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

1. Discrete Dependence-Tree Distributions
   - Dependence-Tree Concept
   - Binary Dependence-Tree Approximation (Chow & Liu)
   - Estimation of Binary Dependence Tree Distribution

2. Dependence-Tree Density Functions
   - Continuous Dependence-Tree Approximation
   - Estimation of Gaussian Dependence Tree

3. Estimation of Gaussian Dependence Tree Mixtures
   - Mixtures of Gaussian Dependence Trees
   - EM Algorithm for Dependence-Tree Mixtures

4. Application of Dependence Tree Mixtures
   - Approximation Accuracy of a Binary Table Distribution
   - Recognition of Numerals by Mixtures of Dependence Trees
   - Evaluation of Mammograms by Gaussian Dependence Tree Mixtures
   - Product Mixtures versus Mixtures of Dependence Trees

5. Conclusion
Dependence-Tree Concept

**Chain expansion formula:**

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

**Dependence-tree expansion:**

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

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

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

in natural ordering:

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

**Marginals:** \( \theta = \{p(x_n, x_{k_n}), n = 2, \ldots, N\} \quad \Rightarrow \quad \{p(x_n), n = 1, \ldots, N\} \)

**Dependence structure:** \( \alpha = (k_2, \ldots, k_N) \)
**Binary Dependence-Tree Approximation (Chow & Liu)**

**minimum Kullback-Leibler information divergence:**

\[ P^*(x) \approx \text{given distribution}, \quad P(x|\alpha, \theta) \approx \text{binary dependence tree} \]

**criterion:**

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

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

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

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

\[ \theta^* = \{p^*(x_n, x_{kn}), n = 2, \ldots, N\} \Rightarrow p(x_1) = p^*(x_1), \quad p(x_n|x_{kn}) = \frac{p^*(x_n, x_{kn})}{p^*(x_{kn})} \]
making substitution for $\theta^*$ we can write:

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

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

**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:**

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

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

$$\Rightarrow \alpha^* = \arg \max_{\alpha} \left\{ \sum_{n=2}^{N} \mathcal{I}(p_n^*, p_{k_n}^*) \right\} \approx \text{maximum-weight spanning tree (MWST)}$$
data set: \( S = \{x^{(1)}, x^{(2)}, \ldots \}, \ x = (x_1, x_2, \ldots, x_N) \in \{0, 1\}^N \)

binary dependence-tree distribution:

\[
P(x|\alpha, \theta) = p(x_1) \prod_{n=2}^{N} p(x_n|x_{kn}) = \prod_{n=1}^{N} p(x_n) \prod_{n=2}^{N} \frac{p(x_n, x_{kn})}{p(x_n)p(x_{kn})}
\]

log-likelihood function:

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

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

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

\[
+ \sum_{n=2}^{N} \sum_{\xi_n=0}^{1} \sum_{\xi_{kn}=0}^{1} \left[ \frac{1}{|S|} \sum_{x \in S} \delta(\xi_n, x_n) \delta(\xi_{kn}, x_{kn}) \right] \log p(\xi_n|\xi_{kn}),
\]
Estimation of Binary Dependence Tree Distribution

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

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

we can write:

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

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

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

$$p(\xi_n) = \hat{p}(\xi_n), \quad p(\xi_n | \xi_{kn}) = \frac{\hat{p}(\xi_n, \xi_{kn})}{\hat{p}(\xi_{kn})}, \quad n \in \mathcal{N}$$
making substitutions $p(\xi_n) = \hat{p}(\xi_n), \ p(\xi_n | \xi_{kn}) = \frac{\hat{p}(\xi_n, \xi_{kn})}{\hat{p}(\xi_{kn})}$ we obtain:

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

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

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

we can write:

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

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

maximum weight spanning tree: \( \hat{\alpha} = \arg \max_{\alpha} \left\{ \sum_{n=2}^{N} \mathcal{I}(\hat{p}_n, \hat{p}_{kn}) \right\} \)
 Continuous Dependence-Tree Approximation

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

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

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

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

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

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 $\vartheta^*$:

$$\vartheta^* = \left\{ f^*(x_n, x_{k_n}), n = 2, \ldots, N \right\} \Rightarrow f(x_1) = f^*(x_1), f(x_n|x_{k_n}) = \frac{f^*(x_n, x_{k_n})}{f^*(x_{k_n})}$$
Continuous Dependence-Tree Approximation

making substitution for $\vartheta^*$ we obtain:

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

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

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

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

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

$\Rightarrow \alpha^* = \arg \max_{\alpha} \left\{ \sum_{n=2}^{N} \mathcal{I}(f_n^*, f_{kn}^*) \right\} \approx \text{maximum-weight spanning tree}$
Estimation of Gaussian Dependence Tree

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

Gaussian dependence-tree:

\[
P(x|\alpha, \mu, \Sigma) = f(x_1|\mu_1, \sigma_1) \prod_{n=2}^{N} f(x_n|x_k, \mu_n, \mu_k, \Sigma_{nk})
\]

\[
P(x|\alpha, \mu, \Sigma) = f(x_1|\mu_1, \sigma_1) \prod_{n=2}^{N} \left[ \frac{f(x_n, x_k|\mu_n, \mu_k, \Sigma_{nk})}{f(x_k|\mu_k, \sigma_k)} \right]
\]

we assume Gaussian densities:

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

\[
f(x_n, x_k|\mu_n, \mu_k, \Sigma_{nk}) = \frac{1}{\sqrt{(2\pi)^2 \det \Sigma_{nk}}} \exp\left\{ -\frac{1}{2} (x_n - \mu_n, x_k - \mu_k)^T \Sigma_{nk}^{-1} (x_n - \mu_n, x_k - \mu_k) \right\}
\]

log-likelihood function:

\[
L(\alpha, \mu, \Sigma) = \frac{1}{|S|} \sum_{x \in S} \log P(x|\alpha, \mu, \Sigma) \to \max
\]
Estimation of Gaussian Dependence Tree

making substitution for \( P(x|\alpha, \mu, \Sigma) \) we obtain:

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

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

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

\[
\Rightarrow \quad f(x_n|x_{kn}, \mu_n, \mu_{kn}, \Sigma_{nk}) = f(x_n, x_{kn}|\hat{\mu}_n, \hat{\mu}_{kn}, \hat{\Sigma}_{nk})/f(x_{kn}|\hat{\mu}_{kn}, \hat{\sigma}_{kn})
\]

m.-l. estimates of parameters:

\[
\hat{\mu}_n = \frac{1}{|S|} \sum_{x \in S} x_n, \quad \hat{\sigma}_n^2 = \frac{1}{|S|} \sum_{x \in S} (x_n - \hat{\mu}_n)^2, \quad \hat{\sigma}_{nk} = \frac{1}{|S|} \sum_{x \in S} (x_n - \hat{\mu}_n)(x_k - \hat{\mu}_k)
\]
Discrete Continuous Mixtures Application Conclusion

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:

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

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

and finally:

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

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

$$I(f(\cdot|\hat{\mu}_n, \hat{\sigma}_n), f(\cdot|\hat{\mu}_{kn}, \hat{\sigma}_{kn})) = -\frac{1}{2} \log \left( 1 - \frac{\hat{\sigma}_{nk}^2}{\hat{\sigma}_n^2\hat{\sigma}_{kn}^2} \right)$$

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

$$\hat{\alpha} = \arg \max_{\alpha} \left\{ \sum_{n=2}^{N} I(f(\cdot|\hat{\mu}_n, \hat{\sigma}_n), f(\cdot|\hat{\mu}_{kn}, \hat{\sigma}_{kn})) \right\}$$
Mixtures of Gaussian Dependence Trees

mixtures of Gaussian dependence-tree components:

$$P(x|w, \alpha, \mu, \Sigma) = \sum_{m \in M} w_m F(x|\alpha_m, \mu_m, \Sigma_m) =$$

$$= \sum_{m \in M} w_m f(x_1|\mu_1^{(m)}, \sigma_1^{(m)}) \prod_{n=2}^N f(x_n|x_k, \mu_n^{(m)}, \mu_k^{(m)}, \Sigma_{nk}^{(m)})$$

weight vector: $w = (w_1, \ldots, w_M)$, structural parameters: $\{\alpha_1, \ldots, \alpha_M\}$, component parameters:

$$\mu = \{\mu_1, \mu_2, \ldots, \mu_M\}, \quad \mu_m = \{\mu_1^{(m)}, \mu_2^{(m)}, \ldots, \mu_N^{(m)}\},$$

$$\Sigma = \{\Sigma_1, \ldots, \Sigma_M\}, \quad \Sigma_m = \{\Sigma_{nk}^{(m)}, n = 2, \ldots, N\}, \quad \Sigma_{nk}^{(m)} = \left( \begin{array}{cc} \sigma_n^{(m)2} & \sigma_{nk}^{(m)} \\ \sigma_{nk}^{(m)} & \sigma_k^{(m)2} \end{array} \right),$$

maximum likelihood criterion:

$$L(w, \alpha, \mu, \Sigma) = \frac{1}{|S|} \sum_{x \in S} \log \left( \sum_{m \in M} w_m F(x|\alpha_m, \mu_m, \Sigma_m) \right) \rightarrow \max$$
**EM Algorithm for Gaussian Dependence-Tree Mixtures**

**EM algorithm:** iterative maximization of weighted likelihood functions

\[
Q_m(\alpha_m, \mu_m, \Sigma_m) = \sum_{x \in S} \frac{q(m|x)}{w'_m |S|} \log F(x|\alpha_m, \mu_m, \Sigma_m), \quad m \in \mathcal{M}
\]

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

\[
q(m|x) = \frac{w_m F(x|\alpha_m, \mu_m, \Sigma_m)}{P(x|w, \alpha, \mu, \Sigma)}, \quad w'_m = \frac{1}{|S|} \sum_{x \in S} q(m|x)
\]

for any fixed structure \(\alpha_m\) the weighted likelihood \(Q_m\) is maximized by weighted maximum-likelihood estimates \(\mu'_n(m), \sigma'_n(m), \sigma'_{nk}\):

\[
\mu'_n(m) = \frac{1}{|S|} \sum_{x \in S} \frac{q(m|x)}{w'_m |S|} x_n, \quad (\sigma'_n(m))^2 = \frac{1}{|S|} \sum_{x \in S} \frac{q(m|x)}{w'_m |S|} (x_n - \mu'_n(m))^2,
\]

\[
\sigma'_{nk} = \frac{1}{|S|} \sum_{x \in S} \frac{q(m|x)}{w'_m |S|} (x_n - \mu'_n(m))(x_k - \mu'_k(m)), \quad n, k \in \mathcal{N}, \ m \in \mathcal{M}
\]
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:

\[
Q_m(\alpha_m, \mu_m^{'}, \Sigma_m^{'}) = \sum_{n=1}^{N} \sum_{x \in S} \frac{q(m|x)}{w_m'|S|} \log f(x_n|\mu_n^{'}(m), \sigma_n^{'}(m)) +
\]

\[
+ \sum_{n=2}^{N} \sum_{x \in S} \frac{q(m|x)}{w_m'|S|} \log \frac{f(x_n, x_k^{}|\mu_n^{'}(m), \mu_k^{'}(m), \sigma_{nk}^{'}(m))}{f(x_n|\mu_n^{'}(m), \sigma_n^{'}(m))f(x_k^{}|\mu_k^{'}(m), \sigma_k^{'}(m))}
\]

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

\[
Q_m(\alpha_m, \mu_m^{'}, \Sigma_m^{'}) = -\sum_{n=1}^{N} \left[ \frac{1 + \log(2\pi\sigma_n^{'}(m)^2)}{2} \right] + \sum_{n=2}^{N} -\frac{1}{2} \log \left( 1 - \frac{\sigma_{nk}^{'}(m)^2}{\sigma_n^{'}(m)^2 \sigma_k^{'}(m)^2} \right)
\]

and therefore \( Q_m(\alpha_m, \mu_m^{'}, \Sigma_m^{'}) \) is maximized by maximizing the last sum of spanning-tree information weights:

\[
\alpha_m^{'} = \arg \max_{\alpha_m} \left\{ \sum_{n=2}^{N} I(f(.|\hat{\mu}_n^{'}(m), \hat{\sigma}_n^{'}(m)), f(.|\hat{\mu}_k^{'}(m), \hat{\sigma}_k^{'}(m))) \right\}
\]
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 $P_7$: $M=2$;

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$P^*(x)$</th>
<th>$P_1(x)$</th>
<th>$P_2(x)$</th>
<th>$P_3(x)$</th>
<th>$P_4(x)$</th>
<th>$P_5(x)$</th>
<th>$P_6(x)$</th>
<th>$P_7(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>·1000</td>
<td>·0456</td>
<td>·1296</td>
<td>·09977</td>
<td>·1037</td>
<td>·1000</td>
<td>·0857</td>
<td>·1000</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>·1000</td>
<td>·0456</td>
<td>·1037</td>
<td>·1000</td>
<td>·1296</td>
<td>·1000</td>
<td>·1143</td>
<td>·1000</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>·0500</td>
<td>·0557</td>
<td>·0370</td>
<td>·04958</td>
<td>·0296</td>
<td>·0500</td>
<td>·0600</td>
<td>·0500</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>·0500</td>
<td>·0557</td>
<td>·0296</td>
<td>·04927</td>
<td>·0370</td>
<td>·0500</td>
<td>·0400</td>
<td>·0500</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>·0000</td>
<td>·0557</td>
<td>·0152</td>
<td>·00051</td>
<td>·0149</td>
<td>·0000</td>
<td>·0000</td>
<td>·0000</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>·0000</td>
<td>·0557</td>
<td>·0121</td>
<td>·00026</td>
<td>·0124</td>
<td>·0000</td>
<td>·0000</td>
<td>·0000</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>·1000</td>
<td>·0681</td>
<td>·0681</td>
<td>·10011</td>
<td>·0669</td>
<td>·0833</td>
<td>·0900</td>
<td>·1000</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>·0500</td>
<td>·0681</td>
<td>·0546</td>
<td>·05035</td>
<td>·0558</td>
<td>·0667</td>
<td>·0600</td>
<td>·0500</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>·0500</td>
<td>·0557</td>
<td>·0530</td>
<td>·05027</td>
<td>·0519</td>
<td>·0600</td>
<td>·0643</td>
<td>·0500</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>·1000</td>
<td>·0557</td>
<td>·0636</td>
<td>·09996</td>
<td>·0648</td>
<td>·0900</td>
<td>·0857</td>
<td>·1000</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>·0000</td>
<td>·0681</td>
<td>·0152</td>
<td>·00039</td>
<td>·0148</td>
<td>·0000</td>
<td>·0000</td>
<td>·0000</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>·0000</td>
<td>·0681</td>
<td>·0182</td>
<td>·00076</td>
<td>·0185</td>
<td>·0000</td>
<td>·0000</td>
<td>·0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>·0500</td>
<td>·0681</td>
<td>·0331</td>
<td>·04945</td>
<td>·0397</td>
<td>·0400</td>
<td>·0500</td>
<td>·0500</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>·0500</td>
<td>·0681</td>
<td>·0397</td>
<td>·04978</td>
<td>·0331</td>
<td>·0600</td>
<td>·0500</td>
<td>·0500</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>·1500</td>
<td>·0832</td>
<td>·1488</td>
<td>·14992</td>
<td>·1785</td>
<td>·1667</td>
<td>·1500</td>
<td>·1500</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>·1500</td>
<td>·0832</td>
<td>·1785</td>
<td>·14962</td>
<td>·1488</td>
<td>·1333</td>
<td>·1500</td>
<td>·1500</td>
</tr>
</tbody>
</table>

Number of param.  
15  4  7  28  9  14  14  15

Number of comp.  
—  1  1  1  2  3  3  2

$H(P^*, P_1)$  
·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

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

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

### Classification Error Matrix for a Multivariate Bernoulli Mixture

<table>
<thead>
<tr>
<th>CLASS</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>false n.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19950</td>
<td>8</td>
<td>43</td>
<td>19</td>
<td>39</td>
<td>32</td>
<td>36</td>
<td>0</td>
<td>38</td>
<td>17</td>
<td>1.1 %</td>
</tr>
<tr>
<td>1</td>
<td>22162</td>
<td>30</td>
<td>4</td>
<td>35</td>
<td>7</td>
<td>18</td>
<td>56</td>
<td>32</td>
<td>6</td>
<td>0.9 %</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19742</td>
<td>43</td>
<td>30</td>
<td>9</td>
<td>8</td>
<td>29</td>
<td>90</td>
<td>16</td>
<td>1.5 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20021</td>
<td>4</td>
<td>137</td>
<td>2</td>
<td>28</td>
<td>210</td>
<td>55</td>
<td>2.6 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>19170</td>
<td>11</td>
<td>31</td>
<td>51</td>
<td>30</td>
<td>247</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>154</td>
<td>4</td>
<td>17925</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>140</td>
<td>19652</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>73</td>
<td>10</td>
<td>73</td>
<td>4</td>
<td>0</td>
<td>20497</td>
<td>22</td>
<td>249</td>
<td>2.1 %</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>100</td>
<td>11</td>
<td>11</td>
<td>19369</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>62</td>
<td>114</td>
<td>22</td>
<td>3</td>
<td>146</td>
<td>93</td>
<td>19274</td>
<td>2.5 %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Classification Error Matrix for a Binary Dependence Tree Mixture

<table>
<thead>
<tr>
<th>CLASS</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>false n.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19979</td>
<td>11</td>
<td>62</td>
<td>21</td>
<td>18</td>
<td>26</td>
<td>25</td>
<td>2</td>
<td>28</td>
<td>10</td>
<td>1.0 %</td>
</tr>
<tr>
<td>1</td>
<td>21981</td>
<td>78</td>
<td>13</td>
<td>74</td>
<td>1</td>
<td>20</td>
<td>155</td>
<td>21</td>
<td>4</td>
<td>1.7 %</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19777</td>
<td>72</td>
<td>26</td>
<td>5</td>
<td>6</td>
<td>35</td>
<td>72</td>
<td>6</td>
<td>1.3 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20169</td>
<td>1</td>
<td>120</td>
<td>1</td>
<td>20</td>
<td>122</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>19245</td>
<td>1</td>
<td>13</td>
<td>52</td>
<td>44</td>
<td>177</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>157</td>
<td>8</td>
<td>17874</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>19575</td>
<td>1</td>
<td>19337</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>71</td>
<td>0</td>
<td>20367</td>
<td>28</td>
<td>299</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>47</td>
<td>27</td>
<td>55</td>
<td>22</td>
<td>17</td>
<td>19337</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>62</td>
<td>89</td>
<td>33</td>
<td>3</td>
<td>144</td>
<td>134</td>
<td>19196</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**false p.**: 0.9% 0.7% 2.7% 2.0% 1.7% 2.3% 0.7% 1.6% 3.3% 3.5% 1.84%

**false p.**: 1.4% 0.7% 2.4% 2.7% 1.8% 1.8% 0.7% 1.6% 3.1% 3.2% 1.97%
decreasing information contribution of the dependence structure (overall spanning-tree weight in iterations 1 ÷ 8 for the numerals 0 ÷ 9)
Preprocessing of Screening Mammograms

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

\[ P(x|w, \alpha, \mu, \Sigma) = \sum_{m \in M} w_m F(x|\alpha_m, \mu_m, \Sigma_m) \]

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**

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

**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


**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.
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

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

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

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

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

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

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

$\Rightarrow$ Minimum information divergence and maximum-likelihood criterion are asymptotically equivalent
Maximization of the Weighted Likelihood Function

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

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

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

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

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

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

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