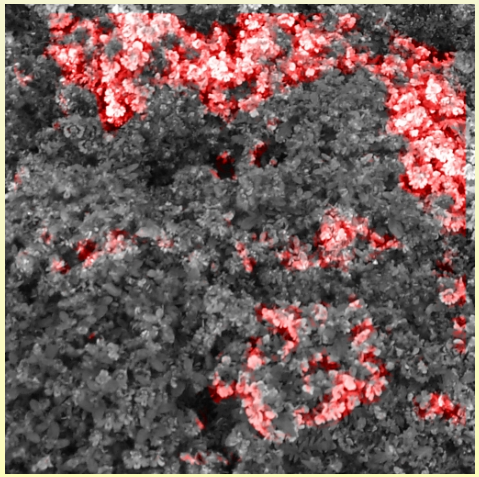
Analýza nepravidelnosti textury "květy":

krajní hodnoty $-\log P(\mathbf{x})$ (střední obrázek) resp. $-\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ (pravý obrázek) jsou zvýrazněny červeným zabarvením centrálního pixelu posuvného okna

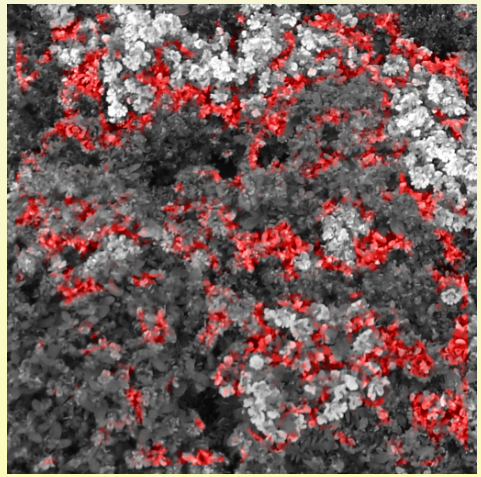
původní obrázek



L-nevěrohodnost



LR-nevěrohodnost

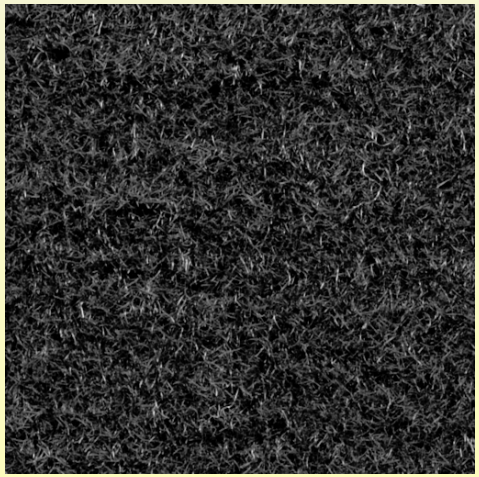


Příklad 6: Vyhledávání poruch a odchylek v textuře

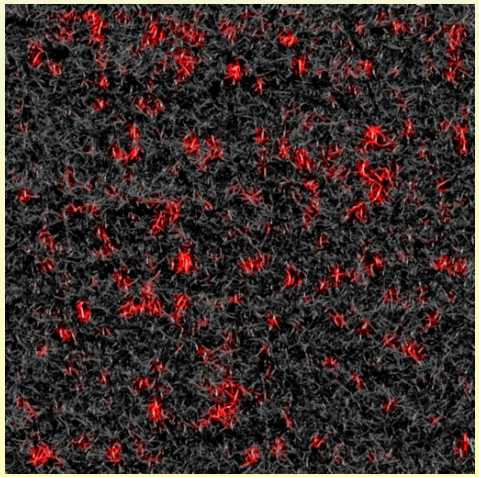
Analýza nepravidelnosti textury "koberec":

krajní hodnoty $-\log P(\mathbf{x})$ (střední obrázek) resp. $-\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ (pravý obrázek) jsou zvýrazněny červeným zabarvením centrálního pixelu posuvného okna

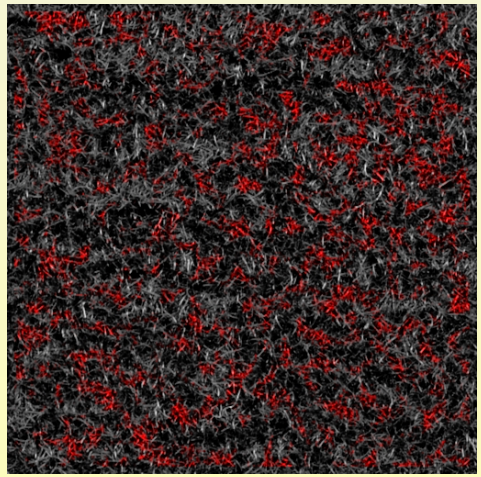
původní obrázek



L-nevěrohodnost



LR-nevěrohodnost



◂ Zpět

Sekvenční rozhodovací schema

statistický problém rozpoznávání (diskrétní proměnné):

$$\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}, \qquad P(\mathbf{x}|\omega)p(\omega), \qquad \omega \in \Omega$$

$$P(\mathbf{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} f(m) \prod_{n=1}^{N} f_n(x_n|m)$$

postupné doplňování příznaků x_n (např. v lékařské diagnostice):

Problém optimálního sekvenčního rozhodování:

volba nejinformativnější proměnné x_n pro danou podmnožinu (subvektor) známých vstupních údajů $\mathbf{x}_D = (x_{j_1}, \ldots, x_{j_l}) \in \mathcal{X}_D$ pomocí kriteria maximální podmíněné informace $I_{\mathbf{x}_D}(\mathcal{X}_n, \Omega), \ n \notin \mathcal{D}$

$$I_{\mathbf{x}_D}(\mathcal{X}_n, \Omega) = H_{\mathbf{x}_D}(\mathcal{X}_n) - H_{\mathbf{x}_D}(\mathcal{X}_n|\Omega), \qquad n \notin \mathcal{D}, \quad \mathcal{D} = \{j_1, \ldots, j_l\} \subset \mathcal{N}$$

$$H_{\mathbf{x}_D}(\mathcal{X}_n|\Omega) = \sum_{\omega \in \Omega} p(\omega) \sum_{x_n \in \mathcal{X}_n} -P_{n|D\omega}(x_n|\mathbf{x}_D, \omega) \log P_{n|D\omega}(x_n|\mathbf{x}_D, \omega)$$

$$P_{n|D\omega}(x_n|\mathbf{x}_D, \omega) = \frac{P_{nD|\omega}(x_n, \mathbf{x}_D|\omega)}{P_{D|\omega}(\mathbf{x}_D|\omega)} = \sum_{m \in \mathcal{M}_\omega} W_m(\mathbf{x}_D, \omega) f_n(x_n|m)$$

Zpět

Výběr nejinformativnějšího podprostoru

Motivace:

- výběr nejinformativnější podmnožiny příznaků
- vícestupňové rozpoznávání (urychlení a zpřesnění klasifikace)
- rychlá lokalizace grafických objektů v rovině (číslice, písmena)

výběr příznaků (proměnných): kritérium maximální informativnosti:

$$\mathcal{D}^* = \arg\max_{\mathcal{D}\subset\mathcal{N}} \{I(\mathcal{X}_D, \Omega)\} = \arg\max_{\mathcal{D}\subset\mathcal{N}} \{H(\mathcal{X}_D) - H(\mathcal{X}_D|\Omega)\}$$

$$P_{D|\omega}(\mathbf{x}_D|\omega) = \sum_{m\in\mathcal{M}_\omega} f(m) \prod_{n\in\mathcal{D}} f_n(x_n|m), \quad \mathcal{D} = \{j_1, \ldots, j_k\} \subset \mathcal{N}, \quad |\mathcal{D}| = k$$

$$H(\mathcal{X}_D) = \sum_{\mathbf{x}_D\in\mathcal{X}_D} -P_D(\mathbf{x}_D) \log P_D(\mathbf{x}_D)$$

$$H(\mathcal{X}_D|\Omega) = \sum_{\omega\in\Omega} p(\omega) \sum_{\mathbf{x}_D\in\mathcal{X}_D} -P_{D|\omega}(\mathbf{x}_D|\omega) \log P_{D|\omega}(\mathbf{x}_D|\omega)$$

optimální podmnožina $\mathcal{D} \subset \mathcal{N}$: úplné prohledání, přibližné metody

Zpět