SOUČINOVÉ DISTRIBUČNÍ SMĚSI

II. část: Příklady použití součinových směsí

Jiří Grim

Ústav teorie informace a automatizace AV ČR

Oddělení rozpoznávání obrazů

Listopad 2014

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

Outline

VLASTNOSTI SOUČINOVÝCH SMĚSÍ

ÚČEL: aproximace neznámého rozložení pravděpodobnosti

Výpočetní vlastnosti součinových distribučních směsí

- možnost aproximace multimodálních rozložení pravděpodobnosti (!)
- efektivní odhad parametrů směsi v mnohorozměrném prostoru (!)
- snadný výpočet marginálních rozložení pravděpodobnosti (!)
- při velkém počtu komponent se vlastnosti součinové směsi blíží obecnosti neparametrického jádrového odhadu
- směsi jsou jednodušší než jádrové odhady (méně komponent) není třeba řešit problém optimalizace vyhlazení
- možnost odhadu parametrů směsi z neúplných datových vektorů
- umožňují sekvenční rozhodování s postupným doplňováním nejinformativnějších příznaků
- strukturní modifikace součinové směsi umožňuje "lokální" výběr příznaků a rozhodování v prostorech s velkou dimenzi

▸ Literatura

EM algoritmus - směs Bernoulliho rozložení

KOMPONENTY: mnohorozměrná Bernoulliho rozložení

binární data: $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}, \quad x_n \in \{0,1\}, \quad \mathcal{X} = \{0,1\}^N$

(např. číslice na binárním rastru, výsledky biochemických testů a pod.)

$$F(\mathbf{x}|m) = F(\mathbf{x}|) = \prod_{n=1}^{N} f_n(x_n|m) = \prod_{n=1}^{N} \theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}$$

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} \log[\sum_{m=1}^{M} w_m F(\mathbf{x}|\boldsymbol{\theta}_m)], \qquad \mathcal{S} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\}$$

iterační rovnice:

▸ Příklad kódu: Bernoulliovská směs

$$q(m|\mathbf{x}) = \frac{w_m F(\mathbf{x}|\boldsymbol{\theta}_m)}{\sum_{j=1}^{M} f(j) F(\mathbf{x}|\boldsymbol{\theta}_j)}, \quad \mathbf{x} \in \mathcal{S}, \quad m = 1, 2, \ldots, M$$

$$w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}), \qquad \boldsymbol{\theta}_m^{'} = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{S}} \mathbf{x} q(m|\mathbf{x})$$

ÚTIA

EM algoritmus - součinová normální směs

KOMPONENTY: normální hustoty s diagonální kovarianční maticí:

$$F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{mn}} \exp\left\{ - \frac{(x_n - \mu_{mn})^2}{2\sigma_{mn}^2} \right\}, \quad \mathbf{x} \in \mathcal{R}^N$$

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} \log P(\mathbf{x}) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} \log \left\{ \sum_{m=1}^{M} w_m F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \right\}$$

iterační rovnice: $\mathbf{x} \in \mathcal{S},\ m = 1, 2, \ldots, M$

▸ Příklad kódu: normální směs

$$q(m|\mathbf{x}) = \frac{w_m F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)}{\sum_{j=1}^{M} f(j) F(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j)}, \qquad w_m' = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})$$

$$\mu_{mn}' = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{S}} x_n q(m|\mathbf{x}), \qquad n = 1, 2, \ldots, N$$

$$(\sigma_{mn}')^2 = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{S}} x_n^2 q(m|\mathbf{x}) \ - \ (\mu_{mn}')^2$$

ÚTIA

Strukturní modely směsi (Grim et al. 1986, 1999, 2002)

binární strukturní parametry: $\phi_m = (\phi_{m1}, \ldots, \phi_{mN}) \in \{0,1\}^N$

$$F(\mathbf{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}, \qquad \text{(obvykle: } f_n(x_n|0) = P_n(x_n)\text{)}$$

$\phi_{mn} = 0 \;\Rightarrow\;$ místo $f_n(x_n|m)$ se v součinu dosadí fixní distribuce $f_n(x_n|0)$

$$P(\mathbf{x}) = \sum_{m \in \mathcal{M}} F(\mathbf{x}|m) w_m = \sum_{m \in \mathcal{M}} F(\mathbf{x}|0) G(\mathbf{x}|m, \phi_m) w_m$$

$$G(\mathbf{x}|m, \phi_m) = \prod_{n \in \mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)} \right]^{\phi_{mn}}, \qquad F(\mathbf{x}|0) = \prod_{n \in \mathcal{N}} f_n(x_n|0)$$

motivace: **"distribuce pozadí"** $F(\mathbf{x}|0)$ **se vykrátí v Bayesově vzorci:**

$$p(\omega|\mathbf{x}) = \frac{P(\mathbf{x}|\omega)p(\omega)}{P(\mathbf{x})} = \frac{\sum_{m \in \mathcal{M}_\omega} G(\mathbf{x}|m, \phi_m) w_m}{\sum_{j \in \mathcal{M}} G(\mathbf{x}|j, \phi_j) f(j)} \approx \sum_{m \in \mathcal{M}_\omega} G(\mathbf{x}|m, \phi_m) w_m$$

POZN. $\approx$ Lokální výběr příznaků bez redukce dimenze.

Strukturní modifikace EM algoritmu - normální směs

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

$$L = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \log\left[\sum_{m\in\mathcal{M}} w_m \prod_{n\in\mathcal{N}} \left(\frac{f_n(x_n|\mu_{mn},\sigma_{mn})}{f_n(x_n|\mu_{0n},\sigma_{0n})} \right)^{\phi_{mn}} \right],$$

iterační rovnice: $(m \in \mathcal{M}, n \in \mathcal{N}, \mathbf{x} \in \mathcal{S})$

$$q(m|\mathbf{x}) = \frac{G(\mathbf{x}|m,\phi_m)w_m}{\sum_{j\in\mathcal{M}} G(\mathbf{x}|j,\phi_j)f(j)}, \qquad w_m^{'} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}),$$

$$\mu_{mn}^{'} = \frac{1}{w_m^{'}|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} x_n q(m|\mathbf{x}), \quad (\sigma_{mn}^{'})^2 = \frac{1}{w_m^{'}|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} x_n^2 q(m|\mathbf{x}) - (\mu_{mn}^{'})^2,$$

strukturní optimalizace: $\phi_{mn}^{'} = 1$ pro r nejvyšších hodnot $\gamma_{mn}^{'}$

$$\gamma_n^{'}(m) = \frac{w_m^{'}}{2}\left[\frac{(\mu_{mn}^{'}-\mu_{0n}^{'})^2 + (\sigma_{mn}^{'})^2}{(\sigma_{0n}^{'})^2} - 1 - 2\log\frac{\sigma_{mn}^{'}}{\sigma_{0n}^{'}} \right]$$

Strukturní modifikace EM algoritmu - diskrétní směs

$f_n(x_n|m),\ x_n \in \mathcal{X}_n,\ n \in \mathcal{N}\ \approx$ **diskrétní rozložení pravděpodobnosti**

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \Big[\sum_{m \in \mathcal{M}} F(\mathbf{x}|0) G(\mathbf{x}|m, \phi_m) w_m \Big], \qquad F(\mathbf{x}|0) = \prod_{n \in \mathcal{N}} f_n(x_n|0)$$

iterační rovnice: $(m \in \mathcal{M}, n \in \mathcal{N}, \mathbf{x} \in \mathcal{S})$

$$q(m|\mathbf{x}) = \frac{G(\mathbf{x}|m, \phi_m) w_m}{\sum_{j \in \mathcal{M}} G(\mathbf{x}|j, \phi_j) f(j)}, \qquad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x})$$

$$f_n^{'}(\xi|m) = \frac{1}{\sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x})} \sum_{\mathbf{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\mathbf{x})$$

strukturní optimalizace: $\phi_{mn}^{'} = 1$ pro r nejvyšších hodnot $\gamma_{mn}^{'}$

$$\gamma_{mn}^{'} = \sum_{\mathbf{x} \in \mathcal{S}} \frac{q(m|\mathbf{x})}{|\mathcal{S}|} \log \Big[\frac{f_n^{'}(x_n|m)}{f_n(x_n|0)} \Big] = w_m^{'} \sum_{x_n \in \mathcal{X}_n} f_n^{'}(x_n|m) \log \frac{f_n^{'}(x_n|m)}{f_n(x_n|0)}$$

POZN. Určení $\phi_{mn}^{'}$ je výhodnější pomocí prahu pro $\gamma_{mn}^{'}$.

ÚTIA

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 s pravděpodobnostmi $p(\omega)$

$P(\mathbf{x}|\omega), \quad \omega \in \Omega$: 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 rozpoznávání (postupné doplňování proměnných) a výběr informativních příznaků (globálně i lokálně)

Příklad 1: Rozpoznávání číslic na binárním rastru

PROBLÉM: rozpoznávání rukou psaných číslic

Databáze NIST SD19 obsahuje celkem cca 400 000 číslic na binárním rastru, tj asi 40 000 pro každou třídu; skládá se ze 7 srovnatelných částí napsaných uředníky z US Bureau of Census s výjimkou 4. části, kterou napsali středoškoláci z Bethesdy (horší kvalita)

DATA

- **liché datové vektory trénovací** ($\sum |\mathcal{S}_\omega| = 201485$ **číslic**)
- **sudé datové vektory testovací** ($\sum |\mathcal{S}_\omega^T| = 201479$ **číslic**)
- normalizace číslic na velikost rastru 32x32 (tj. dimenze $N = 1024$)
- rozšíření: doplněny tři rotace každé číslice (-10,-5,+5 stupňů)
- $\Rightarrow$ asi 80 000 trénovacích resp. testovacích číslic pro každé $\omega \in \Omega$

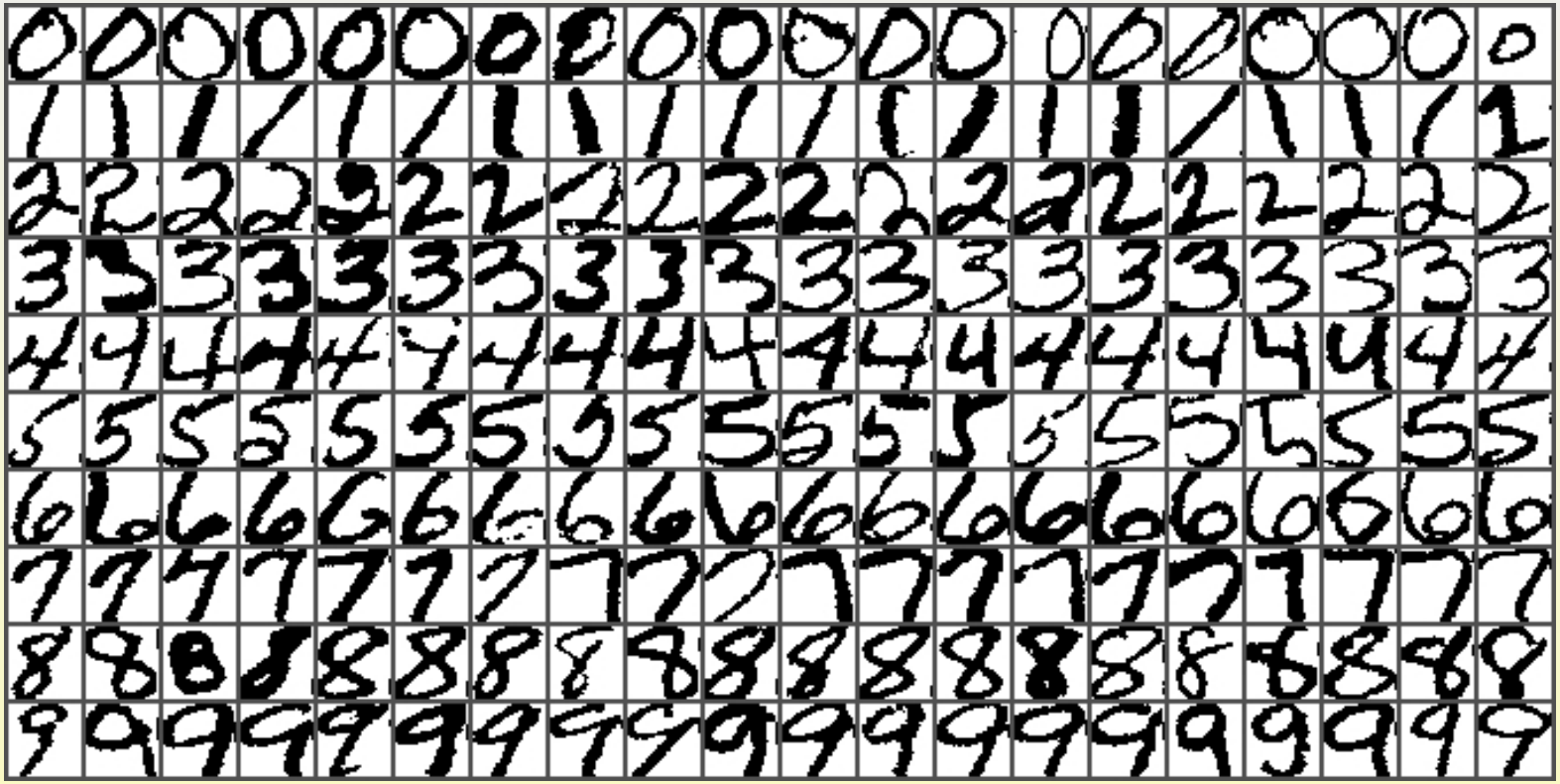
popis číslic: $x_n \in \{0,1\}, \quad \mathbf{x} = (x_1, x_2, \ldots, x_{1024}) \in \mathcal{X}, \quad \mathcal{X} = \{0,1\}^{1024}$

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

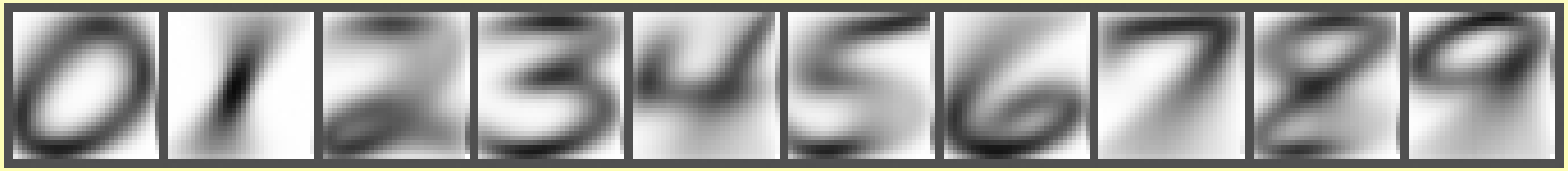
ÚTIA

Příklad 1: Databáze rukou psaných číslic NIST SD19

příklady číslic NIST SD19 normalizovaných na velikost rastru 32x32

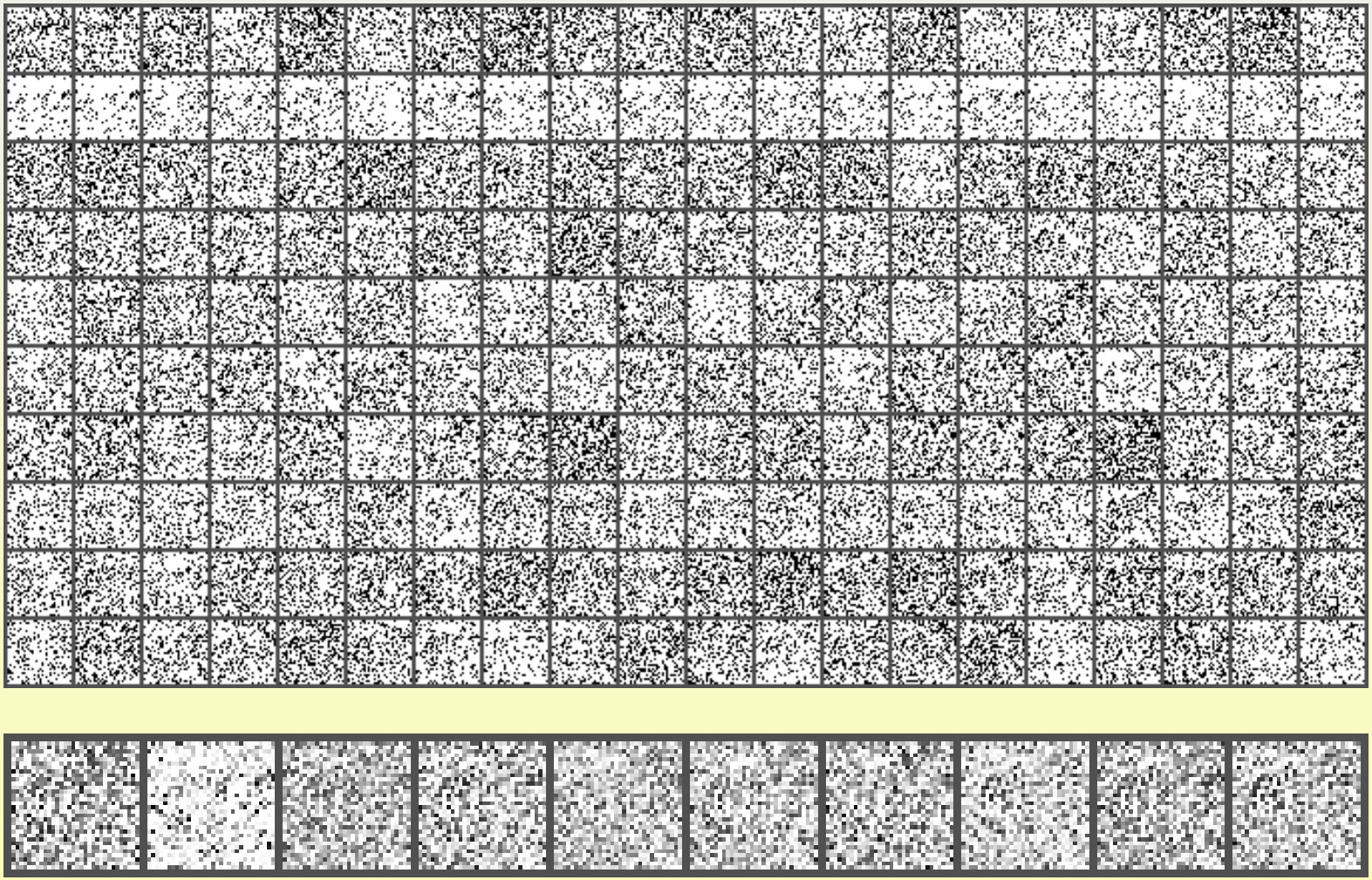


“průměrné číslice” (marginální pravděpodobnosti trénovacích dat)



Příklad 1: Ukázka permutace políček rastru

Příklady permutovaných číslic a příslušných průměrů



POZN. Přesnost rozpoznávání permutovaných číslic je identická.

Příklad 1: Rozpoznávání číslic na binárním rastru

ŘEŠENÍ: (Grim J., Hora J., 2010)

aproximace podmíněných distribucí $P(\mathbf{x}|\omega)$ v původním 1024-rozměrném prostoru pomocí **strukturní Bernoulliovské směsi**

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

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

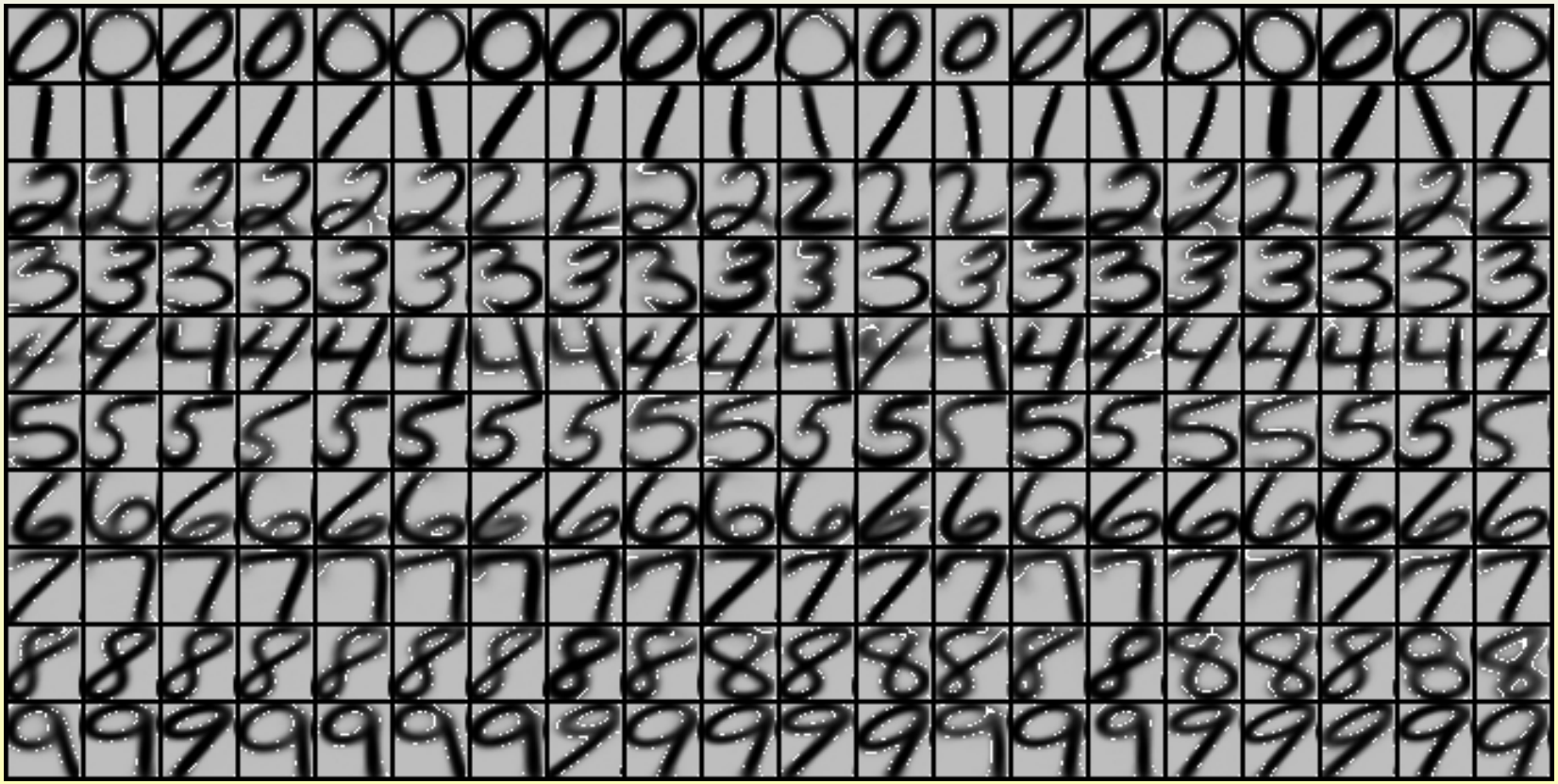
- $F(\mathbf{x}|0)$: konstantní distribuce pozadí ($\boldsymbol{\theta}_0 \approx$ "průměrná" číslice)
- celkový počet komponent: $\sum_\omega |\mathcal{M}_\omega| = 1571$
- počet nenulových parametrů: $\sum_{m,n} \phi_{mn} = 1462373$, tj. asi 90%
- náhodné počáteční hodnoty: $\theta_{mn} \in \langle 0.1, 0.9 \rangle$
- ukončení výpočtu prahem relativního přírůsku kriteria: $(L' - L)/L < 0.0001$)

ÚTIA

Příklad 1: Rozpoznávání číslic na binárním rastru

parametry komponent θ_{mn} jako úrovně šedi v uspořádání rastru

(bílá políčka označují "nepoužité" proměnné určené parametrem $\phi_{mn} = 0$)



Příklad 1: Rozpoznávání číslic na binárním rastru

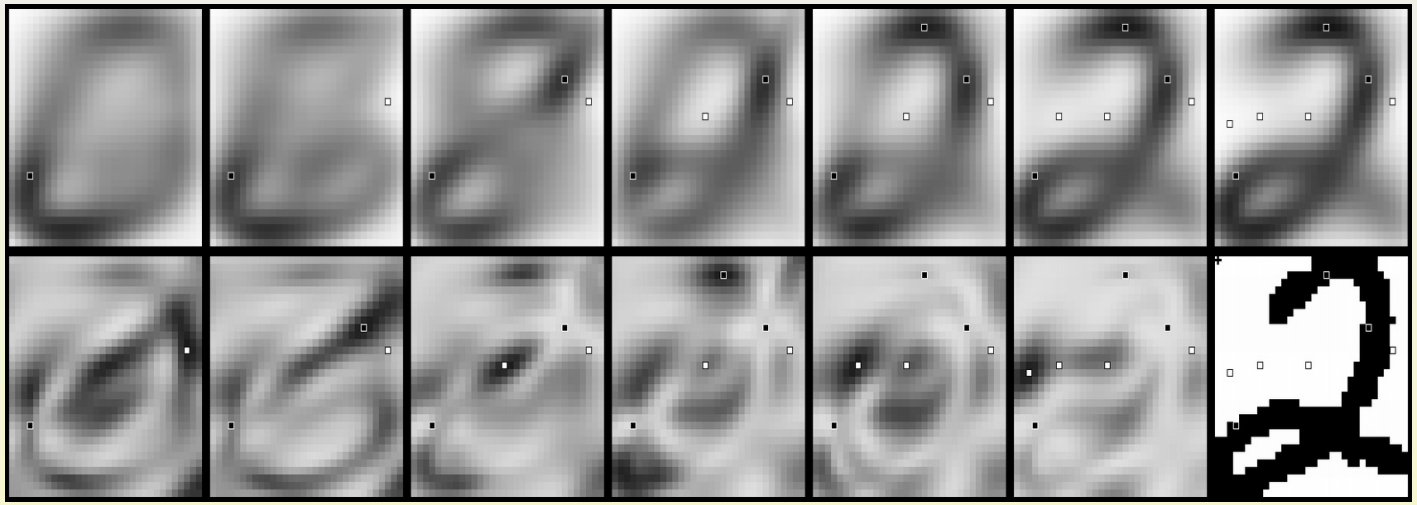
řádky: četnost jednotlivých rozhodnutí pro číslice z dané třídy
poslední sloupec: procento chybně klasifikovaných číslic z dané třídy
poslední řádek: četnost číslic chybně zařazených do dané třídy

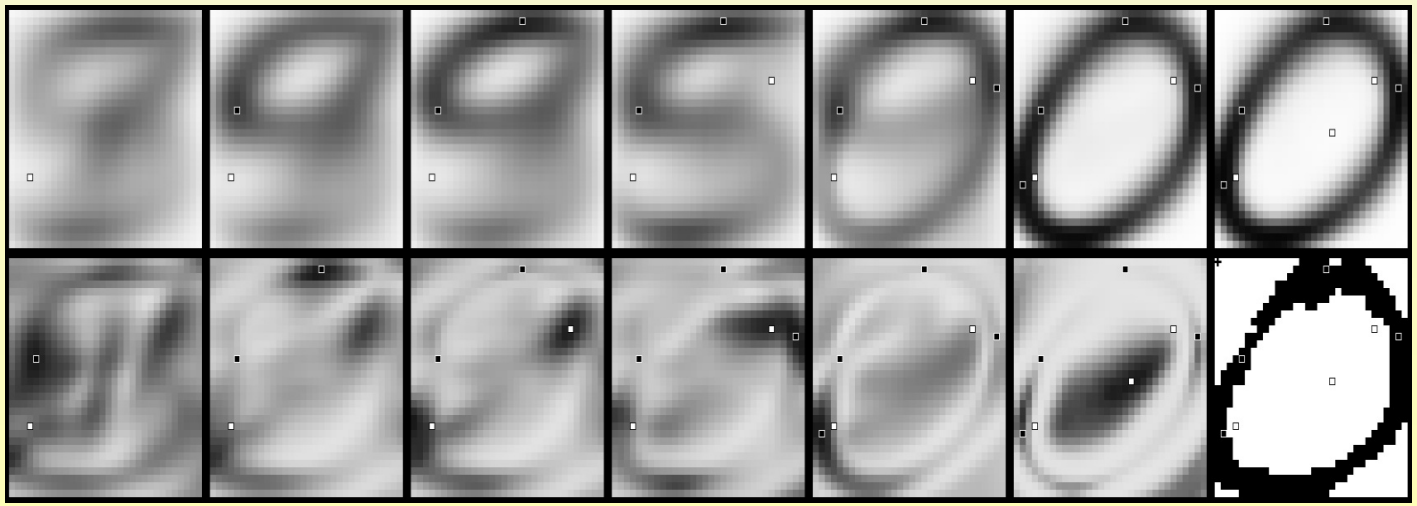
CLASS	0	1	2	3	4	5	6	7	8	9	chybně
testovací:	20182	22352	20038	20556	19577	18303	19969	20947	19790	19767	neg.:
0	19950	8	43	19	39	32	36	0	38	17	1.1%
1	2	22162	30	4	35	7	18	56	32	6	0.9%
2	32	37	19742	43	30	9	8	29	90	16	1.5%
3	20	17	62	20021	4	137	2	28	210	55	2.6%
4	11	6	19	1	19170	11	31	51	30	247	2.1%
5	25	11	9	154	4	17925	39	6	96	34	2.1%
6	63	10	17	6	23	140	19652	1	54	3	1.6%
7	7	12	73	10	73	4	0	20497	22	249	2.1%
8	22	25	53	97	30	100	11	11	19369	72	2.1%
9	15	13	25	62	114	22	3	146	93	19274	2.5%
chybně poz.:	197	139	537	396	352	462	148	328	665	699	1.84%

Celková chyba v procentech: 1.84%

Příklad 1: Optimální sekvenční rozhodování

volba proměnné x_n podle maximální podmíněné informace $I_{\mathbf{x}_D}(\mathcal{X}_n, \Omega)$





Příklad 2: Rozpoznávání obrazců na šachovnici

PROBLÉM: Složitost modelu (počet komponent) a "overfitting"

Problém:

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

rozměry šachovnice: 16 x 16 políček $\Rightarrow$ dimenze vektorů: N = 256
počet náhodných tahů: do obsazení 10 různých políček

$$\mathbf{x} = (x_1, \ldots, x_{256}) \in \{0,1\}^{256}, \quad x_n \in \{0,1\}, \quad \sum_{n=1}^{256} x_n = 10, \quad \Omega = \{\omega_1, \omega_2\}$$

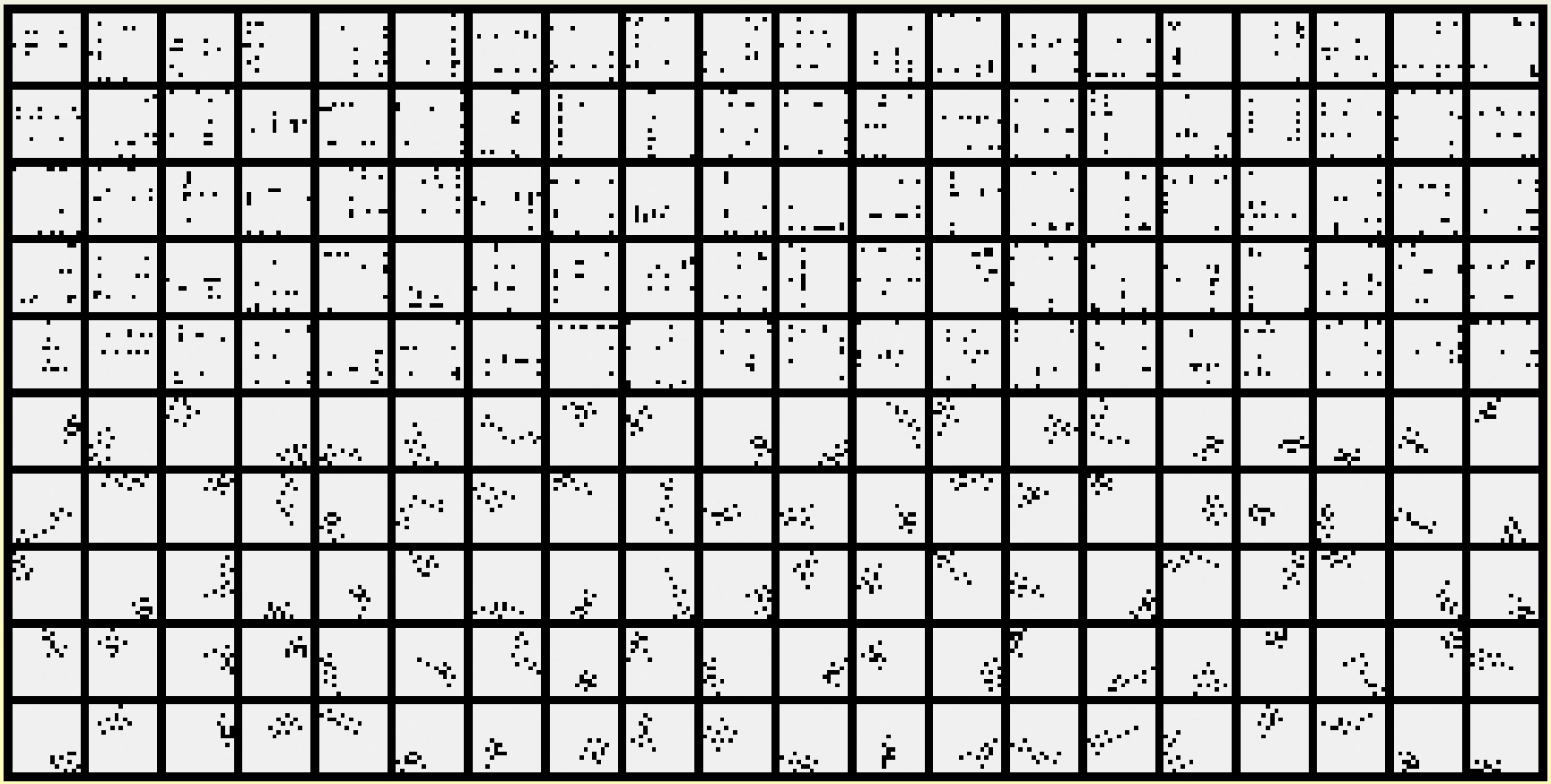
Vlastnosti problému:

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

Příklad 2: Rozpoznávání obrazců na šachovnici

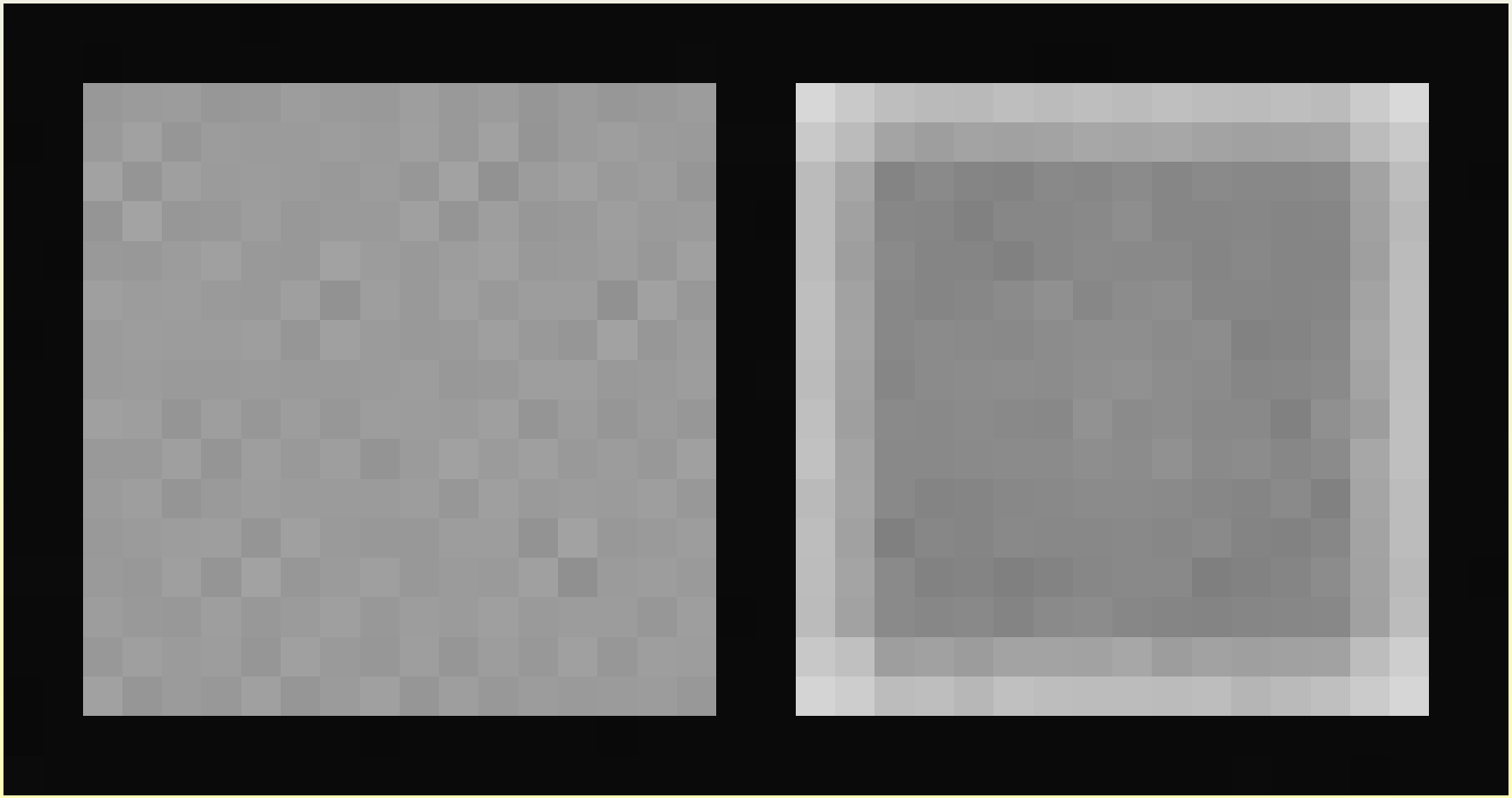
Náhodně generované obrazce

pro třídu "věž" (horní část) resp. "jezdec" (dolní část)



Příklad 2: Rozpoznávání obrazců na šachovnici

Marginální pravděpodobnosti pro třídu "věž"(vlevo) resp. "jezdec"(vpravo) zobrazené na šachovnici pomocí úrovní šedi



Příklad 2: Rozpoznávání obrazců na šachovnici

Řešení: (Grim J., Hora J. 2010)

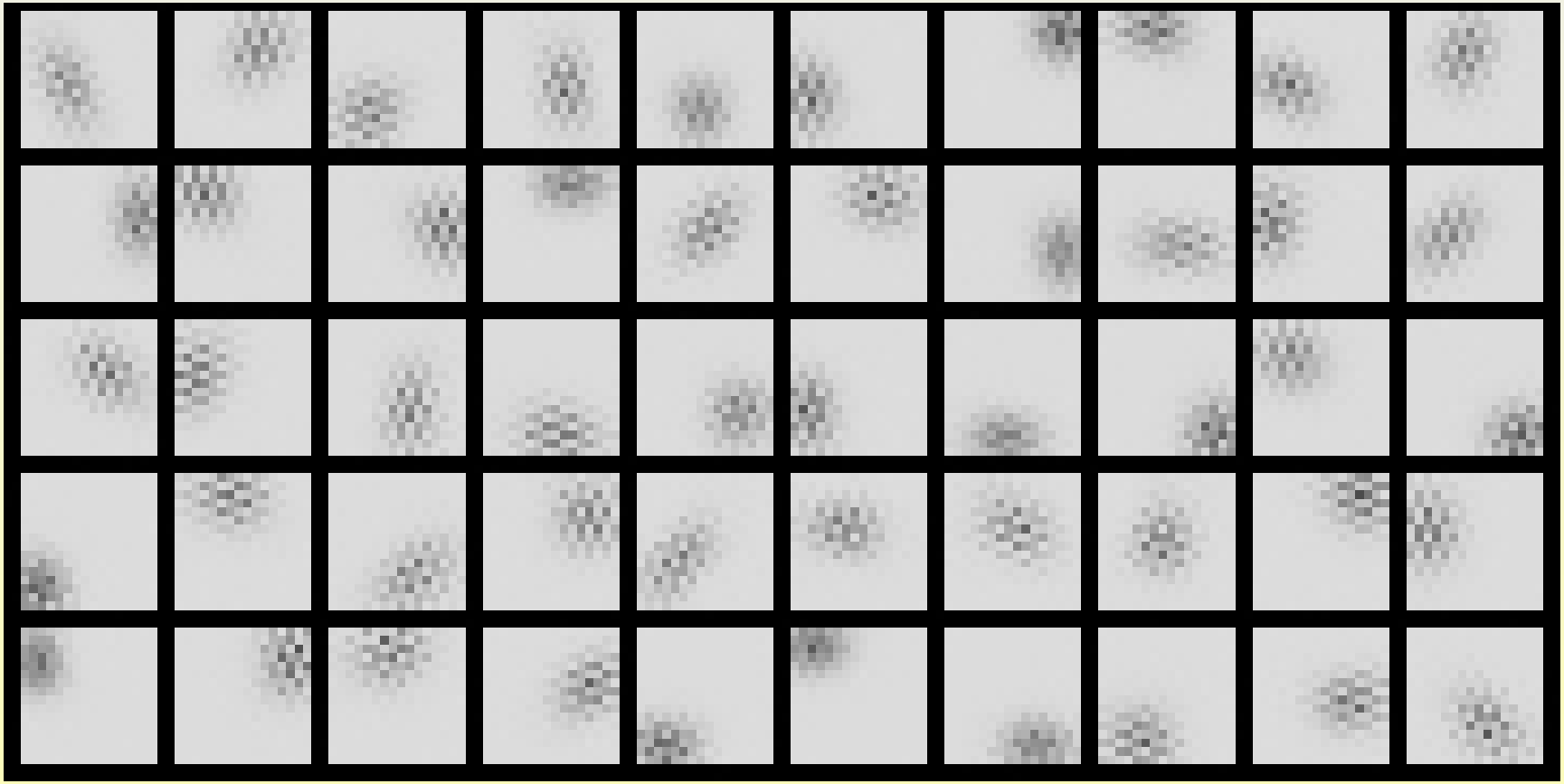
aproximace podmíněných distribucí $P(\mathbf{x}|\omega_1), P(\mathbf{x}|\omega_2)$ pomocí směsí Bernoulliho rozložení s dimenzí $N = 256$

$$P(\mathbf{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m \prod_{n=1}^{256} \theta_{mn}^{x_n} (1 - \theta_{mn})^{1-x_n}, \quad x_n \in \{0, 1\}, \ \omega \in \Omega$$

- počet komponent směsi: $|\mathcal{M}_\omega| = 1, 2, 5, 10, 20, 50, 100, 200, 500$
- identické počáteční váhy komponent: $w_m = 1/|\mathcal{M}_\omega|$
- velikost trénovacího souboru: $|\mathcal{S}_\omega| = 1000, \ 10000, \ 100000$
- počáteční hodnoty θ_{mn} generovány náhodně z intervalu $\langle 0.1, 0.9 \rangle$
- počet iterací EM algoritmu omezen podmínkou $(L' - L)/L < 0.0001$

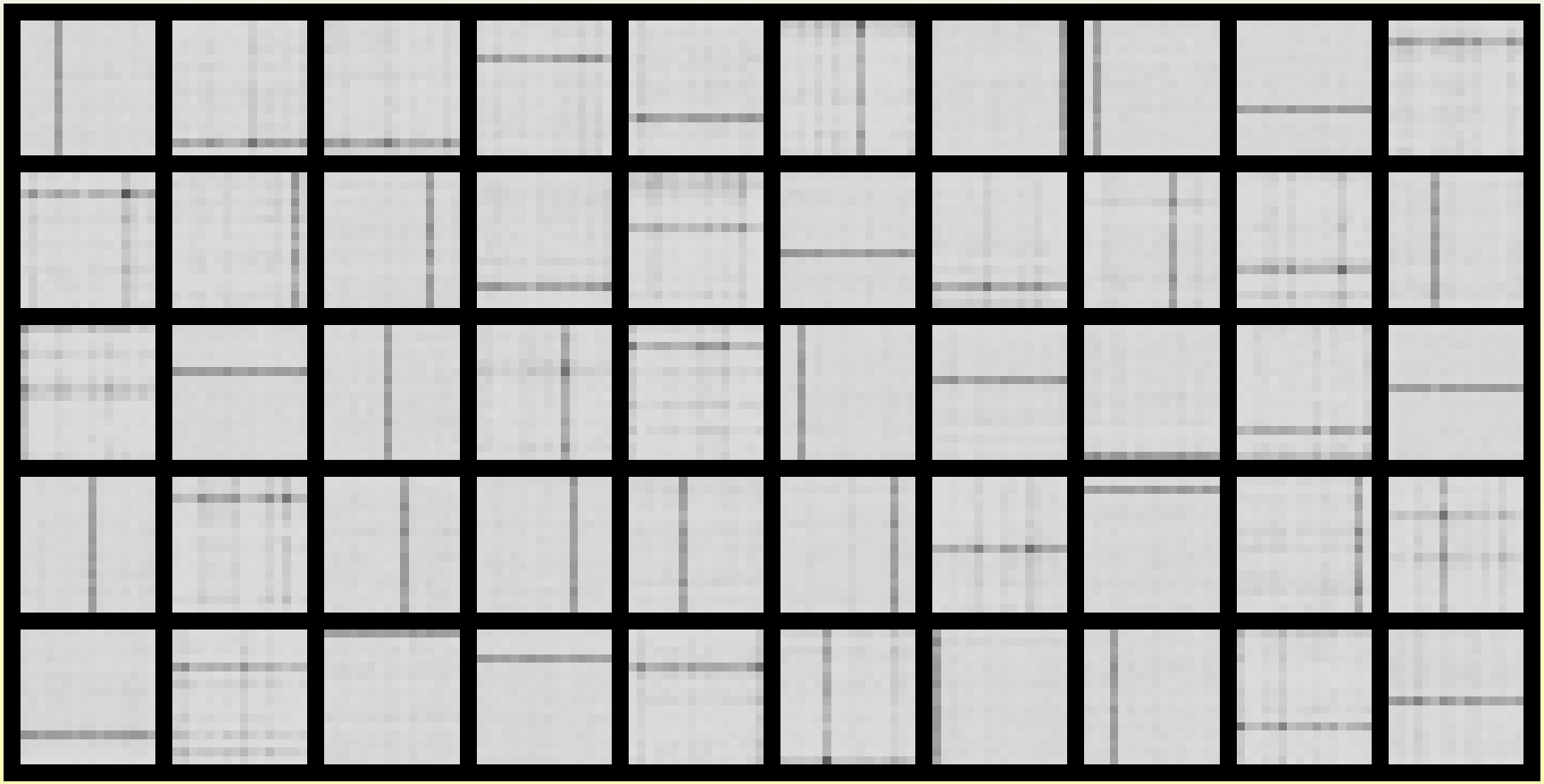
Příklad 2: Rozpoznávání obrazců na šachovnici

Odhady parametrů směsi θ_{mn} pro třídu "jezdec" (50 komponent) zobrazené v šachovnicovém uspořádání pomocí úrovní šedi



Příklad 2: Rozpoznávání obrazců na šachovnici

Odhady parametrů směsi θ_{mn} pro třídu "věž" (50 komponent) zobrazené v šachovnicovém uspořádání pomocí úrovní šedi



Příklad 2: Rozpoznávání obrazců na šachovnici

ROZPOZNÁVÁNÍ OBRAZCŮ NA ŠACHOVNICI (chyba v %)

$\lvert\mathcal{M}_\omega\rvert$	**1 000**	200 000	**10 000**	200 000	**100 000**	200 000
1	**34.70**	41.56	**39.59**	40.38	**39.90**	40.02
2	**13.10**	15.83	**16.54**	16.65	**16.42**	16.48
5	**1.65**	7.72	**6.60**	7.00	**6.49**	6.60
10	**0.95**	9.21	**5.40**	5.90	**4.04**	4.34
20	**0.15**	8.76	**3.91**	4.90	**2.73**	2.89
50	**0.00**	9.35	**2.01**	4.54	**1.37**	1.90
100	**0.00**	11.02	**1.22**	5.57	**0.84**	1.68
200	**0.00**	15.40	**0.69**	8.35	**0.45**	1.92
500	**0.00**	17.77	**0.20**	14.66	**0.14**	3.76

trénovací soubor (tučně): $|\mathcal{S}_\omega^{train}| = 1000,\ 10\,000,\ 100\,000$
nezávislý testovací soubor: $|\mathcal{S}_\omega^{test}| = 200\,000$

POZN. složitost modelu $\times$ velikost trénovacího souboru + "overfitting"
(pro daný trénovací soubor existuje optimální složitost směsi)

ÚTIA

Příklad 3: Klasifikace textových dokumentů (Grim et al. 2008)

PROBLÉM: **automatické třídění dokumentů do předem daných tříd**

textový dokument:
$\mathbf{d} = \langle w_{i_1}, \ldots, w_{i_k} \rangle \approx$ seznam termínů z nějakého slovníku $\mathcal{V}$

slovník termínů: $\mathcal{V} = \{t_1, \ldots, t_N\} \approx$ množina informativních termínů (odvozená z trénovacích dokumentů odstraněním spojek, koncovek a vzácných slov, typicky $N \approx 10^4$)

dokument jako "pytel slov"
(zápis pomocí četnosti slovníkových termínů)

$\mathbf{x} = \mathbf{x}(\mathbf{d}) = (x_1, \ldots, x_N) \in \mathcal{X} \approx$ vektor celých čísel

$x_n \approx$ četnost termínu $t_n \in \mathcal{V}$ $\quad |\mathbf{x}| = \sum_{n=1}^{N} x_n \approx$ délka dokumentu $\mathbf{x}$

POZN.: Zápis dokumentu pomocí "pytle slov" ignoruje pořadí slov.

Příklad 3: Klasifikace textových dokumentů

pravděpodobnostní popis:

$\mathcal{C} = \{c_1, \ldots, c_J\} \approx$ množina tříd dokumentů

$P(\mathbf{x}|c)p(c),\ c \in \mathcal{C} \approx$ podmíněné distribuce dokumentů

$p(c),\ c \in \mathcal{C} \approx$ apriorní pravděpodobnosti tříd

"naivní" Bayesův klasifikátor:

$$p(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)p(c)}{P(\mathbf{x})}, \qquad P(\mathbf{x}) = \sum_{c \in \mathcal{C}} p(c)P(\mathbf{x}|c)$$

předpokládá podmíněnou nezávislost proměnných:

$$P(\mathbf{x}|c) = \prod_{n \in \mathcal{N}} f_n(x_n|c),\ c \in \mathcal{C}, \quad \mathcal{N} = \{1, \ldots, N\}$$

POZN.: Naivní Bayesův klasifikátor ignoruje statistické závislosti slovníkových termínů. Složtější statistické modely však přes četné pokusy nepřinesly podstatné zlepšení.

Příklad 3: Klasifikace textových dokumentů

IDEA: aproximace distribucí $P(\mathbf{x}|c)$ pomocí Poissonovských směsí:

$$P(\mathbf{x}|c) = \sum_{m\in\mathcal{M}_c} w_m F(\mathbf{x}|\boldsymbol{\lambda}_m) = \sum_{m\in\mathcal{M}_c} w_m \prod_{n\in\mathcal{N}} f_n(x_n|\lambda_{mn})$$

$F(\mathbf{x}|\boldsymbol{\lambda}_m) \approx$ součinové Poissonovské distribuce

pravděpodobnost x_n výskytů termínu $t_n \in \mathcal{V}$ v dokumentu o délce $|\mathbf{x}|$:

$$f_n(x_n|\lambda_{mn}) = \frac{(\lambda_{mn})^{x_n}}{x_n!}\mathrm{e}^{-\lambda_{mn}}, \qquad (|\mathbf{x}| = \sum_{n=1}^{N} x_n)$$

$\lambda_{mn} \approx$ střední četnost termínu t_n v dokumentu s délkou $|\mathbf{x}|$

dokumenty mohou mít různou délku: $\Rightarrow \quad \lambda_{mn} = \theta_{mn}|\mathbf{x}|$

$\theta_{mn} \approx$ **pravděpodobnost výskytu termínu t_n v dokumentu**

$$P(\mathbf{x}|c) = \sum_{m\in\mathcal{M}_c} F(\mathbf{x}|\boldsymbol{\theta}_m) w_m = \prod_{n\in\mathcal{N}} f_n(x_n|\theta_{mn}|\mathbf{x}|) = \prod_{n\in\mathcal{N}} \frac{(\theta_{mn}|\mathbf{x}|)^{x_n}}{x_n!}\mathrm{e}^{-\theta_{mn}|\mathbf{x}|}$$

POZN. Směs poissonovských distribucí má $M(N+1)$ parametrů.
($\approx$ velké číslo při vysokém počtu slovníkových termínů N.)

ÚTIA

Příklad 3: Klasifikace textových dokumentů

“strukturní” mnohorozměrné poissonovské směsi: ▸ EM algoritmus

$$P(\mathbf{x}|c) = \sum_{m \in \mathcal{M}_c} F(\mathbf{x}|\boldsymbol{\theta}_0) G(\mathbf{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m) w_m, \quad c \in \mathcal{C}$$

$F(\mathbf{x}|\boldsymbol{\theta}_0) \approx$ distribuce “pozadí” společná pro všechny třídy dokumentů

$$F(\mathbf{x}|\boldsymbol{\theta}_0) = \prod_{n \in \mathcal{N}} f_n(x_n|\theta_{0n}|\mathbf{x}|) = \prod_{n \in \mathcal{N}} \frac{(\theta_{0n}|\mathbf{x}|)^{x_n}}{x_n!} \mathrm{e}^{-\theta_{0n}|\mathbf{x}|}$$

$G(\mathbf{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m) \approx$ komponenty, $\phi_{mn} \in \{0, 1\} \approx$ strukturní parametry

$$G(\mathbf{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m) = \prod_{n \in \mathcal{N}} \left[\frac{f_n(x_n|\theta_{mn}|\mathbf{x}|)}{f_n(x_n|\theta_{0n}|\mathbf{x}|)} \right]^{\phi_{mn}} = \prod_{n \in \mathcal{N}} \left[\left(\frac{\theta_{mn}}{\theta_{0n}} \right)^{x_n} \mathrm{e}^{(\theta_{0n} - \theta_{mn})|\mathbf{x}|} \right]^{\phi_{mn}}$$

distribuce “pozadí” $F(\mathbf{x}|\boldsymbol{\theta}_0)$ se krátí v Bayesově vzorci:

$$p(c|\mathbf{x}) = \frac{p(c) \sum_{m \in \mathcal{M}_c} G(\mathbf{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m) w_m}{\sum_{c \in \mathcal{C}} p(c) \sum_{j \in \mathcal{M}_c} G(\mathbf{x}|\boldsymbol{\theta}_j, \boldsymbol{\phi}_j) f(j)}.$$

ÚTIA

Příklad 3: Výsledky klasifikace dokumentů REUTERS

textové dokumenty REUTERS:

8941 dokumentů v **33** různě velkých třídách

10105 slovníkových termínů (bez spojek, koncovek a vzácných slov)

6431 trénovacích dokumentů, **2510** testovacích dokumentů
(≈ "APTE split" bez malých a vícenásobně řazených dokumentů)

Experiment č.	**1**	**2**	**3**	**4**	**5**
Počet komponent:	33	33	35	35	43
Počet parametrů:	333465	208366	285220	327184	201417
Počet parametrů [v %]:	100.0	62.5	80.6	92.5	46.4
Počet chyb:	155	156	162	152	147
Chybně [v %]:	6.17	6.21	6.45	6.07	5.86

POZN. Nejlepší výsledek klasifikace (experiment 5) je jen nepatrně lepší než chyba "naivního" Bayesova klasifikátoru (experiment 1).

Příklad 3: Výsledky klasifikace dokumentů NEWSGROUPS

textové dokumenty "20 NEWSGROUPS":

19956 dokumentů v **20** různých srovnatelně velkých třídách

31826 slovníkových termínů (bez spojek, koncovek a vzácných slov)

13314 trénovacích dokumentů, **6632** testovacích dokumentů
(≈ náhodný rozklad bez vícenásobně zařazených dokumentů)

Experiment No.	**1**	**2**	**3**	**4**
Počet komponent:	20	40	40	80
Počet parametrů:	636520	1204262	1102073	1024782
Počet parametrů [v %]:	100.0	94.6	86.6	40.2
Počet chyb:	1406	1379	1370	1412
Chybně [v %]:	21.20	20.79	20.66	21.29

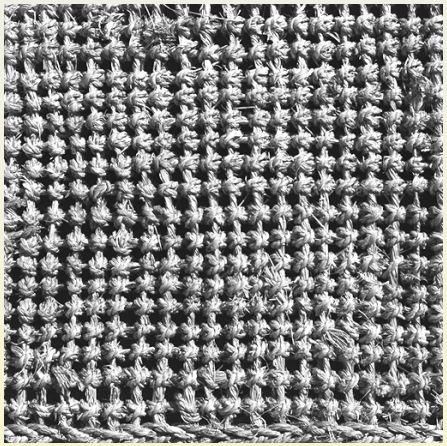
POZN. Výsledky klasifikace se liší jen v desítkách dokumentů, "naivní" Bayesův klasifikátor (experiment 1) je jen nepatrně horší.

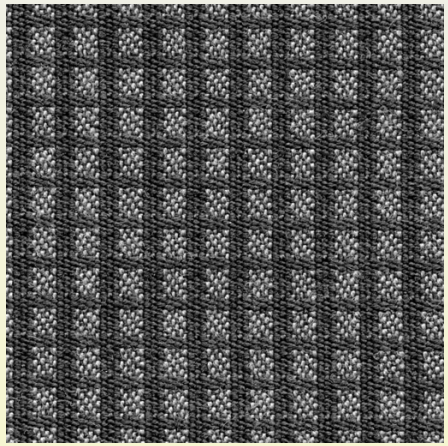
ÚTIA

Příklad 4: 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 ($N \approx 10^2$)

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

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

ÚTIA

Příklad 4: Modelování textur pomocí normální směsi

Princip modelování (Grim et al. 2003, 2004, 2005, 2006):

- 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 $P(\mathbf{x})$
- $\Rightarrow$ **unikátní možnost vizuálního posouzení kvality odhadnuté směsi**

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

$\mathcal{D} = \{j_1, \ldots, j_l\} \subset \mathcal{N} \quad \approx$ definovaná část okna
$\mathcal{C} = \{i_1, \ldots, i_k\} = \mathcal{N} \setminus \mathcal{D} \quad \approx$ nedefinovaná část okna

marginální distribuce:

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

Příklad 4: Modelování textur pomocí normální směsi

podmíněné distribuce:

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

$$W_m(\mathbf{x}_D) = \frac{w_m F(\mathbf{x}_D|\boldsymbol{\mu}_{mD},\boldsymbol{\sigma}_{mD})}{\sum_{j\in\mathcal{M}} f(j)F(\mathbf{x}_D|\boldsymbol{\mu}_{jD},\boldsymbol{\sigma}_{jD})} \quad \approx \quad \text{téměř binární}$$

PREDIKCE: očekávaná hodnota $\bar{\mathbf{x}}_C$ při definované části $\mathbf{x}_D$:

$$\bar{\mathbf{x}}_C = \mathrm{E}_{C|D}\{\mathbf{x}_C|\mathbf{x}_D\} = \int \mathbf{x}_C P_{C|D}(\mathbf{x}_C|\mathbf{x}_D)d\mathbf{x}_C = \sum_{m\in\mathcal{M}} W_m(\mathbf{x}_D)\boldsymbol{\mu}_{mC} \approx \boldsymbol{\mu}_{m_0 C}$$

$\boldsymbol{\mu}_{m_0 C}$ $\approx$ "vyhlazené" dlaždice neobsahující vysoké frekvence

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

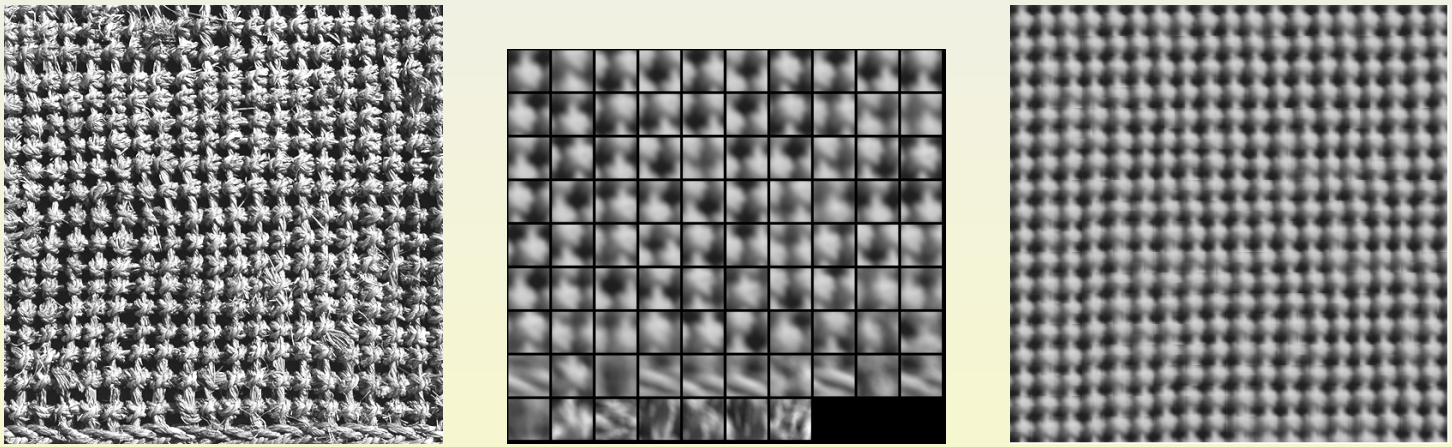
$$\boldsymbol{\mu}^*_m = \arg\min_{\mathbf{x}\in\mathcal{S}}\{\|\boldsymbol{\mu}_m - \mathbf{x}\|^2\}$$

$\Rightarrow$ **"stochastické vzorkování"**

Příklad 4: Modelování textur pomocí normální směsi

model textura "ratan": predikce pomocí průměrů μ_m

původní textura | průměry komp. μ_m | predikce



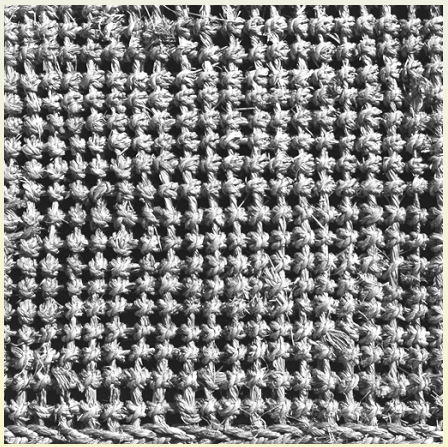
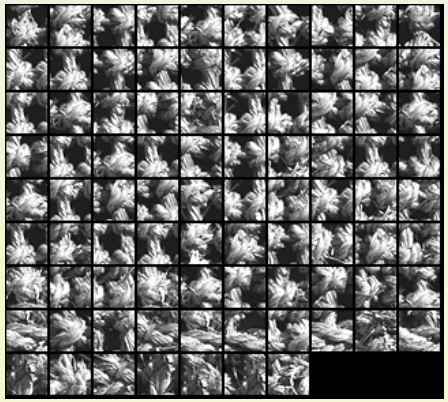
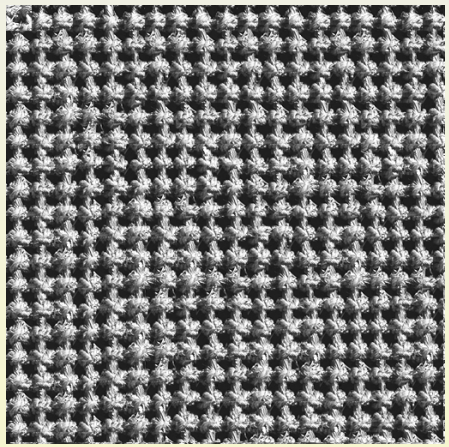
- velikost obrázku: 512x512 pixels $\Rightarrow$ $|\mathcal{S}| \doteq 233000$
- velikost posuvného okna: 30x30 pixelů, dimenze N=900
- počet komponent: $|\mathcal{M}| = 80$
- počet iterací EM algoritmu: $t = 15$

POZN. Průměry μ_m pro jednotlivé komponenty jsou zobrazeny v uspořádání pixelů okna.

Příklad 4: Modelování textur pomocí normální směsi

textura "ratan": predikce pomocí optimálních "dlaždic" μ_m^*

původní textura | optimální "dlaždice" | syntéza vzorkováním

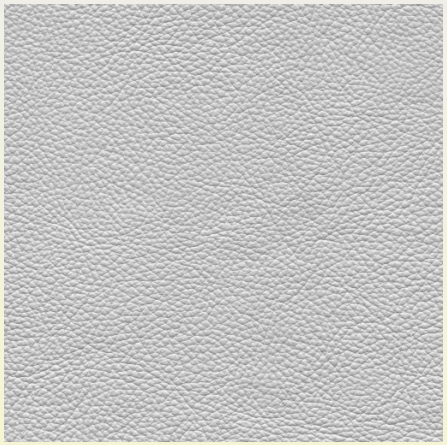
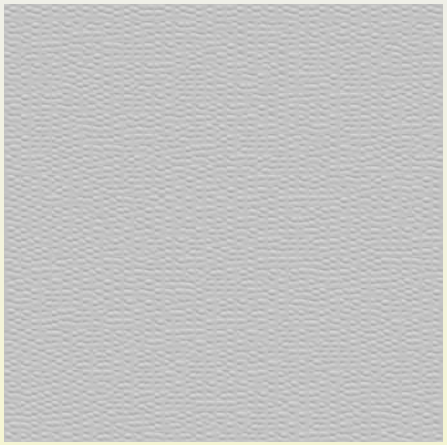
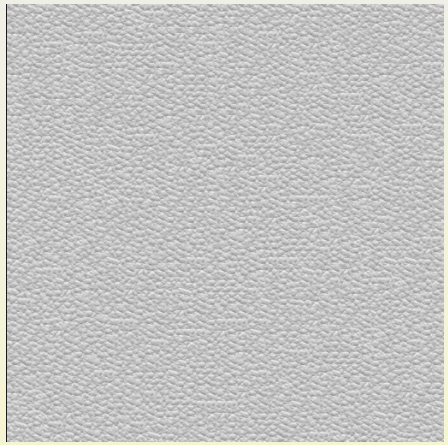
"realistická" syntéza: komponenty μ_m nahrazeny podobnými částmi původní textury μ_m^* optimálně vyhledanými podle kriteria:

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

POZN. Metoda "stochastického vzorkování" je blízká modelování textur kombinováním konečného počtu spojitě navazujících "dlaždic".

Příklad 4: Modelování textur pomocí normální směsi

textura "světlá kůže": predikce pomocí "dlaždic" μ_m^*

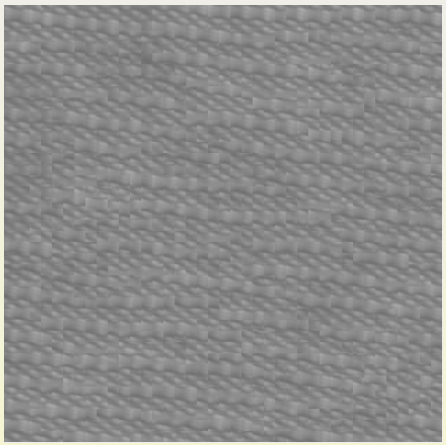
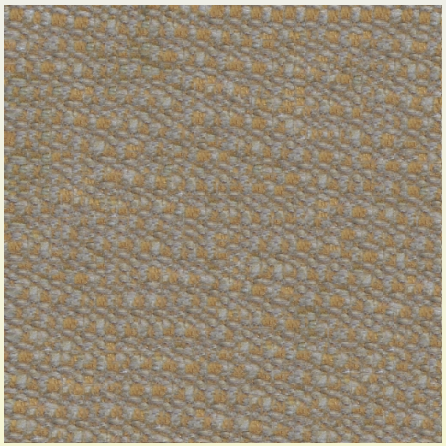
- velikost posuvného okna: 20x20 pixelů, $|\mathcal{S}| \doteq 242000$ vzorků
- dimenze směsi: $N = 20x20 = 400$, počet komponent: $|\mathcal{M}| = 50$
- míra vzdálenosti komponent směsi: $\bar{q}_{max} = 0.959$
- posuv okna při syntéze: 12 pixelů

POZN. Paradoxně: nejmenší krok při syntéze není nejlepší, optimální krok odpovídá přibližně polovině strany okna.
($\approx$ odhad směsi je spolehlivější na podprostoru ?).

ÚTIA

Příklad 4: Modelování textur pomocí normální směsi

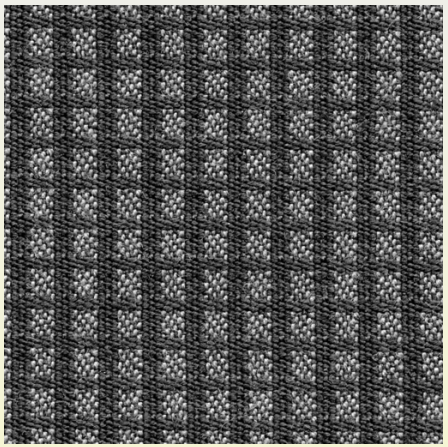
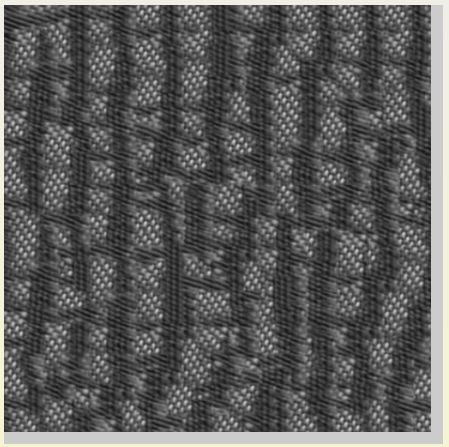
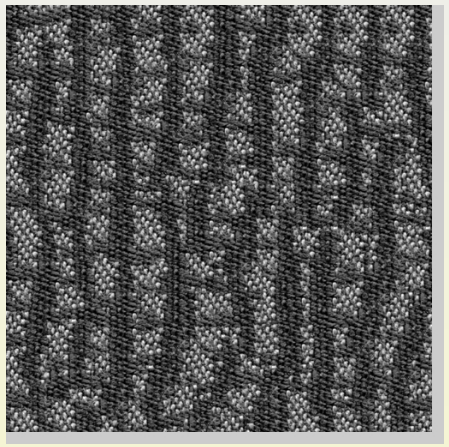
textura "hrubá látka": predikce pomocí barevných "dlaždic" μ_m^*

- velikost posuvného okna: 30*x*30 pixelů
- počet vzorků textury získaných pohybem okna: $|\mathcal{S}| \doteq 232000$
- dimenze směsi: $N = 30x30 = 900$, počet komponent: $|\mathcal{M}| = 128$
- 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 4: Modelování textur pomocí normální směsi

textura "koberec": stochastické vzorkování pomocí "dlaždic"

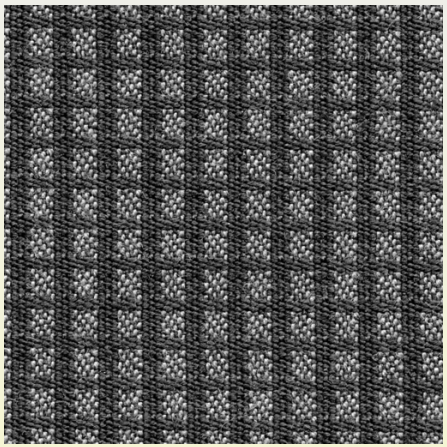
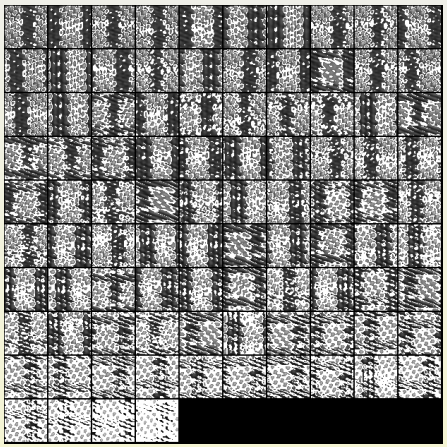
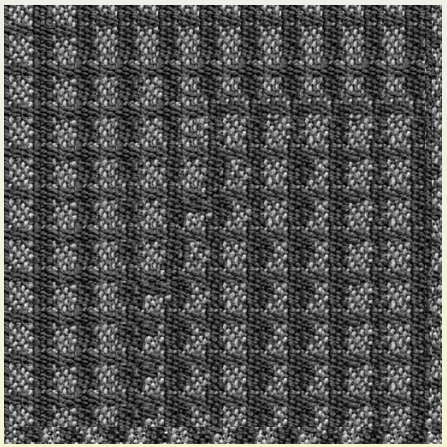




- dimenze směsi: $N = 30x30 = 900$, počet komponent: $|\mathcal{M}| = 90$
- počet vzorků textury získaných posuvem okna: $|\mathcal{S}| \doteq 232000$
- počet iterací EM algoritmu: $t = 20$
- míra vzdálenosti komponent směsi: $\bar{q}_{max} = 0.997$
- posuv okna při syntéze: 18 pixelů

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

Příklad 4: Modelování textury pomocí strukturní směsi

strukturní model textury "koberec"s vysokou dimenzí

- **dimenze směsi:** $N = 60x60 = 3600$, počet komponent: $|\mathcal{M}| = 94$
- počet vzorků textury získaných posuvem okna: $|\mathcal{S}| \doteq 205000$
- míra vzdálenosti komponent směsi: $\bar{q}_{\max} = 0.999$
- posuv okna při syntéze: 24 pixelů
- bílá místa dlaždic se při syntéze nahrazují "pozadím"

POZN. Popis čtvercového vzoru koberce je zřetelně lepší než při velikosti okna $30x30$ pixelů.

Příklad 5: 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: (Grim et al. 2005)

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

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

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

ÚTIA

Příklad 5: 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: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})} \approx$ "strukturní" typičnost okna $\mathbf{x}$
(jmenovatel potlačuje vliv úrovní šedi)

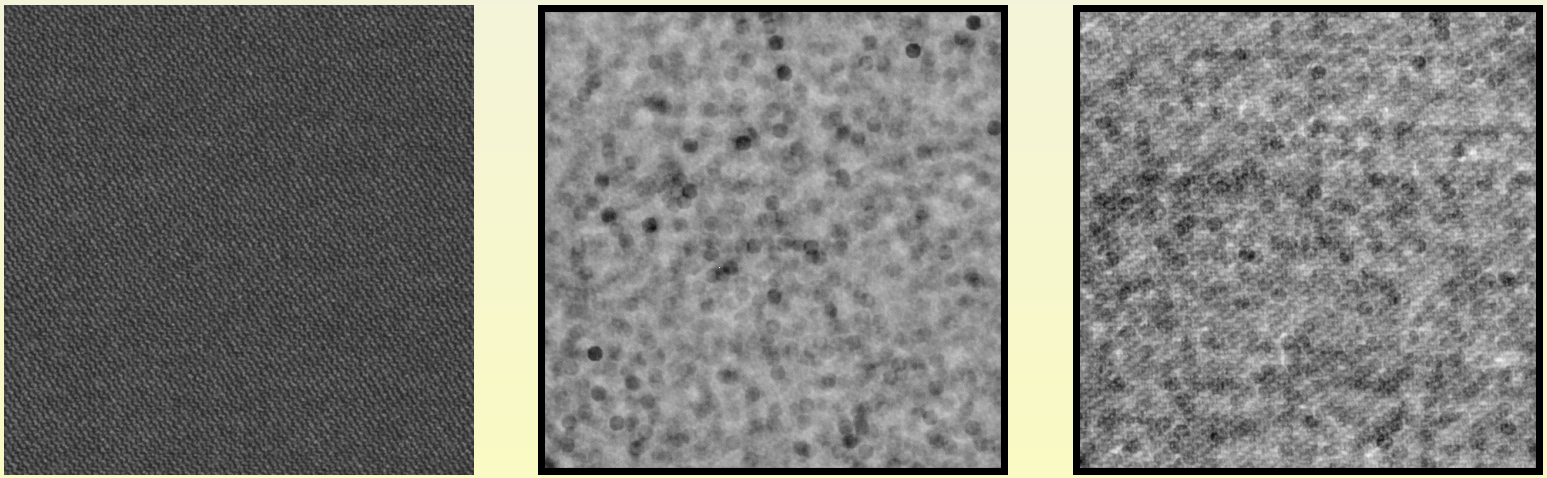
POZN.: Hodnoty průměru a rozptylu μ_{0n}, σ_{0n} jsou téměř identické pro všechna $n \in \mathcal{N} \Rightarrow$ hodnota $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$ méně závisí na změnách úrovní šedi a je více ovlivněna odchylkami struktury.

Příklad 5: 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



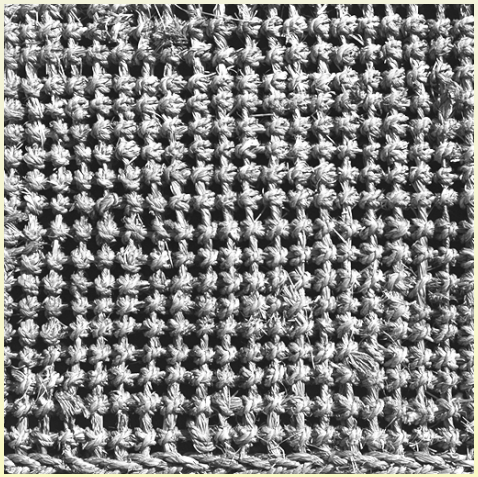
POZN. Hodnoty $\log P(\mathbf{x})$ jsou velmi citlivé na odchylky úrovní šedi. Tak např. stěží viditelné světlejší pixely v textuře "obklad" (levý obrázek) se projeví jako výrazné tmavé skvrny o velikosti okna (střední obrázek).

Příklad 5: Vyhledávání poruch a odchylek v textuře

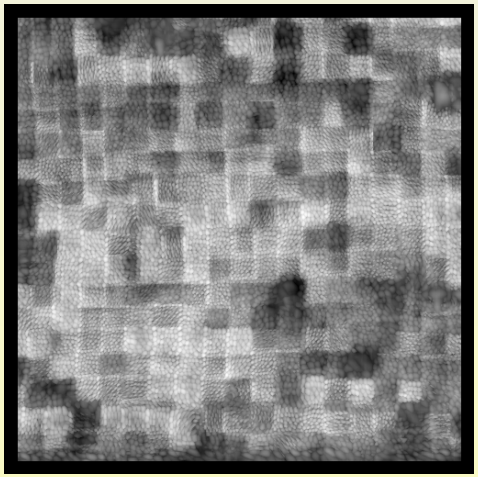
Lokální analýza textury "ratan":

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

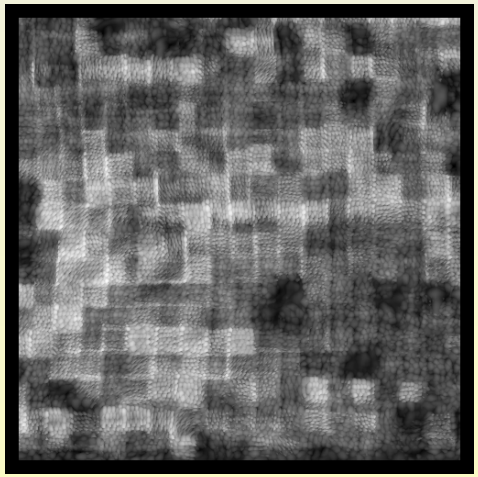
původní obrázek



L-věrohodnost



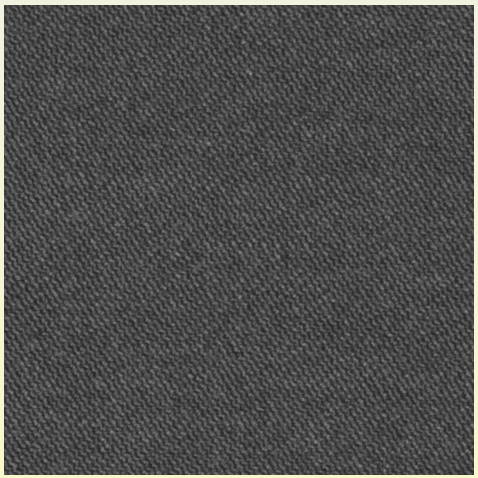
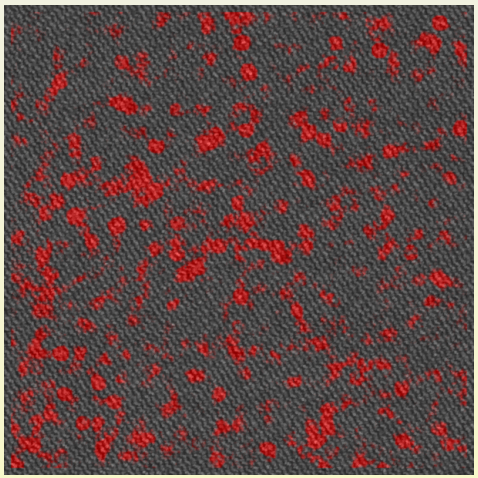
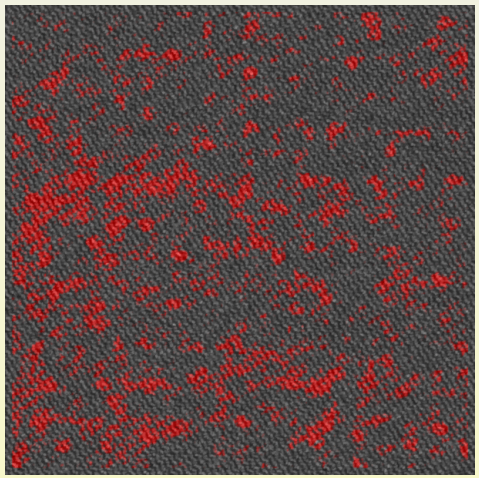
LR-věrohodnost



POZN. Hodnoty $\log(P(\mathbf{x})/P_0(\mathbf{x}))$ 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 $\log(P(\mathbf{x})/P_0(\mathbf{x}))$.

Příklad 5: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "obklad":

původní obrázek	L-nevěrohodnost	LR-nevěrohodnost
		

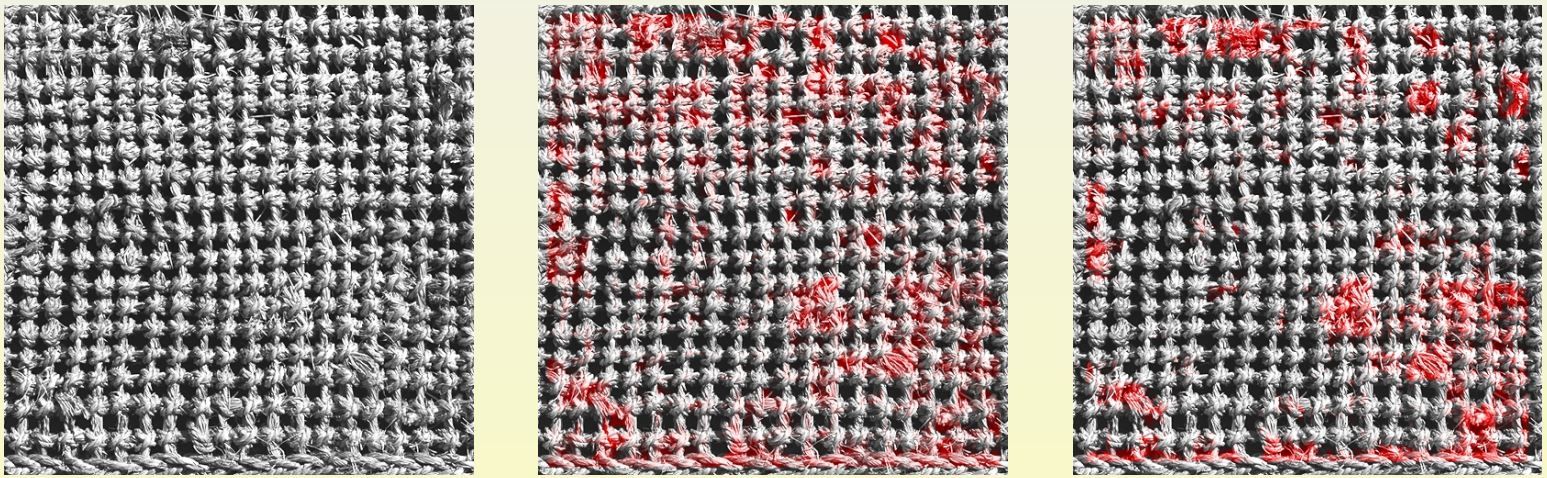
červené zabarvení: $\approx$ neobvyklá (atypická) místa (poruchy) v textuře

- střední obrázek: $\approx$ nízké hodnoty věrohodnosti: $\log P(\mathbf{x})$
- pravý obrázek: $\approx$ nízké hodnoty věrohodnostního poměru: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$

Příklad 5: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "ratan":

původní obrázek	L-nevěrohodnost	LR-nevěrohodnost

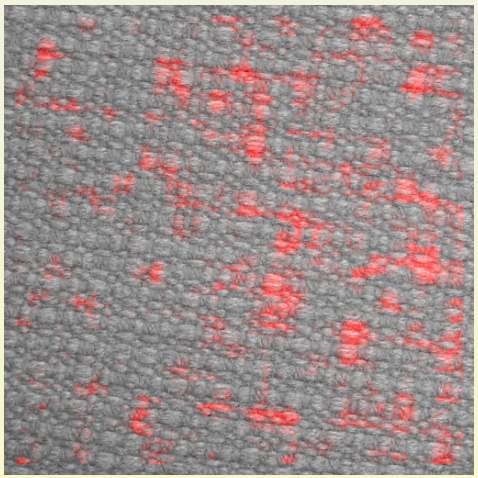
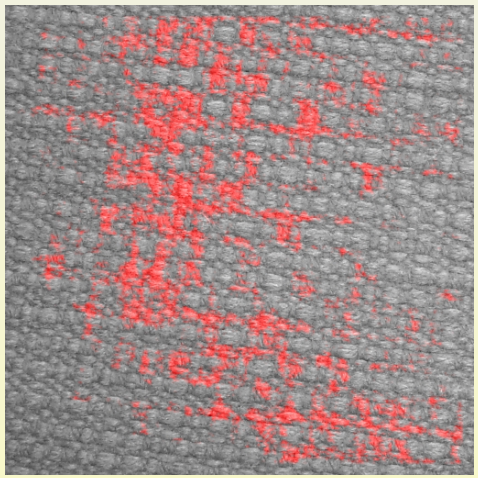


červené zabarvení: $\approx$ neobvyklá (atypická) místa (poruchy) v textuře

- střední obrázek: $\approx$ nízké hodnoty věrohodnosti: $\log P(\mathbf{x})$
- pravý obrázek: $\approx$ nízké hodnoty věrohodnostního poměru: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$

Příklad 5: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "látka":

původní obrázek	L-nevěrohodnost	LR-nevěrohodnost
		

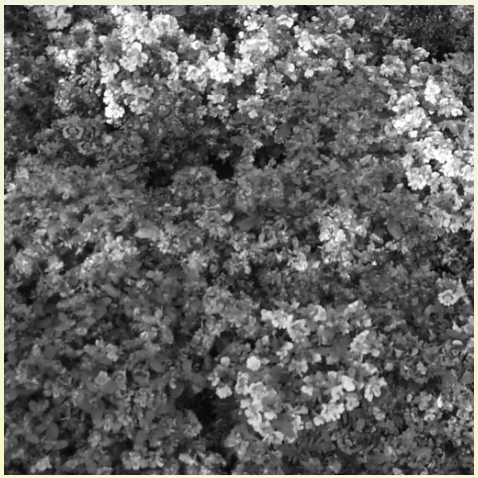
červené zabarvení: $\approx$ neobvyklá (atypická) místa (poruchy) v textuře

- střední obrázek: $\approx$ nízké hodnoty věrohodnosti: $\log P(\mathbf{x})$
- pravý obrázek: $\approx$ nízké hodnoty věrohodnostního poměru: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$

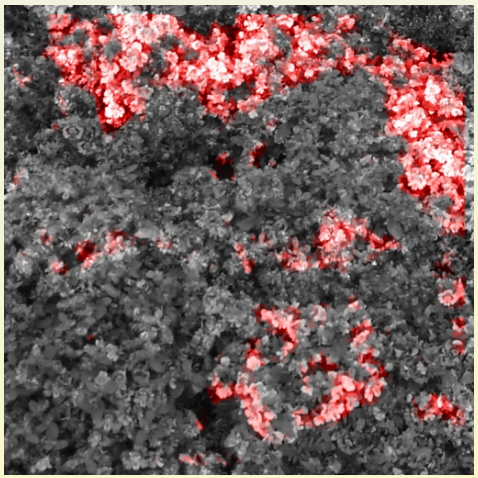
Příklad 5: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "květy":

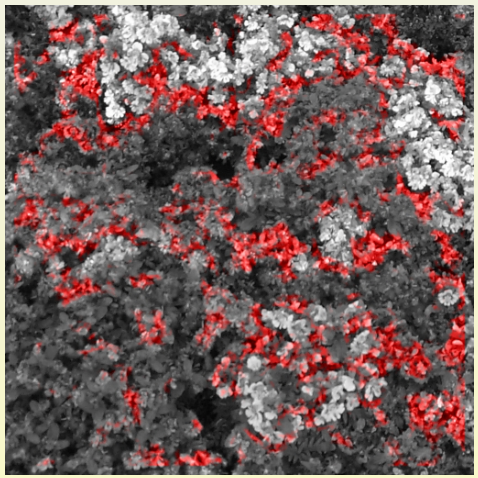
původní obrázek



L-nevěrohodnost



LR-nevěrohodnost



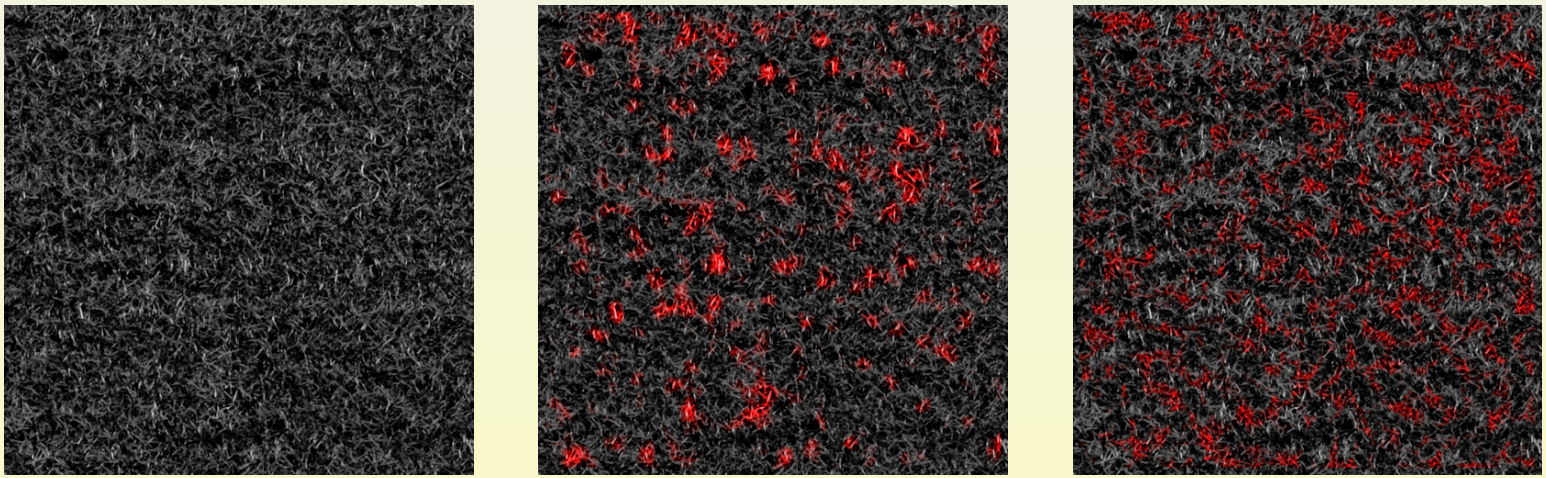
červené zabarvení: $\approx$ neobvyklá (atypická) místa (poruchy) v textuře

- střední obrázek: $\approx$ nízké hodnoty věrohodnosti: $\log P(\mathbf{x})$
- pravý obrázek: $\approx$ nízké hodnoty věrohodnostního poměru: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$

Příklad 5: Vyhledávání poruch a odchylek v textuře

Analýza nepravidelnosti textury "koberec":

původní obrázek	L-nevěrohodnost	LR-nevěrohodnost



červené zabarvení: $\approx$ neobvyklá (atypická) místa (poruchy) v textuře

- střední obrázek: $\approx$ nízké hodnoty věrohodnosti: $\log P(\mathbf{x})$
- pravý obrázek: $\approx$ nízké hodnoty věrohodnostního poměru: $\log \frac{P(\mathbf{x})}{P_0(\mathbf{x})}$

Příklad 5: Statistický model a lokální analýza textury

- (!) datové vektory $\mathbf{x}$ generované posuvem okna se překrývají a proto nejsou nezávislé
- $\Rightarrow$ je porušen základní předpoklad použití metody maximálně věrohodného odhadu
- datový soubor $\mathcal{S}$ odpovídá pouze "trajektorii" v prostoru $\mathcal{X}$ vzniklé posuvem okna ($\Rightarrow$ není reprezentativní)
- **na rozdíl od jiných aplikací (rozpoznávání,syntéza textury, predikce) se odhadnutá směs $P(\mathbf{x})$ aplikuje na původní datový soubor $\mathcal{S}$**, nehrozí "overfitting"
- 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 v souladu s použitou metodou odhadu parametrů směsi
- $\Rightarrow$ 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 6: 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 6: Vyhodnocování screeningových mamogramů

Cíl věrohodnostní analýzy: (Grim et al. 2009)

usnadnit diagnostické vyhodnocování screeningových mamogramů zvýrazněním atypických resp. podezřelých míst

METODA: lokální směsový model $P(\mathbf{x}) = \sum_{m=1}^{M} w_m F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$

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

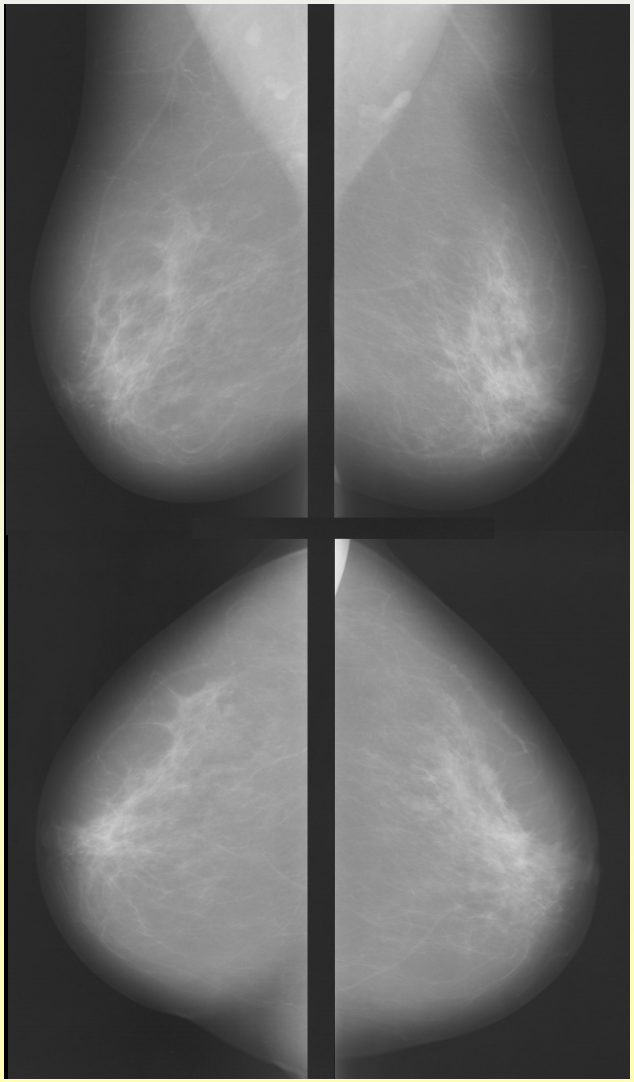
LOKÁLNÍ ZOBRAZENÍ VĚROHODNOSTNÍHO POMĚRU:
$\log P(\mathbf{x})/P_0(\mathbf{x}) \approx$ nebylo použito

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

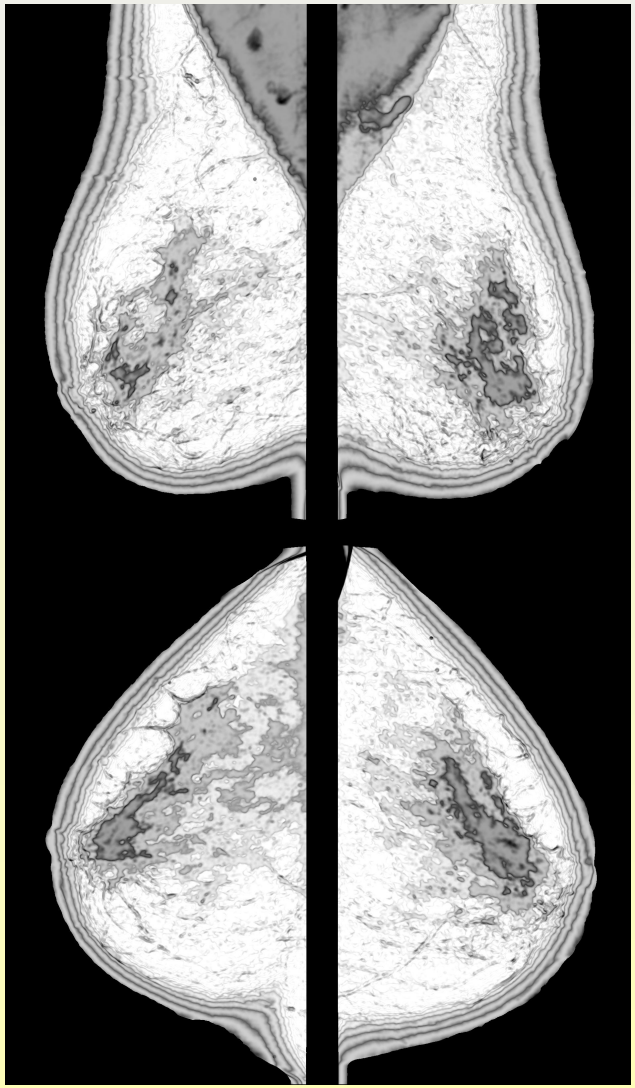
ÚTIA

Příklad 6: Vyhodnocování screeningových mamogramů

původní mamogram



věrohodnostní analýza

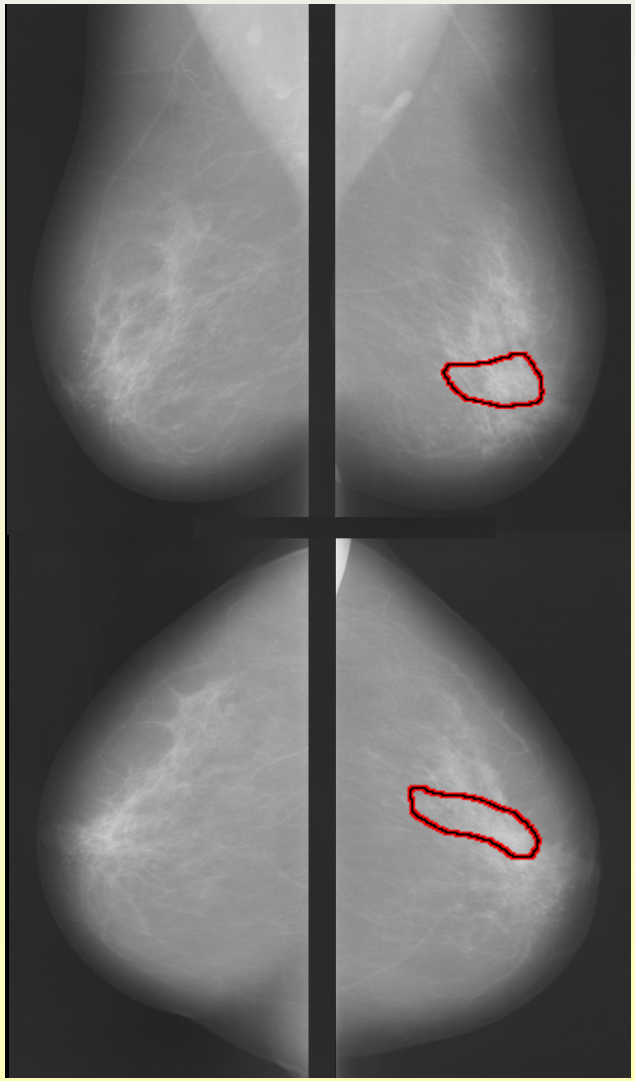


Příklad 6: Výpočetní aspekty vyhodnocování mamogramů

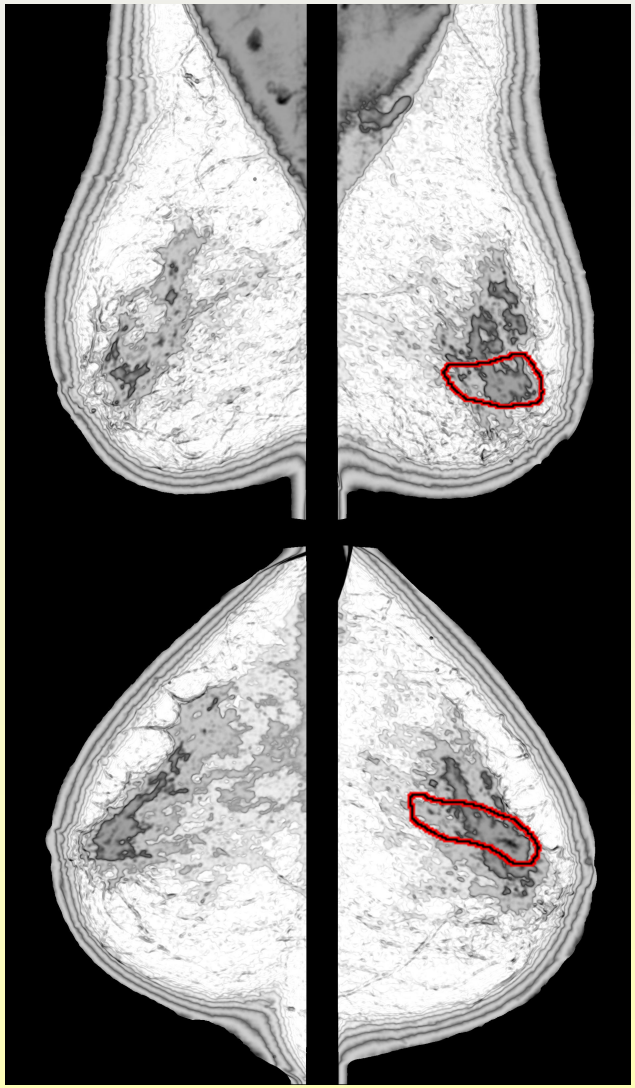
- zdrojová databáze: 2600 tzv. úplných mammogramů
  University of South Florida:
  http://marathon.csee.usf.edu/Mammography/Database.html
- úplný mamogram zahrnuje 4 snímky: dva medio-laterální pohledy a dva cranio-caudální pohledy a je vyhodnocován jako celek
- pravá část mamogramu je před vyhodnocením 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 6: Vyhodnocování screeningových mamogramů

ověřený nález

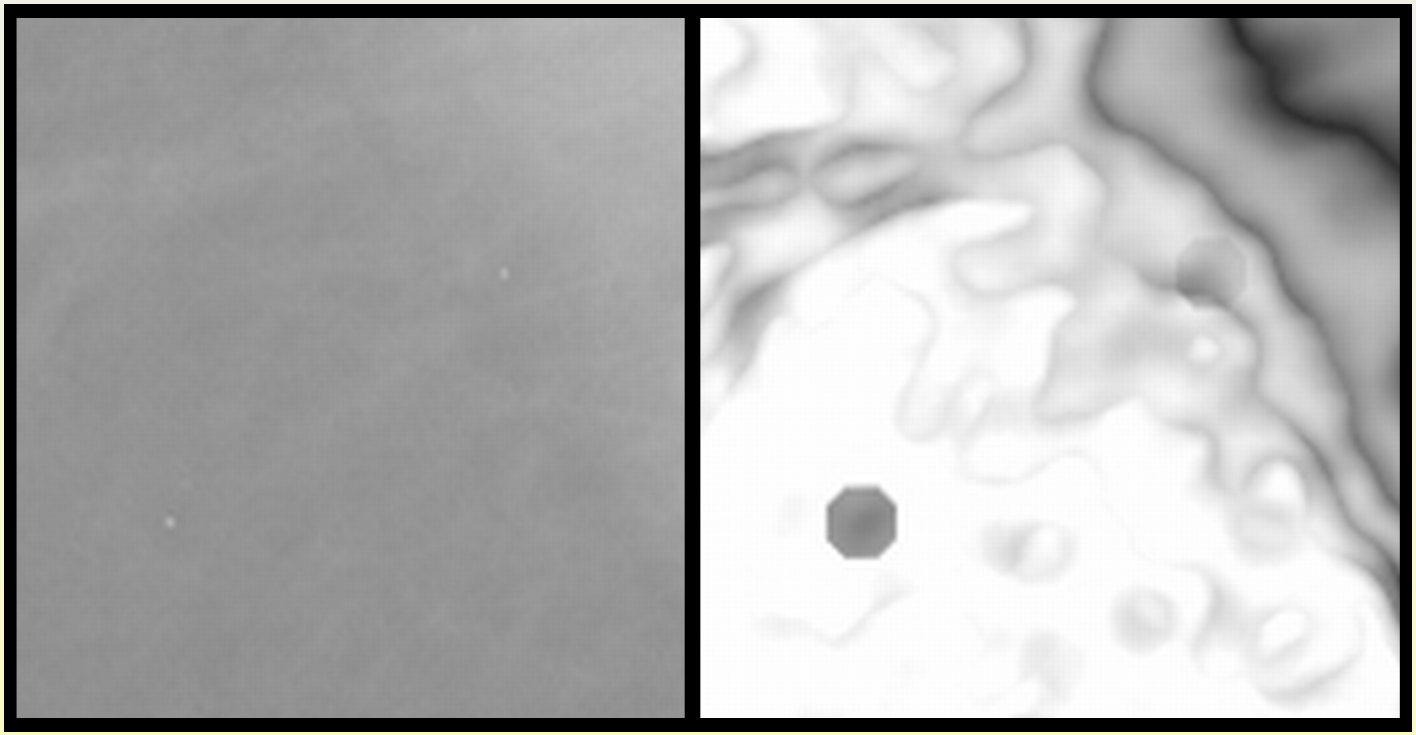


porovnání analýzy a nálezu



Příklad 6: Věrohodnostní analýza mikrokalcifikací

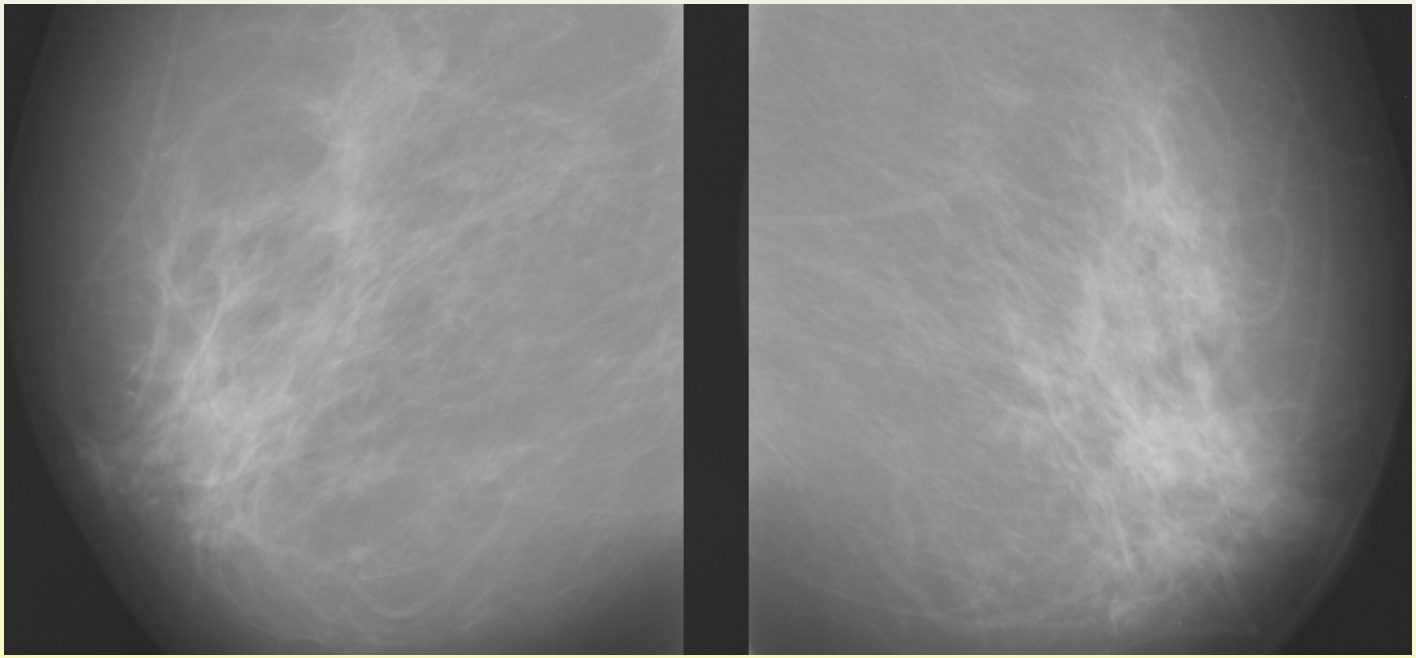
mikrokalcifikace zvýrazněné mikrokalcifikace



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 6: Identifikace "hmot" pomocí "vrstevnic"

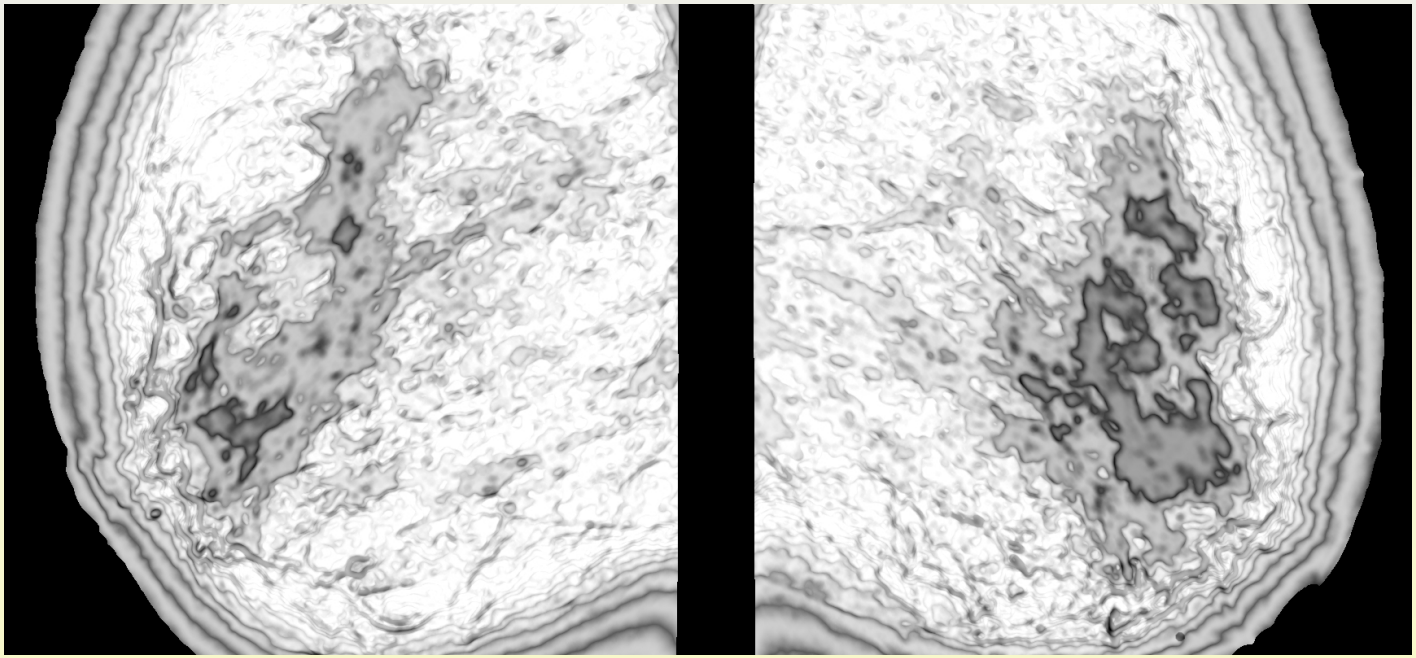
část screeningového mamogramu obsahující podezřelé "hmoty"



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 6: Identifikace "hmot" pomocí "vrstevnic"

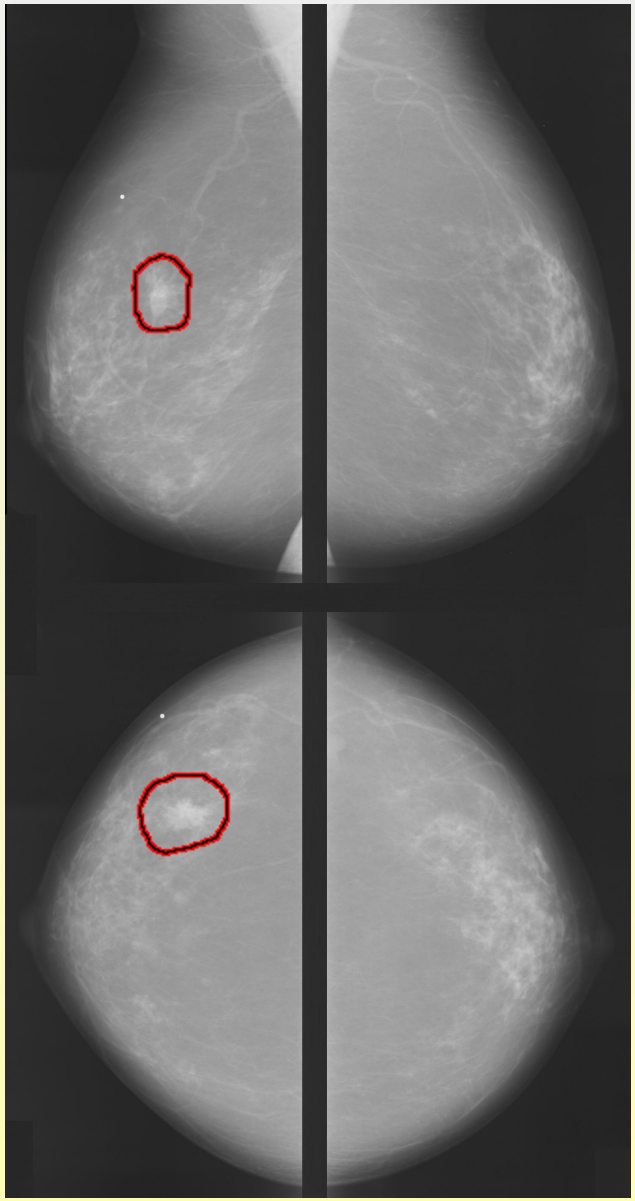
"vrstevnice" zvýrazňující hranice "hmot" a okraje mamogramu



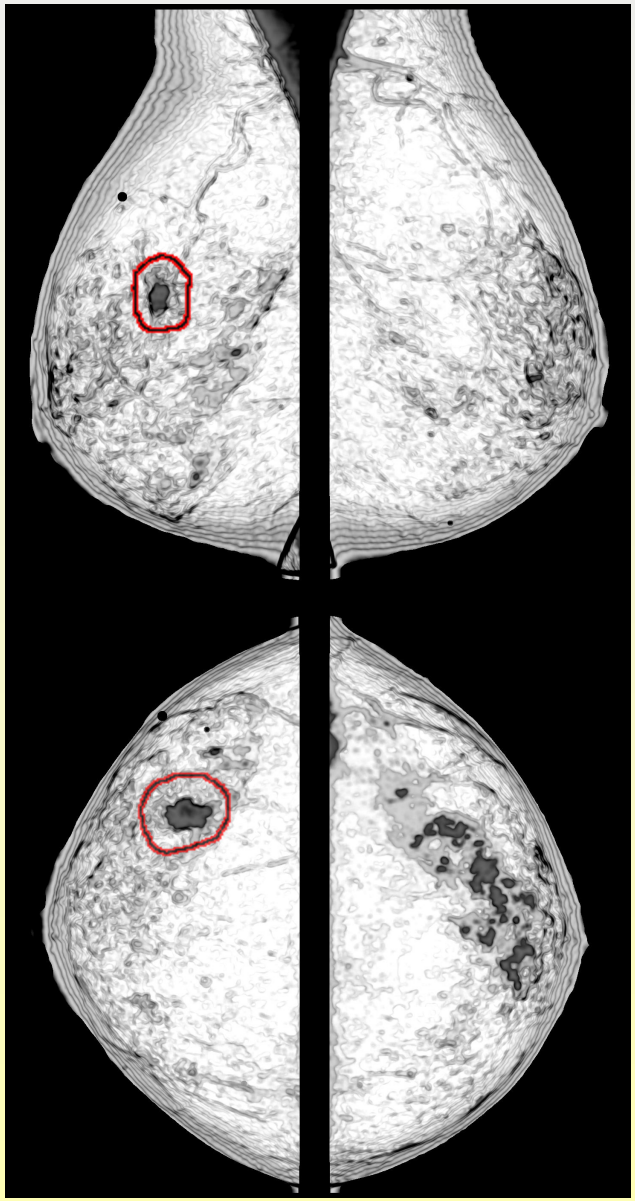
Remark: Jednotlivé věrohodnostní hodnoty $\log P(\mathbf{x})$ jsou typicky určeny jedinou komponentou směsi, která nejlépe odpovídá dané pozici okna. Na okraji různých oblastí mamogramu dochází k záměně komponent, která je spojena s poklesem věrohodnosti $\log P(\mathbf{x})$. Záměna komponent je příčinou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

Příklad 6: Vyhodnocování screeningových mamogramů

ověřený nález

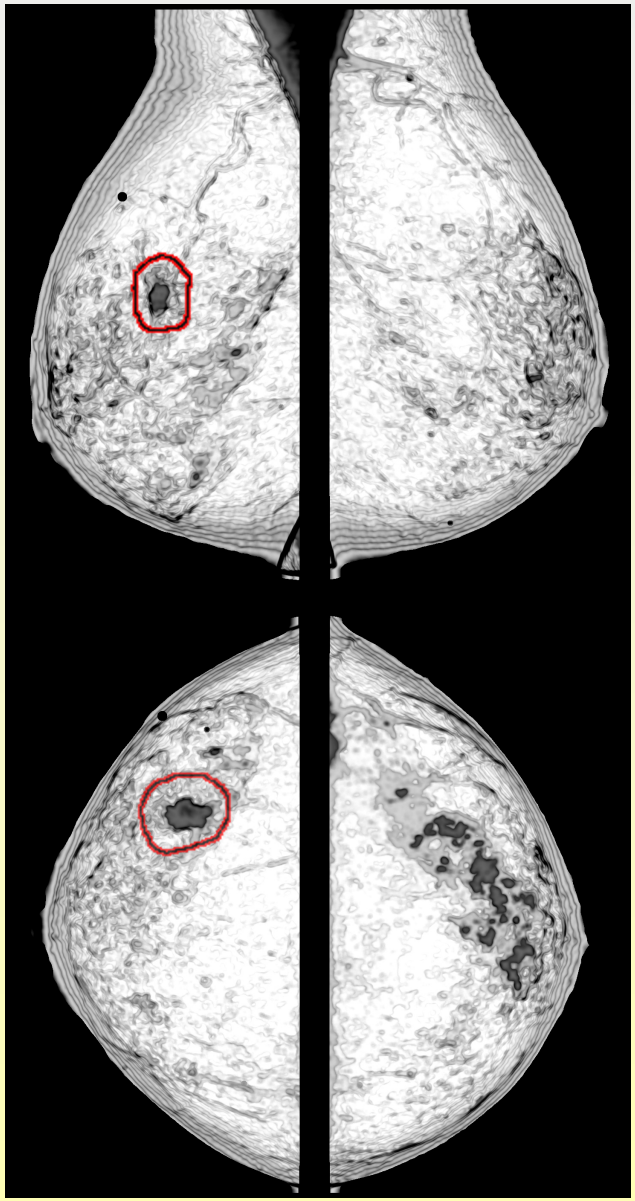


věrohodnostní analýza

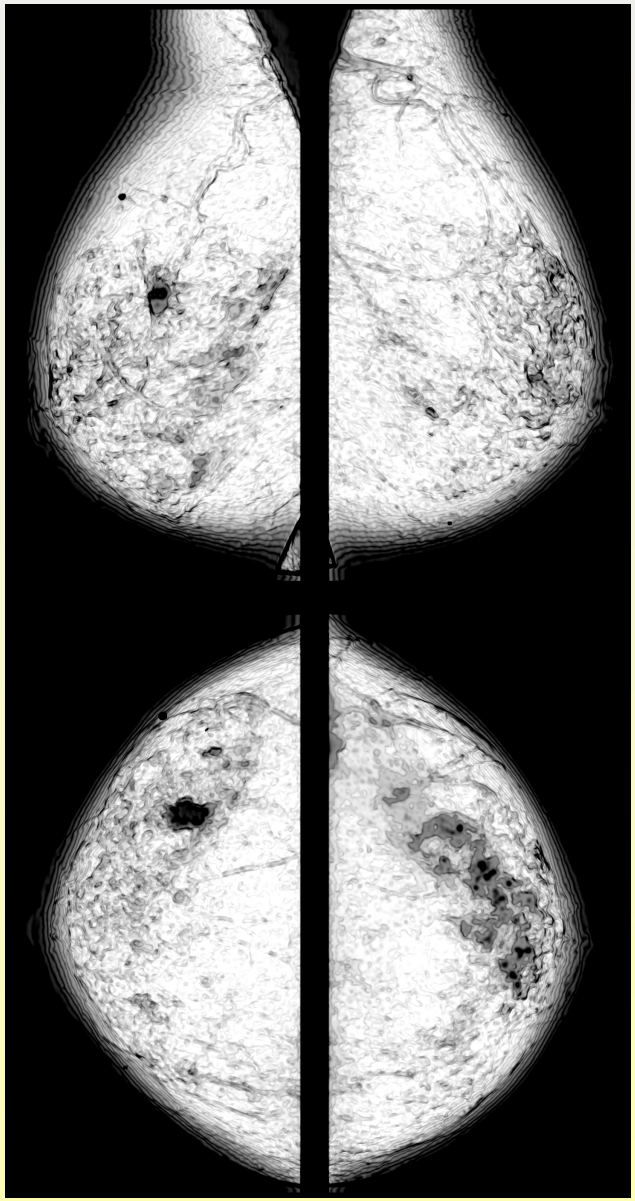


Příklad 6: segmentálně rozložená mikrokalcifikace

normální směsový model

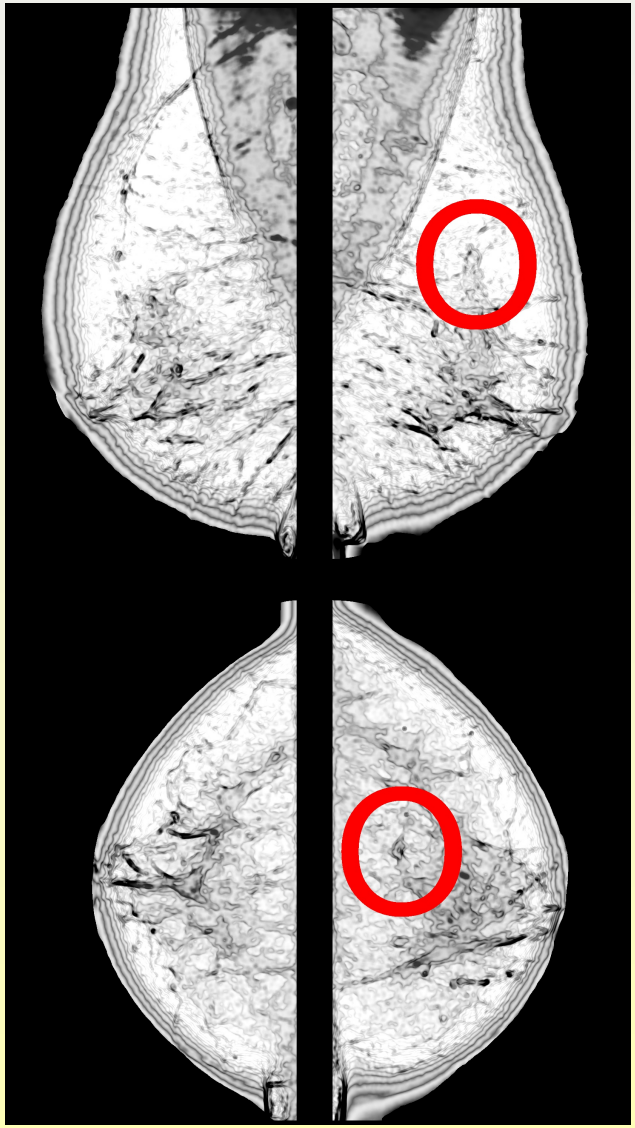


strukturní směsový model

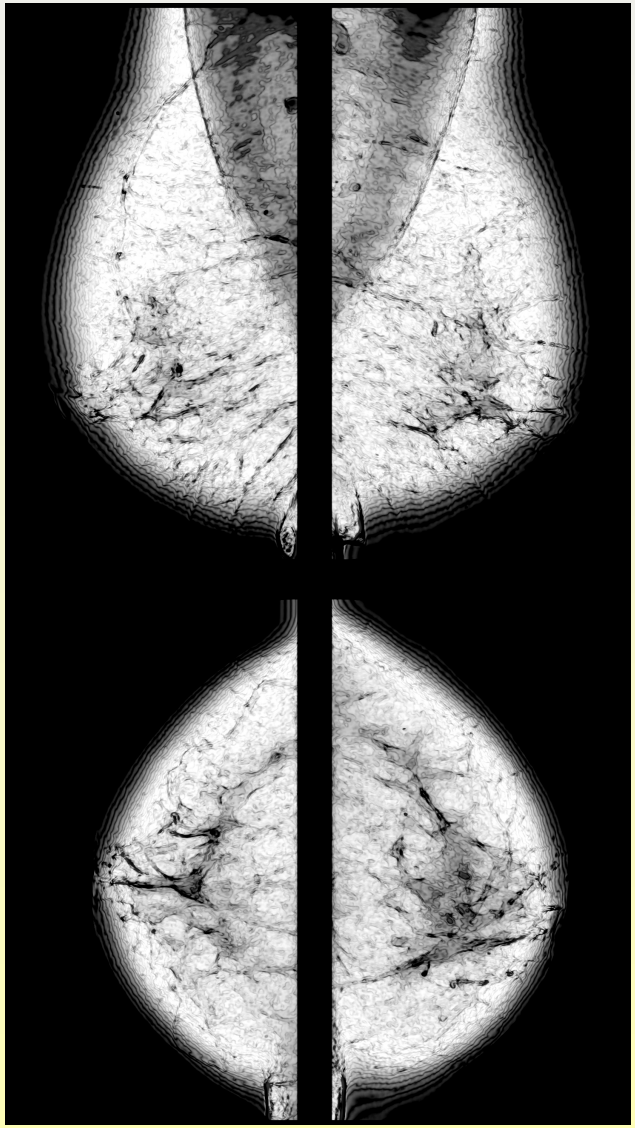


Příklad 6: maligní "hmota"s oválnými okraji

normální směsový model

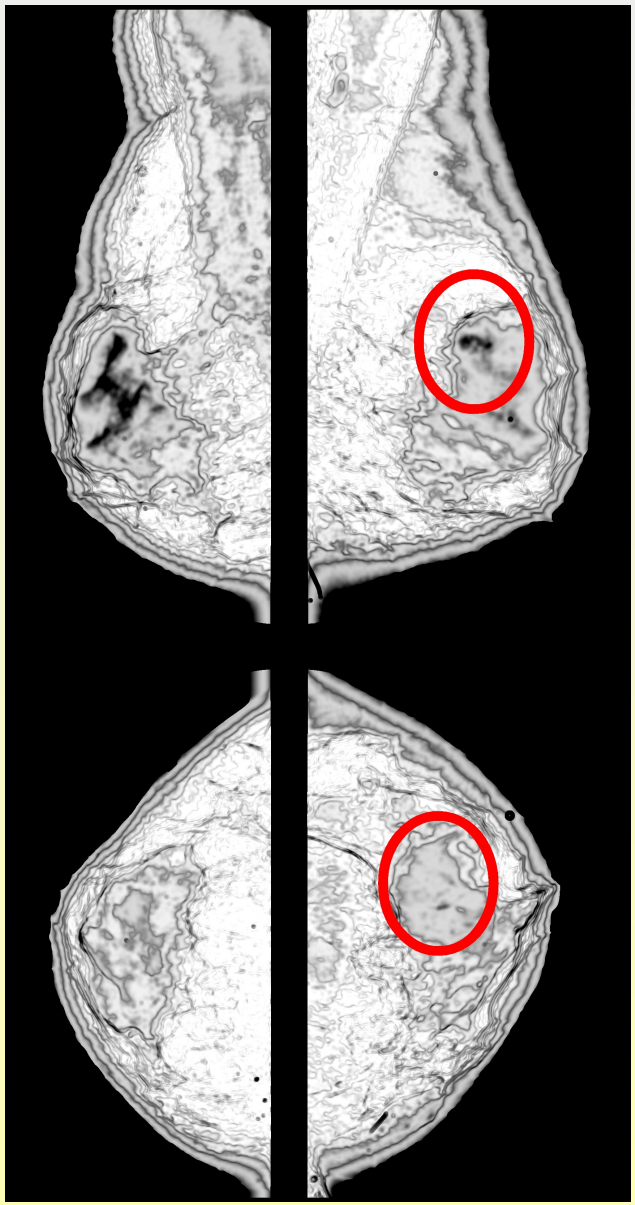


strukturní směsový model

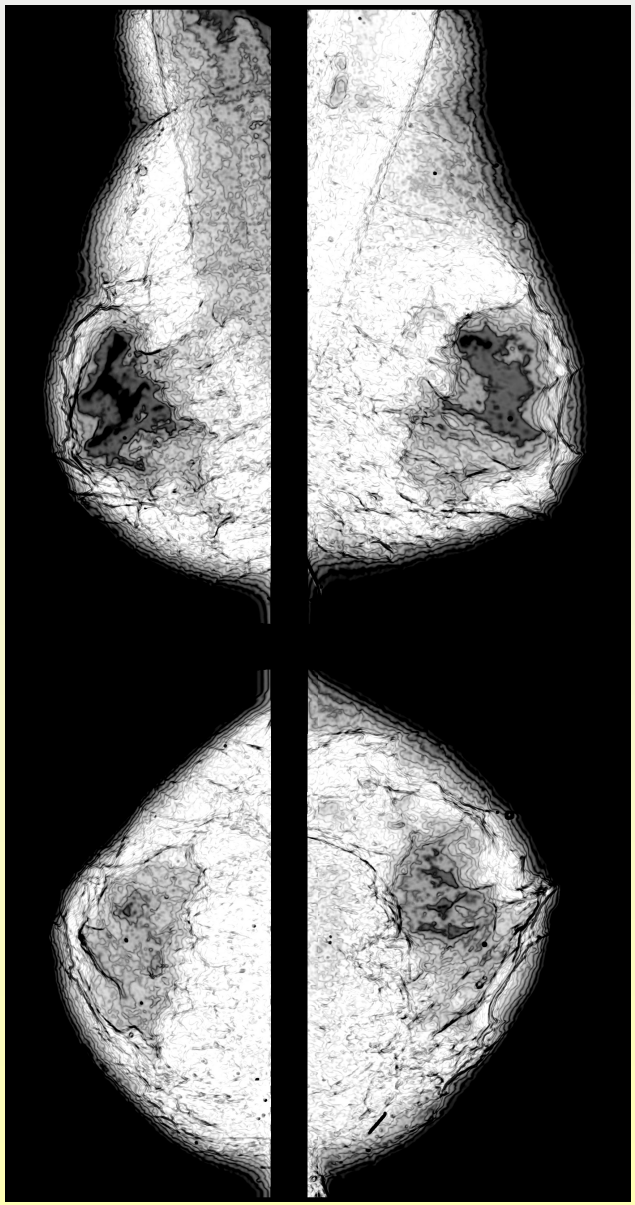


Příklad 6: maligní "hmota" s asymetrickou hustotou

normální směsový model



strukturní směsový model



Příklad 6: Vyhodnocování screeningových mamogramů

SHRNUTÍ:

- **princip:** zvýraznění možného patologického nálezu jako atypické lokality na základě lokálního statistického modelu
- **výhoda:** věrohodnostní obraz má jednoznačnou statistickou interpretaci
- **nevýhoda:** statistický model nevyužívá odborné lékařské znalosti
- "micro-calcifikace" se jeví jako tmavé skvrny ve tvaru posuvného okna
- "hmoty (masses)" se zvýrazňují pomocí obrysových kontur
- úplný mamogram zvýrazňuje diagnosticky významné asymetrie
- výpočet věrohodnostního obrazu pomocí statistického modelu nevyžaduje trénovací data z jiných mamogramů
- nezávislost věrohodnostního obrazu na nezávislých trénovacích datech je důležitá vzhledem k velké variabilitě mamogramů
- statistický model je invariantní vůči lineární transformaci úrovní šedi

▸ Důkaz

Příklad 7: Forenzní analýza obrazových dat (Grim et al. 2010)

Specifický problém forenzní analýzy:

zjišťování stop manipulace obrázků neznámého původu (naslepo)

příklady dostupných metod:

- detekce zkopírovaných částí obrázku
- identifikace nekonzistentního osvětlení
- detekce periodicit způsobených převzorkováním
- detekce artefaktů JPG kódování
- detekce lokálních odchylek statistických vlastností

vlastnosti dostupných metod:

- každá metoda identifikuje pouze určitý typ manipulace
- výsledky forenzní analýzy obvykle nejsou jednoznačné
- přesnost detekce klesá při ztrátovém kódování obrázku

Příklad 7: Forenzní analýza obrazových dat

METODA SMĚSOVÉHO MODELU:

detekce podezřelých lokalit podle odlišných statistických vlastností

IDEA:

některé atypické vlastnosti obrázku (spektrum, textura) mohou být identifikovány lokálně na základě statistických vlastností pixelů uvnitř zvoleného klouzavého okna

digitalizovaný barevný obrázek: $\mathcal{Z} = [\mathbf{z}_{ij}]_{i=1}^{I}{}_{j=1}^{J}$

$\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, z_{ij3}) \in \langle 0, 255 \rangle^3 \quad \approx$ tři spektrální hodnoty pro každý pixel

$\mathbf{x} \approx$ spektrální složky vnitřních pixelů okna v daném fixním uspořádání

$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \langle 0, 255 \rangle^N$

princip metody:

- odhad mnohorozměrné hustoty pravděpodobnosti $P(\mathbf{x})$
- identifikace netypických částí obrázku podle nízké věrohodnosti

Příklad 7: Forenzní analýza obrazových dat

STATISTICKÝ MODEL: **normální směs součinových komponent**

$$P(\mathbf{x}) = \sum_{m=1}^{M} w_m F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) = \sum_{m=1}^{M} w_m \prod_{n=1}^{N} f_n(x_n|\mu_{mn}, \sigma_{mn})$$

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

ODHAD MODELU: **pomocí EM algoritmu**

Invariance statistického modelu:

Věrohodnostní transformace obrázku je invariantní vůči libovolné lineární transformaci spektrálních složek původního obrázku. ▸ Důkaz

POZN. Průměry komponent $\boldsymbol{\mu}_m$ se v EM algoritmu počítají jako vážené průměry vzorků $\mathbf{x} \in \mathcal{S}$ a proto jsou typicky "vyhlazené" bez jemných detailů. Vložená část obrázku s potlačeným vysokofrekvenčním spektrem (např. v důsledku interpolace) se z tohoto důvodu jeví jako pravděpodobnější.

Příklad 7: Forenzní analýza obrazových dat

$\log P(\mathbf{x}) \approx$ **míra typičnosti obsahu okna x**
$\log P(\mathbf{x}) \approx$ **zobrazuje se pomocí vhodně zvolené stupnice úrovní šedi**

INTERPRETACE: **tmavé pixely** odpovídající nízkým hodnotám $\log P(\mathbf{x})$ mohou indikovat "netypické" nebo "podezřelé" části obrázku

Mechanizmus detekce podezřelých částí obrázku:

- malé oblasti s atypickými spektrálními vlastnostmi nebo s atypickou texturou se jeví jako málo pravděpodobné
- rozostřené (zašumněné) části obrázku mají vyšší pravděpodobnost **(!)** v důsledku chybějících vysokofrekvenčních detailů

zobrazovaný interval úrovní šedi: $\log P(\mathbf{x}) \in \langle \mu_0 - 2 * \sigma_0;\ \mu_0 + 2 * \sigma_0 \rangle$

POZN. V mnohorozměrných prostorech se hodnoty $P(\mathbf{x})$ pro dvě okna posunutá o jeden pixel mohou lišit o několik řádů; logaritmické hodnoty věrohodnosti $\log P(\mathbf{x})$ jsou proto vhodnější mírou typičnosti obsahu okna.

ÚTIA

Příklad 7: Forenzní analýza obrazových dat

PARAMETRY NUMERICKÝCH EXPERIMENTŮ:

- čtvercové okno 5x5 pixelů s odříznutými rohy (21 vnitřních pixelů) (zvětšováním okna dochází k vyhlazování detailů)
- 21 vnitřních pixelů okna ve třech spektrálních složkách implikuje dimenzi vektoru $\mathbf{x}$: N=63
- hustota $P(\mathbf{x})$ popisuje statistické vlastnosti 63 složek vektoru $\mathbf{x}$
- trénovací množina dat $\mathcal{S}$ vzniká posouváním okna v mezích obrázku
- vychozí obrázek generuje trénovací soubory o velikosti $|\mathcal{S}| \approx 10^6$
- počet komponent byl volen přibližně v rozsahu $M \approx 20 - 80$
- náhodná inicializace komponent s rovnoměrnými vahami
- ukončení výpočtu prahem relativního přírůstku kriteria $\Delta L \approx 10^{-3}$ což odpovídá cca 10 - 20 iteracím
- **trvání výpočtu:** cca $\approx$ 15 - 30 minut (běžný PC)

Příklad 7: Forenzní analýza obrazových dat

Původní obrázek s vloženou oválnou částí v levém horním rohu

Příklad 7: Forenzní analýza obrazových dat

Oválná část v levém horním rohu je zřetelně světlejší v důsledku odlišných lokálních vlastností textury

Příklad 7: Forenzní analýza obrazových dat

Původní obrázek složený ze dvou částí pomocí softwaru "Autostitch".

Příklad 7: Forenzní analýza obrazových dat

Mírně rozostřená levá část obrázku je po transformaci světlejší.

Příklad 7: Forenzní analýza obrazových dat

Původní obrázek složený ze tří částí pomocí softwaru "Autostitch".

Příklad 7: Forenzní analýza obrazových dat

Špatně zaostřená střední část obrázku je po transformaci světlejší.

Příklad 7: Forenzní analýza obrazových dat

Shrnutí:

Vlastnosti věrohodnostního obrázku:

- průměry komponent μ_m počítané v EM algoritmu jako vážené součty vzorků jsou "vyhlazené"
- věrohodnostní transformace obrázku je invariantní vůči lineární transformaci spektrálních složek
- i malé rozdíly v ostrosti, frekvenčním spektru obrázku nebo v textuře mohou být viditelné po věrohodnostní transformaci

Identifikace podezřelých částí obrázku:

- forenzní analýza pomocí lokálního statistického modelu je "slepá" metoda
- je použitelná k obrázkům neznámého původu bez apriorní informace
- věrohodnostní analýza nepředpokládá konkrétní typ manipulace obrázku
- výsledek forenzní analýzy je rozumně odolný vůči ztrátovému kódování

Příklad 8: Predikce chybějících částí obrázku

Princip metody: (Grim et al. 2008)

- odhad lokálního statistického modelu obrázku pomocí normální součinové směsi $P(\mathbf{x})$
- postupná predikce chybějících částí obrázku na základě podmíněných distribucí odvozených z $P(\mathbf{x})$

doplněk chybějící části okna: $\mathbf{x}_C = (x_{i_1}, \ldots, x_{i_k}), \quad C = \{i_1, \ldots, i_k\} \subset \mathcal{N}$

podmíněná distribuce pro chybějící pixel x_n:

$$P_{n|C}(x_n|\mathbf{x}_C) = \sum_{m \in \mathcal{M}} W_m(\mathbf{x}_C) f_n(x_n|\mu_{mn}, \sigma_{mn}), \quad n \notin C$$

$$W_m(\mathbf{x}_C) = \frac{w_m F_C(\mathbf{x}_C|\boldsymbol{\mu}_{mC}, \boldsymbol{\sigma}_{mC})}{\sum_{j=1}^{M} w_j F_C(\mathbf{x}_C|\boldsymbol{\mu}_{jC}, \boldsymbol{\sigma}_{jC})} \approx \text{přibližně binární}$$

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

$$\bar{x}_n = \mathrm{E}_{n|C}\{x_n|\mathbf{x}_C\} = \int x_n P_{n|C}(x_n|\mathbf{x}_C) dx_n = \sum_{m \in \mathcal{M}} W_m(\mathbf{x}_C)\mu_{mn} \approx \mu_{m_0 n}$$

ÚTIA

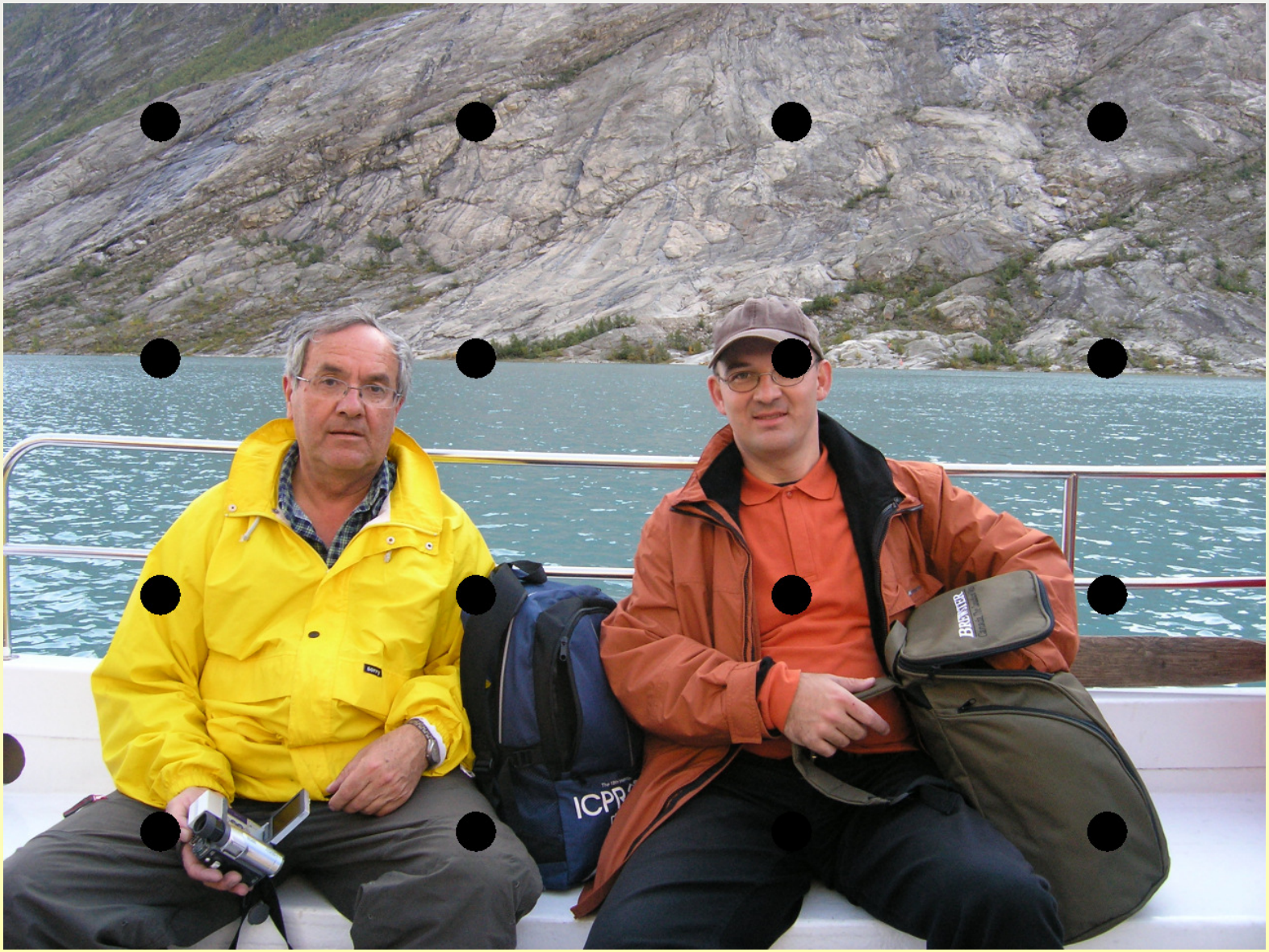
Příklad 8: Predikce chybějících částí obrázku

POPIS NUMERICKÝCH EXPERIMENTŮ:

- obrázky o velikosti 1280 x 960 pixelů
- posouváním okna v mezích obrázku vzniká soubor $|\mathcal{S}| \approx 10^6$ vektorů
- použité čtvercové okno 7x7 pixelů s odříznutými rohy obsahuje 37 pixelů
- 37 vnitřních pixelů okna ve třech spektrálních složkách implikuje dimenzi vektoru $\mathbf{x}$: N=111
- počet komponent byl volen přibližně v rozsahu $M \approx 20 - 80$
- náhodná inicializace parametrů komponent s rovnoměrnými vahami
- ukončení výpočtu prahem relativního přírůstku kriteria $\Delta L \approx 10^{-3}$ což odpovídá cca 10 - 20 iteracím
- **trvání výpočtu modelu:** cca $\approx$ 15 - 30 minut (běžný PC)
- postupná predikce chybějících částí obrázku v několika iteracích

Příklad 8: Predikce chybějících částí obrázku

původní poškozený obrázek

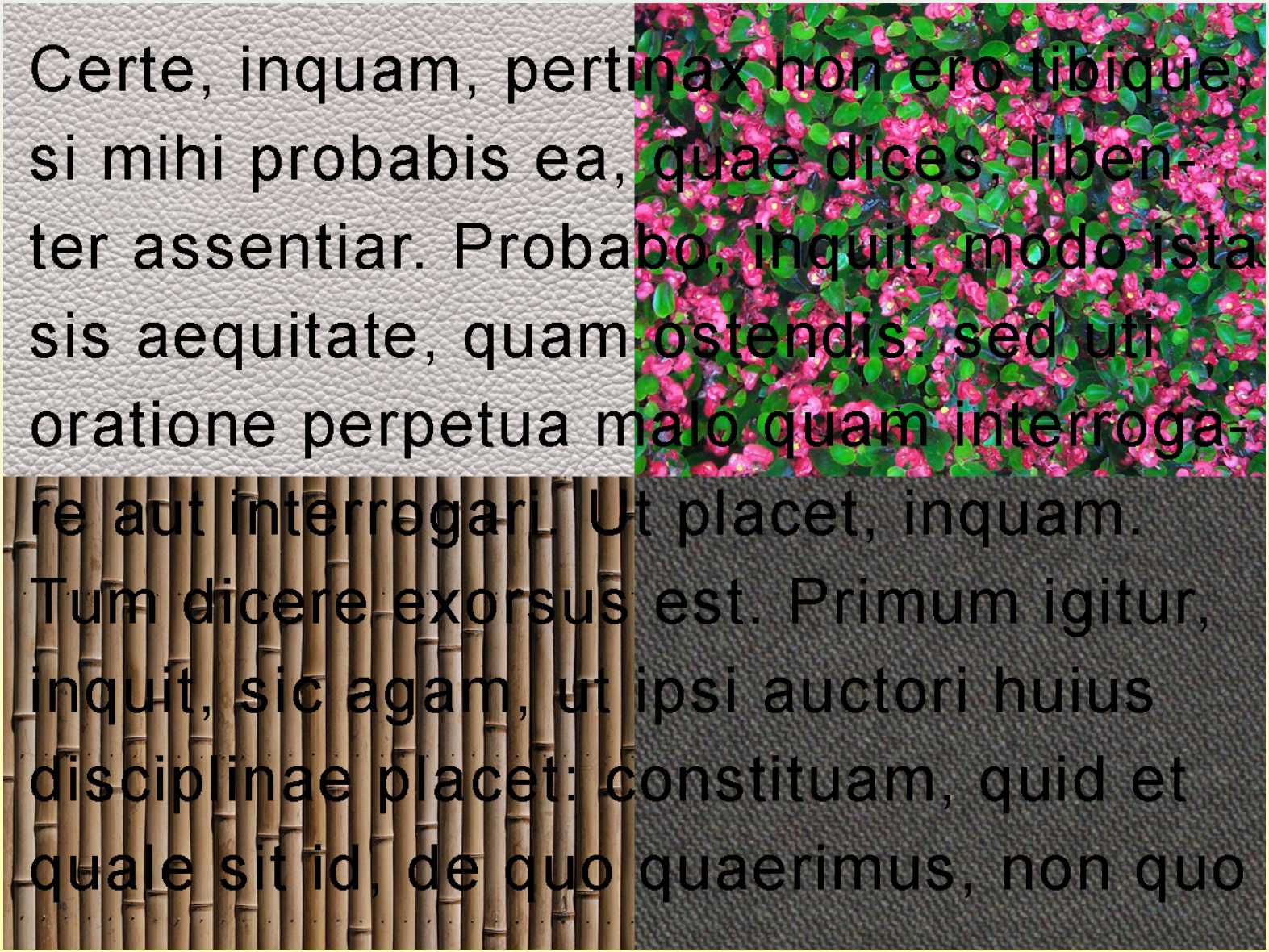


Příklad 5: Predikce chybějících částí obrázku

opravený obrázek

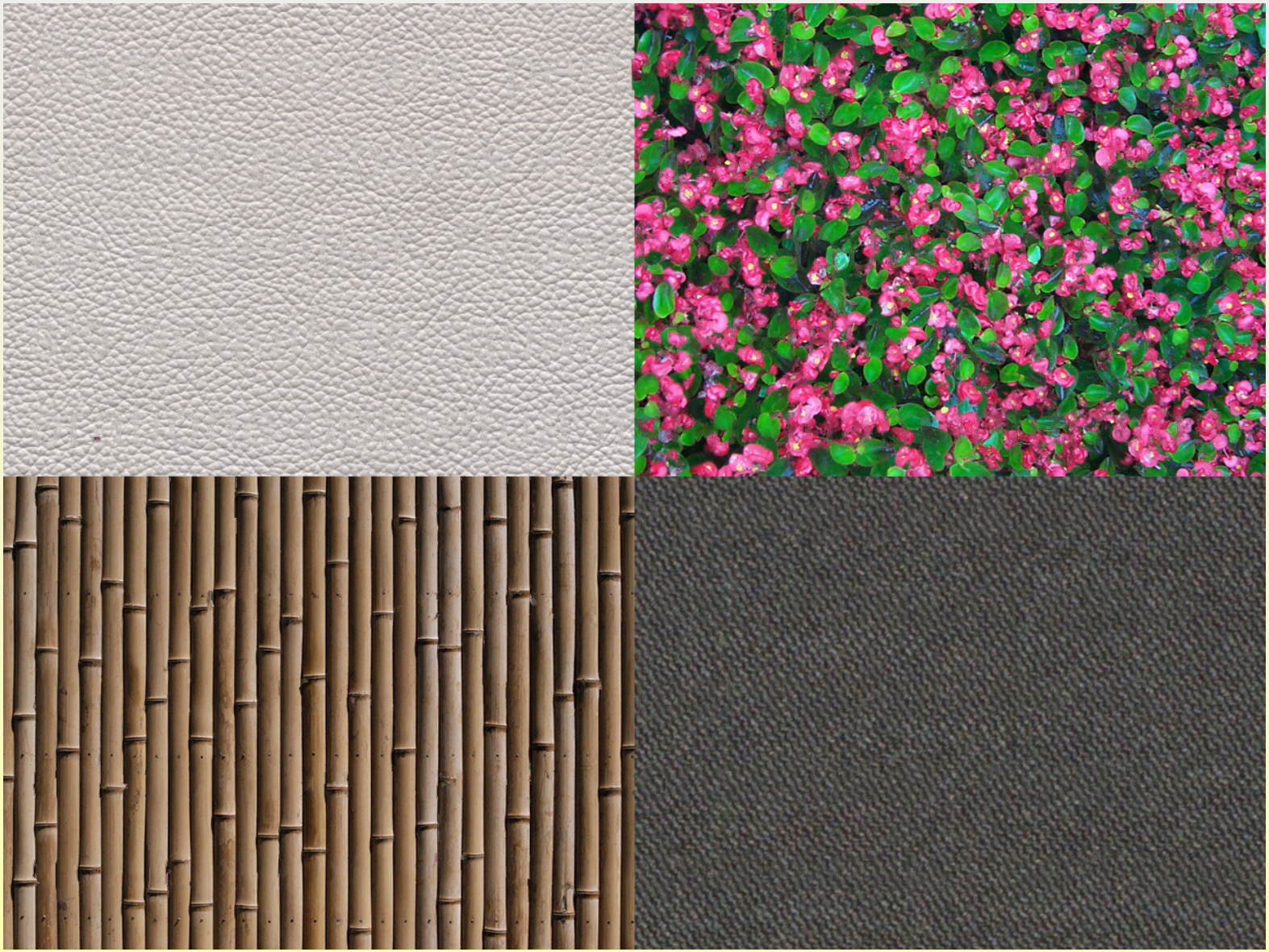
Příklad 8: Predikce chybějících částí obrázku

původní poškozený obrázek



Příklad 8: Predikce chybějících částí obrázku

opravený obrázek



Příklad 8: Predikce chybějících částí obrázku

původní poškozený obrázek

Příklad 8: Predikce chybějících částí obrázku

opravený obrázek

Příklad 9: Interaktivní statistický model dat ze sčítání lidu

První sčítání lidu v českých zemích: 1869

Mf DNES, 17.7.2002

Poslední sčítání lidu domů a bytů v České republice 1.3.2001

- dotazník osob: 25 otázek, bytový dotazník: 17 otázek
- počet obyvatel: 10 230 060
- náklady: 2.4 miliardy Kč

Příští sčítání lidu v České republice: 26.3.2011 (Zákon č. 296/2009 Sb.)

ÚTIA

Příklad 9: Interaktivní statistický model dat ze sčítání lidu

OCHRANA OSOBNÍCH ÚDAJŮ:

Povinnost vyplnit sčítací dotazník: zákon č. 296/2009 Sb., § 7

⇒ Povinnost ochrany osobních údajů: zákon č. 101/2000 Sb.

§ 13: Správce dat a zpracovatel jsou povinni přijmout taková opatření, aby nemohlo dojít k neoprávněnému nebo nahodilému přístupu k osobním údajům

§ 4: Osobním údajem se rozumí jakýkoliv údaj týkající se určeného nebo určitelného subjektu. Subjekt se považuje za určený nebo určitelný, jestliže lze na základě jednoho či více osobních údajů přímo či nepřímo zjistit jeho identitu.

PROBLÉM: dotazníky lze identifikovat kombinací obecně známých údajů

Důsledky:

- **⇒ ČSÚ nesmí volně poskytovat data ze sčítání lidu**
- **⇒ velmi omezená dostupnost výsledků sčítání lidu**

Příklad 9: Interaktivní statistický model dat ze sčítání lidu

Současné metody publikace výsledků sčítání lidu:

- **aggregovaná data** (územně, např. na úrovni sčítacích okrsků)
  výhoda: přesné součty pro malé územní celky
  nevýhoda: zcela se znehodnotí informace o subpopulacích
- **komerční služby statistických úřadů** (písemný dotaz)
  výhoda: přesná odpověď na jednoznačně formulovaný dotaz
  nevýhody: zdlouhavý postup, popř. poplatky
- **publikace tabulek** (tiskem nebo na paměťových médiích)
  výhoda: přesné výsledky censu pro vybrané kombinace proměnných
  nevýhody: omezený počet tabulek pro tisk, nutná anonymizace
- **interaktivní portál** (výsledek požadované analýzy se provádí na vzdáleném serveru na chráněných datech)
  výhoda: přesná odpověď na jednoznačně formulovaný dotaz
  nevýhody: drahý provoz, nutná anonymizace
- **náhodně vybrané podsoubory mikrodat ($\approx 10^6$ dotazníků)**
  výhoda: neomezené možnosti formulace otázek
  nevýhody: omezená distribuce, nutná anonymizace

Příklad 9: Interaktivní statistický model dat ze sčítání lidu

PRINCIP: odvozování informací ze statistického modelu
(Grim et al. 1992, 1995, 2001, 2004, 2009, 2010)

statistický model datového souboru $\mathcal{S}$:

$$P(\mathbf{x}) = \sum_{m=1}^{M} w_m F(\mathbf{x}|m) = \sum_{m=1}^{M} w_m \prod_{n=1}^{N} p_n(x_n|m)$$

podmíněné distribuce $P_{n|C}(x_n|\mathbf{x}_C), n \notin C$ pro dané $\mathbf{x}_C = (x_{i_1}, \ldots, x_{i_k})$:

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

Interaktivní statistický model

- dokonale zabezpečená ochrana dat **(vhodné pro databáze pacientů)**
- distribuce modelu bez omezení (internet, CD)
- možnost výpočtu modelu z neúplných dat **(databáze pacientů)**
- **nevýhoda:** informace odvozené z modelu jsou zatíženy chybou

Příklad 9: Interaktivní statistický model dat ze sčítání lidu

Postup výpočtu modelu: (Grim et al. 2010)

- **výpočet parametrů distribuční směsi z neúplných dat**
- **odhad chybějících údajů na základě modelu**
- **výpočet parametrů z doplněných dat**

Relativní a absolutní chyba modelu pro M=15000 komponent

(porovnání četností pro **26 mil.** kombinací nejvýše pěti hodnot)

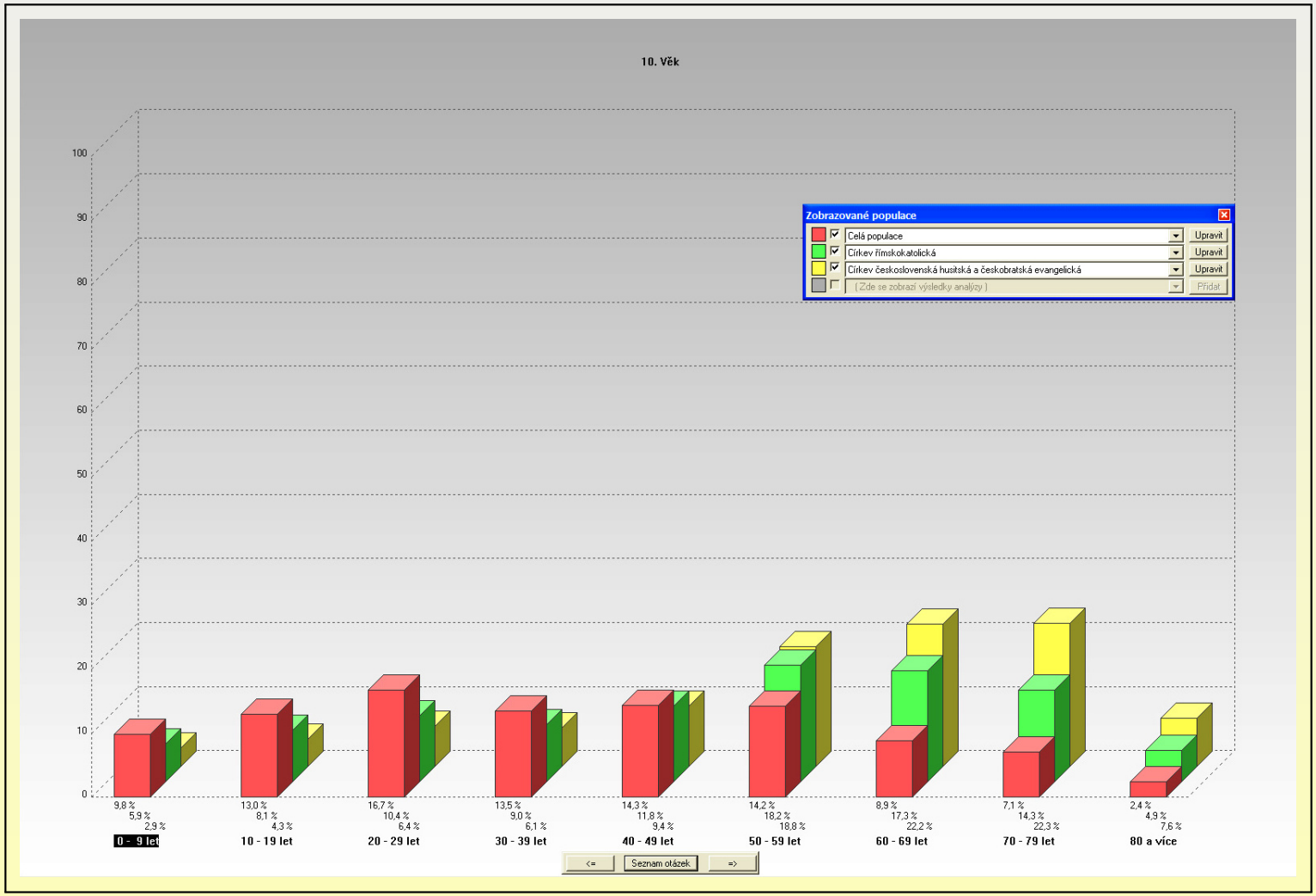
Kriterium přesnosti	**Chyba**
Průměrná absolutní chyba četnosti:	**348**
Průměrná relativní chyba modelu v %:	**4.17**

⇒ **sloupce histogramů se v průměru zobrazují s přesností 4.17%**

interaktivní model a publikace: *http://ro.utia.cas.cz/dem.html*

ÚTIA

Srovnání věkového složení tří různých subpopulací



POZN. Možnost analýzy subpopulací je omezena pouze přesností modelu.

Literatura 1/5

Grim J. (1992): A dialog presentation of census results by means of the probabilistic expert system PES, in *Proceedings of the Eleventh European Meeting on Cybernetics and Systems Research*, Vienna, April 1992, (Ed. R.Trappl), s. 997-1005, World Scientific, Singapore 1992. Paper Award

Grim J., Boček P. (1995): Statistical Model of Prague Households for Interactive Presentation of Census Data, In *SoftStat'95. Advances in Statistical Software 5*, s. 271 - 278, Lucius & Lucius: Stuttgart, 1996.

Grim J., (1996a): Design of multilayer neural networks by information preserving transforms. In: E. Pessa, M.P. Penna, A. Montesanto (Eds.), *Proceedings of the Third European Congress on System Science* (s. 977-982), Roma: Edizzioni Kappa.

Grim J., Pudil P., Somol P. (2000): Recognition of handwritten numerals by structural probabilistic neural networks. In: Proceedings of the Second ICSC Symposium on Neural Computation, Berlin, 2000. (Bothe H., Rojas R. eds.). ICSC, Wetaskiwin, 2000, pp 528-534. Paper Award

Zpět

Literatura 2/5


Grim J. (2000): "Self-organizing maps and probabilistic neural networks". Neural Network World, 3(10): 407-415. Paper Award


Grim J., Boček P., Pudil P. (2001): Safe dissemination of census results by means of interactive probabilistic models. In: *Proceedings of the ETK-NTTS 2001 Conference*, (Hersonissos (Crete), June 18-22, 2001), Vol.2, s. 849-856, European Communities 2001.


Grim J., Haindl M. (2003): Texture Modelling by Discrete Distribution Mixtures. Computational Statistics and Data Analysis, 3-4 **41** 603-615


Grim J., Hora J., Pudil P. (2004): Interaktivní reprodukce výsledků sčítání lidu se zaručenou ochranou anonymity dat. *Statistika*, Vol. 84, No. 5, s. 400-414.


Haindl M., Grim J., Somol P., Pudil P., Kudo M. (2004): A Gaussian mixture-based colour texture model. In: *Proc. of the 17th International Conference on Pattern Recognition.* IEEE, Los Alamitos 2004, s. 177-180.


Grim J., Somol P., Haindl M., Pudil P. (2005): A statistical approach to local evaluation of a single texture image. In: Proc. of the 16-th Annual Symposium PRASA 2005. (Nicolls F. ed.). University of Cape Town, 2005, s. 171-176.

Zpět

Literatura 3/5


Haindl M., Grim J., Pudil P., Kudo M. (2005): A Hybrid BTF Model Based on Gaussian Mixtures. In: Texture 2005. Proceedings of the 4th International Workshop on Texture Analysis. (Chantler M., Drbohlav O. eds.). IEEE, Los Alamitos 2005, s. 95-100.


J. Grim, M. Haindl, P. Somol, and P. Pudil. (2006): A subspace approach to texture modelling by using Gaussian mixtures. In *Proc. of the 18th Int. Conf. ICPR 2006*, Eds. B. Haralick, T.K. Ho), s. 235–238, 2006.


J. Grim, P. Somol, M. Haindl, and P. Pudil, (2006): Color texture segmentation by decomposition of Gaussian mixture model, In *Progress in Pattern Recognition, Image Analysis and Applications*, vol. 19, No. 4225, s. 287–296.


Grim J., Hora J. (2007): Recurrent Bayesian Reasoning in Probabilistic Neural Networks. *Artificial Neural Networks – ICANN 2007*, Ed. Marques de Sá et al., LNCS 4669, s. 129–138, Berlin: Springer


Grim, J. (2007): Neuromorphic features of probabilistic neural networks. *Kybernetika.*, 5 **43** 697–712

◂ Zpět

Literatura 4/5

Grim J., Hora, J. (2008): Iterative principles of recognition in probabilistic neural networks. *Neural Networks*, Special Issue, 6 **21**, 838–846 Paper Award

Grim J. (2008): Extraction of Binary Features by Probabilistic Neural Networks. In: *Artificial Neural Networks - ICANN 2008 Part II*, Springer: Berlin, LNCS **5164** 52–61

Grim J., Novovičová J., Somol P. (2008): Structural poisson mixtures for classification of documents. ICPR 2008: 1-4, http://dx.doi.org/10.1109/ICPR.2008.4761669

Grim J., Somol P., Pudil P., Míková I., Malec M. (2008): Texture Oriented Image Inpainting based on Local Statistical Model. In: Proc. 10th IASTED Conf. on Signal & Image Processing, SIP 2008. Calgary : ACTA Press, 2008 - (Cristea, P.), s. 15-20.

Grim J., Hora J., Somol P., Boček P., Pudil, P. (2009): Interaktivní statistický model dat ze sčítání lidu v ČR v r. 2001. Statistika. Roč. 89, č. 4, s. 285-299

Grim J., Hora J., Somol P., Boček P., Pudil, P. (2010): Statistical Model of the 2001 Czech Census for Interactive Presentation. *Journal of Official Statistics*. Vol. 26, No. 4, pp. 673–694.

Zpět

ÚTIA

Literatura 5/5

Grim J., Somol P., Haindl M., Danes J. (2009): Computer-Aided Evaluation of Screening Mammograms Based on Local Texture Models. *IEEE Transactions on Image Processing 18(4): 765-773* Paper Award

Grim J., Hora J., Boček P., Somol P. and P. Pudil (2010): Statistical Model of the 2001 Czech Census for Interactive Presentation. *Journal of Official Statistics*. Vol. 26, No. 4, pp. 673–694.

Grim J., Hora, J. (2010): Computational Properties of Probabilistic Neural Networks. In: *Artificial Neural Networks - ICANN 2010 Part II*, Springer: Berlin, LNCS **5164** 52–61

Grim J., Somol P., Pudil P. (2010): Digital Image Forgery Detection by Local Statistical Models. In: Proc. 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Los Alamitos, California, IEEE computer society, 2010 - (Eds. Echizen, I. et al.) s. 579-582. Paper Award

Grim, J. (2014). Sequential pattern recognition by maximum conditional informativity. *Pattern Recognition Letters*, Vol. 45C, pp. 39-45. *http:// dx.doi.org/10.1016/j.patrec.2014.02.024* Paper Award

Zpět

EM algoritmus pro Bernoulliovskou směs

základní schema EM algoritmu v C++: **směs Bernoulliho rozložení**

```
//    Odhad parametru smesi Bernoulliho rozlozeni pomoci EM algoritmu
//=========================================================================
//int       NN;                         //  dimenze binarniho vektoru (DNN=NN+1)
//int       MM;                         //  pocet komponent smesi (DMM=MM+1)
//short     X[DNN];                     //  binarni datovy vektor
//double    P[DMM][DNN],SP[DMM][DNN];//  parametry smesi (theta) a scitaci promenne
//double    W[DMM],SW[DMM];             //  vahy komponent a prislusne scitaci promenne
//double    FX[DMM];                    //  hodnoty komponent pro dany vektor X[DNN]
//double    FXM,SWM,Q,SUM,SWM;          //  pomocne promenne
//int       N,M,IT,ITERMAX;             //  pomocne promenne

for(IT=1; IT<=ITERMAX; IT++)
//*************************************************************************
{  for(M=1; M<=MM; M++) {SW[M]=0.0; for(N=1; N<=NN; N++) SP[M][N]=0.0;}
   Q=0.0;
   for(J=1;J<=JJ;J++)                  //  cyklus pres vsechny datove vektory X
   { READ(X);  SUM=0.0;                //  nacteni X ze vstupniho souboru
     for(M=1; M<=MM; M++)
     {  FXM=W[M];
        for(N=1; N<=NN; N++) if(X[N]==1) FXM*=P[M][N]; else FXM*=(1-P[M][N]);
        FX[M]=FXM;  SUM+=FXM;
     }  // end of M-loop
     Q=Q+log(SUM);
     for((M=1; M<=MM; M++)
     {  G=FX[M]/SUM; SW[M]+=G;  for(N=1; N<=NN; N++) if(X[N]==1) SP[M][N]+=G;
     }  //  end of M-loop
   } //  end of J-loop
   Q=Q/JJ;
   for(M=1; M<=MM; M++)               //  vypocet novych parametru komponent
   {  SWM=SW[M];   W[M]=SWM/JJ;  for((N=1; N<=NN; N++) P[M][N]=SP[M][N]/SWM;
   }  // end of M-loop
   print(IT,Q);
}  //  end of IT-loop
//*************************************************************************
printf("\nKonec EM algoritmu\n\n");
```

POZN. Výpočet vhodný pouze pro malou dimenzi NN.

Zpět

EM algoritmus pro součinovou normální směs

EM algoritmus v C++: **součinová normální směs s velkou dimenzí**

```
//    Odhad parametru normalni soucinove smesi pomoci EM algoritmu
//=============================================================================
//int IT,N,M;  long K;  double F,G,FXM,SWM,SUM,FMAX,Q0;    // globalni promenne:
//short    X[DNN];                    // datovy vektor (DNN=NN+1)
//double   FX[DMM],W[DMM],SW[DMM];    // komponenty, vahy a odhady vah komponent
//double   C[DMM][DNN], A[DMM][DNN];  // vektory prumeru a rozptylu (DMM=MM+1)
//double   SC[DMM][DNN],SA[DMM][DNN]; // nove odhady vektory prumeru a rozptylu
for(IT=1; IT<=ITMAX; IT++)
//*****************************************************************************
{ Q=0.0
  for(M=1; M<=MM; M++)                // logaritmicke parametry a nulovani stradacu
  {  SW[M]=RMIN;          F=log(W[M]+RMIN)-NN2LN2PI;
     for(N=1; N<=NN; N++) {F-=log(A[M][N]); SC[M][N]=RMIN; SA[M][N]=RMIN;}
     W[M]=2*F;                        // kvuli deleni pri vypoctu exponentu
  }  // end of M-loop
  for(I=1;I<=K;I++)                   // cyklus pres vsechny datove vektory X
  {  READ(X);  FMAX=-RMAX;
     for(M=1; M<=MM; M++)             // vypocet logaritmu komponent
     {  FXM=W[M]; for(N=1; N<=NN; N++) {F=(X[N]-C[M][N])/A[M][N]; FXM-=F*F;}
        FXM/=2.0f;    FX[M]=FXM;  if(FXM>FMAX) FMAX=FXM;
     } // end of M-loop
     SUM=0.0;
     for(M=1; M<=MM; M++)             // odlogaritmovani komponent a vypocet P(X)
     {  FXM=FX[M]-FMAX;  if(FXM>MINLOG) {FXM=exp(FXM); SUM+=FXM;} else FXM=0.0;
        FX[M]=FXM;
     }  // end of M-loop
     Q=Q+log(SUM)+FMAX;               // vypocet hodnoty verohodnostni funkce
     for(M=1; M<=MM; M++)
     {  G=FX[M]/SUM;  SW[M]+=G;
        for(N=1; N<=NN; N++) {F=X[N]; SC[M][N]+=G*F; SA[M][N]+=G*F*F;}
     }  //  end of M-loop
  }  //  end of K-loop
  Q/=K;
  for(M=1; M<=MM; M++)                 //  vypocet novych parametru komponent
  {  SWM=SW[M];   W[M]=SWM/K;
     for(N=1; N<=NN; N++)
     {  F=SC[M][N]/SWM;   C[M][N]=F;  A[M][N]=sqrt(SA[M][N]/SWM-F*F);
     }  //  end of N-loop
  }   //  end of M-loop
  printf("\nIT=%2d   Q=%15.7lf  \n",IT,Q);
  //***************************************************************************
} //  end of IT-loop
```

◀ Zpět

Příklad 3: Klasifikace textových dokumentů

věrohodnostní funkce:

$$L = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} \log[\sum_{m \in \mathcal{M}_c} G(\mathbf{x}|\theta_m, \phi_m) w_m], \qquad \mathcal{S}_c = \{\mathbf{x}_1, \ldots, \mathbf{x}_{K_c}\}$$

EM algoritmus:

$$q(m|\mathbf{x}) = \frac{G(\mathbf{x}|\theta_m, \phi_m) w_m}{\sum_{j \in \mathcal{M}_c} G(\mathbf{x}|\theta_j, \phi_j) f(j)}, \qquad m \in \mathcal{M}_c,\ n \in \mathcal{N},\ \mathbf{x} \in \mathcal{S}_c$$

$$\tilde{x}_n^{(m)} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} x_n q(m|\mathbf{x}), \qquad |\bar{\mathbf{x}}|^{(m)} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} |\mathbf{x}| q(m|\mathbf{x})$$

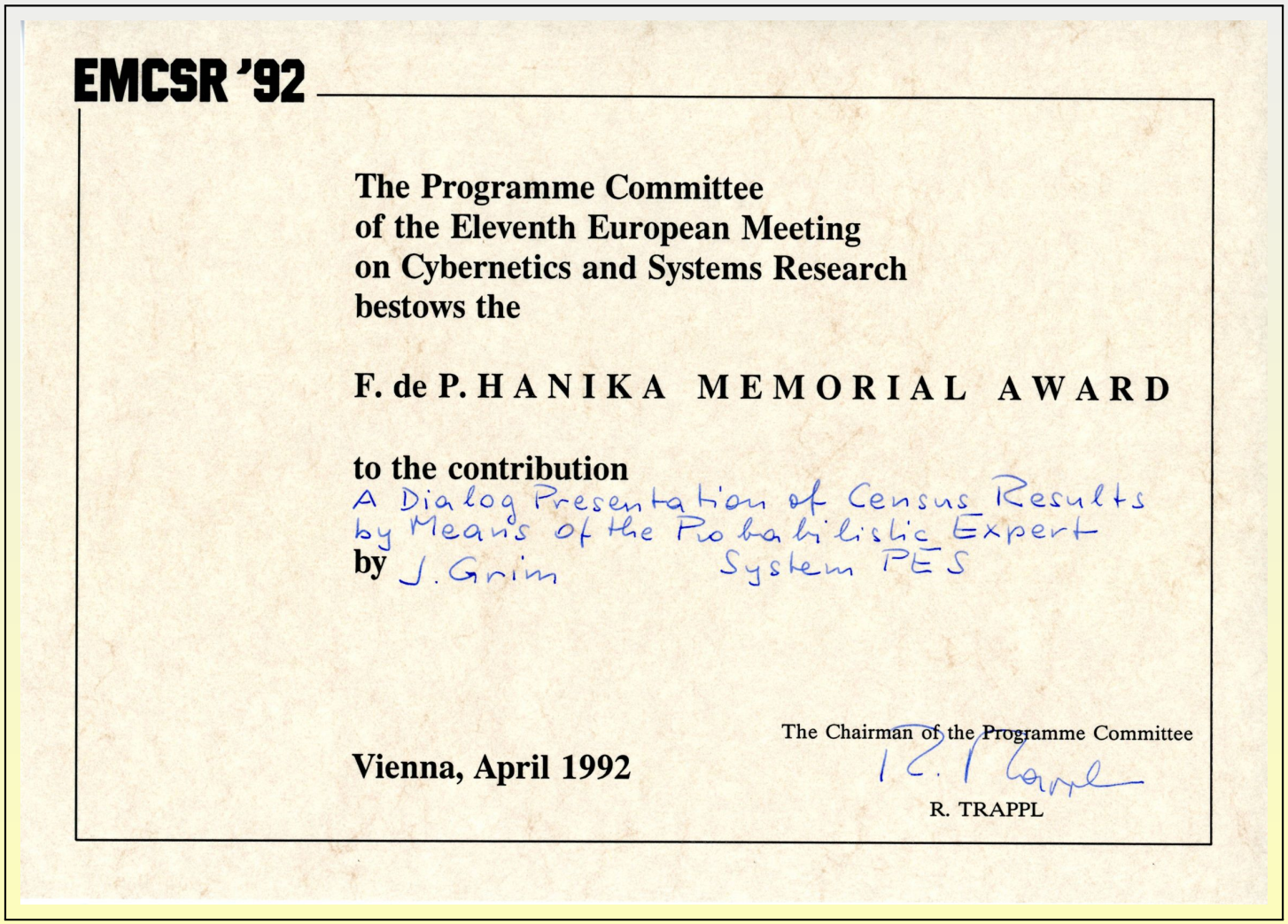
$$w_m^{'} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} q(m|\mathbf{x}), \qquad \theta_{mn}^{'} = \frac{\tilde{x}_n^{(m)}}{|\bar{\mathbf{x}}|^{(m)}}$$

$$\phi_{mn}^{'} = \begin{cases} 1, & \gamma_{mn}^{'} \in \Gamma_r^{'}, \\ 0, & \gamma_{mn}^{'} \notin \Gamma_r^{'}, \end{cases}, \qquad \gamma_{mn}^{'} = \tilde{x}_n^{(m)} \log \frac{\theta_{mn}^{'}}{\theta_{0n}} + |\bar{\mathbf{x}}|^{(m)} (\theta_{0n}^{'} - \theta_{mn})$$

$\Gamma_r^{'}$ označuje množinu r nejvyšších hodnot $\gamma_{mn}^{'}$

◂ Zpět

Paper Award



EMCSR '92

The Programme Committee
of the Eleventh European Meeting
on Cybernetics and Systems Research
bestows the

F. de P. HANIKA MEMORIAL AWARD

to the contribution
A Dialog Presentation of Census Results
by Means of the Probabilistic Expert
by J. Grim System PES

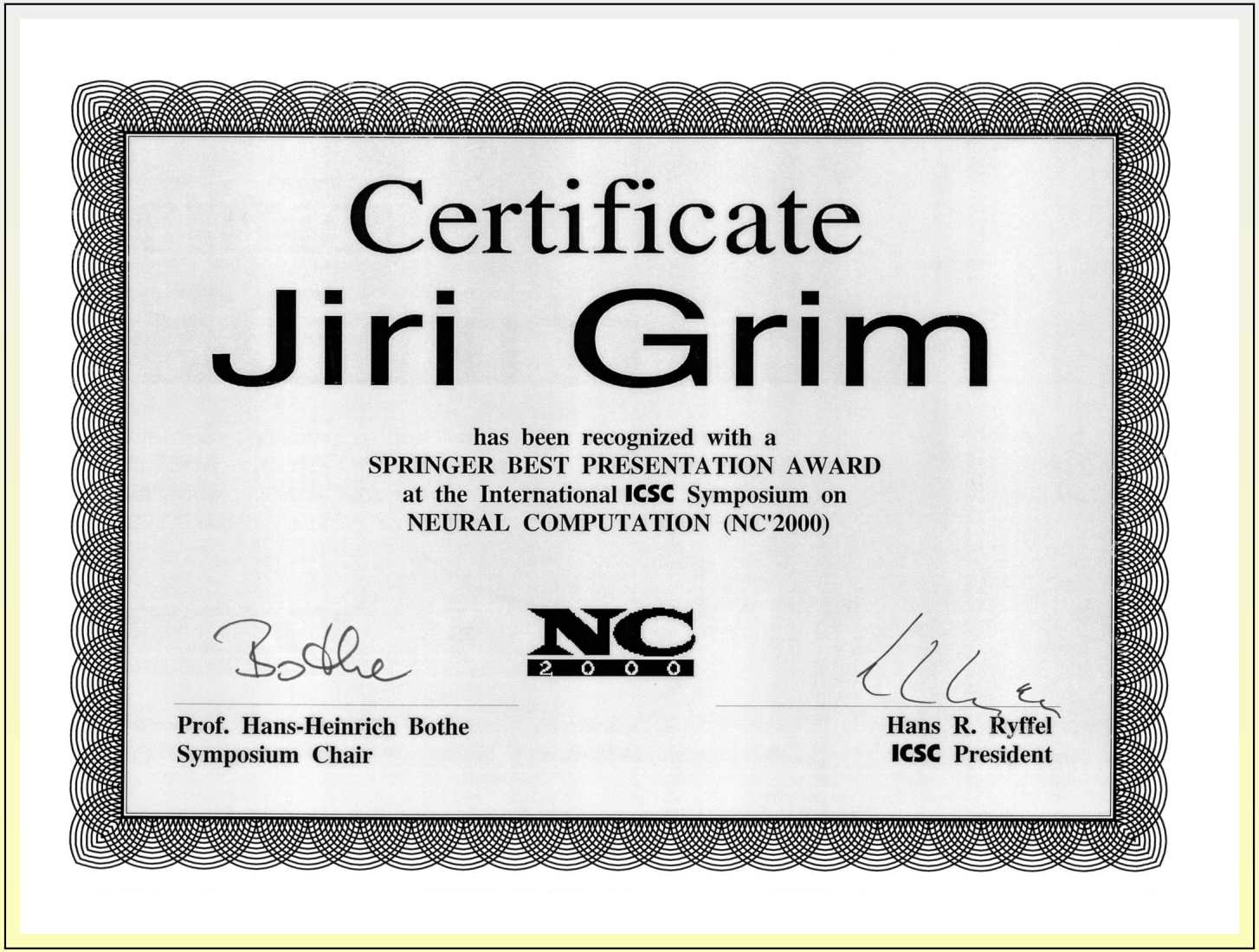
Vienna, April 1992

The Chairman of the Programme Committee
R. TRAPPL

Eleventh European Meeting on Cybernetics and Systems Research, Vienna, April 1992

◂ Zpět

Paper Award



Certificate

Jiri Grim

has been recognized with a
SPRINGER BEST PRESENTATION AWARD
at the International ICSC Symposium on
NEURAL COMPUTATION (NC'2000)

NC 2000

Prof. Hans-Heinrich Bothe
Symposium Chair

Hans R. Ryffel
ICSC President

Second ICSC Symposium on Neural Computation, Berlin, 2000

Zpět

Paper Award

ÚTIA

Ústav teorie informace a automatizace
Akademie věd České republiky v.v.i.

Uděluje v soutěži o Nejlepší publikaci 2009

2. cenu

Jiřímu Grimovi

za práci

Computer-Aided Evaluation of Screening Mammograms Based on Local Texture Models

IEEE Trans. on Image Processing, vol. 18, No. 4, pp. 765–773, ISSN 1057–7149, April 2009.

(spoluautoři Somol P., Haindl M., Daneš J.)

ředitel ústavu

IEEE Transactions on Image Processing 18(4): 765-773, 2009

◀ Zpět

ÚTIA

Paper Award

Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing
IIH-MSP-2010
October 15-17, 2010, Darmstadt, Germany

Best Paper Award Certificate

in recognition of the professional excellence of the paper

Digital Image Forgery Detection by Local Statistical Models

by

Jiri Grim, Petr Somol

&

Pavel Pudil

Xiamu Niu
General Chair
Harbin Institute of Technology
China

Alexander Nouak ,
Conference executive Chair
Fraunhofer IGD
Germany

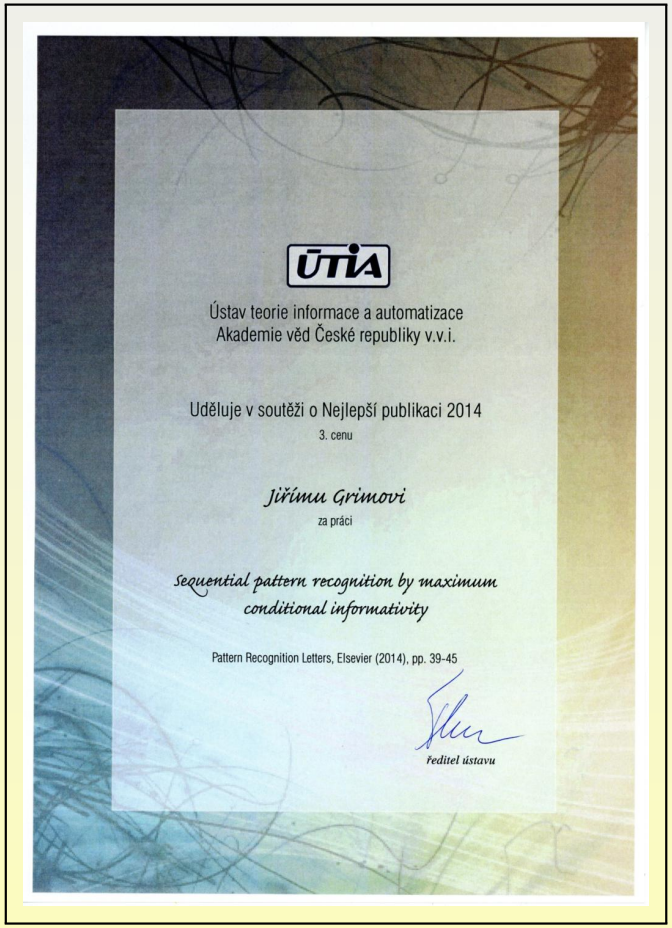
Jeng-Shyang Pan
Program Committee Chair
National Kaohsiung University of Applied Sciences
Taiwan

Isao Echizen,
Program Committee Chair
National Institute of Informatics
Japan

Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP Darmstadt, 2010

◀ Zpět

Paper Award



Pattern Recognition Letters, Vol. 45C, pp. 39-45, 2014

◀ Zpět

Paper Award

ÚTIA

Ústav teorie informace a automatizace
Akademie věd České republiky

Uděluje v soutěži o Nejlepší publikaci 2008

2. cenu

Jiřímu Grimovi

za práci

Iterative principles of recognition in probabilistic neural networks

Neural Networks 21 (2008), 838-846.
(spoluautor Hora J.)

ředitel ústavu

Neural Networks, 21(6): 838–846, 2008

◀ Zpět

Paper Award

ÚTIA

Ústav teorie informace a automatizace
Akademie věd České republiky

Cena za nejlepší článek

udělena v roce 2000

Jiřímu Grimovi

za práci

Self-Organizing Maps
and Probabilistic Neural Networks

Neural Network World No. 3, 2000, 407-415

ředitel ústavu

Neural Network World, 3(10): 407-415, 2000

◀ Zpět

ÚTIA