Approximating Probability Densities by Mixtures of Gaussian Dependence Trees

Jiří Grim

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Prague
Department of Pattern Recognition
http://www.utia.cas.cz/RO

SPMS 2014, June 23-28, 2014, Malá Skála, Czech Republic

Outline

Dependence-Tree Concept

chain expansion formula:

$$P(\mathbf{x}) = p(x_1)\prod_{n=2}^{N} p(x_n|x_{n-1},\ldots,x_1), \quad \mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbf{X},$$

dependence-tree expansion:

$\pi = (i_1, i_2, \ldots, i_N) \;\approx\;$ permutation of the index set $\;\mathcal{N} = \{1, 2, \ldots, N\}$

$$P(\mathbf{x}|\pi) = p(x_{i_1})\prod_{n=2}^{N} p(x_{i_n}|x_{j_n}), \quad j_n \in \{i_1, \ldots, i_{n-1}\} \;\approx\; \textbf{spanning tree of } \mathcal{N}$$

$$P(\mathbf{x}|\pi) \;=\; p(x_{i_1})\prod_{n=2}^{N}\frac{p(x_{i_n}, x_{j_n})}{p(x_{j_n})} \;=\; \left[\prod_{n=1}^{N} p(x_{i_n})\right]\left[\prod_{n=2}^{N}\frac{p(x_{i_n}, x_{j_n})}{p(x_{i_n})p(x_{j_n})}\right],$$

in natural ordering:

$$P(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i)\prod_{n=2}^{N}\frac{p(x_n, x_{k_n})}{p(x_n)p(x_{k_n})} = p(x_1)\prod_{n=2}^{N} p(x_n|x_{k_n})$$

marginals: $\boldsymbol{\theta} = \{p(x_n, x_{k_n}), n = 2, .., N\} \quad // \Rightarrow \; \{p(x_n), n = 1, .., N\}$

dependence structure: $\boldsymbol{\alpha} = (k_2, .., k_N)$

Binary Dependence-Tree Approximation (Chow & Liu)

minimum Kullback-Leibler information divergence: Properties:

$P^*(\mathbf{x}) \approx$ given distribution, $P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) \approx$ binary dependence tree

criterion: $$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = \sum_{x \in X} P^*(\mathbf{x}) \log \frac{P^*(\mathbf{x})}{P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta})} =$$

$$-H(P^*) - \sum_{x_1=0}^{1} p^*(x_1) \log p(x_1) - \sum_{n=2}^{N} \sum_{x_n=0}^{1} \sum_{x_{k_n}=0}^{1} p^*(x_n, x_{k_n}) \log p(x_n|x_{k_n}) =$$

$$-H(P^*) - \sum_{x_1=0}^{1} p^*(x_1) \log p(x_1) - \sum_{n=2}^{N} \sum_{x_{k_n}=0}^{1} p^*(x_{k_n}) \left[\sum_{x_n=0}^{1} \frac{p^*(x_n, x_{k_n})}{p^*(x_{k_n})} \log p(x_n|x_{k_n}) \right]$$

for any fixed dependence structure $\alpha = (k_2, \ldots, k_N)$ the criterion $I(P^*(\cdot)||P(\cdot|\alpha,\boldsymbol{\theta}))$ is minimized by the two-dimensional marginals θ^*:

$$\boldsymbol{\theta}^* = \{p^*(x_n, x_{k_n}), n = 2, .., N\} \Rightarrow p(x_1) = p^*(x_1), \quad p(x_n|x_{k_n}) = \frac{p^*(x_n, x_{k_n})}{p^*(x_{k_n})}$$

Binary Dependence-Tree Approximation (Chow & Liu)

making substitution for $\boldsymbol{\theta}^*$ we can write:

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta}^*)) = \sum_{x \in X} P^*(\mathbf{x}) \log \frac{P^*(\mathbf{x})}{P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}^*)} =$$

$$= -H(P^*) + \sum_{n=1}^{N} H(p_n^*) - \sum_{n=2}^{N} \sum_{x_n=0}^{1} \sum_{x_{k_n}=0}^{1} p^*(x_n, x_{k_n}) \log \frac{p^*(x_n, x_{k_n})}{p^*(x_n) p^*(x_{k_n})}$$

Shannon information: $\mathcal{I}(p_n^*, p_{k_n}^*) = \sum_{x_n=0}^{1} \sum_{x_{k_n}=0}^{1} p^*(x_n, x_{k_n}) \log \frac{p^*(x_n, x_{k_n})}{p^*(x_n) p^*(x_{k_n})}$

Shannon entropies: $H(P^*),\ \sum_{n=1}^{N} H(p_n^*) \approx$ **structure independent**

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta}^*)) = -H(P^*) + \sum_{n=1}^{N} H(p_n^*) - \sum_{n=2}^{N} \mathcal{I}(p_n^*, p_{k_n}^*) \ \rightarrow \ \textbf{min}$$

$\Rightarrow \ \boldsymbol{\alpha}^* = \arg\max_{\alpha} \left\{ \sum_{n=2}^{N} \mathcal{I}(p_n^*, p_{k_n}^*) \right\} \approx$ **maximum-weight spanning tree**

MWST

Estimation of Binary Dependence Tree (Chow & Liu)

data set: $\mathcal{S} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}, \quad \mathbf{x} = (x_1, x_2, \ldots, x_N) \in \{0,1\}^N$

binary dependence-tree distribution:

$$P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = p(x_1)\prod_{n=2}^{N} p(x_n|x_{k_n}) = \prod_{n=1}^{N} p(x_n) \prod_{n=2}^{N} \frac{p(x_n, x_{k_n})}{p(x_n)p(x_{k_n})}$$

log-likelihood function:

$$L(\boldsymbol{\alpha},\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \log P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \left[\log p(x_1) + \sum_{n=2}^{N} \log p(x_n|x_{k_n})\right]$$

using δ-function notation $\sum_{\xi_n=0}^{1} \delta(\xi_n, x_n) = 1$ we can write

$$L(\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{\xi_1=0}^{1} \left[\frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \delta(\xi_1, x_1)\right] \log p(\xi_1) +$$

$$+ \sum_{n=2}^{N} \sum_{\xi_n=0}^{1} \sum_{\xi_{k_n}=0}^{1} \left[\frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \delta(\xi_n, x_n)\delta(\xi_{k_n}, x_{k_n})\right] \log p(\xi_n|\xi_{k_n}),$$

Estimation of Binary Dependence Tree Distribution

denoting $\hat{p}(\xi_n)$, $\hat{p}(\xi_n, \xi_{k_n})$ the estimates of marginal probabilities:

$$\hat{p}(\xi_n) = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}\delta(\xi_n, x_n), \quad \hat{p}(\xi_n, \xi_{k_n}) = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}\delta(\xi_n, x_n)\delta(\xi_{k_n}, x_{k_n}), \quad n \in \mathcal{N},$$

we can write:

$$L(\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{\xi_1=0}^{1}\hat{p}(\xi_1)\log p(\xi_1) + \sum_{n=2}^{N}\sum_{\xi_n=0}^{1}\sum_{\xi_{k_n}=0}^{1}\hat{p}(\xi_n, \xi_{k_n})\log p(\xi_n|\xi_{k_n})$$

$$L(\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{\xi_1=0}^{1}\hat{p}(\xi_1)\log p(\xi_1) + \sum_{n=2}^{N}\sum_{\xi_{k_n}=0}^{1}\hat{p}(\xi_{k_n})\sum_{\xi_n=0}^{1}\frac{\hat{p}(\xi_n, \xi_{k_n})}{\hat{p}(\xi_{k_n})}\log p(\xi_n|\xi_{k_n})$$

$\Rightarrow$ for any fixed dependence structure α the log-likelihood criterion $L(\alpha, \theta)$ is maximized by the distributions:

$$p(\xi_n) = \hat{p}(\xi_n), \quad p(\xi_n|\xi_{k_n}) = \frac{\hat{p}(\xi_n, \xi_{k_n})}{\hat{p}(\xi_{k_n})}, \quad n \in \mathcal{N}$$

Estimation of Binary Dependence Tree Distribution

making substitutions $p(\xi_n) = \hat{p}(\xi_n)$, $p(\xi_n|\xi_{k_n}) = \frac{\hat{p}(\xi_n,\xi_{k_n})}{\hat{p}(\xi_{k_n})}$ **we obtain:**

$$L(\alpha, \hat{\theta}) = \sum_{n=1}^{N} \sum_{\xi_n=0}^{1} \hat{p}(\xi_n) \log \hat{p}(\xi_n) + \sum_{n=2}^{N} \sum_{\xi_n=0}^{1} \sum_{\xi_{k_n}=0}^{1} \hat{p}(\xi_n, \xi_{k_n}) \log \frac{\hat{p}(\xi_n, \xi_{k_n})}{\hat{p}(\xi_n)\hat{p}(\xi_{k_n})},$$

and using the mutual Shannon information formula $\mathcal{I}(\hat{p}_n, \hat{p}_{k_n})$**:**

$$\mathcal{I}(p^*_n, p^*_{k_n}) = \sum_{x_n=0}^{1} \sum_{x_{k_n}=0}^{1} p^*(x_n, x_{k_n}) \log \frac{p^*(x_n, x_{k_n})}{p^*(x_n)p^*(x_{k_n})}$$

we can write:

$$L(\alpha, \hat{\theta}) = \sum_{n=1}^{N} -H(\hat{p}_n) + \sum_{n=2}^{N} \mathcal{I}(\hat{p}_n, \hat{p}_{k_n})$$

$\Rightarrow$ **the dependence structure α is optimized by maximizing the last structure-dependent sum:** $(\mathcal{I}(\hat{p}_n, \hat{p}_{k_n}) \approx$ edge weight$)$

maximum weight spanning tree: $\hat{\alpha} = \arg\max_{\alpha} \left\{ \sum_{n=2}^{N} \mathcal{I}(\hat{p}_n, \hat{p}_{k_n}) \right\}$

MWST

ÚTIA

Continuous Dependence-Tree Approximation

multivariate probability density functions $P^*(\mathbf{x}),\ P(\mathbf{x}|\alpha,\vartheta)$

$$P(\mathbf{x}|\alpha,\vartheta) = f(x_1)\prod_{n=2}^{N} f(x_n|x_{k_n}),\ \mathbf{x}\in\mathcal{R}^N,$$

criterion: $$I(P^*(\cdot)||P(\cdot|\alpha,\vartheta) = \int_{R^N} P^*(\mathbf{x})\log\frac{P^*(\mathbf{x})}{P(\mathbf{x}|\alpha,\vartheta)}d\mathbf{x} =$$

$$\int P^*(\mathbf{x})\log P^*(\mathbf{x})d\mathbf{x} - \int P^*(\mathbf{x})\left[\log f(x_1) + \sum_{n=2}^{N}\log f(x_n|x_{k_n})\right]d\mathbf{x} = -H(P^*) -$$

$$-\int_R f^*(x_1)\log f(x_1)dx_1 - \sum_{n=2}^{N}\int_R f^*(x_{k_n})\left[\int_R \frac{f^*(x_n,x_{k_n})}{f^*(x_{k_n})}\log f(x_n|x_{k_n})dx_n\right]dx_{k_n}$$

for any fixed dependence structure $\alpha = (k_2,\ldots,k_N)$ **the criterion** $I(P^*(\cdot)||P(\cdot|\alpha,\vartheta))$ **is minimized by the two-dimensional marginals** ϑ^***:**

$$\vartheta^* = \{f^*(x_n,x_{k_n}), n=2,\ldots,N\} \Rightarrow f(x_1) = f^*(x_1), f(x_n|x_{k_n}) = \frac{f^*(x_n,x_{k_n})}{f^*(x_{k_n})}$$

Continuous Dependence-Tree Approximation

making substitution for ϑ^* we obtain:

$$I(P^*(\cdot)||P(\cdot|\alpha,\vartheta^*)) = \int_{R^N} P^*(\mathbf{x}) \log \frac{P^*(\mathbf{x})}{P(\mathbf{x}|\alpha,\vartheta^*)} =$$

$$= -H(P^*) + \sum_{n=1}^{N} H(f_n^*) - \sum_{n=2}^{N} \int_{R^2} f^*(x_n, x_{k_n}) \log \frac{f^*(x_n, x_{k_n})}{f^*(x_n) f^*(x_{k_n})} \, dx_n dx_{k_n}$$

Shannon information: $\mathcal{I}(f_n^*, f_{k_n}^*) = \int_{R^2} f^*(x_n, x_{k_n}) \log \frac{f^*(x_n, x_{k_n})}{f^*(x_n) f^*(x_{k_n})} \, dx_n dx_{k_n}$

$$I(P^*(\cdot)||P(\cdot|\alpha,\vartheta^*)) = -H(P^*) + \sum_{n=1}^{N} H(f_n^*) - \sum_{n=2}^{N} \mathcal{I}(f_n^*, f_{k_n}^*) \ \rightarrow \ \textbf{min}$$

Shannon entropies: $H(P^*)$, $\sum_{n=1}^{N} H(f_n^*) \ \approx$ **structure independent**

$\Rightarrow \ \alpha^* = \arg\max_{\alpha} \left\{ \sum_{n=2}^{N} \mathcal{I}(f_n^*, f_{k_n}^*) \right\} \approx$ **maximum-weight spanning tree**

▸ MWST

Estimation of Gaussian Dependence Tree

data set: $\mathcal{S} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}, \quad \mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{R}^N$

Gaussian dependence-tree:

$$P(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(x_1|\mu_1, \sigma_1) \prod_{n=2}^{N} f(x_n|x_{k_n}, \mu_n, \mu_{k_n}, \Sigma_{nk_n})$$

$$P(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(x_1|\mu_1, \sigma_1) \prod_{n=2}^{N} \left[\frac{f(x_n, x_{k_n}|\mu_n, \mu_{k_n}, \Sigma_{nk_n})}{f(x_{k_n}|\mu_{k_n}, \sigma_{k_n})} \right]$$

we assume Gaussian densities:

$$f(x_n|\mu_n, \sigma_n) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left\{-\frac{(x_n-\mu_n)^2}{2\sigma_n^2}\right\}, \quad \Sigma_{nk} = \begin{pmatrix} \sigma_n^2 & \sigma_{nk} \\ \sigma_{nk} & \sigma_k^2 \end{pmatrix}, \quad n, k \in \mathcal{N},$$

$$f(x_n, x_k|\mu_n, \mu_k, \Sigma_{nk}) =$$

$$= \frac{1}{\sqrt{(2\pi)^2 \det \Sigma_{nk}}} \exp\{-\tfrac{1}{2}(x_n - \mu_n, x_k - \mu_k)^T \Sigma_{nk}^{-1} (x_n - \mu_n, x_k - \mu_k)\}$$

log-likelihood function:

$$L(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log P(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \to \ \textbf{max}$$

Estimation of Gaussian Dependence Tree

making substitution for $P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma})$ we obtain:

$$L(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}\left[\log f(x_1|\mu_1,\sigma_1) - \sum_{n=2}^{N}\log f(x_{k_n}|\mu_{k_n},\sigma_{k_n})\right] +$$

$$+\sum_{n=2}^{N}\frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}\log f(x_n,x_{k_n}|\mu_n,\mu_{k_n},\Sigma_{nk_n})$$

for any fixed dependence structure $\boldsymbol{\alpha}$ the criterion $L(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma})$ is maximized by m.-l. estimates of the two-dimensional marginals:

$$f(x_n,x_k|\mu_n,\mu_k,\Sigma_{nk}) = f(x_n,x_k|\hat{\mu}_n,\hat{\mu}_k,\hat{\Sigma}_{nk}), \quad \Rightarrow \; f(x_n|\mu_n,\sigma_n) = f(x_n|\hat{\mu}_n,\hat{\sigma}_n)$$

$$\Rightarrow \; f(x_n|x_{k_n},\mu_n,\mu_{k_n},\Sigma_{nk_n}) = f(x_n,x_{k_n}|\hat{\mu}_n,\hat{\mu}_{k_n},\hat{\Sigma}_{nk_n})/f(x_{k_n}|\hat{\mu}_{k_n},\hat{\sigma}_{k_n})$$

m.-l. estimates of parameters:

$$\hat{\mu}_n = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}x_n, \quad \hat{\sigma}_n^2 = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}(x_n-\hat{\mu}_n)^2, \quad \hat{\sigma}_{nk} = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}(x_n-\hat{\mu}_n)(x_k-\hat{\mu}_k)$$

Estimation of Gaussian Dependence Tree

making substitutions $\mu_n = \hat{\mu}_n, \sigma_n = \hat{\sigma}_n, \sigma_{nk} = \hat{\sigma}_{nk}$ **we can write:**

$$L(\boldsymbol{\alpha}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sum_{n=1}^{N} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log f(x_n|\hat{\mu}_n, \hat{\sigma}_n) +$$

$$+ \sum_{n=2}^{N} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \frac{f(x_n, x_{k_n}|\hat{\mu}_n, \hat{\mu}_{k_n}, \hat{\Sigma}_{nk_n})}{f(x_n|\hat{\mu}_n, \hat{\sigma}_n) f(x_{k_n}|\hat{\mu}_{k_n}, \hat{\sigma}_{k_n})}$$

and finally:

$$L(\boldsymbol{\alpha}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sum_{n=1}^{N} \frac{1}{2} \left[1 + \log(2\pi\hat{\sigma}_n^2)\right] + \sum_{n=2}^{N} -\frac{1}{2} \log\left(1 - \frac{\hat{\sigma}_{nk_n}^2}{\hat{\sigma}_n^2 \hat{\sigma}_{k_n}^2}\right)$$

last term is the Shannon information between the variables x_n, x_{k_n}

$$\mathcal{I}(f(\cdot|\hat{\mu}_n, \hat{\sigma}_n), f(\cdot|\hat{\mu}_{k_n}, \hat{\sigma}_{k_n})) = -\tfrac{1}{2} \log\left(1 - \tfrac{\hat{\sigma}_{nk_n}^2}{\hat{\sigma}_n^2 \hat{\sigma}_{k_n}^2}\right)$$

$\Rightarrow$ **the structure is optimized by the maximum-weight spanning tree:**

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\alpha} \left\{ \textstyle\sum_{n=2}^{N} \mathcal{I}(f(\cdot|\hat{\mu}_n, \hat{\sigma}_n), f(\cdot|\hat{\mu}_{k_n}, \hat{\sigma}_{k_n})) \right\}$$

Mixtures of Gaussian Dependence Trees

mixtures of Gaussian dependence-tree components:

$$P(\mathbf{x}|\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{m\in\mathcal{M}} w_m F(\mathbf{x}|\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) =$$

$$= \sum_{m\in\mathcal{M}} w_m f(x_1|\mu_1^{(m)},\sigma_1^{(m)}) \prod_{n=2}^{N} f(x_n|x_{k_n},\mu_n^{(m)},\mu_{k_n}^{(m)},\Sigma_{nk_n}^{(m)})$$

weight vector: $\mathbf{w} = (w_1,\ldots,w_M)$, structural parameters: $\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_M\}$, component parameters:

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_M\}, \quad \boldsymbol{\mu}_m = \{\mu_1^{(m)},\mu_2^{(m)},\ldots,\mu_N^{(m)}\},$$

$$\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1,..,\boldsymbol{\Sigma}_M\}, \;\; \boldsymbol{\Sigma}_m = \{\Sigma_{nk_n}^{(m)}, n=2,..,N\}, \quad \Sigma_{nk_n}^{(m)} = \begin{pmatrix} \sigma_n^{(m)2} & \sigma_{nk_n}^{(m)} \\ \sigma_{nk_n}^{(m)} & \sigma_k^{(m)2} \end{pmatrix},$$

maximum likelihood criterion:

$$L(\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \log\left[\sum_{m\in\mathcal{M}} w_m F(\mathbf{x}|\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)\right] \to \max$$

EM Algorithm for Gaussian Dependence-Tree Mixtures

EM algorithm: iterative maximization of weighted likelihood functions

$$Q_m(\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \sum_{x \in \mathcal{S}} \frac{q(m|\mathbf{x})}{w'_m |\mathcal{S}|} \log F(\mathbf{x}|\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \;\; m \in \mathcal{M}$$

conditional weights $q(m|\mathbf{x})$ and the new component weights w'_m:

$$q(m|\mathbf{x}) = \frac{w_m F(\mathbf{x}|\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{P(\mathbf{x}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}, \;\; w'_m = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\mathbf{x})$$

for any fixed structure α_m the weighted likelihood Q_m is maximized by weighted maximum-likelihood estimates $\mu_n'^{(m)}, \sigma_n'^{(m)}, \sigma_{nk}'^{(m)}$:

$$\mu_n'^{(m)} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \frac{q(m|\mathbf{x})}{w'_m |\mathcal{S}|} x_n, \quad (\sigma_n'^{(m)})^2 = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \frac{q(m|\mathbf{x})}{w'_m |\mathcal{S}|} (x_n - \mu_n'^{(m)})^2,$$

$$\sigma_{nk}'^{(m)} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \frac{q(m|\mathbf{x})}{w'_m |\mathcal{S}|} (x_n - \mu_n'^{(m)})(x_k - \mu_k'^{(m)}), \quad n, k \in \mathcal{N}, \;\; m \in \mathcal{M}$$

Proof

EM Algorithm for Gaussian Dependence-Tree Mixtures

making substitutions $\mu_n^{(m)} = \mu_n^{'(m)}, \sigma_n^{(m)} = \sigma_n^{'(m)}, \sigma_{nk}^{(m)} = \sigma_{nk}^{'(m)}$ **we get:**

$$Q_m(\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m^{'}, \boldsymbol{\Sigma}_m^{'}) = \sum_{n=1}^{N} \sum_{x \in \mathcal{S}} \frac{q(m|\mathbf{x})}{w_m^{'}|\mathcal{S}|} \log f(x_n|\mu_n^{'(m)}, \sigma_n^{'(m)}) +$$

$$+ \sum_{n=2}^{N} \sum_{x \in \mathcal{S}} \frac{q(m|\mathbf{x})}{w_m^{'}|\mathcal{S}|} \log \frac{f(x_n, x_{k_n}|\mu_n^{'(m)}, \mu_{k_n}^{'(m)}, \sigma_{nk_n}^{'(m)})}{f(x_n|\mu_n^{'(m)}, \sigma_n^{'(m)}) f(x_{k_n}|\mu_{k_n}^{'(m)}, \sigma_{k_n}^{'(m)})}$$

the weighted log-likelihood can be transformed to the form:

$$Q_m(\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m^{'}, \boldsymbol{\Sigma}_m^{'}) = -\sum_{n=1}^{N} \left[\frac{1 + \log(2\pi\sigma_n^{'(m)2})}{2} \right] + \sum_{n=2}^{N} -\frac{1}{2} \log \left(1 - \frac{\sigma_{nk_n}^{'(m)2}}{\sigma_n^{'(m)2} \sigma_{k_n}^{'(m)2}} \right)$$

and therefore $Q_m(\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m^{'}, \boldsymbol{\Sigma}_m^{'})$ is maximized by maximizing the last sum of spanning-tree information weights: ▸ MWST

$$\boldsymbol{\alpha}_m^{'} = \arg\max_{\boldsymbol{\alpha}_m} \left\{ \sum_{n=2}^{N} \mathcal{I}(f(\cdot|\hat{\mu}_n^{'(m)}, \hat{\sigma}_n^{'(m)}), f(\cdot|\hat{\mu}_{k_n}^{'(m)}, \hat{\sigma}_{k_n}^{'(m)})) \right\}$$

Approximation Accuracy of a Binary Table Distribution

P^*: original; P_1: product of marginals; P_2: Chow & Liu; P_3: Ku & Kullback; product mixtures P_4: M=2; P_5, P_6: M=3; dependence tree mixture P_7: M=2;

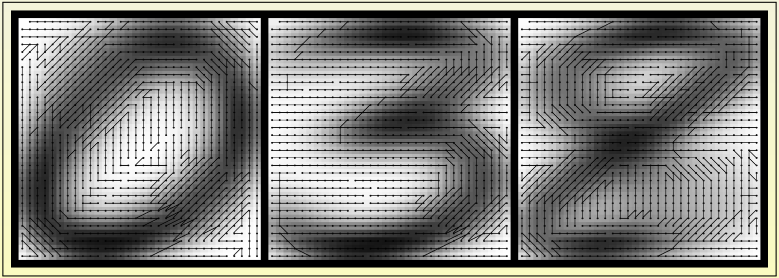
x_1 x_2 x_3 x_4	$P^*(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_5(x)$	$P_6(x)$	$P_7(x)$
0 0 0 0	·1000	·0456	·1296	·09977	·1037	·1000	·0857	·1000
0 0 0 1	·1000	·0456	·1037	·10000	·1296	·1000	·1143	·1000
0 0 1 0	·0500	·0557	·0370	·04958	·0296	·0500	·0600	·0500
0 0 1 1	·0500	·0557	·0296	·04927	·0370	·0500	·0400	·0500
0 1 0 0	·0000	·0557	·0152	·00051	·0149	·0000	·0000	·0000
0 1 0 1	·0000	·0557	·0121	·00026	·0124	·0000	·0000	·0000
0 1 1 0	·1000	·0681	·0681	·10011	·0669	·0833	·0900	·1000
0 1 1 1	·0500	·0681	·0546	·05035	·0558	·0667	·0600	·0500
1 0 0 0	·0500	·0557	·0530	·05027	·0519	·0600	·0643	·0500
1 0 0 1	·1000	·0557	·0636	·09996	·0648	·0900	·0857	·1000
1 0 1 0	·0000	·0681	·0152	·00039	·0148	·0000	·0000	·0000
1 0 1 1	·0000	·0681	·0182	·00076	·0185	·0000	·0000	·0000
1 1 0 0	·0500	·0681	·0331	·04945	·0397	·0400	·0500	·0500
1 1 0 1	·0500	·0681	·0397	·04978	·0331	·0600	·0500	·0500
1 1 1 0	·1500	·0832	·1488	·14992	·1785	·1667	·1500	·1500
1 1 1 1	·1500	·0832	·1785	·14962	·1488	·1333	·1500	·1500
Number of param.	15	4	7	28	9	14	14	15
Number of comp.	—	1	1	1	2	3	3	2
$H(P^*, P_i)$	·0000	·3687	·0952	·0098	·0952	·0092	·0084	·0000

(J. Grim, Kybernetika, Vol. 20, No. 1, pp. 1-17, 1984)

Recognition of Numerals by Mixtures of Dependence Trees

$$P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{m\in\mathcal{M}} w_m p^{(m)}(x_1) \prod_{n=2}^{N} p^{(m)}(x_n|x_{k_n})$$

number of components M=400, number of parameters: 819200



examples of dependence-tree components (numerals: 0, 3, 8)

Error Matrix: Product Mixture x Dependence Tree Mixture

CLASS	0	1	2	3	4	5	6	7	8	9	false n.
0	19950	8	43	19	39	32	36	0	38	17	1.1 %
1	2	22162	30	4	35	7	18	56	32	6	0.9 %
2	32	37	19742	43	30	9	8	29	90	16	1.5 %
3	20	17	62	20021	4	137	2	28	210	55	2.6 %
4	11	6	19	1	19170	11	31	51	30	247	2.1 %
5	25	11	9	154	4	17925	39	6	96	34	2.1 %
6	63	10	17	6	23	140	19652	1	54	3	1.6 %
7	7	12	73	10	73	4	0	20497	22	249	2.1 %
8	22	25	53	97	30	100	11	11	19369	72	2.1 %
9	15	13	25	62	114	22	3	146	93	19274	2.5 %
false p.	0.9%	0.7%	2.7%	2.0%	1.7%	2.3%	0.7%	1.6%	3.3%	3.5%	1.84%

classification error matrix for a multivariate Bernoulli mixture (M=1571 components, 1462373 parameters)

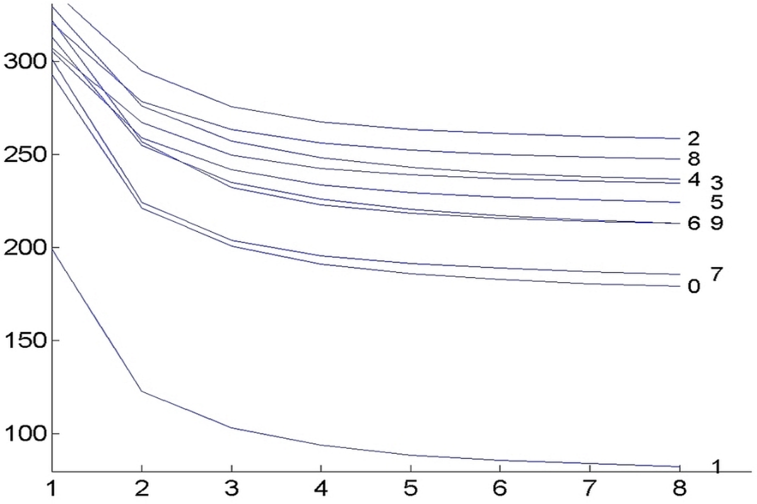
CLASS	0	1	2	3	4	5	6	7	8	9	false n.
0	19979	11	62	21	18	26	25	2	28	10	1.0 %
1	5	21981	78	13	74	1	20	155	21	4	1.7 %
2	22	15	19777	72	26	5	6	35	72	6	1.3 %
3	20	10	66	20169	1	120	1	20	122	27	1.9 %
4	12	16	13	4	19245	1	13	52	44	177	1.7 %
5	25	5	15	157	8	17874	45	9	129	36	2.3 %
6	100	19	38	25	43	90	19575	1	75	3	2.0 %
7	17	33	108	24	71	0	0	20367	28	299	2.8 %
8	18	30	47	167	27	55	22	17	19337	70	2.3 %
9	12	20	62	74	89	33	3	144	134	19196	2.9 %
false p.	1.4%	0.7%	2.4%	2.7%	1.8%	1.8%	0.7%	1.6%	3.1%	3.2%	1.97%

classification error matrix for a binary dependence tree mixture (M=400 components, 819200 parameters)

Recognition of Numerals by Mixtures of Dependence Trees

decreasing information contribution of the dependence structure

(overall spanning-tree weight in iterations $1 \div 8$ for the numerals $0 \div 9$)

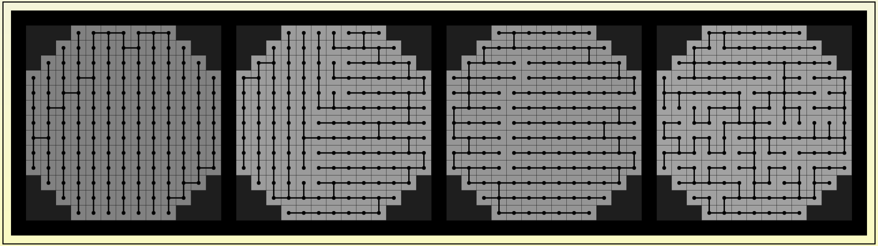


Preprocessing of Screening Mammograms

local statistical model of a screening mammogram based on a mixture of Gaussian dependence trees:

$$P(\mathbf{x}|\mathbf{w}, \alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{m \in \mathcal{M}} w_m F(\mathbf{x}|\alpha_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

examples of dependence-tree components (window: 13x13, N=145, M=36)

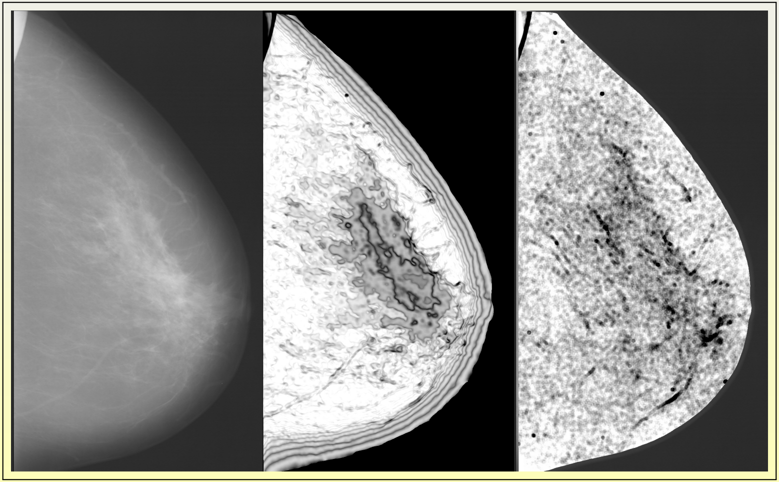


example of changing overall spanning-tree weight in iterations 1 ÷ 6
(window: 23x23, dimension: N=445, dependence-tree components: M=5)

1. 2438.90 **2.** 2739.22 **3.** 2781.51 **4.** 2777.5 **5.** 2782.35 **6.** 2792.63

Comparison of Different Log-Likelihood Images

log-likelihood images for a digital mammogram



original mammogram product mixture model dependence tree mixture

Product Mixtures versus Mixtures of Dependence Trees

$$P(\mathbf{x}) = \sum_m w_m \prod_{n=1}^{N} p(x_n|m) \quad \otimes \quad P(\mathbf{x}) = \sum_m w_m p(x_1|m) \prod_{n=2}^{N} p(x_n|x_{k_n}, m)$$

product mixtures:

- ⊕ marginals simply available by omitting superfluous product terms
- ⊕ computationally efficient implementation of EM algorithm
- ⊕ EM algorithm directly applicable to incomplete data
- ⊕ support "subspace" modification (component specific features)
- ⊖ restrictive assumption: conditional independence of variables

mixtures of dependence trees

- ⊕ statistical relationship between two variables by a single component
- ⊕ structural optimization by maximum weight spanning tree
- ⊖ difficult evaluation of marginal distributions
- ⊖ computationally demanding implementation of EM algorithm

Conclusion

large number of components:

- intuitively: mixture of dependence trees is similar to nonparametric Parzen estimate, the form of the kernels is less relevant
- dependence structure of components does not improve the approximation power of the product mixture essentially
- information contribution of the dependence structure decreases in the course of EM iterations
- optimal estimate of the dependence tree mixture tends to approach a simple product mixture model

small number of multidimensional components:

- single dependence tree describes statistical relations between variables
- information contribution of the dependence structure can increase in the course of EM iterations
- dependence structure of components can essentially improve the approximation quality

Literatura 1/3

- C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees", *IEEE Trans. on Information Theory*, Vol. IT-14, No.3, pp. 462- 467, 1968.
- J. Grim, "On structural approximating multivariate discrete probability distributions", *Kybernetika*, Vol. 20, No. 1, pp. 1-17, 1984. *http://dml.cz/dmlcz/125676*
- M. Meila and M.I. Jordan, "Learning with mixtures of trees", *Journal of Machine Learning Research*, Vol. 1, No. 9, pp. 1-48, 2001.
- I.J. Kim and J.H. Kim, "Statistical Character Structure Modeling and Its Application to Handwritten Chinese Character Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 11, pp. 1422-1436, 2003.
- S. Kirshner and P. Smyth, "Infinite mixtures of trees", *Proc. of the 24th International Conference on Machine Learning (ICML'07)*, Ed. Zoubin Ghahramani, ACM, New York, USA, pp. 417-423, 2007.

Back

Literatura 2/3

- B. Behsaz and M. Rahmati. "Estimation of Probability Density Function by Dependence Tree Methods for Pattern Recognition Systems." *Tech. Rep. U. Alberta*, 1, 2006.
- A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Statist. Soc., B*, Vol. 39, pp. l-38, 1977.
- J. Grim, "On numerical evaluation of maximum - likelihood estimates for finite mixtures of distributions", *Kybernetika*, Vol. l8, No.3, pp.173-190, 1982. *http://dml.cz/dmlcz/124132*
- J. Grim, "Multivariate statistical pattern recognition with nonreduced dimensionality", *Kybernetika*, Vol. 22, No. 2, pp. 142-157, 1986. *http://dml.cz/dmlcz/125022*
- J. Grim, "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, 2011

Literatura 3/3

- O. Boruvka, "On a minimal problem", *Transaction of the Moravian Society for Natural Sciences* (in Czech), No. 3, 1926.
- J.B Kruskal, "On the shortest spanning sub-tree of a graph", *Proc. Amer. Math. Soc.*, No. 7, pp. 48-50, 1956.
- R.C. Prim, "Shortest connection networks and some generalizations", *Bell System Tech. J.*, Vol. 36 , pp. 1389-1401, 1957.
- E. Parzen, "On estimation of a probability density function and its mode," *Annals of Mathematical Statistics*, Vol. 33., pp. 1065-1076, 1962.

Remark: Both J.B. Kruskal and R.C. Prim refer to an "... obscure Czech paper of O. Boruvka ..." describing construction of the minimum-weight spanning tree and the corresponding proof of uniqueness. Back (Outline)

Appendix: Maximum-Weight Spanning-Tree Construction

```
//*****************************************************************************
// Maximum-weight spanning tree construction (Prim, 1957)
//*****************************************************************************
//  NN........ number of nodes, N=1,2,...,NN
//  T[N]...... characteristic function of the known part of spanning tree
//  E[N][K]... positive weight of the edge <N,K>
//  A[K]...... index of the heaviest neighbor of node K in the known subtree
//  GE[K]..... greatest edge weight between the node K and the known subtree
//  K0........ index of the most heavy neighbor of the defined part of tree
//  SUM....... total weight of the spanning tree: {<2,A[2]>,...,<NN,A[NN]>}
//*****************************************************************************
    for(N=1; N<=NN; N++) {GE[N]=-1; T[N]=0; A[N]=0;}  //  initial values
    N0=1; T[N0]=1; K0=0;
    for(I=2; I<=NN; I++)
    {  FMAX=-1E0;
        for(N=2; N<=NN; N++) if(T[N]<1)
        {  F=E[N0][N];
        if(F>GE[N]) {GE[N]=F; A[N]=N0;}  else F=GE[N];
        if(F>FMAX)  {FMAX=F; K0=N;}
        }  //  end of N-loop
        N0=K0;  T[N0]=1;  SUM+=FMAX;
    }  // end of spanning tree construction
```

Minimum Information Divergence and Maximum Likelihood

By expanding the formula for $I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta}))$ we obtain

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = \sum_{x\in X} P^*(\mathbf{x}) \log P^*(\mathbf{x}) - \sum_{x\in X} P^*(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) \geq 0$$

$$\Rightarrow \quad \sum_{x\in X} P^*(\mathbf{x}) \log P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) \leq \sum_{x\in X} P^*(\mathbf{x}) \log P^*(\mathbf{x})$$

$\Rightarrow$ **The left-hand sum is uniquely maximized by** $P(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = P^*(\mathbf{x})$

Denoting $\gamma(\mathbf{x}) \geq 0$ the relative frequency of the vector $\mathbf{x}$ in the sequence $\mathcal{S}$ and P^* the true probability distribution, we can write

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

$\Rightarrow$ **Minimum information divergence and maximum-likelihood criterion are asymptotically equivalent**

Back

Maximization of the Weighted Likelihood Function

Let the parameter $\mathbf{b}$ of the probability distribution $F(\mathbf{x}|\mathbf{b})$ has a maximum-likelihood estimate defined as an additive function of $\mathbf{x} \in \mathcal{S}$:

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

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

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

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

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

$\Rightarrow$ **The weighted likelihood function is maximized by the weighted maximum-likelihood estimate** *(for details cf. Grim, 1982)* Back