SOUČINOVÉ DISTRIBUČNÍ SMĚSI

I. část: EM algoritmus

Jiří Grim

Ústav teorie informace a automatizace AV ČR

Oddělení rozpoznávání obrazů

Září 2010

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

ÚTIA

Outline

ÚTIA

Aproximace neznámého rozložení pravděpodobnosti

Výchozí informace:

soubor nezávislých pozorování $\mathcal{S}$ (trénovací data) s nějakým neznámým rozložením pravděpodobnosti $P^*(\mathbf{x})$

$$\mathcal{S} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(K)}\}, \qquad \mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_N^{(k)}) \in \mathcal{X}$$

Princip metody distribučních směsí:

aproximace neznámého rozložení pravděpodobnosti $P^*(\mathbf{x})$ pomocí distribuční směsi; komponenty směsi: hustoty nebo diskrétní distribuce

$$P(\mathbf{x}) = \sum_{m=1}^{M} f(m) F(\mathbf{x}|m), \qquad \sum_{m=1}^{M} f(m) = 1, \qquad \sum_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}|m) = 1$$

Příklady aplikací:

▸ rozpoznávání obrazů, problém predikce, modelování textur, analýza obrazů, klasifikace textových dokumentů, statistické modely dat, ...

Směsi jako kompromis

parametrický přístup: např. odhad normální hustoty pravděpodobnosti

$$P(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det A}} \exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{c})^T A^{-1}(\mathbf{x}-\mathbf{c})\},\ \mathbf{x} \in \mathcal{X}$$

průměr: $\mathbf{c} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} \mathbf{x}$, kovarianční matice: $A = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} (\mathbf{x}-\mathbf{c})(\mathbf{x}-\mathbf{c})^T$

---

neparametrický přístup: jádrový (Parzenův) odhad ▸ Theorem (Parzen, 1962)

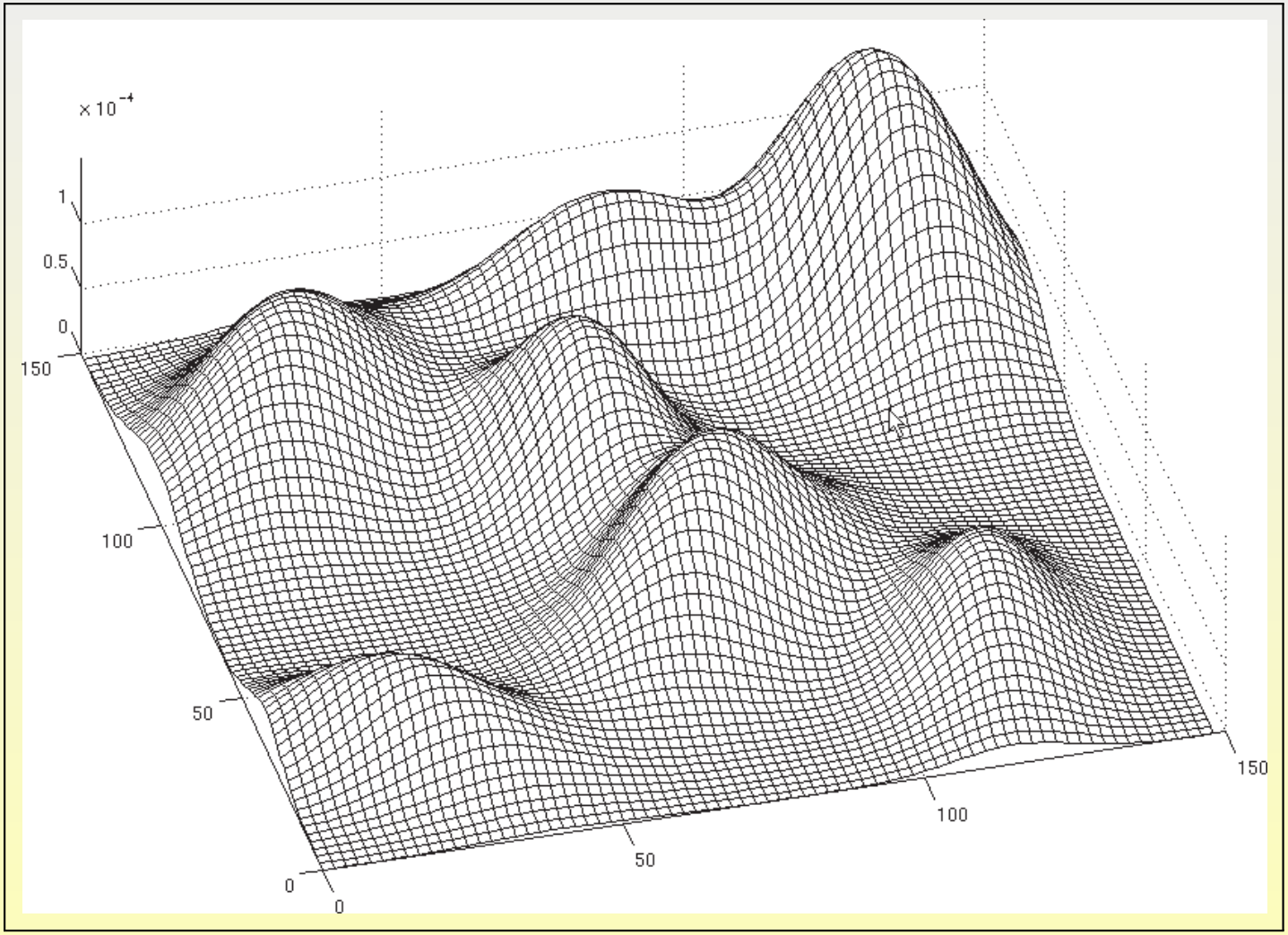
$$P(\mathbf{x}) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{y}\in\mathcal{S}} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left\{\frac{(x_n - y_n)^2}{2\sigma_n^2}\right\},\ \mathbf{x} \in \mathcal{X}$$

problém: ▸ Optimální vyhlazení (volba parametrů σ_n)

Distribuční směsi ≈ "semiparametrický přístup"

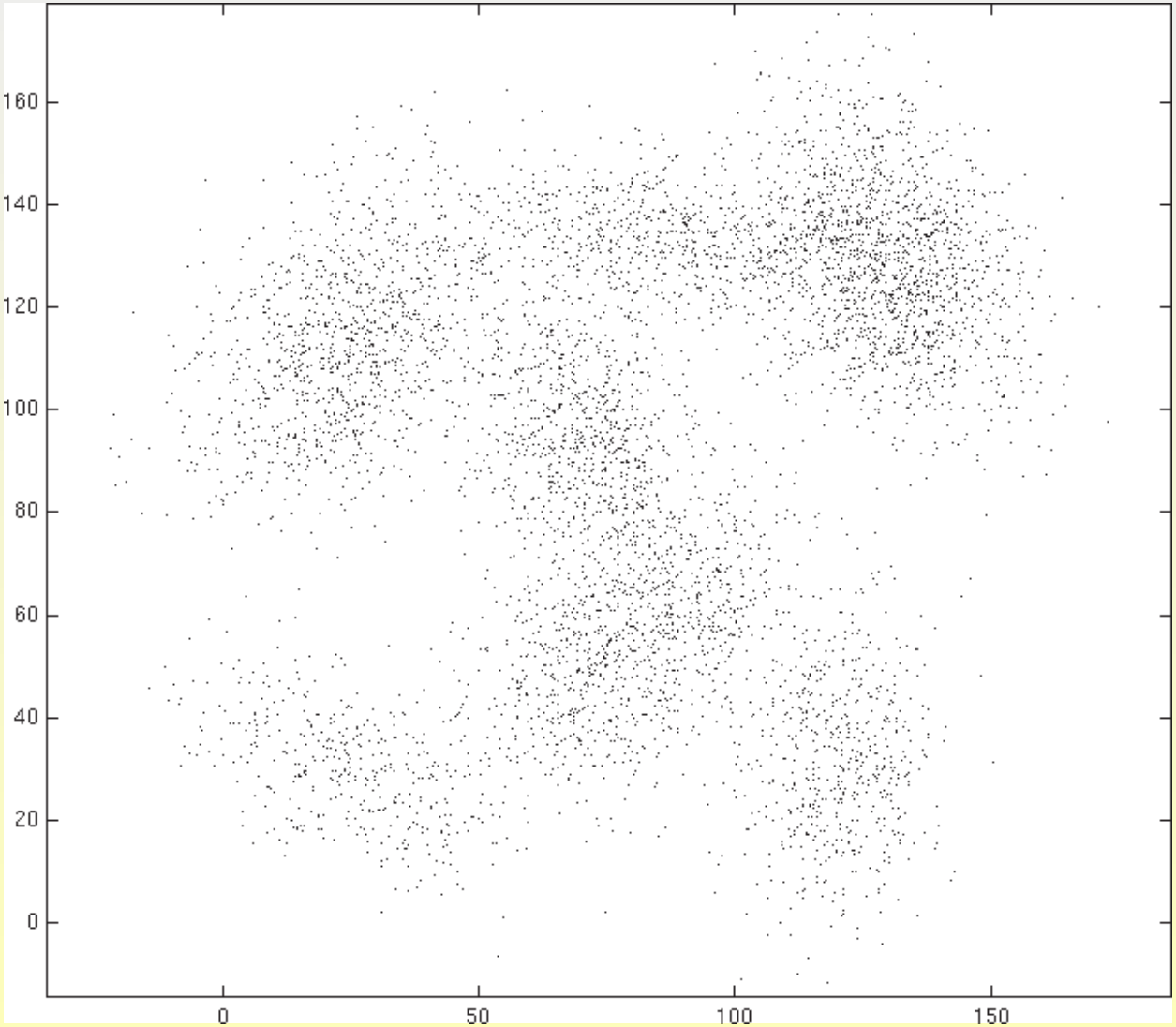
- kompromis mezi jednoduchostí parametrického modelu a obecností neparametrických odhadů
- efektivní odhad parametrů směsi (+"vyhlazení") pomocí EM algoritmu

Příklad normální směsi



dimenze distribuční směsi $N = 2$, počet komponent směsi $M = 7$

Data náhodně generovaná podle normální směsi (M=7)



počet bodů: $|\mathcal{S}| = 6000$

EM algoritmus pro směs normálních hustot

výpočet maximálně věrohodného odhadu parametrů směsi:

$$F(\mathbf{x}|\mathbf{c}_m, A_m) = \frac{1}{\sqrt{(2\pi)^N \det A_m}} \exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_m)^T A_m^{-1}(\mathbf{x}-\mathbf{c}_m)\},\ \mathbf{x} \in \boldsymbol{R}^N$$

$$L = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} \log\left[\sum_{m=1}^{M} F(\mathbf{x}|\mathbf{c}_m, A_m) f(m)\right], \qquad \mathcal{S} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\}$$

Iterační rovnice: $\approx$ maximalizace věrohodnostní funkce

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

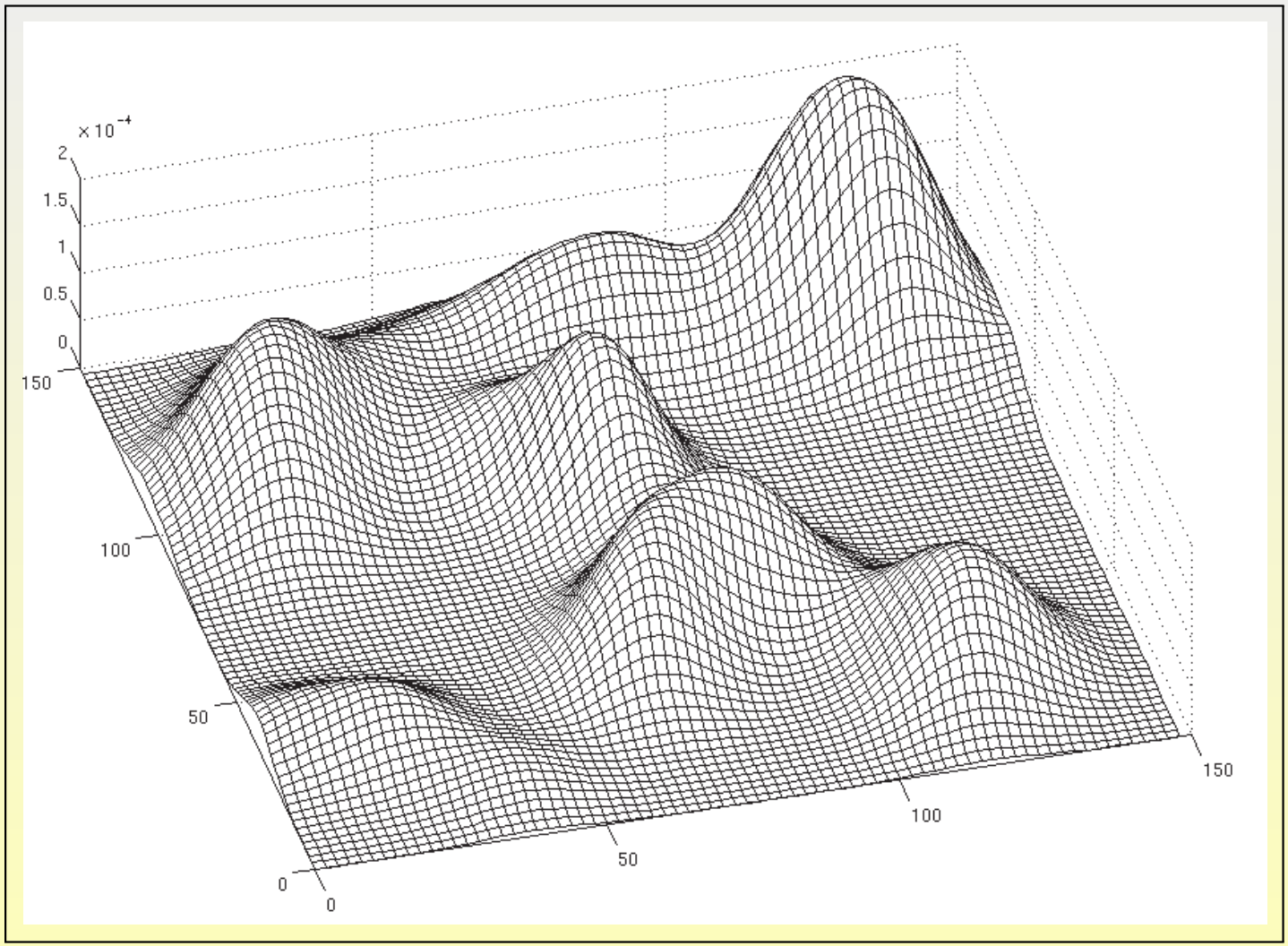
$$f'(m) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}), \qquad \mathbf{c}'_m = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})}\sum_{\mathbf{x}\in\mathcal{S}} \mathbf{x}\ q(m|\mathbf{x})$$

$$A'_m = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})}\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})\ (\mathbf{x}-\mathbf{c}'_m)(\mathbf{x}-\mathbf{c}'_m)^T$$

POZN. **Počet komponent směsi je nutné zvolit předem.**

ÚTIA

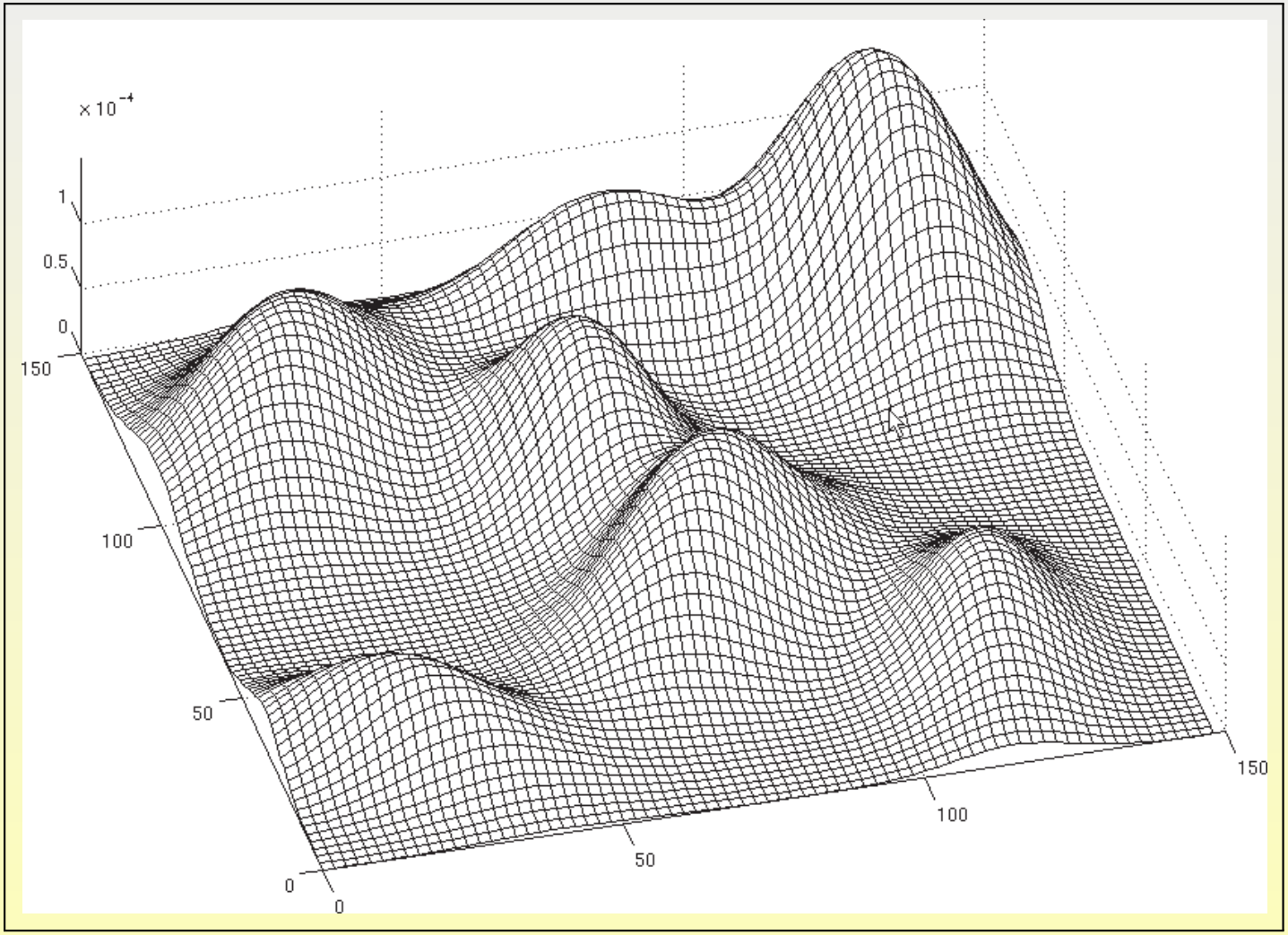
Příklad odhadu normální směsi (M=28)



použitý počet komponent směsi $M = 28$ ($\neq 7$)

▸ (SROVNÁNÍ: jádrový odhad)

Původní směs normálních hustot (M=7)



dimenze distribuční směsi $N = 2$, počet komponent směsi $M = 7$

Obecná verze EM algoritmu

EM algoritmus: maximalizuje věrohodnostní funkci (kritérium)

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log P(\mathbf{x}) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \left[\sum_{m=1}^{M} f(m) F(\mathbf{x}|m) \right]$$

Iterační rovnice: $(m = 1, 2, \ldots, M, \quad \mathbf{x} \in \mathcal{S}, \quad \mathcal{S} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\})$

E-krok: $q(m|\mathbf{x}) = \dfrac{f(m) F(\mathbf{x}|m)}{\sum_{j=1}^{M} f(j) F(\mathbf{x}|j)}, \qquad f'(m) = \dfrac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x})$

M-krok: $F'(.|m) = \arg \max_{F(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log F(\mathbf{x}|m) \right\}$ ▸ Explicitní řešení

---

pro součinové komponenty: $F(\mathbf{x}|m) = \prod_{n=1}^{N} f_n(x_n|m)$

$$\Rightarrow \quad f_n'(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log f_n(x_n|m) \right\}, \quad n = 1, 2, \ldots, N$$

POZN. **V kroku M stačí nerovnost $\Rightarrow$ zobecněný EM algoritmus.**

Monotónní vlastnost EM algoritmu (Schlesinger, 1968)

Posloupnost hodnot věrohodnostní funkce $\{L^{(t)}\}_{t=0}^{\infty}$ je neklesající:

$$L^{(t+1)} - L^{(t)} \geq 0, \quad t = 0, 1, 2, \ldots$$

a pokud je shora omezená, konverguje monotónně k lokálnímu nebo globálnímu maximu (nebo sedlovému bodu):

$$\lim_{t \to \infty} L^{(t)} = L^* < \infty.$$

Z existence konečné limity $L^* < \infty$ plynou následující nutné podmínky konvergence: ▸ Důkaz

$$\lim_{t \to \infty} (L^{(t+1)} - L^{(t)}) = 0 \quad \Rightarrow$$

$$\Rightarrow \quad \lim_{t \to \infty} ||f^{(t+1)}(\cdot) - f^{(t)}(\cdot)|| = 0, \qquad \lim_{t \to \infty} ||q^{(t+1)}(\cdot|\mathbf{x}) - q^{(t)}(\cdot|\mathbf{x})|| = 0$$

POZN. Z konvergence posloupnosti $\{L^{(t)}\}_{t=0}^{\infty}$ neplyne automaticky konvergence posloupností odhadovaných parametrů směsi !

ÚTIA

Důkaz monotónní vlastnosti

Lemma

Kullback-Leiblerova informační divergence $I(q(\cdot|\mathbf{x})||q'(\cdot|\mathbf{x}))$ *je nezáporná pro libovolné dvě podmíněné distribuce* $q(\cdot|\mathbf{x}), q'(\cdot|\mathbf{x})$ *a rovná se nule právě když jsou obě distribuce identické.* ▸ Důkaz

$$\Rightarrow \quad \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} I(q(\cdot|\mathbf{x})||q'(\cdot|\mathbf{x})) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\left[\sum_{m=1}^{M} q(m|\mathbf{x})\log\frac{q(m|\mathbf{x})}{q'(m|\mathbf{x})}\right] \geq 0$$

Dosazením za $q(m|\mathbf{x}), q'(m|\mathbf{x})$ podle kroku E dostaneme nerovnost:

$$\frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\sum_{m=1}^{M} q(m|\mathbf{x})\log\frac{P'(\mathbf{x})}{P(\mathbf{x})} - \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\sum_{m=1}^{M} q(m|\mathbf{x})\log\left[\frac{f'(m)F'(\mathbf{x}|m)}{f(m)F(\mathbf{x}|m)}\right] \geq 0$$

přičemž první člen na levé straně je roven přírůstku kritéria L:

$$\frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\sum_{m=1}^{M} q(m|\mathbf{x})\log\frac{P'(\mathbf{x})}{P(\mathbf{x})} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\log\frac{P'(\mathbf{x})}{P(\mathbf{x})} = L' - L.$$

ÚTIA

Důkaz monotónní vlastnosti

Z upravené nerovnosti dále plyne po dosazení z předchozí rovnice:

$$(*) \quad L^{'} - L \geq \sum_{m=1}^{M} \left[\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \right] \log \frac{f^{'}(m)}{f(m)} + \sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log \frac{F^{'}(\mathbf{x}|m)}{F(\mathbf{x}|m)}$$

S využitím substituce podle kroku M

$$f^{'}(m) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}), \quad m = 1, 2, \ldots, M$$

můžeme napsat nerovnost:

$$\sum_{m=1}^{M} \left[\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \right] \log \frac{f^{'}(m)}{f(m)} = \sum_{m=1}^{M} f^{'}(m) \log \frac{f^{'}(m)}{f(m)} \geq 0.$$

tzn. první suma na pravé straně výchozí nerovnosti (*) je nezáporná.

POZN. Definice vah $f^{'}(m)$ maximalizuje výraz na levé straně

Důkaz monotónní vlastnosti

Podle definice v kroku M funkce $F'(\cdot|m)$ maximalizuje levou stranu, tzn.:

$$\sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \log F'(\mathbf{x}|m) \geq \sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \log F(\mathbf{x}|m).$$

Předchozí nerovnost lze upravit na tvar

$$\sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \log \frac{F'(\mathbf{x}|m)}{F(\mathbf{x}|m)} \geq 0$$

tj. přírůstek věrohodnostní funkce je nezáporný:

$$L' - L \geq \sum_{m=1}^{M} f'(m) \log \frac{f'(m)}{f(m)} + \sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \log \frac{F'(\mathbf{x}|m)}{F(\mathbf{x}|m)} \geq 0$$

$$\Rightarrow \quad L' \geq L$$

▸ Alternativní důkaz podle Schlesingera

POZN. **Důkaz monotonie je návodem na odvození EM algoritmu!**

Identifikace směsi × aproximace pomocí směsi

POČET KOMPONENT ? **POČÁTEČNÍ PARAMETRY ?**

Problém identifikace směsi (např. shluková analýza)

- cílem je zjistit počet komponent a odhadnout parametry směsi
- je třeba aby odhadovaná směs byla identifikovatelná ▸ Definice
- POTÍŽ: věrohodnostní funkce směsi má téměř vždy lokální maxima (zvláště při velké dimenzi a malém počtu dat)
- $\Rightarrow$ nalezené lokální maximum závisí na volbě počátečních parametrů
- POTÍŽ: kvalita odhadu směsi závisí na zvoleném počtu komponent a počátečních parametrech

× **Problém aproximace neznámé pravděpodobnostní distribuce**

- cílem je co nejpřesnější aproximace neznámého rozložení pravděpodobnosti pomocí součinové distribuční směsi ▸ Problém aproximace
- směs nemusí být identifikovatelná
- konkrétní počet komponent směsi není důležitý
- počáteční parametry směsi je možné generovat náhodně

Výpočetní vlastnosti EM algoritmu

Typická aproximační úloha: velký počet dat + velký počet komponent

- efektivní aproximace multimodálních distribucí s velkou dimenzí
- existence lokálních extrémů není důležitá protože při velkém počtu komponent je hodnota kriteria v různých lokálních maximech srovnatelná
- při velkém počtu komponent ($M \approx 10^1 - 10^2$) lze zanedbat komponenty s řádově nižší váhou (konkrétní počet není důležitý)
- $\Rightarrow$ inicializace parametrů směsi nemá podstatný vliv, počáteční parametry komponent je možné generovat náhodně
- iterace EM algoritmu v závěrečné pomalé fázi výpočtu mají obvykle malý vliv na hodnotu kritéria a přesnost aproximace a lze je vynechat
- závěrečná fáze výpočtu obvykle zvyšuje riziko "přeurčení" parametrů (overfitting), tzn. včasné ukončení iterací může zlepšit kvalitu řešení
- EM algoritmus lze použít na "vážená" data ▸ Modifikace EM algoritmu

POZN. Uvedené vlastnosti neplatí obecně, nelze je dokázat, závisí na konkrétních datech.

Z historie problému odhadování směsí

v období 1895 - 1965 bylo publikováno asi 80 prací o identifikaci směsí

- **Pearson (1894):** "Contributions to the mathematical theory of evolution. 1. Dissection of frequency curves." Philosophical Transactions of the Royal Society of London **185**, 71-110. *(odhady směsi dvou jednorozměrných normálních hustot metodou momentů)*

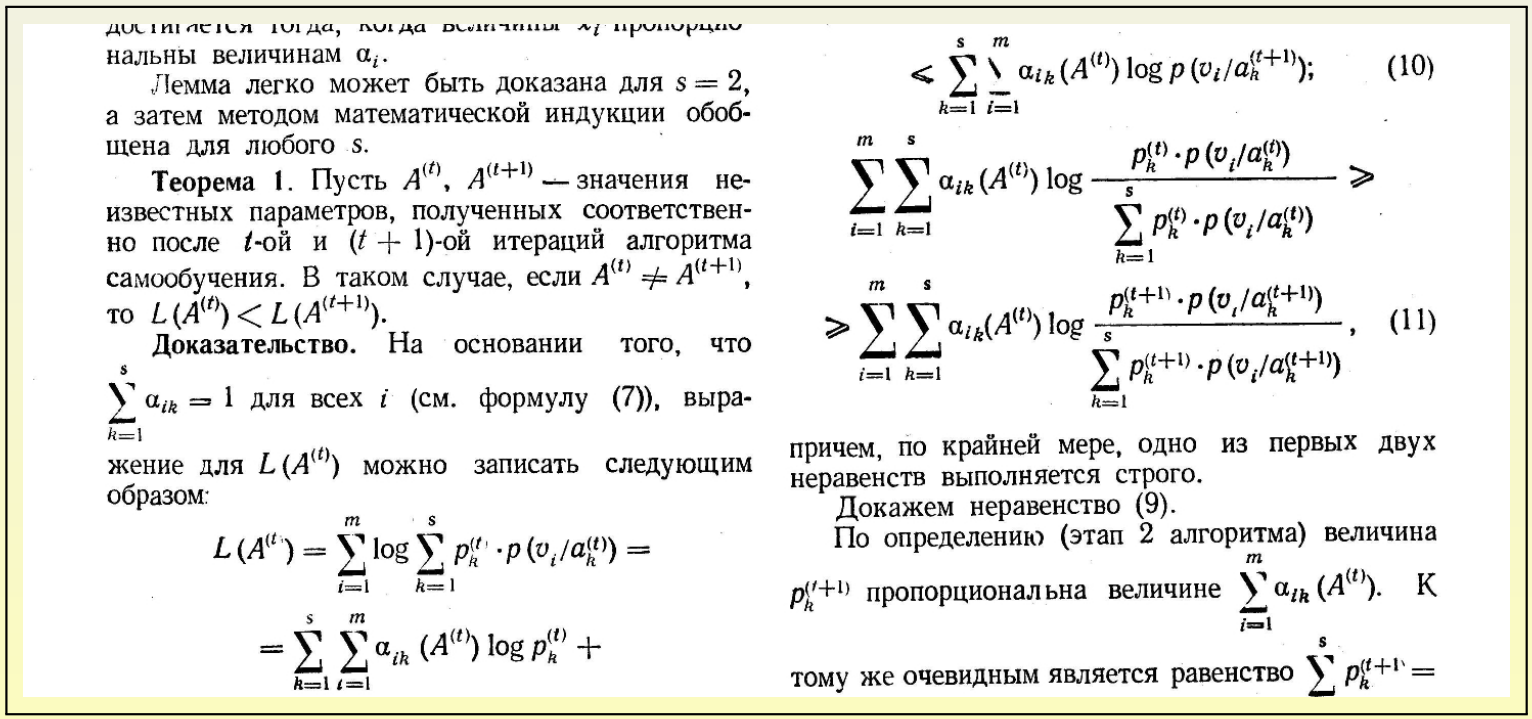
efektivní řešení problému odhadu směsí až s příchodem počítačů:

- **Kale (1962), Hasselblad (1966), Day (1969), Wolfe (1970):** *jednoduché iterační schema (EM algoritmus) původně odvozené intuitivně algebraickou úpravou věrohodnostních rovnic pro normální směsi, metoda snadno použitelná v mnoharozměrných případech, v každé iteraci zvyšuje hodnotu věrohodnostní funkce*
- **Hosmer (1973):** "Iterative m.-l. estimates were proposed by Hasselblad and subsequently have been looked at by Day, Hosmer and Wolfe"
- **Peters a Walker (1978):** "... we have observed in experiments that the convergence is monotone, i.e. that the likelihood function is actually increased in each iteration, but we have been unable to prove it."

Z historie problému odhadování směsí

první důkaz monotonie EM algoritmu:

- **Schlesinger M.I. (1968):** "Relation between learning and self learning in pattern recognition", **(in Russian)**, *Kibernetika*, (Kiev), No. 2, 81-88.



достигается тогда, когда величины x_i пропорциональны величинам α_i.

Лемма легко может быть доказана для $s = 2$, а затем методом математической индукции обобщена для любого s.

Теорема 1. Пусть $A^{(t)}$, $A^{(t+1)}$ — значения неизвестных параметров, полученных соответственно после t-ой и $(t+1)$-ой итераций алгоритма самообучения. В таком случае, если $A^{(t)} \neq A^{(t+1)}$, то $L(A^{(t)}) < L(A^{(t+1)})$.

Доказательство. На основании того, что $\sum_{k=1}^{s} \alpha_{ik} = 1$ для всех i (см. формулу (7)), выражение для $L(A^{(t)})$ можно записать следующим образом:

$$L(A^{(t)}) = \sum_{i=1}^{m} \log \sum_{k=1}^{s} p_k^{(t)} \cdot p(v_i/a_k^{(t)}) =$$

$$= \sum_{k=1}^{s} \sum_{i=1}^{m} \alpha_{ik}(A^{(t)}) \log p_k^{(t)} +$$

$$\leqslant \sum_{k=1}^{s} \sum_{i=1}^{m} \alpha_{ik}(A^{(t)}) \log p(v_i/a_k^{(t+1)}); \qquad (10)$$

$$\sum_{i=1}^{m} \sum_{k=1}^{s} \alpha_{ik}(A^{(t)}) \log \frac{p_k^{(t)} \cdot p(v_i/a_k^{(t)})}{\sum_{k=1}^{s} p_k^{(t)} \cdot p(v_i/a_k^{(t)})} \geqslant$$

$$\geqslant \sum_{i=1}^{m} \sum_{k=1}^{s} \alpha_{ik}(A^{(t)}) \log \frac{p_k^{(t+1)} \cdot p(v_i/a_k^{(t+1)})}{\sum_{k=1}^{s} p_k^{(t+1)} \cdot p(v_i/a_k^{(t+1)})}, \qquad (11)$$

причем, по крайней мере, одно из первых двух неравенств выполняется строго.

Докажем неравенство (9).

По определению (этап 2 алгоритма) величина $p_k^{(t+1)}$ пропорциональна величине $\sum_{i=1}^{m} \alpha_{ik}(A^{(t)})$. К тому же очевидным является равенство $\sum_{k=1}^{s} p_k^{(t+1)} =$

Foto

- **Ajvazjan et al. (monografie, 1974):** cituje Schlesingerův výsledek
- **Isaenko a Urbach (1976):** cituje Schlesingerův výsledek

Z historie problému odhadování směsí

název EM algoritmus pochází z často citované práce:

- **Dempster et al. (1977):** "Maximum likelihood from incomplete data via the EM algorithm." *J. Roy. Statist. Soc., B, Vol. 39, pp.l-38.*

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

Harvard University and Educational Testing Service

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

for all $p \geqslant p(\varepsilon)$ and all $r \geqslant 1$, where each term in the sum is non-negative.

Applying assumption (2) in the theorem for $p, p+1, p+2, \ldots, p+r-1$ and summing, we obtain from (3.12)

$$\varepsilon > \lambda \sum_{j=1}^{r} (\boldsymbol{\phi}^{(p+j)} - \boldsymbol{\phi}^{(p+j-1)})(\boldsymbol{\phi}^{(p+j)} - \boldsymbol{\phi}^{(p+j-1)})^{\mathrm{T}}, \tag{3.13}$$

whence

$$\varepsilon > \lambda (\boldsymbol{\phi}^{(p+r)} - \boldsymbol{\phi}^{(p)})(\boldsymbol{\phi}^{(p+r)} - \boldsymbol{\phi}^{(p)})^{\mathrm{T}}, \tag{3.14}$$

as required to prove convergence of $\boldsymbol{\phi}^{(p)}$ to some $\boldsymbol{\phi}^*$.

Theorem 1 implies that $L(\boldsymbol{\phi})$ is non-decreasing on each iteration of a GEM algorithm, and is strictly increasing on any iteration such that $Q(\boldsymbol{\phi}^{(p+1)} | \boldsymbol{\phi}^{(p)}) > Q(\boldsymbol{\phi}^{(p)} | \boldsymbol{\phi}^{(p)})$. The corollaries

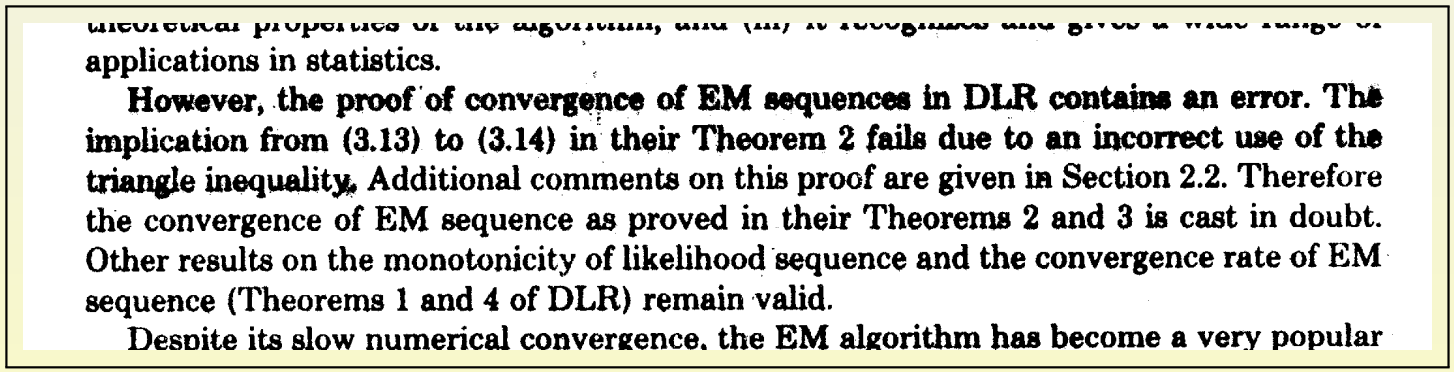
Scholar Google, září 2010: **22 390 citací článku Dempster et al.**

ÚTIA

Z historie problému odhadování směsí

chyba v důkazu konvergence posloupnosti parametrů:

- **Boyles R.A. (1983):** "On the convergence of the EM algorithm." *J. Roy. Statist. Soc., B, Vol. 45, pp. 47-50.*
- **Wu C.F.J. (1983):** "On the convergence properties of the EM algorithm." *Ann. Statist., Vol. 11, pp. 95-103.*



theoretical properties of the algorithm, and (iii) it recognizes and gives a wide range of applications in statistics.

However, the proof of convergence of EM sequences in DLR contains an error. The implication from (3.13) to (3.14) in their Theorem 2 fails due to an incorrect use of the triangle inequality. Additional comments on this proof are given in Section 2.2. Therefore the convergence of EM sequence as proved in their Theorems 2 and 3 is cast in doubt. Other results on the monotonicity of likelihood sequence and the convergence rate of EM sequence (Theorems 1 and 4 of DLR) remain valid.

Despite its slow numerical convergence, the EM algorithm has become a very popular

Monografie:

- **Everitt, B.S. and D.J. Hand (1981)**
- **Titterington et al. (1985)**
- **McLachlan and Peel (2000)**
- **Scholar Google (2010):** heslo "EM algorithm"- 2 340 000 výsledků

ÚTIA

SOUČINOVÉ SMĚSI

směs součinových komponent (model podmíněné nezávislosti):

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

Základní výhody použití součinových komponent

- zjednodušení výpočtu (odpadá např. inverze kovaričních matic)
- stabilita výpočtu (odpadá riziko špatně podmíněných matic)
- snadný výpočet marginálních rozložení pravděpodobnosti ▸ Odvození
- možnost odhadu parametrů směsi z neúplných datových vektorů

Nevýhody součinových směsí (+ mýty a pověry)

- skrytý předpoklad nezávislosti proměnných (?!) (platí pouze pro M=1)
- předpoklad součinových komponent je omezující (?) ▸ Jádrový (Parzenův) odhad
- vhodná normalizace dat před použitím součinové směsi (?) ▸ Invariance

Příklad EM algoritmu - součinová normální směs

KOMPONENTY: **normální hustoty s diagonální kovarianční maticí:**

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

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

iterační rovnice:

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

$$f'(m) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}), \qquad \mu'_{mn} = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{S}} x_n q(m|\mathbf{x})$$

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

$\Rightarrow$ **Jednodušší a stabilnější výpočet**

▸ Bernoulliovská směs ▸ Diskrétní součinová směs

Poznámky k implementaci EM algoritmu

- implementace EM algoritmu: jako cyklus přes data (pro $|\mathcal{S}| >> 1$)

$$\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \rightarrow f^{'}(m), \qquad \sum_{\mathbf{x}\in\mathcal{S}} x_n \; q(m|\mathbf{x}) \rightarrow \mu^{'}_{mn}, \theta^{'}_{mn}$$

- podmínka ukončení výpočtu: $(L^{'} - L)/L < \epsilon, \quad (\epsilon \approx 10^{-3} - 10^{-5})$
  (v konečné fázi výpočtu obvykle dochází k nadměrnému přizpůsobení odhadovaných parametrů k datům, tzv. "overpeaking")
- EM algoritmus spontánně potlačuje váhy přebytečných komponent, z rozložení vah lze posoudit potřebný počet komponent
- užitečný údaj:

$$q_{max}(\mathbf{x}) = \max_{m\in\mathcal{M}}\{q(m|\mathbf{x})\}, \qquad \bar{q}_{max} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} q_{max}(\mathbf{x})$$

- pro data s velkou dimenzí ($N >> 1$) je typický malý "překryv" komponent, tj. poměrně vysoké hodnoty $\bar{q}_{max} \approx 0.85 - 0.99$
- nutná podmínka správnosti programu: $L^{'} \geq L$

Ověření algoritmu: generování umělých dat + reidentifikace parametrů

ÚTIA

Implementace EM algoritmu v prostoru s velkou dimenzí

PROBLÉM: **numerická nestabilita v prostoru s velkou dimenzí**

- komponenty $F(\mathbf{x}|m)$ "podtékají" již při dimenzi $N \approx 30 - 40$
  $\Rightarrow$ výpočet komponent je třeba provést v logaritmickém tvaru:

$$\log[F(\mathbf{x}|m)f(m)] = \log f(m) + \sum_{n \in \mathcal{N}} \log f_n(x_n|m)$$

maximum: $\log C_0 = \max_m \{ \log[F(\mathbf{x}|m)f(m)] \} \Rightarrow C_0$

- "odlogaritmované" hodnoty $F(\mathbf{x}|m)$ a $P(\mathbf{x})$ nutné pro výpočet $q(m|\mathbf{x})$:

$$\exp\{-\log C_0 + \log f(m) + \sum_{n \in \mathcal{N}} \log f_n(x_n|m)\} = C_0^{-1} F(\mathbf{x}|m)f(m)$$

$$q(m|\mathbf{x}) = \frac{C_0^{-1} F(\mathbf{x}|m)f(m)}{\sum_{j=1}^{M} C_0^{-1} F(\mathbf{x}|j)f(j)} = \frac{F(\mathbf{x}|m)f(m)}{\sum_{j=1}^{M} F(\mathbf{x}|j)f(j)}$$

Příklady zdrojového kódu v C++: ▸ Bernoulliovská směs ▸ Normální součinová směs

ÚTIA

Modifikace EM algoritmu - neúplné datové vektory

neúplná data: $\mathbf{x} = (x_1, -, x_3, x_4, -, -, x_7, \ldots, x_N) \in \mathcal{X}$

$\mathcal{N}(\mathbf{x}) = \{n \in \mathcal{N} : \text{souřadnice } x_n \text{ je definovaná v } \mathbf{x}\}, \quad \mathbf{x} \in \mathcal{X}$

$\mathcal{S}_n = \{\mathbf{x} \in \mathcal{S} : n \in \mathcal{N}(\mathbf{x})\}, \quad \approx$ vektory $\mathbf{x} \in \mathcal{S}$ s definovanou souřadnicí x_n

předpoklad: součinové komponenty $\Rightarrow$ ▸ Snadný výpočet marginál

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log[\sum_{m=1}^{M} f(m) \bar{F}(\mathbf{x}|m)], \qquad \bar{F}(\mathbf{x}|m) = \prod_{n \in \mathcal{N}(x)} f_n(x_n|m)$$

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

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

$$f_n'(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}_n} q(m|\mathbf{x}) \log f_n(x_n|m) \right\}$$

POZN. Nahrazování chybějících údajů pomocí odhadů ovlivňuje data.

ÚTIA

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

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

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

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

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

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

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

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

POZN. Není nutná redukce dimenze, tj. výběr příznaků.

Strukturní modifikace EM algoritmu (diskrétní proměnné)

strukturní optimalizace je součástí odhadu parametrů směsi

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

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

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

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

strukturní optimalizace: $\phi'_{mn} = 1$ pro r nejvyšších hodnot γ'_{mn}

$$\gamma'_{mn} = \sum_{\mathbf{x} \in \mathcal{S}} \frac{q(m|\mathbf{x})}{|\mathcal{S}|} \log \left[\frac{f_n'(x_n|m)}{f_n(x_n|0)} \right] = f'(m) \sum_{x_n \in \mathcal{X}_n} f_n'(x_n|m) \log \frac{f_n'(x_n|m)}{f_n(x_n|0)}$$

▸ Odvození kritéria

Vlastnosti strukturního modelu směsi

strukturní model směsi realizuje "podprostorový" přístup:

- **mechanizmus:** málo informativní jednorozměrné distribuce $f_n(x_n|m)$ nahrazuje model příslušným fixním "pozadím" $f_n(x_n|0)$
- řeší obecný problém výběru příznaků individuálně pro každou komponentu - jako součást EM algoritmu
- umožňuje bayesovské rozhodování nezávisle na dimenzi prostoru (řešení rozhodovacích problémů bez redukce dimenze prostoru)
- snižuje počet parametrů modelu (i komponent) a tím omezuje riziko "nadměrného" přizpůsobení modelu datům (overpeaking)
- kritériem strukturní optimalizace odvozeným z EM algoritmu je Kullback-Leiblerova informační divergence (pro diskrétní směs)
- umožňuje statisticky korektní řešení problému neúplného propojení vstupních proměnných a neuronové vrstvy
- umožňuje strukturní optimalizaci neuronové sítě při zachování monotónní vlastnosti EM algoritmu: $L' - L \geq 0$

Sekvenční rozhodovací schema

Problém: postupné doplňování příznaků (př. lékařská diagnostika)

dané hodnoty: $\mathbf{x}_D = (x_{j_1}, \ldots, x_{j_l}) \in \mathcal{X}_D, \quad \mathcal{D} = \{j_1, \ldots, j_l\} \subset \mathcal{N}$

$$P(\mathbf{x}_D|\omega) = \sum_{m \in \mathcal{M}_\omega} f(m) \prod_{n \in D} f_n(x_n|m,\omega), \qquad P(\mathbf{x}_D) = \sum_{\omega \in \Omega} P(\mathbf{x}_D|\omega) p(\omega)$$

Optimální sekvenčního rozhodování: $p(\omega|x_n, \mathbf{x}_D),\ \omega \in \Omega$

Volba nejinformativnější proměnné $x_n,\ n \notin D$ při dané podmnožině známých vstupních údajů $\mathbf{x}_D = (x_{j_1}, \ldots, x_{j_l}) \in \mathcal{X}_D$ podle kriteria **maximální podmíněné informace:** $I_{\mathbf{x}_D}(\mathcal{X}_n, \Omega)$

$$I_{\mathbf{x}_D}(\mathcal{X}_n, \Omega) = H_{\mathbf{x}_D}(\mathcal{X}_n) - H_{\mathbf{x}_D}(\mathcal{X}_n|\Omega), \qquad n^* = \arg\max_{n \notin D} \{I_{\mathbf{x}_D}(\mathcal{X}_n, \Omega)\}$$

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

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

$$P_{n|D\omega}(x_n|\mathbf{x}_D, \omega) = P_{nD|\omega}(x_n, \mathbf{x}_D|\omega) / P_{D|\omega}(\mathbf{x}_D|\omega)$$

ÚTIA

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:

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

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

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

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

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

ÚTIA

VLASTNOSTI SOUČINOVÝCH SMĚSÍ

SOUHRN: výpočetní vlastnosti součinových distribučních směsí

- 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
- vhodné pro aproximaci multimodálních rozložení pravděpodobnosti
- možnost odhadu parametrů směsi z neúplných datových vektorů
- umožňují sekvenční rozhodování s postupným doplňováním nejinformativnějších příznaků
- existuje strukturní modifikace součinové směsi, která umožňuje "lokální" výběr příznaků a rozhodování v prostorech s velkou dimenzi
- součinové směsi lze interpretovat jako neuronovou síť

▸ Literatura

A1: Statistické řešení problému rozpoznávání

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

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\} \quad : \quad \text{konečný počet tříd}$$

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

$$\mathcal{S}_\omega = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K_\omega)}\} \quad : \quad \text{trénovací data}$$

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

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

$\Rightarrow$ **ŘEŠENÍ:** odhad neznámých distribucí $P(\mathbf{x}|\omega)$ na základě trénovacích datových souborů $\mathcal{S}_\omega, \omega \in \Omega$

◂ Zpět

ÚTIA

A2: Vlastnosti neparametrického jádrového odhadu

Theorem (Parzen, 1962; Cacoullos, 1966)

Nechť $\mathcal{S}_K$ je posloupnost K nezávislých realizací N-rozměrného náhodného vektoru s nějakou neznámou hustotou pravděpodobnosti $P^*(\mathbf{x})$. Neparametrický jádrový odhad $P(\mathbf{x})$ s vyhlazovacím parametrem σ_K

$$P(\mathbf{x}) = \frac{1}{K} \sum_{y \in \mathcal{S}_K} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left\{ \frac{(x_n - y_n)^2}{2\sigma_K^2} \right\}$$

je v každém bodě spojitosti $P^*(\mathbf{x})$ asymptoticky nestranný, tj. platí

$$\lim_{K \to \infty} \mathrm{E}_{\mathcal{S}_K}\{P(\mathbf{x})\} = P^*(\mathbf{x}),$$

pokud $\lim_{K \to \infty} \sigma_K = 0$. *Platí-li navíc* $\lim_{K \to \infty} K\sigma_K^N = \infty$, *potom $P(\mathbf{x})$ je také asymptoticky konzistentní v kvadratickém průměru:*

$$\lim_{K \to \infty} \mathrm{E}_{\mathcal{S}_K}\{[P^*(\mathbf{x}) - P(\mathbf{x})]^2\} = 0.$$

Výpočetní náročnost + problém vyhlazení

◂ Zpět: Kompromis ◂ Součinové směsi

A3: Optimalizace vyhlazení Parzenova odhadu

Parzenův odhad s normálním jádrem:

$$P(\mathbf{x}) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{y}\in\mathcal{S}} f(\mathbf{x}|\mathbf{y},\boldsymbol{\sigma}) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{y}\in\mathcal{S}}\left[\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left\{\frac{(x_n-y_n)^2}{2\sigma_n^2}\right\}\right]$$

optimalizace vyhlazení metodou cross-validace:

$\approx$ maximalizace upravené věrohodnostní funkce pomocí EM algoritmu:

$$L(\boldsymbol{\sigma}) = \sum_{\mathbf{x}\in\mathcal{S}}\log\left[\frac{1}{(|\mathcal{S}|-1)}\sum_{\mathbf{y}\in\mathcal{S},\mathbf{y}\neq\mathbf{x}}\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left\{\frac{(x_n-y_n)^2}{2\sigma_n^2}\right\}\right]$$

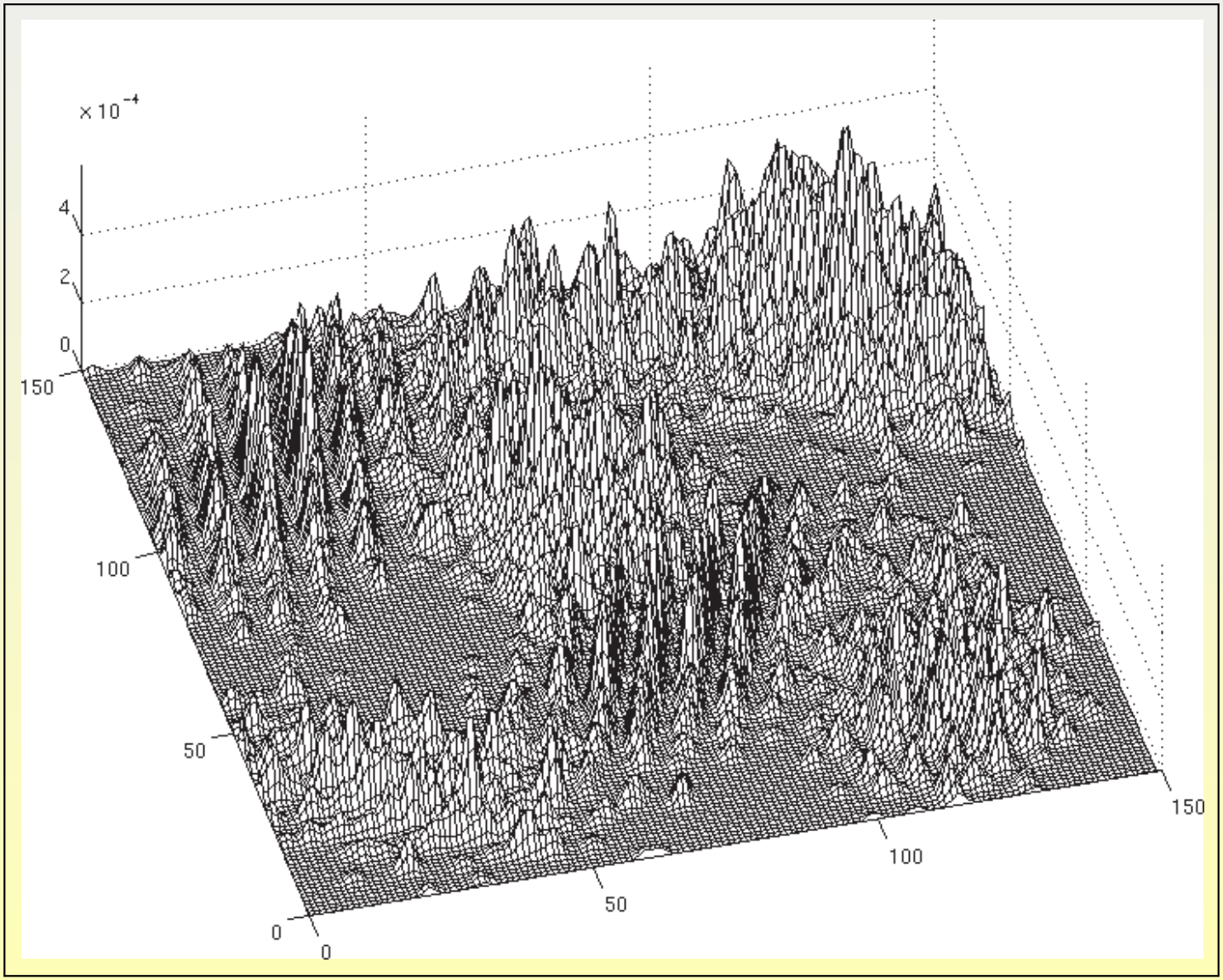
$$q(\mathbf{y}|\mathbf{x}) = \frac{f(\mathbf{x}|\mathbf{y},\boldsymbol{\sigma})}{\sum_{\mathbf{u}\in\mathcal{S},\mathbf{u}\neq\mathbf{x}} f(\mathbf{x}|\mathbf{u},\boldsymbol{\sigma})}, \quad \mathbf{y}\in\mathcal{S}$$

$$(\sigma_n')^2 = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\sum_{\mathbf{y}\in\mathcal{S},\mathbf{y}\neq\mathbf{x}}(x_n-y_n)^2 q(\mathbf{y}|\mathbf{x})$$

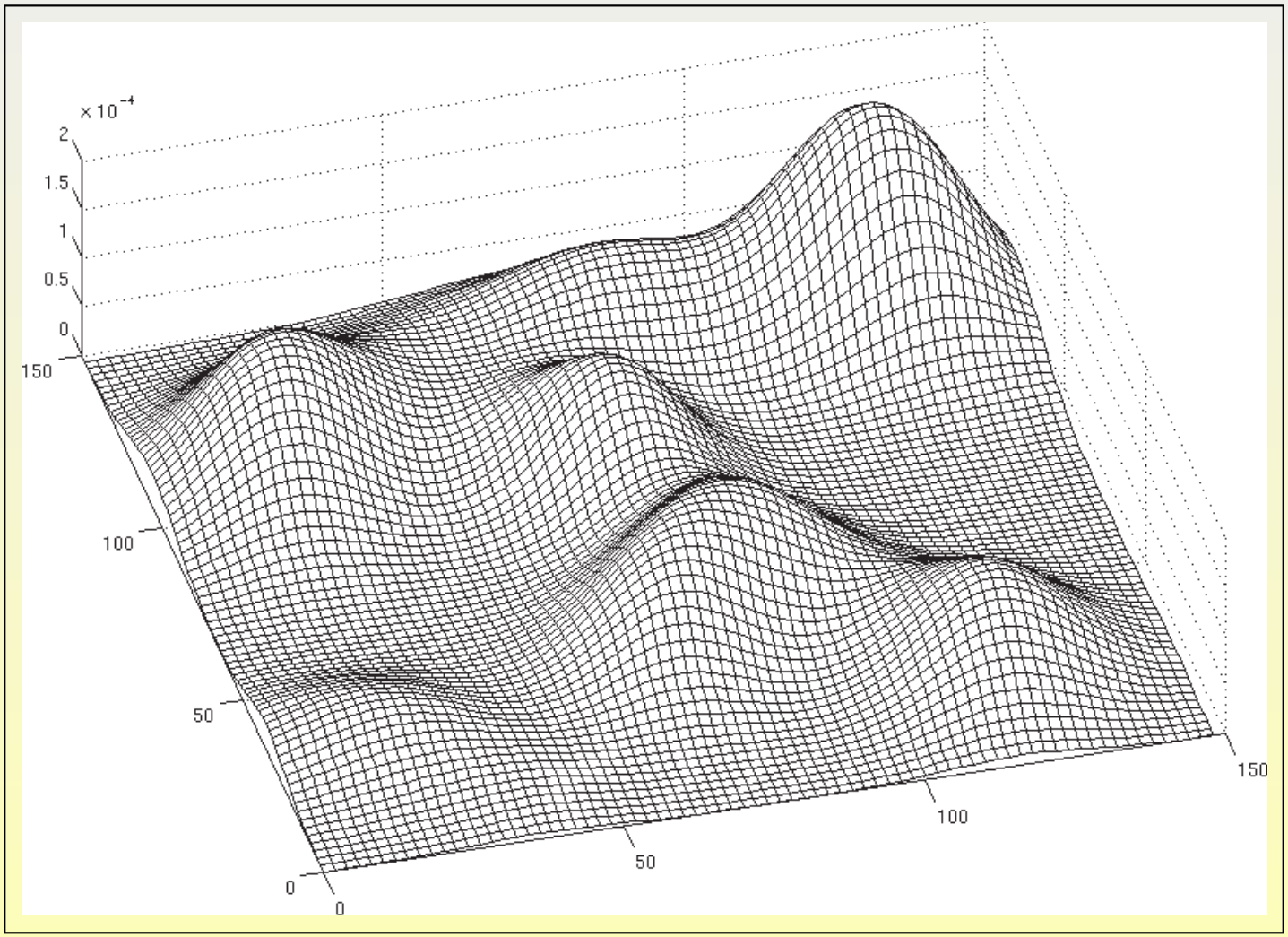
POZN. **Časově náročný výpočet!**

◂ Zpět: Kompromis ◂ Součinové směsi

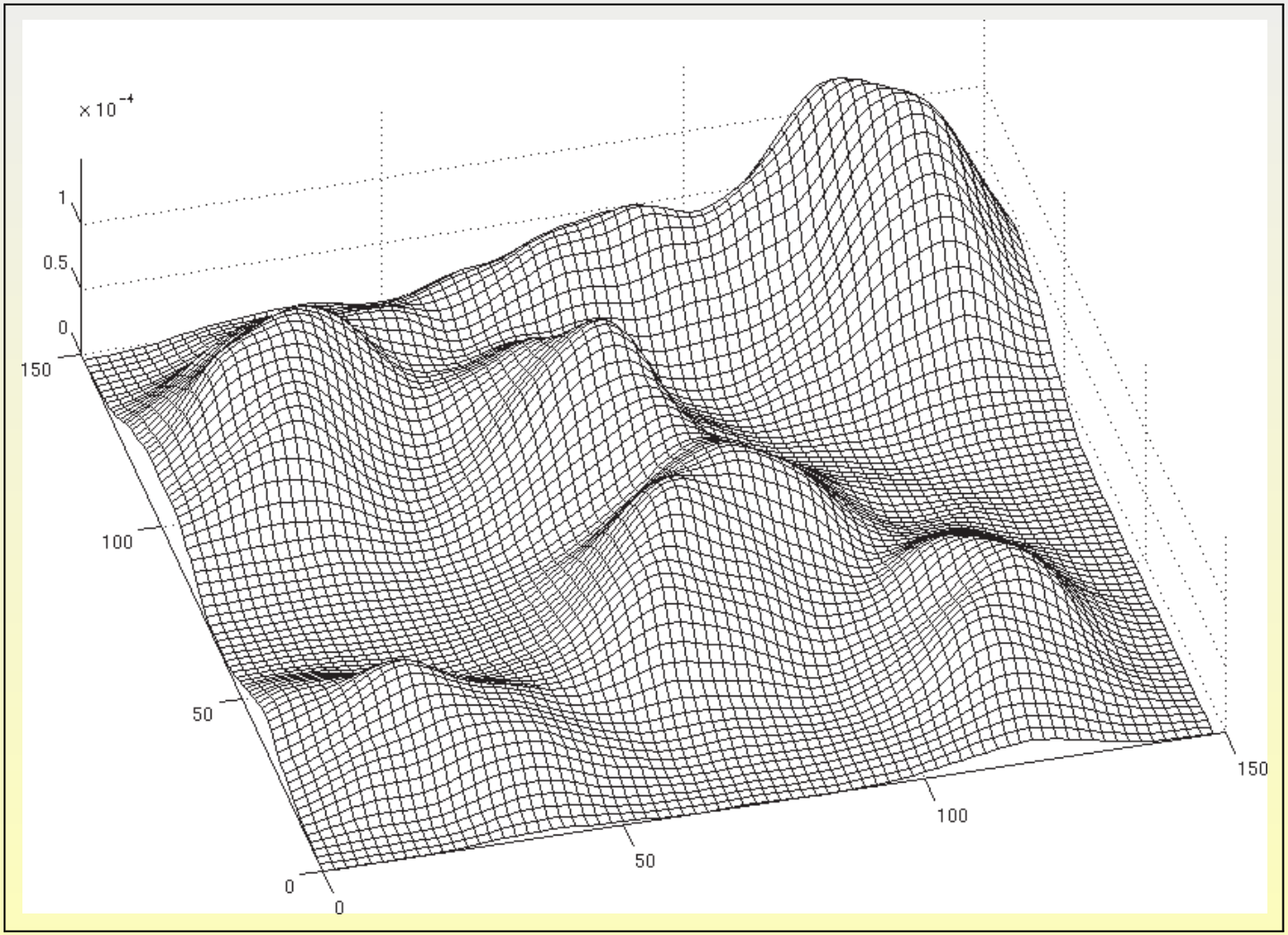
"Podhlazený "jádrový odhad



Příliš vyhlazený jádrový odhad



Optimálně vyhlazený jádrový odhad



(normální jádro s obecnou kov. maticí)

◂ Zpět: Norm. směs

A4: Odvození marginálních distribucí ze součinové směsi

$$P(\mathbf{x}) = \sum_{m=1}^{M} f(m) F(\mathbf{x}|m) = \sum_{m=1}^{M} f(m) \prod_{n=1}^{N} f_n(x_n|m), \qquad \mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}$$

$$\sum_{x_i \in \mathcal{X}_i} P(\mathbf{x}) = \sum_{m=1}^{M} f(m) \Big(\sum_{x_i \in \mathcal{X}_i} f_i(x_i|m) \Big) \prod_{n \in \mathcal{N} \setminus i} f_n(x_n|m) = \sum_{m=1}^{M} f(m) \prod_{n \in \mathcal{N} \setminus i} f_n(x_n|m)$$

$$\mathbf{x}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathcal{X}_C, \quad \mathcal{X}_C = \mathcal{X}_{i_1} \times \cdots \times \mathcal{X}_{i_k}, \quad C = \{i_1, \ldots, i_k\} \subset \mathcal{N}$$

$$P_C(\mathbf{x}_C) = \sum_{m=1}^{M} f(m) F_C(\mathbf{x}_C|m), \qquad F_C(\mathbf{x}_C|m) = \prod_{n \in C} f_n(x_n|m)$$

$$P_{n|C}(x_n|\mathbf{x}_C) = \frac{P_{nC}(x_n, \mathbf{x}_C)}{P_C(\mathbf{x}_C)} = \sum_{m=1}^{M} \frac{f(m) F_C(\mathbf{x}_C|m)}{P_C(\mathbf{x}_C)} f_n(x_n|m)$$

$$P_{n|C}(x_n|\mathbf{x}_C) = \sum_{m=1}^{M} W_m(\mathbf{x}_C) f_n(x_n|m), \qquad W_m(\mathbf{x}_C) = \frac{f(m) F_C(\mathbf{x}_C|m)}{P_C(\mathbf{x}_C)}$$

◂ Zpět - Nekompletní data ◂ Zpět - Výhody součinových směsí

A5: Invariance vůči lineární transformaci proměnných

Invariance součinové směsi vůči lineární transformaci

Nechť parametry normální součinové směsi $\{w_m, \mu_{mn}, \sigma_{mn}, m \in \mathcal{M}, n \in \mathcal{N}\}$ jsou stacionárním bodem EM algoritmu, tj. splňují iterační rovnice. Dále nechť $\mathbf{y} = T(\mathbf{x})$ je lineární transformace dat $\mathbf{x} \in \mathcal{X}$ a parametrů směsi:

$$y_n = a_n x_n + b_n, \ \mathbf{x} \in \mathcal{S}, \quad \tilde{w}_m = w_m, \quad \tilde{\mu}_{mn} = a_n \mu_{mn} + b_n, \quad \tilde{\sigma}_{mn} = a_n \sigma_{mn}.$$

Potom transformované parametry $\{\tilde{w}_m, \tilde{\mu}_{mn}, \tilde{\sigma}_{mn}, m \in \mathcal{M}, n \in \mathcal{N}\}$ jsou rovněž stacionárním bodem EM algoritmu v transformovaném prostoru $\mathcal{Y}$.

Důkaz: Substitucí lze ověřit platnost rovnic: Zpět

$$F(\mathbf{y}|\tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{\sigma}}_m) = \frac{1}{\prod_{n=1}^{N} a_n} F(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m), \quad \tilde{P}(\mathbf{y}) = \frac{1}{\prod_{n=1}^{N} a_n} P(\mathbf{x})$$

$$q(m|\mathbf{y}) = q(m|\mathbf{x}), \quad \mathbf{y} = T(\mathbf{x}), \quad \mathbf{x} \in \mathcal{S}, \quad m \in \mathcal{M}$$

$$\tilde{\mu}_{mn} = \frac{1}{\tilde{w}_m |\mathcal{S}|} \sum_{\mathbf{y} \in \tilde{\mathcal{S}}} y_n q(m|\mathbf{y}), \quad (\tilde{\sigma}_{mn})^2 = \frac{1}{\tilde{w}_m |\mathcal{S}|} \sum_{\mathbf{y} \in \tilde{\mathcal{S}}} (y_n - \tilde{\mu}_{mn})^2 q(m|\mathbf{y})$$

A6: Explicitní řešení kroku M

Lemma (podrobněji viz Grim, 1982)

Nechť maximálně věrohodný odhad parametru $\mathbf{b}$ *hustoty pravděpodobnosti* $F(\mathbf{x}|\mathbf{b})$ *je aditivní funkcí dat* $\mathbf{x} \in \mathcal{S}$*:*

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log F(\mathbf{x}|\mathbf{b}), \quad \mathbf{x} \in \mathcal{X}, \qquad \mathbf{b} \approx \textit{parametr}$$

$$\mathbf{b}^* = \arg\max_{\mathbf{b}} \left\{ \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log F(\mathbf{x}|\mathbf{b}) \right\} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \mathbf{a}(\mathbf{x})$$

Jestliže $\gamma(\mathbf{x}) = N(\mathbf{x})/|\mathcal{S}|$ *je relativní četnost vektoru* $\mathbf{x}$ *v* $\mathcal{S}$*, platí ekvivalentně:*

$$L = \sum_{\mathbf{x} \in \bar{\mathcal{X}}} \gamma(\mathbf{x}) \log F(\mathbf{x}|\mathbf{b}), \qquad \bar{\mathcal{X}} = \{\mathbf{x} \in \mathcal{X} : \gamma(\mathbf{x}) > 0\}, \qquad (\sum_{\mathbf{x} \in \bar{\mathcal{X}}} \gamma(\mathbf{x}) = 1)$$

$$\mathbf{b}^* = \sum_{\mathbf{x} \in \bar{\mathcal{X}}} \gamma(\mathbf{x})\ \mathbf{a}(\mathbf{x}) = \arg\max_{\mathbf{b}} \left\{ \sum_{\mathbf{x} \in \bar{\mathcal{X}}} \gamma(\mathbf{x}) \log F(\mathbf{x}|\mathbf{b}) \right\}$$

Důsledek: Maximum vážené věrohodnostní funkce lze vyjádřit jako vážený maximálně věrohodný odhad.

Explicitní řešení kroku M - normální směs

normální směs s obecnou kovarianční maticí:

$$P(\mathbf{x}) = \sum_{m=1}^{M} f(m) F(\mathbf{x}|\mathbf{c}_m, A_m)$$

$$F(\mathbf{x}|\mathbf{c}_m, A_m) = \frac{1}{\sqrt{(2\pi)^N \det A_m}} \exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_m)^T A_m^{-1}(\mathbf{x}-\mathbf{c}_m)\}$$

implicitní tvar kroku M:

$$(\mathbf{c}_m^{'}, A_m^{'}) = \arg \max_{(\mathbf{c}_m, A_m)} \left\{ \sum_{\mathbf{x}\in\mathcal{S}} \frac{q(m|\mathbf{x})}{\sum_{\mathbf{y}\in\mathcal{S}} q(m|\mathbf{y})} \log F(\mathbf{x}|\mathbf{c}_m, A_m) \right\}$$

explicitní řešení:

$$\mathbf{c}_m^{'} = \sum_{\mathbf{x}\in\mathcal{S}} \frac{q(m|\mathbf{x})}{\sum_{\mathbf{y}\in\mathcal{S}} q(m|\mathbf{y})}\ \mathbf{x}, \qquad \left(\gamma(\mathbf{x}) = \frac{q(m|\mathbf{x})}{\sum_{\mathbf{y}\in\mathcal{S}} q(m|\mathbf{y})}\right)$$

$$A_m^{'} = \sum_{\mathbf{x}\in\mathcal{S}} \frac{q(m|\mathbf{x})}{\sum_{\mathbf{y}\in\mathcal{S}} q(m|\mathbf{y})} (\mathbf{x}-\mathbf{c}_m^{'})(\mathbf{x}-\mathbf{c}_m^{'})^T$$

◂ Zpět: EM algoritmus

A7: Příklad EM algoritmu - diskrétní součinová směs

$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$
$x_n \in \mathcal{X}_n, \ |\mathcal{X}_n| < \infty \ \approx$ diskrétní proměnné s konečným počtem hodnot
(např. dotazníky, kvalitativní data, popisy herních situací)

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \left[\sum_{m=1}^{M} f(m) F(\mathbf{x}|m) \right], \quad F(\mathbf{x}|m) = \prod_{n=1}^{N} f_n(x_n|m)$$

iterační rovnice: $(\mathbf{x} \in \mathcal{S}, \ \mathcal{S} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\})$

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

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

▸ Podrobnější odvození

POZN. 1 Diskrétní součinová směs není identifikovatelná. ▸ Důkaz
(problém při shlukové analýze × výhoda při aproximaci) ◂ Zpět: Příklad EM

POZN. 2 Každé diskrétní rozložení psti ▸ lze zapsat jako součinovou směs.

Odvození M-kroku pro diskrétní součinovou směs

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log f_n(x_n|m) \right\}, \quad n \in \mathcal{N}, \quad m \in \mathcal{M}$$

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} \left(\sum_{\xi \in \mathcal{X}_n} \delta(\xi, x_n) \right) q(m|\mathbf{x}) \log f_n(x_n|m) \right\}$$

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\xi \in \mathcal{X}_n} \sum_{\mathbf{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\mathbf{x}) \log f_n(\xi|m) \right\}$$

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\xi \in \mathcal{X}_n} \left(\sum_{\mathbf{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\mathbf{x}) \right) \log f_n(\xi|m) \right\}$$

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \sum_{\xi \in \mathcal{X}_n} \left(\frac{\sum_{\mathbf{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\mathbf{x})}{\sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x})} \right) \log f_n(\xi|m) \right\}$$

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

◂ Zpět: EM algoritmus

A8: Příklad EM algoritmu - směs Bernoulliho rozložení

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

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

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

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

iterační rovnice:

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

$$f'(m) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}), \qquad \theta'_{mn} = \frac{1}{\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{S}} x_n q(m|\mathbf{x})$$

POZN. Problém přesnosti součinů velkého počtu parametrů θ_{mn}.

◂ Zpět

A9: Důkaz neidentifikovatelnosti diskrétní součinové směsi

Definice identifikovatelnosti směsi (Teicher, 1963)

Třída směsí $\mathcal{P} = \{P(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ je identifikovatelná, jestliže parametry $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ libovolných dvou ekvivalentních směsí

$$P(\mathbf{x}, \boldsymbol{\theta}) = P(\mathbf{x}, \boldsymbol{\theta}'), \quad \forall\, \mathbf{x} \in \mathcal{X}$$

se mohou lišit pouze pořadím komponent.

◂ Zpět: identifikace x aproximace

Theorem (srv. Teicher, 1963, 1968; Gyllenberg et al., 1994; Grim, 2001)

Libovolná diskrétní součinová směs $(x_n \in \mathcal{X}_n, \quad |\mathcal{X}_n| < \infty)$

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

může být ekvivalentně vyjádřena nekonečně mnoha různými způsoby (tj. pomocí různých množin parametrů), pokud alespoň jedna z podmíněných distribucí $f_i(x_i|m)$ *splňuje podmínku*

◂ Zpět: Diskrétní směs

$$0 < f_i(x_i|m) < 1, \quad \textit{pro nějaké} \quad x_i \in \mathcal{X}_i.$$

Důkaz neidentifikovatelnosti diskrétní součinové směsi

Důkaz:

Jestliže pro nějaké $i \in \mathcal{N}, x_i \in \mathcal{X}_i$ a $m \in \mathcal{M}$ platí $0 < f_i(x_i|m) < 1$ potom jednorozměrné rozložení pravděpodobnosti $f_i(\cdot|m)$ můžeme nekonečně mnoha způsoby vyjádřit jako konvexní kombinaci dvou různých rozložení $f_i^{'}(\cdot|m), f_i^{''}(\cdot|m)$, např. $(0 < \alpha < 1,\ \beta = 1 - \alpha)$:

$$f_i(\xi|m) = \alpha f_i^{'}(\xi|m) + \beta f_i^{''}(\xi|m), \quad \xi \in \mathcal{X}_i$$

S využitím předchozí substituce můžeme napsat

$$f(m)F(\mathbf{x}|m) = f^{'}(m)F^{'}(\mathbf{x}|m) + f^{''}(m)F^{''}(\mathbf{x}|m)$$

kde

$$f^{'}(m) = \alpha f(m), \quad f^{''}(m) = \beta f(m)$$

$$F^{'}(\mathbf{x}|m) = f^{'}(x_i|m) \prod_{n \in \mathcal{N}, n \neq i} f_n(x_n|m), \quad F^{''}(\mathbf{x}|m) = f^{''}(x_i|m) \prod_{n \in \mathcal{N}, n \neq i} f_n(x_n|m)$$

a po dosazení za komponentu $f(m)F(\mathbf{x}|m)$ můžeme původní směs $P(\mathbf{x})$ vyjádřit pomocí netriviálně odlišných parametrů, cbd.

◂ Zpět: EM algoritmus

A10: Modifikace EM algoritmu - vážená data

$\gamma(\mathbf{x}) > 0$: relativní četnost výskytu vektoru $\mathbf{x}$ ($\approx$ váha) v posloupnosti $\mathcal{S}$

$\mathcal{M} = \{1, \ldots, M\}, \quad \mathcal{N} = \{1, \ldots, N\} \approx$ indexové množiny

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log[\sum_{m \in \mathcal{M}} f(m) F(\mathbf{x}|m)] = \sum_{\mathbf{x} \in \mathcal{X}} \gamma(\mathbf{x}) \log[\sum_{m \in \mathcal{M}} f(m) F(\mathbf{x}|m)]$$

$\bar{\mathcal{S}} = \{\mathbf{x} \in \mathcal{S} : \gamma(\mathbf{x}) > 0\}$: sčítání lze omezit na vektory $\mathbf{x} \in \bar{\mathcal{S}}$:

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

$$q(m|\mathbf{x}) = \frac{f(m) F(\mathbf{x}|m)}{\sum_{j \in \mathcal{M}} f(j) F(\mathbf{x}|j)}, \qquad F(\mathbf{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m)$$

$$f'(m) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) = \sum_{\mathbf{x} \in \bar{\mathcal{S}}} \gamma(\mathbf{x}) q(m|\mathbf{x})$$

$$f_n'(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \bar{\mathcal{S}}} \gamma(\mathbf{x}) q(m|\mathbf{x}) \log f_n(x_n|m) \right\}$$

POUŽITÍ: agregace dat, "nekonečná" data: $\gamma(\mathbf{x}) = P^*(\mathbf{x})$

◂ Zpět: Vlastnosti

ÚTIA

A11: Alternativní důkaz monotónie EM algoritmu: $L' \geq L$

Kullback-Leiblerova informační divergence je nezáporná, tj. platí:

$$I(q(\cdot|\mathbf{x}), q'(\cdot|\mathbf{x})) = \sum_{m=1}^{M} q(m|\mathbf{x}) \log \frac{q(m|\mathbf{x})}{q'(m|\mathbf{x})} \geq 0,$$

▸ Důkaz

---

následující postup vychází z původního Schlesingerova důkazu

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log[\sum_{m=1}^{M} f(m) F(\mathbf{x}|m)], \qquad q(m|\mathbf{x}) = \frac{f(m) F(\mathbf{x}|m)}{\sum_{j=1}^{M} f(j) F(\mathbf{x}|j)}$$

Věrohodnostní funkci L resp. L' lze zapsat ekvivalentně pomocí $q(m|\mathbf{x})$:

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \Big\{ \sum_{m=1}^{M} q(m|\mathbf{x}) \log[f(m) F(\mathbf{x}|m)] - \sum_{m=1}^{M} q(m|\mathbf{x}) \log q(m|\mathbf{x}) \Big\}$$

$$L' = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \Big\{ \sum_{m=1}^{M} q(m|\mathbf{x}) \log[f'(m) F'(\mathbf{x}|m)] - \sum_{m=1}^{M} q(m|\mathbf{x}) \log q'(m|\mathbf{x}) \Big\}$$

Alternativní důkaz monotónie EM algoritmu: $L' \geq L$

Předchozí vzorce použijeme k vyjádření přírůstku věrohodnostní funkce:

$$L' - L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left\{ \sum_{m=1}^{M} q(m|\mathbf{x}) \log \left[\frac{f'(m) F'(\mathbf{x}|m)}{f(m) F(\mathbf{x}|m)} \right] + \sum_{m=1}^{M} q(m|\mathbf{x}) \log \frac{q(m|\mathbf{x})}{q'(m|\mathbf{x})} \right\}$$

Druhý člen na pravé straně představuje Kullback-Leiblerovu divergenci:

$$L' - L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left\{ \sum_{m=1}^{M} q(m|\mathbf{x}) \log \left[\frac{f'(m) F'(\mathbf{x}|m)}{f(m) F(\mathbf{x}|m)} \right] + I(q(\cdot|\mathbf{x}), q'(\cdot|\mathbf{x})) \right\}$$

Po vynechání nezáporné informační divergence dostaneme nerovnost

$$L' - L \geq \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \left\{ \sum_{m=1}^{M} q(m|\mathbf{x}) \log \left[\frac{f'(m) F'(\mathbf{x}|m)}{f(m) F(\mathbf{x}|m)} \right] \right\}$$

$$L' - L \geq \sum_{m=1}^{M} \left[\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \right] \log \frac{f'(m)}{f(m)} + \frac{1}{|\mathcal{S}|} \sum_{m=1}^{M} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log \frac{F'(\mathbf{x}|m)}{F(\mathbf{x}|m)}$$

Alternativní důkaz monotónie EM algoritmu: $L' \geq L$

S využitím substituce za $f'(m)$ podle kroku M obdržíme nerovnost

$$\sum_{m=1}^{M} \left[\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \right] \log \frac{f'(m)}{f(m)} = \sum_{m=1}^{M} f'(m) \log \frac{f'(m)}{f(m)} \geq 0$$

Podle definice v kroku M: $F'(.|m) = \arg \max_{F(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log F(\mathbf{x}|m) \right\}$

tzn. pro libovolnou funkci $F(\mathbf{x}|m)$ platí nerovnost:

$$(*) \quad \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log F'(\mathbf{x}|m) \geq \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log F(\mathbf{x}|m), \quad m \in \mathcal{M}$$

Z uvedených nerovností plyne monotónní vlastnost EM algoritmu:

$$L' - L \geq \sum_{m=1}^{M} f'(m) \log \frac{f'(m)}{f(m)} + \frac{1}{|\mathcal{S}|} \sum_{m=1}^{M} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log \frac{F'(\mathbf{x}|m)}{F(\mathbf{x}|m)} \geq 0$$

POZN. Definice M-kroku je zbytečně silná, stačí aby nové parametry splňovaly nerovnost (*) $\Rightarrow$ GEM algoritmus

Zpět

A12: Důsledky monotónní vlastnosti EM algoritmu

Neklesající shora omezená posloupnost hodnot kritéria $\{L^{(t)}\}_{t=0}^{\infty}$ má konečnou limitu $L^* < \infty$ a proto splňuje nutnou podmínku konvergence:

$$\lim_{t\to\infty} L^{(t)} = L^* < \infty \quad \Rightarrow \quad \lim_{t\to\infty} (L^{(t+1)} - L^{(t)}) = 0$$

Stejnou podmínku splňují i posloupnosti $\{f^{(t)}(\cdot)\}_{t=0}^{\infty}$, $\{q^{(t)}(\cdot|\mathbf{x})\}_{t=0}^{\infty}$:

$$\lim_{t\to\infty} ||f^{(t+1)}(\cdot) - f^{(t)}(\cdot)|| = 0, \qquad \lim_{t\to\infty} ||q^{(t+1)}(\cdot|\mathbf{x}) - q^{(t)}(\cdot|\mathbf{x})|| = 0.$$

Předchozí limity plynou z nerovnosti

$$L^{(t+1)} - L^{(t)} \geq I(f^{(t+1)}(\cdot)||f^{(t)}(\cdot)) + \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} I(q^{(t)}(\cdot|\mathbf{x})||q^{(t+1)}(\cdot|\mathbf{x}))$$

s použitím následující obecné nerovnosti (viz Kullback (1966)):

$$\sum_{\mathbf{x}\in\mathcal{X}} P^*(\mathbf{x}) \log \frac{P^*(\mathbf{x})}{P(\mathbf{x})} \geq \frac{1}{4}\Big(\sum_{\mathbf{x}\in\mathcal{X}} |P^*(\mathbf{x}) - P(\mathbf{x})|\Big)^2 \geq \frac{1}{4}\|P^*(\cdot) - P(\cdot)\|^2$$

A13: Maximálně věrohodné odhady a problém aproximace

Lemma

Maximalizace věrohodnostní funkce je asymptoticky ekvivalentní minimalizaci horní meze euklidovské vzdálenosti mezi skutečnou diskrétní distribucí P^* a její aproximací P.

Důkaz: Asymptoticky, pro $|\mathcal{S}| \to \infty$, platí

$$\lim_{|\mathcal{S}|\to\infty} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}\in\mathcal{S}} \log P(\mathbf{x}) = \lim_{|\mathcal{S}|\to\infty} \sum_{\mathbf{x}\in\mathcal{S}} \gamma(\mathbf{x}) \log P(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}} P^*(\mathbf{x}) \log P(\mathbf{x})$$

kde $\gamma(\mathbf{x}) \geq 0$ je relativní četnost výskytu diskrétního vektoru $\mathbf{x}$ v posloupnosti $\mathcal{S}$ a P^* je skutečné rozložení pravděpodobnosti. Tvrzení věty plyne z nerovnosti (viz Kullback, 1966):

$$\sum_{\mathbf{x}\in\mathcal{X}} P^*(\mathbf{x}) \log \frac{P^*(\mathbf{x})}{P(\mathbf{x})} \geq \frac{1}{4} \Big(\sum_{\mathbf{x}\in\mathcal{X}} |P^*(\mathbf{x}) - P(\mathbf{x})| \Big)^2 \geq \frac{1}{4} \|P^*(\cdot) - P(\cdot)\|^2$$

◂ Zpět

A14: Důkaz nezápornosti Kullback-Leiblerovy divergence

Theorem (viz např. Vajda, 1992)

Pro libovolná dvě diskrétní rozložení pravděpodobnosti $\{q_1, q_2, \ldots, q_M\}$, $\{q_1', q_2', \ldots, q_M'\}$ *platí nerovnost*

$$I(\mathbf{q} \| \mathbf{q}') = \sum_{m=1}^{M} q_m \log \frac{q_m}{q_m'} \geq 0$$

přičemž rovnost nastane právě tehdy, je-li $q_m' = q_m$, *pro všechna* $m \in \mathcal{M}$.

Důkaz: Bez ztráty obecnosti můžeme předpokládat $q_m > 0$ pro všechna $m \in \mathcal{M}$ (protože $0 \log 0 = 0$). Podle Jensenovy nerovnosti platí:

$$\sum_{m=1}^{M} q_m \log \frac{q_m'}{q_m} \leq \log \left(\sum_{m=1}^{M} q_m \frac{q_m'}{q_m} \right) = \log \left(\sum_{m=1}^{M} q_m' \right) = \log 1 = 0,$$

přičemž rovnost nastane právě tehdy, je-li $q_1'/q_1 = \cdots = q_M'/q_M$, cbd.

Důsledek: následující suma na levé straně je maximální pro $\mathbf{q}' = \mathbf{q}$

$$\sum_{m=1}^{M} q_m \log q_m' \leq \sum_{m=1}^{M} q_m \log q_m$$

◂ Zpět - Důkaz ◂ Zpět (alternativní důkaz) ◂ Zpět (M-krok)

A15: Diskrétní součinová směs univerzálně aproximuje

Lemma (viz např. Grim, 2006)

Nechť $p^{(k)}, k = 1, \ldots, K,\ K = |\mathcal{X}|$ *jsou tabulkou definované hodnoty libovolného diskrétního rozložení pravděpodobnosti* $P(\mathbf{x})$ *na prostoru* $\mathcal{X}$*:*

$$P(\mathbf{x}^{(k)}) = p^{(k)}, \quad \mathbf{x}^{(k)} \in \mathcal{X}, \quad k = 1, \ldots, K, \quad \mathcal{X} = \cup_{k=1}^{K} \{\mathbf{x}^{(k)}\}$$

Potom diskrétní rozložení pravděpodobnosti $P(\mathbf{x})$ *může být vyjádřeno ve tvaru součinové distribuční směsi*

$$P(\mathbf{x}) = \sum_{k=1}^{K} w_k F(\mathbf{x}|k) = \sum_{k=1}^{K} p^{(k)} \prod_{n \in \mathcal{N}} \delta(x_n, x_n^{(k)}), \quad \mathbf{x} \in \mathcal{X}.$$

Důkaz: Je zřejmý z uvedeného vzorce, kde komponenty směsi definované pomocí delta-funkcí jsou umístěny v jednotlivých bodech prostoru $\mathbf{x}^{(k)} \in \mathcal{X}$ a váha komponenty je rovna příslušné tabulkové hodnotě $p^{(k)}$:

$$F(\mathbf{x}|k) = \prod_{n \in \mathcal{N}} \delta(x_n, x_n^{(k)}), \quad w_k = p^{(k)}, \quad k = 1, \ldots, K.$$

POZN. Uvedený konstruktivní důkaz má pouze formální význam, aproximace s využitím EM algoritmu je numericky výhodnější.

◀ Zpět - ("lze zapsat")

A16: Odvození kritéria strukturní optimalizace

v implicitní rovnici kroku M můžeme vynechat konstantní "pozadí" $F(\mathbf{x}|0)$:

$$G^{'}(.|m,\phi_m^{'}) = \arg \max_{G(.|m,\phi_m)} \left\{ \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\mathbf{x}) \log G(\mathbf{x}|m,\phi_m) \right\}$$

po dosazení za $G(\mathbf{x}|m,\phi_m)$ upravíme výraz v závorce:

$$\sum_{\mathbf{x} \in \mathcal{S}} \frac{q(m|\mathbf{x})}{|\mathcal{S}|} \log \prod_{n \in \mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)} \right]^{\phi_{mn}} = \sum_{n \in \mathcal{N}} \phi_{mn} \sum_{\mathbf{x} \in \mathcal{S}} \frac{q(m|\mathbf{x})}{|\mathcal{S}|} \log \left[\frac{f_n(x_n|m)}{f_n(x_n|0)} \right]$$

$\Rightarrow$ výchozí implicitní vztah lze rozepsat odděleně pro jednorozměrné distribuce $f_n^{'}(x_n|m)$ a strukturní parametry $\phi_{mn}^{'}$:

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \left\{ \sum_{\mathbf{x} \in \mathcal{S}} \frac{q(m|\mathbf{x})}{|\mathcal{S}|} \log f_n(x_n|m) \right\}$$

$$\boldsymbol{\phi}_m^{'} = \arg \max_{\boldsymbol{\phi}_m} \left\{ \phi_{mn} \sum_{\mathbf{x} \in \mathcal{S}} \frac{q(m|\mathbf{x})}{|\mathcal{S}|} \log \left[\frac{f_n^{'}(x_n|m)}{f_n(x_n|0)} \right] \right\} = \arg \max_{\boldsymbol{\phi}_m} \left\{ \phi_{mn} \gamma_{mn}^{'} \right\}$$

Zpět

A17: 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");
```

◀ Zpět

A18: 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
```

Prof. M.I. Schlesinger se svou ženou

Při výletu na Karlštejn během pobytu v Praze v roce 1995.

◂ Zpět

Literatura 1/10

Ajvazjan S.A., Bezhaeva Z.I., Staroverov O.V. (1974): *Classification of Multivariate Observations*, (in Russian). Moscow: Statistika.

Boyles R.A. (1983): On the convergence of the EM algorithm. *J. Roy. Statist. Soc., B*, Vol. 45, pp. 47-50.

Cacoullos I. (1966): Estimation of a multivariate density. *Ann. Inst. Stat. Math.*, Vol. 18, pp. 179-190.

Carreira-Perpignan M.A., Renals S. (2000): Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, Vol. 12, pp. 141-152.

Day N.E. (1969): Estimating the components of a mixture of normal distributions. *Biometrika*, Vol. 56, pp. 463-474.

Dempster A.P., Laird N.M. and Rubin D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., B*, Vol. 39, pp.l-38.

◂ Zpět

Literatura 2/10


Duda R.O., Hart P.E. (1973): Pattern Classification and Scene Analysis. New York: Wiley-Interscience.


Everitt, B.S. and D.J. Hand (1981): *Finite Mixture Distributions*. Chapman & Hall: London, 1981.


Grim J. (1982): On numerical evaluation of maximum - likelihood estimates for finite mixtures of distributions. *Kybernetika*, Vol.l8, No.3, pp.173-190. *http://www.kybernetika.cz/content.html*


Grim J. (1982): Design and optimization of multilevel homogeneous structures for multivariate pattern recognition. In *Fourth FORMATOR Symposium 1982*, Academia, Prague 1982, pp. 233-240.


Grim J. (1986): Multivariate statistical pattern recognition with nonreduced dimensionality, *Kybernetika*, Vol. 22, pp. 142-157. *http://www.kybernetika.cz/content.html*


Grim J. (1994): Knowledge representation and uncertainty processing in the probabilistic expert system PES, *International Journal of General Systems*, Vol. 22, No. 2, p. 103 - 111.

Literatura 3/10

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

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

Grim J., (1996): Maximum Likelihood Design of Layered Neural Networks. In: *Proceedings of the 13th International Conference on Pattern Recognition* **IV** (pp. 85-89), Los Alamitos: IEEE Computer Society Press.

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* (pp. 977-982), Roma: Edizzioni Kappa.

◂ Zpět

Literatura 4/10

Grim J. (1998): A sequential modification of EM algorithm. In *Studies in Classification, Data Analysis and Knowledge Organization*, Gaul W., Locarek-Junge H., (Eds.), pp. 163 - 170, Springer, 1999.

Grim J., Somol P., Novovičová J., Pudil P., Ferri F., (1998b): Initializing normal mixture of densities. In *Proc. 14th Int. Conf. on Pattern Recognition ICPR'98*, A.K. Jain, S. Venkatesh, B.C. Lovell (Eds.), pp. 886-890, IEEE Computer Society: Los Alamitos, California, 1998

Grim J. (1999): Information approach to structural optimization of probabilistic neural networks. In proceedings of: 4th System Science European Congress, L. Ferrer et al. (Eds.), (pp: 527-540), Valencia: Sociedad Espanola de Sistemas Generales, 1999.

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, pp. 849-856, European Communities 2001.

◂ Zpět

Literatura 5/10


Grim J. (2001): Latent Structure Analysis for Categorical Data. Research Report No. 2019. ÚTIA AV ČR, Praha 2001, 13 pp.


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


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


Grim J., Just P., Pudil P. (2003): Strictly modular probabilistic neural networks for pattern recognition. *Neural Network World*, Vol. 13 , No. 6, pp. 599-615.


Grim J., Kittler J., Pudil P., Somol P. (2000): Combining multiple classifiers in probabilistic neural networks, In *Multiple Classifier Systems*, Eds. Kittler J., Roli F., Springer, 2000, pp. 157 - 166.

◂ Zpět

Literatura 6/10

Grim J., Kittler J., Pudil P., Somol P. (2001): Information analysis of multiple classifier fusion. In: *Multiple Classifier Systems 2001*, Kittler J., Roli F., (Eds.), Lecture Notes in computer Science, Vol. 2096, Springer-Verlag, Berlin, Heidelberg, New York 2001, pp. 168 - 177.

Grim J., Kittler J., Pudil P., Somol P., (2002): Multiple classifier fusion in probabilistic neural networks. *Pattern Analysis & Applications* Vol. 5, No. 7, pp. 221-233.

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.

Grim J., Somol P., Novovičová J., Pudil P., Ferri F. (1998): Initializing normal mixtures of densities. In *Proceedings of the 14th International Conference on Pattern Recognition ICPR'98*, Brisbane, August 16 - 20, 1998, Eds. A.K. Jain, S. Venkatesh, B.C. Lovell, pp. 886-890, IEEE Computer Society: Los Alamitos, California, 1998

◂ Zpět

Literatura 7/10


Grim J., Somol P., Pudil P., Just P. (2003): Probabilistic neural network playing a simple game. In *Artificial Neural Networks in Pattern Recognition.* (Marinai S., Gori M. Eds.). University of Florence, Florence 2003, pp. 132-138.


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


Grim J.: EM cluster analysis for categorical data. In: *Structural, Syntactic and Statistical Pattern Recognition.* (Yeung D. Y., Kwok J. T., Fred A. eds.), (LNCS 4109). Springer, Berlin 2006, pp. 640-648.


Gyllenberg M., T. Koski, E. Reilink, M. Verlaan (1994): Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, Vol. 31, pp. 542–548.


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, pp. 177-180.

◂ Zpět

Literatura 8/10

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, pp. 95-100.

Hasselblad V. (1966): Estimation of prameters for a mixture of normal distributions. *Technometrics*, Vol. 8, pp. 431-444.

Hasselblad V. (1969): Estimation of finite mixtures of distributions from the exponential family. *Journal of Amer. Statist. Assoc.*, Vol. 58, pp. 1459-1471.

Isaenko O.K., Urbakh K.I. (1976): Decomposition of probability distribution mixtures into their components (in Russian). In: *Theory of probability, mathematical statistics and theoretical cybernetics*, Vol. 13, Moscow: VINITI.

Kullback S. (1966): An information-theoretic derivation of certain limit relations for a stationary Markov Chain. *SIAM J. Control*, Vol. 4, No. 3, pp. 454-459.

◂ Zpět

Literatura 9/10

McLachlan G.J. and Peel D. (2000): *Finite Mixture Models*, John Wiley & Sons, New York, Toronto, (2000)

Parzen E. (1962): On estimation of a probability density function and its mode. *Annals of Mathematical Statistics*, Vol. 33., pp. 1065-1076.

Pearson C. (1894): Contributions to the mathematical theory of evolution. 1. Dissection of frequency curves." *Philosophical Transactions of the Royal Society of London* **185**, 71-110.

Peters B.C., Walker H.F. (1978): An iterative procedure for obtaining maximumlikelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal Appl. Math.*, Vol. 35, No. 2, pp. 362-378.

Schlesinger M.I. (1968): Relation between learning and self learning in pattern recognition (in Russian), *Kibernetika*, (Kiev), No. 2, 81-88.

Teicher H. (1963): Identifiability of finite mixtures. *Ann. Math. Statist.*, Vol. 34, pp. 1265-1269.

◂ Zpět

Literatura 10/10

Teicher H. (1968): Identifiability of mixtures of product measures. *Ann. Math. Statist.*, Vol. 39, pp. 1300-1302.

Titterington D.M., Smith A.F.M. and Makov U.E. (1985): *Statistical analysis of finite mixture distributions*, John Wiley & Sons: Chichester, New York.

Vajda I., Grim J. (1998): About the maximum information and maximum likelihood principles in neural networks, *Kybernetika*, Vol. 34, No. 4, pp. 485-494.

Wolfe J.H. (1970): Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, Vol. 5, pp. 329-350.

Wu C.F.J. (1983): On the convergence properties of the EM algorithm. *Ann. Statist.*, Vol. 11, pp. 95-103.

Xu L. and Jordan M.I. (1996): On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, Vol. 8. pp. 129-151.

◂ Zpět