Distribution Mixtures of Product Components

Part I: EM Algorithm & Modifications

Jiří Grim

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

January 2017

Available at: http://www.utia.cas.cz/people/grim

Cf. J. Grim: "Approximation of Unknown Multivariate Probability Distributions by Using Mixtures of Product Components: A Tutorial." *International Journal of Pattern Recognition and Artificial Intelligence,* (2017).
DOI: 10.1142/S0218001417500288

Outline

Method of Distribution Mixtures

Information Source:

training data $\mathcal{S}$: independent observations of a random vector identically distributed (i.i.d.) according to an unknown probability distribution $P^*(\boldsymbol{x})$

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

Principle of the Method of Mixtures:

approximation of unknown multidimensional multimodal distribution $P^*(\boldsymbol{x})$ by means of a linear combination of component distributions $F(\boldsymbol{x}|m)$

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m), \quad \sum_{m \in \mathcal{M}} w_m = 1, \quad \sum_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}|m) = 1 \ \left(= \int_{\mathcal{X}} F(\boldsymbol{x}|m) d\boldsymbol{x} \right)$$

Application examples:

pattern recognition, image analysis, prediction problems, texture modeling, statistical models, classification of text documents, ...

Mixtures as a "Semiparametric" Model

parametric approach: e.g. assuming multivariate normal density

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

mean: $\boldsymbol{c} = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} \boldsymbol{x}$, covariance matrix: $A = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}}(\boldsymbol{x}-\boldsymbol{c})(\boldsymbol{x}-\boldsymbol{c})^T$

---

nonparametric approach: general kernel estimate Theorem (Parzen, 1962)

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

problem: optimal smoothing (choice of the smoothing parameters σ_n)

Mixtures as a Compromise: Semiparametric Multimodal Model

- not so limiting as parametric models
- almost as general as nonparametric model, without smoothing
- efficient estimation of parameters by EM algorithm

Example - EM algorithm for mixtures of Gaussian densities

computation of parameter estimates from data: $\mathcal{S} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}$

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

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log P(\boldsymbol{x}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} F(\boldsymbol{x}|\boldsymbol{c}_m, A_m) w_m \right]$$

Iteration equations: $\approx$ to maximize log-likelihood function

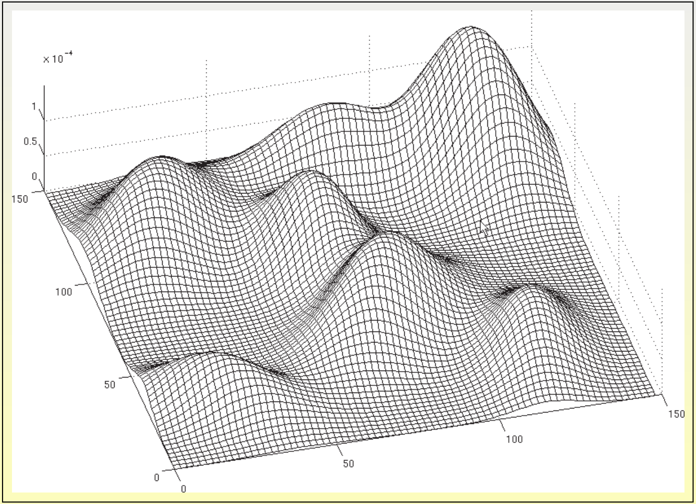
E-step: $$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|\boldsymbol{c}_m, A_m)}{\sum_{j=1}^{M} w_j F(\boldsymbol{x}|\boldsymbol{c}_j, A_j)}, \;\; \boldsymbol{x} \in \mathcal{S}, \quad m = 1, 2, \ldots, M$$

M-Step: $$w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}), \qquad \boldsymbol{c}_m^{'} = \frac{1}{\sum_{x \in \mathcal{S}} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}} \boldsymbol{x} \; q(m|\boldsymbol{x})$$

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

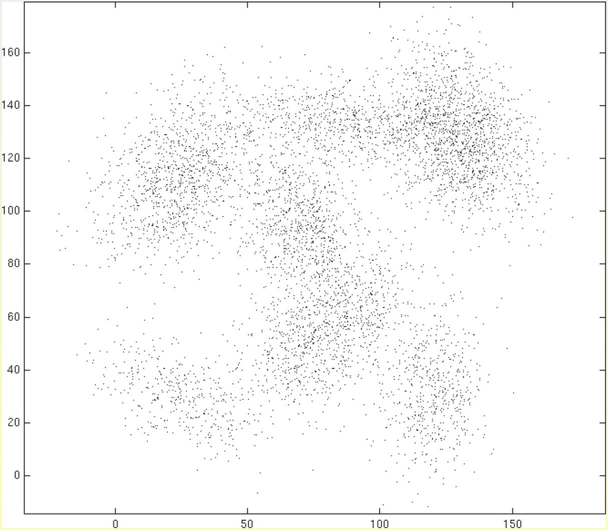
Remark: **The number of components has to be given.**

Example: reconstruction of a Gaussian mixture from data



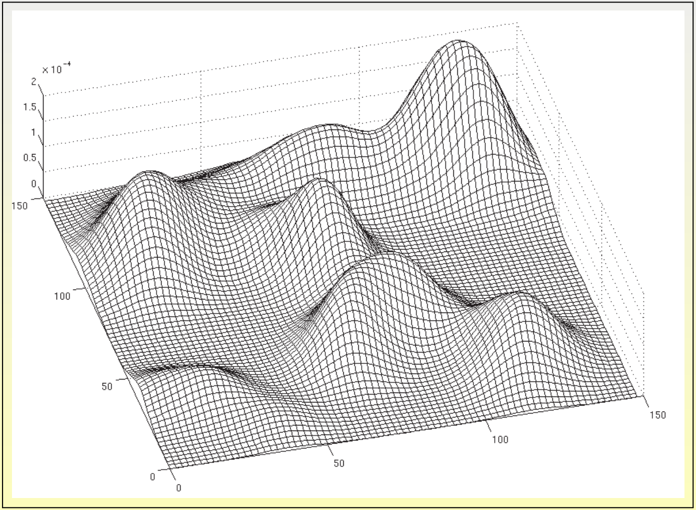
dimension of data: $N = 2$, number of mixture components: $M = 7$

Random sampling from a Gaussian mixture (M=7)



6000 data points (test of the correct implementation of EM algorithm)

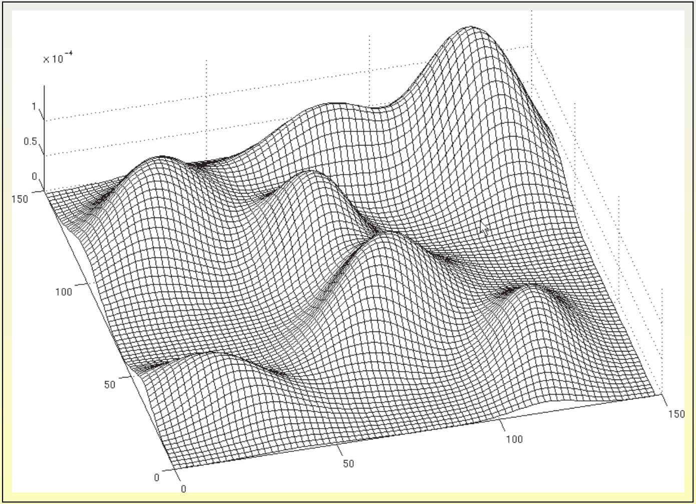
Example of the mixture estimate (M=28)



number of mixture components $M = 28$ ($\neq 7$)

(COMPARISON: kernel estimate)

Original mixture of Gaussian densities (M=7)



dimension of data $N = 2$, number of mixture components $M = 7$

General Version of EM Algorithm

EM algorithm: to maximize log-likelihood function

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

Iteration Equations: $\quad (m = 1, 2, \ldots, M, \quad \boldsymbol{x}\in\mathcal{S}, \quad \mathcal{S} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\})$

E-step:
$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j=1}^{M} w_j F(\boldsymbol{x}|j)}, \qquad w_m^{'} = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x})$$

M-Step:
$$F^{'}(.|m) = \arg\max_{F(.|m)} \left\{\frac{1}{\sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x})}\sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) \log F(\boldsymbol{x}|m)\right\}$$

---

for product components: $F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m), \ \mathcal{N} = \{1, 2, \ldots, N\}$

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

Remark: **Only inequality is sufficient in the M-Step instead of maximum $\Rightarrow$ generalized EM (GEM) algorithm.**

Explicit Solution of the M-Step (Grim,1982)

Let $F(\boldsymbol{x}|\boldsymbol{b}), \boldsymbol{x} \in \mathcal{X}$ be a probability density function and
let $\boldsymbol{b}^*$ be the maximum-likelihood estimate of the parameter $\boldsymbol{b}$:

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

Further let $\boldsymbol{b}^*$ be an additive function of the data vectors $\boldsymbol{x} \in \mathcal{S}$:

$$\boldsymbol{b}^* = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \boldsymbol{a}(\boldsymbol{x}).$$

Denoting $\gamma(\boldsymbol{x}) = N(\boldsymbol{x})/|\mathcal{S}|$ the relative frequency of $\boldsymbol{x}$ in $\mathcal{S}$ we can write:

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

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

Consequence: **Weighted likelihood function is maximized by the weighted analogy of the related m.-l. estimate.**

▸ Example: Gaussian mixture

Monotonic Property of EM Algorithm (Schlesinger, 1968)

The sequence of log-likelihood values $\{L^{(t)}\}_{t=0}^{\infty}$ is non-decreasing:

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

and, if bounded above, converges to a local or global maximum (or a saddle-point) of the log-likelihood function:

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

The existence of a finite limit $L^* < \infty$ implies the related necessary conditions: Proof

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

$$\Rightarrow \quad \lim_{t \to \infty} |w^{(t+1)}(m) - w^{(t)}(m)| = 0, m \in \mathcal{M}, \quad \lim_{t \to \infty} ||q^{(t+1)}(\cdot|\boldsymbol{x}) - q^{(t)}(\cdot|\boldsymbol{x})|| = 0$$

Remark: **The convergence of the sequence $\{L^{(t)}\}_{t=0}^{\infty}$ does not imply the convergence of the corresponding parameter estimates!**

Proof of the Monotonic Property of EM Algorithm

Lemma

Kullback-Leibler information divergence $I(q(\cdot|\mathbf{x})||q'(\cdot|\mathbf{x}))$ is non-negative for any two distributions $q(\cdot|\mathbf{x}), q'(\cdot|\mathbf{x})$ and it is zero if and only if the two distributions are identical. Proof

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

Substitution for $q(m|\mathbf{x}), q'(m|\mathbf{x})$ from the **E-Step** implies the inequality:

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

where the first term is equal to the increment of the criterion L:

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

Proof of the Monotonic Property of EM Algorithm

Making substitution from the last equation we obtain:

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

and by using substitution from the **M-Step**

$$(**) \qquad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}), \qquad m = 1, 2, \ldots, M$$

we can write the inequality:

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

Consequently, the first sum on the right-hand side of the inequality (*) is non-negative.

Remark: The definition (**) of the weights $w_m^{'}$ maximizes the first sum in Eq. (***).

Proof of the Monotonic Property of EM Algorithm

In view of the **M-Step** definition, the function $F^{'}(\cdot|m)$ maximizes the left-hand side, i.e. we can write:

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

The last inequality can be rewritten in the form

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

i.e. the increment of the log-likelihood function L is non-negative:

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

$$\Rightarrow \quad L^{'} \geq L$$

Alternative proof

Remark: **Any statistical interpretation of the proof is unnecessary!**

ÚTIA

Mixture Identification × Approximating by Mixtures

Problem of mixture identification (e.g. cluster analysis)

- **GOAL: to identify the true number of components and to estimate the true mixture parameters**
- the estimated mixture must be identifiable Definition
- PROBLEM: the log-likelihood function has local maxima nearly always (especially in case of small data sets in high dimensional spaces)
- $\Rightarrow$ the resulting local maximum is starting-point dependent
- PROBLEM: the mixture estimate is strongly influenced by the chosen number of components and by the initial parameters

Problem of approximating unknown probability distributions

- **GOAL: precise approximation of the unknown probability distribution by using mixture distributions** Approximation Problem × MLE
- the approximating mixture need not be identifiable
- the exact number of components is irrelevant
- the approximating mixture can be initialized randomly

Computational properties of EM Algorithm

real-life approximation problems $\Rightarrow$

large data sets + large number of components:

- in case of large mixtures ($M \approx 10^1 - 10^2$) the low-weight components may be neglected ($\Rightarrow$ the exact number of components is irrelevant)
- the existence of local log-likelihood maxima of large mixtures is less relevant because the related maximum values are comparable
- $\Rightarrow$ the influence of initial parameters is less relevant, the mixtures can be initialized randomly
- the EM iterations can be stopped e.g. by a relative increment threshold because of limited influence on the achieved log-likelihood value
- a reasonable stopping rule may decrease the risk of overfitting (excessive adaptation to training data)
- the EM algorithm is applicable to weighted data

Remark: **The computational properties are data-dependent and therefore not generally valid.**

From the History of the Mixture Estimation Problem

Computation of m.-l. estimates of mixture parameters by setting partial derivatives to zero cannot be solved analytically. **SOLUTION?**

- **First paper:** **Pearson (1894): "Contributions to the mathematical theory of evolution. 1. Dissection of frequency curves." Philosophical Trans. of the Royal Society of London 185**, 71-110.
  Subject: *mixture of two univariate Gaussian densities estimated by the method of moments. (about 80 papers in the years 1895-1965)*

efficient estimation of mixtures was enabled only by computers:

- **Hasselblad (1966), Day (1969), Wolfe (1970):** *derived simple iteration scheme by algebraic rearrangement of the likelihood equations (at present known as EM algorithm) which was converging and easily applicable to large mixtures in multidimensional spaces*
- **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."

From the History of the Mixture Estimation Problem

the first proof of the monotonic property of EM algorithm:

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

достигается тогда, когда величины x_i пропорцио-нальны 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 $\sum^{s} p_k^{(t+1)} =$

- **Ajvazjan et al. (1974, in Russian):** cite Schlesinger (1968)
- **Isaenko & Urbach (1976, in Russian):** cite Schlesinger (1968)

ÚTIA

From the History of the Mixture Estimation Problem

the standard reference to EM algorithm:

- **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]

- **Dempster et al.** introduced the name EM algorithm and described its wide application possibilities (main subject: problem of incomplete data)
- **Google Scholar (2017): 48 500 citations of the above paper ("all time top 10" in statistics)**
- : the term "EM algorithm" used in 340 000 papers
- : the terms "EM algorithm & mixture" used in 103 000 papers

From the History of the Mixture Estimation Problem

erroneous proof of the convergence of parameter estimates:
(does not concern the monotonic property of EM algorithm)

- **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

Monographs on Mixtures:

- **Titterington et al. (1985):** *Statistical analysis of finite mixture distributions*, John Wiley & Sons: Chichester, New York.
- **McLachlan and Peel (2000):** *Finite Mixture Models*, John Wiley & Sons, New York, Toronto.

PRODUCT MIXTURES

mixtures of product components (conditional independence model):

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}} f_n(x_n|m), \quad \boldsymbol{x} \in \mathcal{X}$$

Examples:

Gaussian mixtures with diagonal covariance matrices (real variables)
mixtures of multivariate Bernoulli distributions (binary variables)

ADVANTAGES:

- **do not imply the assumption of independence of variables**
- $\Rightarrow$ **do not imply the "naive Bayes" assumption**
- the mixture parameters can be efficiently estimated by EM algorithm
- any discrete distribution can be expressed as product mixture Proof
- Gaussian product mixtures approach the asymptotic accuracy of non-parametric Parzen estimates for $M >> 1$ Parzen estimates
- no risk of ill-conditioned covariance matrices in Gaussian components
- marginal distributions: by omitting superfluous terms in the products
- any conditional distributions easily computed
- product mixtures support the subspace (structural) modification

EM Estimation of Gaussian Product Mixtures

COMPONENTS: Gaussian densities with diagonal covariance matrices

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

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

EM iteration equations: $(m\in\mathcal{M},\ n\in\mathcal{N})$ Unnecessary norming of variables

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

$$w_m^{'} = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}), \qquad \mu_{mn}^{'} = \frac{1}{w_m^{'}|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} x_n q(m|\boldsymbol{x})$$

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

no matrix inversion $\Rightarrow$ no risk of ill-conditioned matrices

EM Estimation of Discrete Product Mixtures

COMPONENTS: **products of univariate discrete distributions**

$$F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m), \quad \boldsymbol{x} = (x_1,\ldots,x_N) \in \mathcal{X},\ x_n \in \mathcal{X}_n,\ |\mathcal{X}_n| < \infty$$

$$L = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} \log P(\boldsymbol{x}) = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} \log\left[\sum_{m\in\mathcal{M}} w_m \prod_{n\in\mathcal{N}} f_n(x_n|m)\right], \quad \boldsymbol{x}\in\mathcal{X}$$

EM iteration equations: $(\boldsymbol{x}\in\mathcal{S}, \quad \mathcal{S} = \{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(K)}\})$

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

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

More details:

Remark 1 Discrete product mixture is not identifiable. Proof
($\Rightarrow$ problem in cluster analysis $\times$ advantage in approximation)

Remark 2 Any discrete distribution is representable as a product mixture.

EM Estimation of Multivariate Bernoulli Mixtures

COMPONENTS: products of univariate Bernoulli distributions

binary data: numerals on a binary raster, results of biochemical tests ...

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

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

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

EM iteration equations:

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

$$w_m' = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} q(m|\boldsymbol{x}), \qquad \theta_{mn}' = \frac{1}{w_m'|\mathcal{S}|}\sum_{x\in\mathcal{S}} x_n q(m|\boldsymbol{x})$$

Remark: Product of a large number of parameters θ_{mn} may underflow.

Implementation Comments on EM Algorithm

- implementation of EM algorithm as a data cycle (for $|\mathcal{S}| >> 1$)

$$\sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) \to w_m^{'}, \qquad \sum_{\boldsymbol{x}\in\mathcal{S}} x_n\, q(m|\boldsymbol{x}) \to \mu_{mn}^{'}, \theta_{mn}^{'}$$

- basic condition to verify the correct implementation: $L^{'} \geq L$
- relative increment threshold ϵ to stop iterations: $(L^{'} - L)/L < \epsilon, \quad (\epsilon \approx 10^{-3} - 10^{-5})$
- ϵ is useful to avoid "overpeaking" in final stages of convergence
- EM algorithm suppresses the weights of "superfluous" components (large number of low-weight components $\Rightarrow$ to many components M)
- global information about overlapping components:

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

- in multi-dimensional spaces ($N >> 1$) the criterion $\bar{q}_{max}$ is usually high ($\approx 0.85 \div 0.99$) $\Rightarrow$ the overlap of components is small

Remark: Correct implementation of EM algorithm can be reliably verified by re-identification of mixture parameters from large artificial data.

Implementation of EM Algorithm in High Dimensions

PROBLEM: **numerical instability of the E-step**

- the components $F(\boldsymbol{x}|m)$ may "underflow" at dimensions $N \approx 30 - 40$
- $\Rightarrow$ the "lost" values cannot be "recovered" by norming in Eq. for $q(m|\boldsymbol{x})$
- $\Rightarrow$ inaccurate evaluation of the conditional weights $q(m|\boldsymbol{x})$

SOLUTION:

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

maximum component: $\log C(\boldsymbol{x}) = \max_m\{\ \log[F(\boldsymbol{x}|m)w_m]\ \}$

NORMING of $F(\boldsymbol{x}|m)$ a $P(\boldsymbol{x})$ for evaluation of $q(m|\boldsymbol{x})$:

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

$$q(m|\boldsymbol{x}) = \frac{C(\boldsymbol{x})^{-1}F(\boldsymbol{x}|m)w_m}{\sum_{j=1}^{M} C(\boldsymbol{x})^{-1}F(\boldsymbol{x}|j)w_j} = \frac{F(\boldsymbol{x}|m)w_m}{\sum_{j=1}^{M} F(\boldsymbol{x}|j)w_j}$$

Examples of C-pseudocode: Bernoulli Mixture Gaussian Mixture

Structural Mixture Model (Grim et al. 1986, 1999, 2002)

binary structural parameters: $\phi_m = (\phi_{m1}, \ldots, \phi_{mN}) \in \{0,1\}^N$

$$F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}},$$

$f_n(x_n|0)$: **fixed "background" distributions**, usually $f_n(x_n|0) = P_n^*(x_n)$
$\phi_{mn} = 0 \;\Rightarrow\;$ $f_n(x_n|m)$ is replaced by $f_n(x_n|0)$

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

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

"the background distribution" $F(\boldsymbol{x}|0)$ **reduces in the Bayes formula:**

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

MOTIVATION: Local, component-specific feature selection, "dimensionless" computation, structural neural networks.

Structural Modification of EM Algorithm

structural optimization can be included into EM algorithm:

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} F(\boldsymbol{x}|0) G(\boldsymbol{x}|m, \phi_m) w_m \right]$$

EM iteration equations: $(m \in \mathcal{M}, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S})$

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

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

structural optimization:

$\phi_{mn}^{'} = 1$ for a fixed number R of largest values of the criterion $\gamma_{mn}^{'}$:

$$\gamma_{mn}^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}) \log \left[\frac{f_n^{'}(x_n|m)}{f_n(x_n|0)} \right]$$

Proof

Remark: The background distribution $F(\boldsymbol{x}|0)$ can be included into optimization too (Grim, 1999).

ÚTIA

Structural EM Algorithm - Discrete Mixture

$f_n(x_n|m),\ x_n \in \mathcal{X}_n,\ n \in \mathcal{N}\ \approx$ **discrete probability distributions**

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

EM iteration equations: $(m \in \mathcal{M}, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S})$

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

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

Details

structural optimization: $\phi_{mn}^{'} = 1$ for the R largest values $\gamma_{mn}^{'}$:

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

Proof

Remark: The last sum is the Kullback-Leibler information divergence.

Structural EM Algorithm - Gaussian Mixture

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

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

EM iteration equations: $(m\in\mathcal{M}, n\in\mathcal{N}, \boldsymbol{x}\in\mathcal{S})$

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

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

structural optimization: $\phi_{mn}^{'} = 1$ for the R largest values $\gamma_{mn}^{'}$:

$$\gamma_{mn}^{'} = \frac{w_m^{'}}{2}\left[\frac{(\mu_{mn}^{'}-\mu_{0n})^2}{(\sigma_{0n})^2} + \frac{(\sigma_{mn}^{'})^2}{(\sigma_{0n})^2} - \log\frac{(\sigma_{mn}^{'})^2}{(\sigma_{0n})^2} - 1\right] = w_m^{'} I(f_n^{'}(\cdot|m), f_n(\cdot|0))$$

Remark: $\gamma_{mn}^{'}$ is the Kullback-Leibler information divergence. Proof

ÚTIA

Properties of Structural Mixture Model

STRUCTURAL MIXTURES $\approx$ statistically correct subspace approach:

- **PRINCIPLE:** the less informative univariate distributions $f_n(x_n|m)$ are replaced by fixed "background" distributions $f_n(x_n|0)$
- reduces the number of mixture parameter (and components) $\Rightarrow$ reduces the risk of overpeaking
- suppresses the influence of unreliable (less informative) variables
- the EM algorithm performs feature selection for each component independently (it is not necessary to exclude variables globally)
- Bayesian decision-making based on structural mixtures is dimension independent (Grim 2016)
- the structural optimization implied by EM algorithm is controlled by the Kullback-Leibler information divergence
- avoids the biologically unnatural connection of probabilistic neurons with all input variables (Grim et al. 2000)
- enables the structural optimization of probabilistic neural networks by EM algorithm (Grim 2007)

Modification of EM Algorithm for Incomplete Data

INCOMPLETE DATA: $\boldsymbol{x} = (x_1, -, x_3, x_4, -, -, x_7, \ldots, x_N) \in \mathcal{X}$

$\mathcal{N}(\boldsymbol{x}) = \{n \in \mathcal{N} : \text{variable } x_n \text{ is defined in } \boldsymbol{x}\}, \quad \boldsymbol{x} \in \mathcal{X}$

$\mathcal{S}_n = \{\boldsymbol{x} \in \mathcal{S} : n \in \mathcal{N}(\boldsymbol{x})\}, \quad \approx$ vectors $\boldsymbol{x} \in \mathcal{S}$ with the defined variable x_n

Assumption: components in product form $\Rightarrow$ Easily available marginals

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

EM iteration equations: $(m \in \mathcal{M}, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S})$

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

$$f_n^{'}(.|m) = \arg \max_{f_n(.|m)} \Big\{ \frac{1}{\sum_{x \in \mathcal{S}_n} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}_n} q(m|\boldsymbol{x}) \log f_n(x_n|m) \Big\}$$

Remark: **The likelihood criterion depends on available values only.**

ÚTIA

Modification of EM algorithm for Weighted Data

NOTATION: $\gamma(\boldsymbol{x}) > 0$: relative frequency of $\boldsymbol{x}$ in $\mathcal{S}$, $(\sum_{\boldsymbol{x}\in\mathcal{X}} \gamma(\boldsymbol{x}) = 1)$

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

$\bar{\mathcal{X}} = \{\boldsymbol{x} \in \mathcal{X} : \gamma(\boldsymbol{x}) > 0\}$: the sum can be confined to $\boldsymbol{x} \in \bar{\mathcal{X}}$:

"weighted" EM iteration equations: $(m \in \mathcal{M},\ n \in \mathcal{N},\ \boldsymbol{x} \in \bar{\mathcal{X}})$

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

$$w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\bar{\mathcal{X}}} \gamma(\boldsymbol{x}) q(m|\boldsymbol{x})$$

$$F^{'}(.|m) = \arg\max_{F(.|m)} \left\{ \sum_{\boldsymbol{x}\in\bar{\mathcal{X}}} \frac{\gamma(\boldsymbol{x}) q(m|\boldsymbol{x})}{w_m^{'}} \log F(\boldsymbol{x}|m) \right\}$$

Applications: **relevance of data, aggregation of data, discrete data weighted by table values:** $\gamma(\boldsymbol{x}) = P^*(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X}$

Sequential Decision Scheme (Grim 1986, 2014)

INFORMATION CONTROLLED SEQUENTIAL DECISION-MAKING

Given the observations $\boldsymbol{x}_D = (x_{j_1}, \ldots, x_{j_l}) \in \mathcal{X}_D$, $\mathcal{D} = \{j_1, \ldots, j_l\} \subset \mathcal{N}$ we have to choose the next most informative variable x_n, $n \notin \mathcal{D}$ to maximize the **conditional information** $I_{\boldsymbol{x}_D}(\mathcal{X}_n, \Omega)$ **about the classes** $\Omega = \{\omega_1, \ldots, \omega_K\}$.

SOLUTION: explicit evaluation of the criterion $I_{\boldsymbol{x}_D}(\mathcal{X}_n, \Omega)$

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

$$H_{\boldsymbol{x}_D}(\mathcal{X}_n) = \sum_{x_n \in \mathcal{X}_n} -P_{n|D}(x_n|\boldsymbol{x}_D) \log P_{n|D}(x_n|\boldsymbol{x}_D), \quad P_{n|D}(x_n|\boldsymbol{x}_D) = \frac{P_{nD}(x_n, \boldsymbol{x}_D)}{P_D(\boldsymbol{x}_D)}$$

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

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

$$P_{nD|\omega}(x_n, \boldsymbol{x}_D|\omega) = \sum_{m \in \mathcal{M}} w_m f_n(x_n|m, \omega) \prod_{i \in \mathcal{D}} f_i(x_i|m, \omega),$$

Feature Selection: the Most Informative Subspace

special case of the sequential decision scheme:

INFORMATION CRITERION for the optimal feature subset

ASSUMPTION: class-conditional product mixtures $P(\boldsymbol{x}|\omega), \omega \in \Omega$

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

$$P_{D|\omega}(\boldsymbol{x}_D|\omega) = \sum_{m \in \mathcal{M}_\omega} w_m \prod_{n \in \mathcal{D}} f_n(x_n|m), \qquad \boldsymbol{x}_D \in \mathcal{X}_D,$$

$$H(\mathcal{X}_D) = \sum_{\boldsymbol{x}_D \in \mathcal{X}_D} -P_D(\boldsymbol{x}_D) \log P_D(\boldsymbol{x}_D), \qquad \mathcal{D} = \{j_1, \ldots, j_k\} \subset \mathcal{N}, \qquad |\mathcal{D}| = k$$

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

optimal subset $\mathcal{D} \subset \mathcal{N}$**: complete search, approximate methods**

APPLICATION: informative feature selection for pattern recognition

PROPERTIES OF PRODUCT MIXTURES

SURVEY: computational properties of product mixtures

- efficient estimation of multivariate distribution mixtures **(!)**
- suitable to approximate multi-modal, real-life probability distributions
- with increasing number of components the Gaussian mixtures approach the asymptotic accuracy of Parzen (kernel) estimates
- unlike Parzen estimates the product mixtures are optimally "smoothed" by the efficient EM algorithm
- directly available marginal probability distributions **(!)**
- the mixture parameters can be estimated from incomplete data
- product components enable the information controlled sequential decision-making in multi-dimensional spaces
- product mixtures can be interpreted as probabilistic neural networks
- enable the structural optimization of probabilistic neural networks
- provide information criterion for the optimal feature subset

Literature

A1: Asymptotic Properties of Parzen Estimates

Theorem (Parzen, 1962; Cacoullos, 1966)

Let $\mathcal{S}_K$ be a sequence of K independent observations of an N-dimensional random vector distributed with the probability density function $P^*(\boldsymbol{x})$. The non-parametric density estimate $P(\boldsymbol{x})$ with the soothing parameter σ_K

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

is asymptotically unbiased in each continuity point of $P^*(\boldsymbol{x})$, i.e. it holds

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

if $\lim_{K\to\infty} \sigma_K = 0$. In addition, if $\lim_{K\to\infty} K\sigma_K^N = \infty$, then the unbiased estimate $P(\boldsymbol{x})$ is asymptotically consistent in the quadratic mean sense:

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

◂ Back: Compromise ◂ Back: Product mixtures

A2: Optimal Smoothing of Parzen (Kernel) Estimates

Parzen estimate with Gaussian kernel:

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

optimization by cross-validation (leaving-one-out) method:

$\approx$ to maximize the modified log-likelihood function by EM algorithm:

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

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

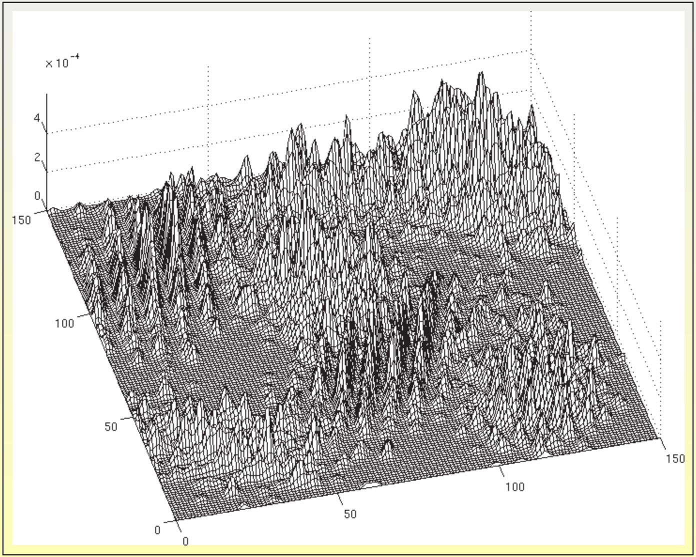
$$(\sigma_n^{'})^2 = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}}\sum_{\boldsymbol{y}\in\mathcal{S},\boldsymbol{y}\neq\boldsymbol{x}} (x_n - y_n)^2 q(\boldsymbol{y}|\boldsymbol{x})$$

Remark: **Optimal smoothing is crucial in high-dimensional spaces!**

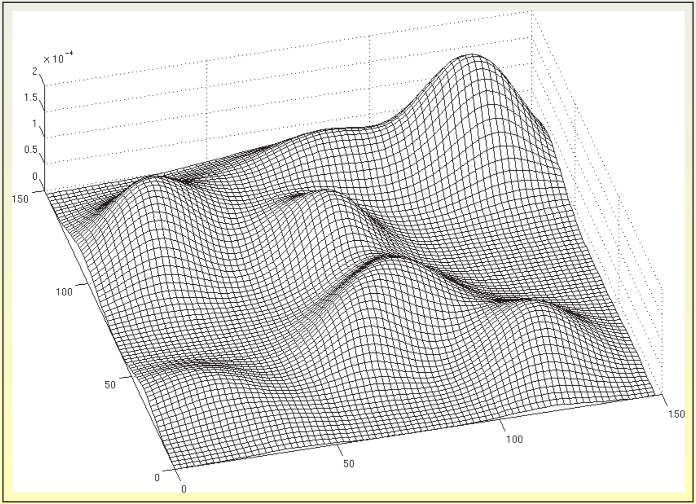
Back: Product mixtures Back: Compromise

ÚTIA

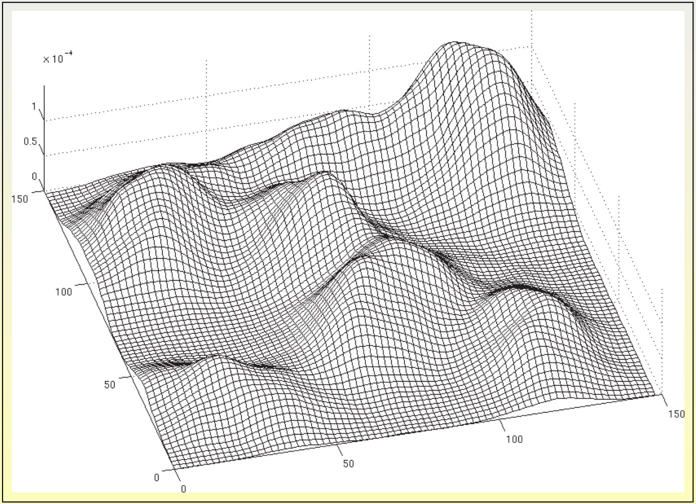
"Under-smoothed" Kernel Estimate



"Over-smoothed" Kernel Estimate



Optimally Smoothed Kernel Estimate



(general Gaussian kernel)

Back: Norm. mixture

A3: Marginal Distributions of a Product Mixture

easily obtained by omitting superfluous terms in products:

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|m) = \sum_{m\in\mathcal{M}} w_m \prod_{n\in\mathcal{N}} f_n(x_n|m), \qquad \boldsymbol{x} = (x_1,\dots,x_N)\in\mathcal{X}$$

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

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

$$P_C(\boldsymbol{x}_C) = \sum_{m\in\mathcal{M}} w_m F_C(\boldsymbol{x}_C|m), \qquad F_C(\boldsymbol{x}_C|m) = \prod_{n\in C} f_n(x_n|m)$$

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \frac{P_{nC}(x_n,\boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m\in\mathcal{M}} \frac{w_m F_C(\boldsymbol{x}_C|m)}{P_C(\boldsymbol{x}_C)} f_n(x_n|m)$$

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \sum_{m\in\mathcal{M}} W_m(\boldsymbol{x}_C) f_n(x_n|m), \qquad W_m(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{P_C(\boldsymbol{x}_C)}$$

◂ Back - Incomplete data   ◂ Back: Product mixtures

A4: Solution of the **M-Step** - Gaussian Mixture

Gaussian Mixture with a General Covariance Matrix:

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

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|\boldsymbol{c}_m, A_m)$$

implicit form of the M-Step:

$$(\boldsymbol{c}_m^{'}, A_m^{'}) = \arg\max_{(\boldsymbol{c}_m, A_m)} \left\{ \sum_{\boldsymbol{x}\in\mathcal{S}} \gamma(\boldsymbol{x}) \log F(\boldsymbol{x}|\boldsymbol{c}_m, A_m) \right\}$$

explicit solution:

$$\boldsymbol{c}_m^{'} = \sum_{\boldsymbol{x}\in\mathcal{S}} \gamma(\boldsymbol{x})\ \boldsymbol{x}, \quad \gamma(\boldsymbol{x}) = \frac{q(m|\boldsymbol{x})}{\sum_{\boldsymbol{y}\in\mathcal{S}} q(m|\boldsymbol{y})}$$

$$A_m^{'} = \sum_{\boldsymbol{x}\in\mathcal{S}} \gamma(\boldsymbol{x})\ (\boldsymbol{x}-\boldsymbol{c}_m^{'})(\boldsymbol{x}-\boldsymbol{c}_m^{'})^T = \sum_{\boldsymbol{x}\in\mathcal{S}} \gamma(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{c}_m^{'}(\boldsymbol{c}_m^{'})^T$$

Back: Gaussian Mixture

A5: Solution of the **M-Step** - Discrete Product Mixture

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

$$\sum_{\xi \in \mathcal{X}_n} \delta(\xi, x_n) = 1, \quad x_n \in \mathcal{X}_n,$$

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

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

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

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

Back: EM algorithm Back: Structural EM

A5: Invariance of EM Algorithm Under Linear Transform

EM estimate of a Gaussian mixture is invariant under linear transform

Let the parameters $\{w_m, \mu_{mn}, \sigma_{mn}, m \in \mathcal{M}, n \in \mathcal{N}\}$ of a Gaussian product mixture define a stationary point of EM algorithm, i.e. they satisfy the EM iteration equations. Further let $\boldsymbol{y} = T(\boldsymbol{x})$ be a linear transform of the vectors $\boldsymbol{x} \in \mathcal{X}$ a of the mixture parameters :

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

Then the transformed parameters $\{\tilde{w}_m, \tilde{\mu}_{mn}, \tilde{\sigma}_{mn}, m \in \mathcal{M}, n \in \mathcal{N}\}$ also define a stationary point of EM algorithm in the transformed space $\mathcal{Y}$.

Proof: The following equations can be verified by related substitutions:

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

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

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

◂ Back: Gaussian Product Mixture

A6: Monotonic Property of Structural EM Algorithm

structural mixture is a special case of product mixture model, i.e.

$$w_m^{'} = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}), \quad f_n^{'}(.|m) = \arg\max_{f_n(.|m)} \left\{ \sum_{\boldsymbol{x}\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|} \log f_n(x_n|m) \right\}$$

It is necessary to prove, that the monotonic property holds for the optimized structural parameters ϕ_{mn}. We use the inequality :

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

and, making substitution for $F^{'}(\boldsymbol{x}|m), F(\boldsymbol{x}|m)$, we obtain:

$$L^{'} - L \geq \sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} \left\{ \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log\left[\frac{G^{'}(\boldsymbol{x}|m,\phi_m^{'})}{G(\boldsymbol{x}|m,\phi_m)}\right]\right\}$$

$$L^{'} - L \geq \sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} \left\{ \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log\left[\frac{f_n^{'}(x_n|m)}{f_n(x_n|0)}\right]^{\phi_{mn}^{'}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)}\right]^{\phi_{mn}}\right\}$$

Monotonic Property of Structural EM Algorithm

The last inequality can be rewritten in the form:

$$(*) \quad L^{'} - L \geq \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} (\phi^{'}_{mn} - \phi_{mn})\gamma^{'}_{mn} + \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \frac{\phi_{mn}}{|\mathcal{S}|} q(m|\boldsymbol{x}) \log \frac{f^{'}_n(x_n|m)}{f_n(x_n|m)}$$

where $\gamma^{'}_{mn}$ is the structural optimization criterion:

$$\gamma^{'}_{mn} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) \log \frac{f^{'}_n(x_n|m)}{f_n(x_n|0)}, \quad n \in \mathcal{N}, m \in \mathcal{M}$$

In view of the above definition of $f^{'}_n(.|m)$ we can write for arbitrary $f_n(\cdot|m)$:

$$\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) \log f^{'}_n(x_n|m) \geq \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) \log f_n(x_n|m)$$

Therefore, the last sum in the inequality $(*)$ is non-negative and, for the same reason, we have $\gamma^{'}_{mn} \geq 0$ for all $n \in \mathcal{N}, m \in \mathcal{M}$;

By setting $\phi^{'}_{mn} = 1$ for the R highest values $\gamma^{'}_{mn}$, we obtain

$$L^{'} - L \geq \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} (\phi^{'}_{mn} - \phi_{mn}) \, \gamma^{'}_{mn} \geq 0 \quad \textbf{q.e.d.}$$

Back: Structural EM

ÚTIA

Interpretation of Structural Criterion - Discrete Mixture

$f_n(x_n|m),\ x_n \in \mathcal{X}_n,\ n \in \mathcal{N}\ \approx$ **discrete probability distribution**

$$\gamma'_{mn} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x}) \log \frac{f'_n(x_n|m)}{f_n(x_n|0)}, \quad n \in \mathcal{N}, m \in \mathcal{M}$$

$$\sum_{\xi \in \mathcal{X}_n} \delta(\xi, x_n) = 1, \quad x_n \in \mathcal{X}_n,$$

$$\gamma'_{mn} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x}) \Big[\sum_{\xi \in \mathcal{X}_n} \delta(\xi, x_n) \Big] \log \frac{f'_n(x_n|m)}{f_n(x_n|0)},$$

$$\gamma'_{mn} = \frac{1}{|\mathcal{S}|} \sum_{\xi \in \mathcal{X}_n} \Big[\sum_{\boldsymbol{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\boldsymbol{x}) \Big] \log \frac{f'_n(\xi|m)}{f_n(\xi|0)},$$

$$\gamma'_{mn} = w'_m \sum_{\xi \in \mathcal{X}_n} f'_n(\xi|m) \log \frac{f'_n(\xi|m)}{f_n(\xi|0)} = w'_m I(f'_n(\cdot|m), f_n(\cdot|0)),$$

$\gamma'_{mn}\ \approx$ **Kullback-Leibler information divergence**

Back: Structural EM

Interpretation of Structural Criterion - Gaussian Mixture

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

$$\gamma'_{mn} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} q(m|\boldsymbol{x}) \log \frac{f_n(x_n|\mu'_{mn},\sigma'_{mn})}{f_n(x_n|\mu_{0mn},\sigma_{0n})}, \quad n \in \mathcal{N}, m \in \mathcal{M},$$

$$\gamma'_{mn} = w'_m \sum_{\boldsymbol{x}\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w'_m|\mathcal{S}|} \left[-\log \frac{\sigma'_{mn}}{\sigma_{0n}} - \frac{(x_n-\mu'_{mn})^2}{2(\sigma'_{mn})^2} + \frac{(x_n-\mu_{0n})^2}{2(\sigma_{0n})^2} \right],$$

$$\gamma'_{mn} = \frac{w'_m}{2} \left[\frac{(\mu'_{mn}-\mu_{0n})^2}{(\sigma_{0n})^2} + \frac{(\sigma'_{mn})^2}{(\sigma_{0n})^2} - 1 - \log \frac{(\sigma'_{mn})^2}{(\sigma_{0n})^2} \right] =$$

it is easily verified: Back: Structural EM

$$= w'_m \int_{\mathcal{X}_n} f_n(x_n|\mu'_{mn},\sigma'_{mn}) \log \frac{f_n(x_n|\mu'_{mn},\sigma'_{mn})}{f_n(x_n|\mu_{0n},\sigma_{0n})} dx_n = w'_m I(f'_n(\cdot|m), f_n(\cdot|0))$$

$\Rightarrow \gamma'_{mn} \approx$ **"continuous" Kullback-Leibler information divergence**

ÚTIA

A7: Non-Identifiability of Discrete Product Mixtures

Definition of Identifiability of Mixtures (Teicher, 1963)

The class of Mixtures $\mathcal{P} = \{P(\boldsymbol{x},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is identifiable, if the parameters $\boldsymbol{\theta}, \boldsymbol{\theta}^{'} \in \Theta$ of any two equivalent mixtures

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

may differ only by the order of components.

Back: identification x aproximation

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

Arbitrary discrete product mixture $(x_n \in \mathcal{X}_n, \quad |\mathcal{X}_n| < \infty)$

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|m) = \sum_{m\in\mathcal{M}} w_m \prod_{n\in\mathcal{N}} f_n(x_n|m)$$

has infinitely many equivalent forms with different parameters, if at least one of the univariate component distributions $f_i(x_i|m)$ *is nonsingular, i.e. satisfies the condition*

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

Back: Discrete mixture

Proof: Non-Identifiability of Discrete Product Mixtures

Proof: Let $0 < f_i(x_i|m) < 1$ for some $i \in \mathcal{N}, x_i \in \mathcal{X}_i$ and $m \in \mathcal{M}$. Then, for any $0 < \alpha < 1,\ \beta = 1 - \alpha$, we can construct two different probability distributions $f_i^{'}(\cdot|m), f_i^{''}(\cdot|m)$ in such a way that the distribution $f_i(\cdot|m)$ represents an internal point of the abscise $\langle f_i^{'}(\cdot|m), f_i^{''}(\cdot|m)\rangle$ in the $|\mathcal{X}_i|$-dimensional space in the sense of the following condition:

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

Consequently, the nonsingular probability distribution $f_i(\cdot|m)$ can be expressed as a convex combination of two distributions $f_i^{'}(\cdot|m), f_i^{''}(\cdot|m)$ in infinitely many ways. By using the above substitution (*) we can write

$$(**) \quad w_m F(\boldsymbol{x}|m) = w_m^{'} F^{'}(\boldsymbol{x}|m) + w_m^{''} F^{''}(\boldsymbol{x}|m),$$

where

$$w_m^{'} = \alpha w_m, \quad w_m^{''} = \beta w_m, \quad (w_m^{'} + w_m^{''} = w_m),$$

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

Finally, making substitution (**) for $w_m F(\boldsymbol{x}|m)$, we obtain a non-trivially different equivalent of the original distribution $P(\boldsymbol{x})$, q.e.d.

Back: EM algorithm

A8: Alternative Proof of the EM Monotonic Property

Kullback-Leibler information divergence is non-negative, i.e. :

$$I(q(\cdot|\boldsymbol{x}), q'(\cdot|\boldsymbol{x})) = \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log \frac{q(m|\boldsymbol{x})}{q'(m|\boldsymbol{x})} \geq 0,$$

Proof

The following proof follows the original idea of Schlesinger. Using notation

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

We can express the log-likelihood functions L and L' equivalently by means of the conditional weights $q(m|\boldsymbol{x}), q'(m|\boldsymbol{x})$:

$$L = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \left\{ \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log[w_m F(\boldsymbol{x}|m)] - \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log q(m|\boldsymbol{x}) \right\}$$

$$L' = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \left\{ \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log[w'_m F'(\boldsymbol{x}|m)] - \sum_{m\in\mathcal{M}} q(m|\boldsymbol{x}) \log q'(m|\boldsymbol{x}) \right\}$$

Alternative Proof of the EM Monotonic Property

Using the above equations we can express the increment $L^{'} - L$ as follows:

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

where the second sum on the right-hand side is the non-negative Kullback-Leibler divergence:

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

and therefore, we can write the inequality:

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

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

Alternative Proof of the EM Monotonic Property

Making substitution for $w_m^{'}$ from the **M-Step** we obtain the inequality

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

Further, in view of the **M-Step** definition

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

we can write for any component $F(\boldsymbol{x}|m)$ the inequality:

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

The monotonic property of EM algorithm follows from the above inequalities:

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

Remark: The **M-Step** definition is redundantly strong, the new parameters need to satisfy only the inequalities (*) $\Rightarrow$ GEM algorithm

Back

A9: Monotonic Property of EM Algorithm - Implications

Nondecreasing and above bounded sequence $\{L^{(t)}\}_{t=0}^{\infty}$ has a finite limit $L^* < \infty$ and therefore the following necessary condition is satisfied:

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

Analogous conditions hold for the sequences $\{w^{(t)}(m)\}_{t=0}^{\infty}$ and $\{q^{(t)}(\cdot|\boldsymbol{x})\}_{t=0}^{\infty}$, $m \in \mathcal{M}$, too:

$$\lim_{t\to\infty} ||w^{(t+1)}(m) - w^{(t)}(m)|| = 0, \quad \lim_{t\to\infty} ||q^{(t+1)}(m|\boldsymbol{x}) - q^{(t)}(m|\boldsymbol{x})|| = 0.$$

The last limits follow from the inequality

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

and from the following general inequality (cf. Kullback (1966)): Back

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

A10: M.-L. Estimates versus Approximation Problems

Lemma

Maximum-likelihood estimate asymptotically minimizes the upper bound of the Euklidean distance between the true discrete distribution $P^*(\cdot)$ and its approximating estimate $P(\cdot)$.

Proof: Asymptotically, for $|\mathcal{S}| \to \infty$, we can write

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

where $\gamma(\boldsymbol{x}) \geq 0$ is the relative frequency of the discrete vector $\boldsymbol{x}$ in the i.i.d. sequence $\mathcal{S}$ and P^* is the true probability distribution. The assertion follows from the inequality (cf. Kullback, 1966):

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

Remark: The m.-l. estimate $P(\cdot)$ is justified as approximation of $P^*(\cdot)$.

A11: Kullback-Leibler Divergence is Non-Negative

Theorem (cf. e.g. Vajda, 1992)

Any two discrete probability distributions $\{q_1, q_2, \dots, q_M\}$, $\{q_1^{'}, q_2^{'}, \dots, q_M^{'}\}$ *satisfy the following inequality*

$$I(\boldsymbol{q} \| \, \boldsymbol{q}^{'}) = \sum_{m \in \mathcal{M}} q_m \log \frac{q_m}{q_m^{'}} \geq 0$$

where the equality holds only if $q_m^{'} = q_m$, *for all* $m \in \mathcal{M}$.

Proof: Without any loss of generality we can assume $q_m > 0$ for all $m \in \mathcal{M}$ (since $0 \log 0 = 0$ asymptotically). By Jensens inequality we have:

$$\sum_{m \in \mathcal{M}} q_m \log \frac{q_m^{'}}{q_m} \leq \log \Big(\sum_{m \in \mathcal{M}} q_m \frac{q_m^{'}}{q_m} \Big) = \log \Big(\sum_{m \in \mathcal{M}} q_m^{'} \Big) = \log 1 = 0,$$

where the equality occurs only if $q_1^{'}/q_1 = \cdots = q_M^{'}/q_M$, q.e.d.

Consequence: The following left-hand sum is maximized by $\boldsymbol{q}^{'} = \boldsymbol{q}$:

$$\sum_{m \in \mathcal{M}} q_m \log q_m^{'} \leq \sum_{m \in \mathcal{M}} q_m \log q_m$$

Back - Proof Back (Alternative Proof) Back (M-Step)

A12: Universality of Discrete Product Mixtures

Lemma (see e.g. Grim, 2006)

Let the table values $p^{(k)}, k = 1, \ldots, K,\ K = |\mathcal{X}|$ define a probability distribution $P(\boldsymbol{x})$ on a discrete space $\mathcal{X}$:

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

Then the discrete probability distribution $P(\boldsymbol{x})$ can be expressed as a product distribution mixture by using δ-functions in the product components:

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

Proof: The products of δ-functions in the components uniquely define the points $\boldsymbol{x}^{(k)} \in \mathcal{X}$ corresponding to the respective probabilistic table values $p^{(k)}$:

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

Remark: The proof has only formal meaning, the mixture approximation based on EM algorithm is numerically more efficient.

Back - ("representable")

Back - Advantages

A13: EM algorithm for Multivariate Bernoulli Mixtures

example of EM algorithm: **multivariate Bernoulli mixture**

```
//    Estimation of Multivariate Bernoulli Mixture by means of EM algoritmu
//===========================================================================
//short     X[NN];                    //  binary data vector
//int       NN;                       //  dimension of binary vectors
//int       MM;                       //  number of mixture components
//double    P[MM][NN],SP[MM][NN];     //  mixture parameters and related estimates
//double    W[MM],SW[MM];             //  component weights and related estimates
//double    FX[MM];                   //  component values for a given vector X[NN]
//double    FXM,SWM,Q,SUM,SWM;        //  auxiliary variables
//int       N,M,IT,ITERMAX;           //  auxiliary variables

for(IT=1; IT<=ITERMAX; IT++)
//***************************************************************************
{  for(M=0; M<MM; M++) {SW[M]=0.0; for(N=0; N<NN; N++) SP[M][N]=0.0;}
   Q=0.0;
   for(J=1;J<=JJ;J++)                 //  cycle over all data vectors X
   { READ(X);  SUM=0.0;               //  to read X from the input data set
     for(M=0; M<MM; M++)
     {  FXM=W[M];
        for(N=0; 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=0; 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=0; M<MM; M++)               //  to compute the new parameter estimates
   {  SWM=SW[M];   W[M]=SWM/JJ;  for(N=0; N<NN; N++) P[M][N]=SP[M][N]/SWM;
   }  // end of M-loop
   print(IT,Q);
}  //  end of IT-loop
//***************************************************************************
printf("\n End of the EM algorithm\n\n");
```

A14: EM algorithm for Gaussian Product Mixtures

example of EM algorithm: **multivariate Gaussian product mixture**

```
//   Estimation of the Gaussian product mixture by means of EM algorithm
//=====================================================================================
//int IT,N,M;  long K;  double F,G,FXM,SWM,SUM,FMAX,Q0;    // global variables
//double   X[DNN];                        // real data vector
//double   FX[DMM],W[DMM],SW[DMM];        // components, weights, weight estimates
//double   C[DMM][DNN], A[DMM][DNN];      // component means and variances
//double   SC[DMM][DNN],SA[DMM][DNN];     // new estimates of means and variances
for(IT=1; IT<=ITMAX; IT++)                // iteration loop
//*************************************************************************************
{ Q=0.0
  for(M=1; M<=MM; M++)                    // logarithmic parameters and initial values
  {  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;                            // to simplify the evaluation of exponents
  }  // end of M-loop
  for(I=1;I<=K;I++)                       // cycle over all data vectors X
  {  READ(X);  FMAX=-RMAX;
     for(M=1; M<=MM; M++)                 // evaluation of the logarithm of components
     {  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++)                 // to compute the component values and 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;                   // to compute the log-likelihood criterion
     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++)                    //  to compute the new parameter estimates
  {  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
```

Remark: Possible solution of the “underflow” problem.

Prof. M.I. Schlesinger with his wife

At Karlštejn castle during his visit in Prague in 1995.

Back Back - Literature

Literature 1/12

- 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.

Back

Literature 2/12

- 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.18, No.3, pp.173-190. *http://dml.cz/dmlcz/124132*
- 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. (1984): On structural approximating multivariate discrete probability distributions. *Kybernetika*, Vol. 20, No. 1, pp. 1-17, 1984. *http://dml.cz/dmlcz/125676*
- Grim J. (1986): Multivariate statistical pattern recognition with nonreduced dimensionality, *Kybernetika*, Vol. 22, pp. 142-157. *http://dml.cz/dmlcz/125022*

Back

Literature 3/12

- Grim, J. (1986): Sequential decision-making in pattern recognition based on the method of independent subspaces. In: *Proceedings of the DIANA II Conference on Discriminant Analysis*, (Ed. F. Zitek), Mathematical Institute of the AS CR, Prague 1986, pp. 139-149.

- 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.

- 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*, (Ed. R.Trappl), Vienna, April 1992, World Scientific, Singapore 1992, pp. 997-1005. Paper Award

- Grim J. and 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.

Back

Literature 4/12

- 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.

- 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. and 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 the 4th System Science European Congress*, L. Ferrer et al. (Eds.), (pp: 527-540), Valencia: Sociedad Espanola de Sistemas Generales, 1999.

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

Paper Award Back

Literature 5/12

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

Back

ÚTIA

Literature 6/12

- Grim J. (2001): Latent Structure Analysis for Categorical Data. Research Report No. 2019. ÚTIA AV ČR, Praha 2001, 13 pp. 23
- Grim J., Kittler J., Pudil P. and Somol P. (2002): Multiple classifier fusion in probabilistic neural networks. *Pattern Analysis & Applications* Vol. 5, No. 7, pp. 221-233.
- Grim J. and Haindl M. (2003): Texture Modelling by Discrete Distribution Mixtures. Computational Statistics and Data Analysis, 3-4 **41**, pp. 603-615.
- Grim J., Just P. and Pudil P. (2003): Strictly modular probabilistic neural networks for pattern recognition. *Neural Network World*, Vol. 13 , No. 6, pp. 599-615.
- Grim J., Somol P., Pudil P. and 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.

Back

Literature 7/12

- Grim J., Hora J. and Pudil P. (2004): Interaktivní reprodukce výsledků sčítání lidu se zaručenou ochranou anonymity dat. *Statistika*, Vol. 84, No. 5, pp. 400-414.
- Grim J., Haindl M., Somol P., Pudil P. and 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.
- Grim J., Somol P., Haindl M. and 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., Haindl M., Pudil P. and 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.
- Grim J. (2006): 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.

Back

Literature 8/12

- J. Grim (2007): Neuromorphic features of probabilistic neural networks. *Kybernetika*, Vol. 43, No. 5, pp.697-712. *http://dml.cz/dmlcz/135807*
- Grim J. and Hora, J. (2008): Iterative principles of recognition in probabilistic neural networks. *Neural Networks*, Special Issue, 6 **21**, 838–846 Paper Award
- Grim J., Novovičová J. and Somol P. (2008): Structural Poisson Mixtures for Classification of Documents , *Proceedings of the 19th International Conference on Pattern Recognition*, Tampa (Florida), US, p. 1324-1327.
- Grim J., Somol P., Haindl M. and J. Daneš (2009): Computer-Aided Evaluation of Screening Mammograms Based on Local Texture Models. *IEEE Trans. on Image Processing*, Vol. 18, No. 4, pp. 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. Paper Award

Back

Literature 9/12

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

- J. Grim (2011): Preprocessing of Screening Mammograms Based on Local Statistical Models. *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2011*, Barcelona, ACM, pp. 1-5

- 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

- Grim, J. (2017). Approximation of unknown multivariate probability distributions by using mixtures of product components: a tutorial. *International Journal of Pattern Recognition and Artificial Intelligence*, to appear.

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

Literature 10/12

- 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. and 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.
- McLachlan, G.J. and Krishnan, T. (1997): The EM algorithm and extensions, John Wiley & Sons, New York.
- McLachlan G.J. and Peel D. (2000): *Finite Mixture Models*, John Wiley & Sons, New York, Toronto, (2000)

Back

Literature 11/12

- Meng X.L. and Van Dyk D. (1997): The EM Algorithm—an Old Folk-song Sung to a Fast New Tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 59, No. 3, pp. 511-567.
- 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. and 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.
- Teicher H. (1963): Identifiability of finite mixtures. *Ann. Math. Statist.*, Vol. 34, pp. 1265-1269.
- Teicher H. (1968): Identifiability of mixtures of product measures. *Ann. Math. Statist.*, Vol. 39, pp. 1300-1302.

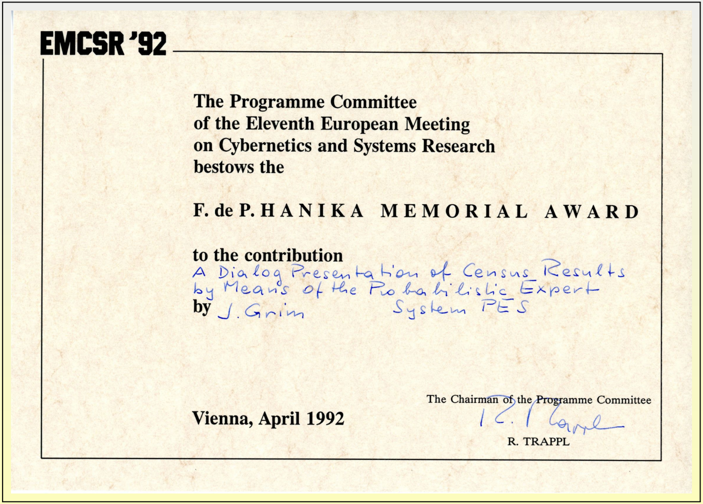
Back

Literature 12/12

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

Back

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
System PES
by J. Grim

Vienna, April 1992

The Chairman of the Programme Committee
R. Trappl
R. TRAPPL

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

Back

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

ÚTIA

Paper Award

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

Back

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

Back

Paper Award

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

Back

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

Back

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

Back