Stochastic and Physical Monitoring Systems
SPMS 2014
June 23 - June 28, 2014
Malá Skála, Czech Republic

Approximating Probability Densities by Mixtures of Gaussian Dependence Trees

Jiří Grim

Institute of Information Theory and Automation, Czech Academy of Sciences
P.O. BOX 18, Pod vodárenskou věží 4, 18208 Prague 8, Czech Republic

Email: grim@utia.cas.cz

Abstract. Considering the probabilistic approach to practical problems we are increasingly confronted with the need to estimate unknown multivariate probability density functions from large high-dimensional databases produced by electronic devices. The underlying densities are usually strongly multimodal and therefore mixtures of unimodal density functions suggest themselves as a suitable approximation tool. In this respect the product mixture models are preferable because they can be efficiently estimated from data by means of EM algorithm and have some advantageous properties. However, in some cases the simplicity of product components could appear too restrictive and a natural idea is to use a more complex mixture of dependence-tree densities. The dependence tree densities can explicitly describe the statistical relationships between pairs of variables at the level of individual components and therefore the approximation power of the resulting mixture may essentially increase.

Key words: Multivariate statistics; Mixtures of dependence trees; EM algorithm; Pattern recognition; Medical image analysis.

1 Introduction

The probabilistic description of data is known to be a powerful tool to solve practical problems. Recall that the true probability distribution represents a complete description of all statistical properties of the underlying random vector. Having estimated a multivariate probability density function from a set of independent data vector observations, we can derive very general and theoretically justified solutions in many fields like pattern recognition, prediction, image analysis, statistical modeling and others.

Naturally, in the last years there is an increasing need for methods which are efficient and practically applicable to estimating multivariate probability density functions from large sets of multidimensional data. Such data sets usually arise as a by-product of different information technologies in various areas like medicine, image processing, monitoring systems, communication networks and others. A typical feature of these databases is a high dimensionality of data and a large number of measurements. The unknown underlying probability density functions are nearly always multimodal and cannot be assumed

in a simple parametric form. A natural way to approximate the underlying multidimensional density functions is to use mixtures of product components. Here we use the term approximation to emphasize the approximation accuracy as a primary goal. We recall that, unlike statistical estimation problems, the approximating mixtures need not be identifiable and the chosen number of components may influence only the approximation accuracy.

The mixtures of product components can be efficiently estimated from data by means of EM algorithm and have some specific advantages as approximation tools, e.g., marginal distributions of product mixtures are directly available by omitting superfluous terms in product components, the approximation "power" of product mixtures can be increased by including additional components, product mixtures can be estimated directly from incomplete data without estimating the missing values and, moreover, they support a subspace modification for the sake of component-specific feature selection. In recent years, product mixtures have been applied to multidimensional problems in different areas like pattern recognition, texture evaluation, preprocessing of screening mammograms, texture modeling and others (cf. [8] - [12]).

Despite the advantageous properties of product mixtures, the simplicity of product components may become a limiting feature in some respects. As mentioned earlier, the approximation potential of the product mixtures can be increased by including additional product components but, in some cases, it could be advantageous to consider the mixture components in a more specific form. In this paper we propose to use dependence-tree densities as components [6], [18]. The structural optimization of tree dependence proposed by Chow and Liu [3] is compatible with the EM algorithm, and thus the estimation of dependence-tree mixtures from data is computationally feasible even in multidimensional spaces. By using the concept of a dependence tree we can explicitly describe statistical relationships between pairs of variables at the level of individual components; therefore, the approximation power of the resulting mixture model may be fundamentally increased. Of course, marginal distributions of dependence-tree mixtures are not easily obtained and we lose some of the above-mentioned advantageous properties of product mixtures. On the other hand, in some cases such properties need not be indispensable and the increased approximation potential may become essential.

2 Dependence-Tree Distributions

The idea of the dependence-tree distribution refers to the known paper of Chow and Liu [3] who proposed approximation of multivariate discrete probability distribution $P^*(\boldsymbol{x})$ by the product distribution

$$P(\boldsymbol{x}|\pi,\beta) = p(x_{i_1}) \prod_{n=2}^{N} p(x_{i_n}|x_{j_n}), \quad j_n \in \{i_1,\ldots,i_{n-1}\}. \tag{1}$$

Here $\pi = (i_1, i_2, \ldots, i_N)$ is a permutation of the index set $\mathcal{N}$ and β is the tree-dependence structure

$$\beta = \{(i_1,-),(i_2,j_2),\ldots,(i_N,j_N)\}, \quad j_n \in \{i_1,..,i_{n-1}\}.$$

Note that, here and in the sections that follow, we use a simplified notation of marginal distributions whenever tolerable. Thus, we write, e.g.,

$$p(x_n) = p_n(x_n), \quad p(x_n|x_k) = p_{n|k}(x_n|x_k).$$

The approximation model (1) is defined by the conditional marginals $p(x_{i_k}|x_{j_k})$ and by the dependence structure β. It can be shown that any dependence structure β uniquely defines a connected graph without circuits, in other words a spanning tree over the vertices $\mathcal{N}$ (cf. [6], p. 7, Theorem 3.2). The probability distribution (1) can be rewritten in the form

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

Here the first product is permutation-invariant and the second product can always be naturally ordered by the indices of variables:

$$P(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = \left[\prod_{n=1}^{N} p(x_n)\right]\left[\prod_{n=2}^{N}\frac{p(x_n,x_{k_n})}{p(x_n)p(x_{k_n})}\right] = p(x_1)\prod_{n=2}^{N} p(x_n|x_{k_n}). \tag{3}$$

In this sense the indices $\boldsymbol{\alpha} = (k_2,\ldots,k_N)$ briefly describe the ordered edges of the spanning tree $\tilde{\beta} = \{(2,k_2),\ldots,(N,k_N)\}$ and $\boldsymbol{\theta} = \{p(x_n,x_{k_n}), n = 2,\ldots,N\}$ stands for the related set of two-dimensional marginals. Note that all univariate marginals can uniquely be derived from the bivariate ones.

An essential advantage of the tree-dependence model (3) is a simple solution of the underlying structural optimization problem. For this purpose Chow and Liu first introduced a measure of approximation closeness. In particular, the optimal approximation $P(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\theta})$ of a probability distribution $P^*(\boldsymbol{x})$ should minimize the Kullback-Leibler information divergence (cf. [17])

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = \sum_{x\in X} P^*(\boldsymbol{x})\log\frac{P^*(\boldsymbol{x})}{P(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\theta})} \geq 0. \tag{4}$$

The information function (4) is not a metric but it is non-negative and equals zero if and only if $P^*(\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\theta})$ for all $\boldsymbol{x}\in\boldsymbol{X}$. By using substitution (3) we can write

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = -H(P^*) - \sum_{x_1=0}^{1} p^*(x_1)\log p(x_1) -$$

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

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = -H(P^*) - \sum_{x_1=0}^{1} p^*(x_1)\log p(x_1) - \tag{5}$$

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

In the previous equation the Shannon entropy $H(P^*)$ is a constant

$$H(P^*) = \sum_{x \in X} -P^*(\boldsymbol{x}) \log P^*(\boldsymbol{x}) \tag{6}$$

and, in order to minimize the information divergence, we have to maximize the last two terms which correspond to the log-likelihood function for the dependence-tree distribution (3). It can be seen (cf. [6], Eq. (46)) that, for any fixed dependence structure $\boldsymbol{\alpha}$, the Kullback-Leibler information divergence $I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha}, \boldsymbol{\theta}))$ is minimized by the two-dimensional marginals $\boldsymbol{\theta}^* = \{p^*(x_n, x_{k_n}), n = 2, \ldots, N\}$:

$$p(x_1) = p^*(x_1), \quad p(x_n|x_{k_n}) = \frac{p^*(x_n, x_{k_n})}{p^*(x_{k_n})}. \tag{7}$$

Making substitution (7) into (5), we obtain

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha}, \boldsymbol{\theta})) = -H(P^*) + \sum_{n=1}^{N} H(p_n^*) - \sum_{n=2}^{N} \sum_{x_n=0}^{1} \sum_{x_{k_n}=0}^{1} p^*(x_n, x_{k_n}) \log \frac{p^*(x_n, x_{k_n})}{p^*(x_n) p^*(x_{k_n})}.$$

In the last formula $H(p_n^*)$ are the respective marginal Shannon entropies (cf. (6))

$$H(p_n^*) = \sum_{x_n=0}^{1} -p^*(x_n) \log p^*(x_n),$$

the sum of which is structure independent. Thus the Kullback-Leibler information divergence $I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha}, \boldsymbol{\theta}))$ is minimized by maximizing the sum of Shannon mutual information values $\mathcal{I}(p_n^*, p_{k_n}^*)$ between the respective variables $x_n, x_{k_n},\ n = 2, \ldots, N$

$$\mathcal{I}(p_n^*, p_{k_n}^*) = \sum_{x_n=0}^{1} \sum_{x_{k_n}=0}^{1} p^*(x_n, x_{k_n}) \log \frac{p^*(x_n, x_{k_n})}{p^*(x_n) p^*(x_{k_n})}. \tag{8}$$

In other words, the optimal dependence structure $\boldsymbol{\alpha}^*$ has to satisfy the condition

$$\boldsymbol{\alpha}^* = \arg\max_{\alpha} \Big\{ \sum_{n=2}^{N} \mathcal{I}(p_n^*, p_{k_n}^*) \Big\}. \tag{9}$$

As shown by Chow and Liu the optimal tree-dependence structure can be found as a maximum weight spanning tree over the complete graph of vertices $\mathcal{N}$ with edge-weights $\mathcal{I}(p_n^*, p_{k_n}^*)$. The maximum-weight spanning tree $\boldsymbol{\alpha}^*$ can be constructed, e.g., by the algorithm of Boruvka-Kruskal [1], [16] as proposed by Chow and Liu. Nevertheless, from the computational point of view, the algorithm of Prime [20] could be preferable, because it does not need any ordering of edge weights.

Let us remark that, in the past, the approximation model (3) has been studied in more general forms including higher-order marginals (cf. [6]).

$$P(\boldsymbol{x}) = p(x_{i_1}) \prod_{n=2}^{N} p(x_{i_n}|\boldsymbol{x}_{B_{i_n}}), \quad B_{i_n} \subset \{i_1, \ldots, i_{n-1}\}.$$

There are also related papers from the area of knowledge-based systems and Bayesian networks (cf. detailed references in [18]). However, the problems arising from the underlying structural optimization become exceedingly difficult from the computational point of view.

2.1 Approximating Densities by Dependence Trees

The original paper of Chow and Liu applies to discrete distributions but, from the formal point of view, the idea of dependence-tree approximation is applicable to continuous data as well [6]. Considering real data vectors $\boldsymbol{x} \in \boldsymbol{X} \equiv \mathcal{R}^{\mathcal{N}}$ we have to approximate a given probability density function $P^*(\boldsymbol{x})$ by the dependence-tree density function

$$P(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = f(x_1)\prod_{n=2}^{N} f(x_n|x_{k_n}). \tag{10}$$

To measure the closeness between the given probability density $P^*(\boldsymbol{x})$ and its dependence-tree approximation we can use the continuous version of Kullback-Leibler information divergence (cf. [15], [22]):

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = \int_{R^N} P^*(\boldsymbol{x}) \log \frac{P^*(\boldsymbol{x})}{P(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\theta})} d\boldsymbol{x} \geq 0. \tag{11}$$

In analogy with (5) we can write

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = \int P^*(\boldsymbol{x}) \log P^*(\boldsymbol{x}) d\boldsymbol{x} - \int P^*(\boldsymbol{x}) \left[\log f(x_1) + \sum_{n=2}^{N} \log f(x_n|x_{k_n})\right] d\boldsymbol{x},$$

and further

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta})) = -H(P^*) - \int_R f^*(x_1) \log f(x_1) dx_1 - \tag{12}$$

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

In the last equation the entropy $H(P^*)$ is a constant and, in order to minimize the information divergence, we have to maximize the last two terms. It can be seen that, for a fixed dependence structure $\boldsymbol{\alpha}$, the information divergence $I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta}))$ is minimized by the two-dimensional marginal densities $\boldsymbol{\theta}^* = \{f^*(x_n, x_{k_n}), n = 2, \ldots, N\}$ which imply the involved univariate marginals:

$$f(x_1) = f^*(x_1), \quad f(x_n|x_{k_n}) = \frac{f^*(x_n, x_{k_n})}{f^*(x_{k_n})}. \tag{13}$$

Again, substituting in (12) according to (13), we obtain

$$I(P^*(\cdot)||P(\cdot|\boldsymbol{\alpha},\boldsymbol{\theta}^*)) = -H(P^*) + \sum_{n=1}^{N} H(f_n^*) - \sum_{n=2}^{N} \mathcal{I}(f_n^*, f_{k_n}^*) \tag{14}$$

where $H(P^*)$ is a constant entropy (cf. (6)) and the second sum is structure-independent

$$\sum_{n=1}^{N} H(f_n^*) = \sum_{n=1}^{N} \int_R -f^*(x_n) \log f^*(x_n) dx_n.$$

Consequently, the considered information divergence is minimized by maximizing the sum of mutual information terms

$$\mathcal{I}(f_n^*, f_{k_n}^*) = \int_{R^2} f^*(x_n, x_{k_n}) \log \frac{f^*(x_n, x_{k_n})}{f^*(x_n) f^*(x_{k_n})}\, dx_n dx_{k_n} = H(f_n^*) + H(f_{k_n}^*) - H(f_{nk_n}^*) \quad (15)$$

as a function of the dependence structure $\boldsymbol{\alpha}$. In other words, the optimal dependence structure $\boldsymbol{\alpha}^*$

$$\boldsymbol{\alpha}^* = \arg\max_{\alpha} \Big\{ \sum_{n=2}^{N} \mathcal{I}(f_n^*, f_{k_n}^*) \Big\}$$

of the approximation model $P(\cdot|\boldsymbol{\alpha}, \boldsymbol{\theta}^*)$ is defined by the maximum-weight spanning tree over the complete graph of vertices $\mathcal{N}$ with the edge-weights $\mathcal{I}(f_n^*, f_{k_n}^*)$. We note that, assuming Gaussian densities with the variances σ_n and covariances σ_{nk_n}, we obtain the Shannon mutual information formula:

$$\mathcal{I}(f_n^*, f_{k_n}^*) = -\frac{1}{2} \log \left(1 - \frac{\sigma_{nk_n}^2}{\sigma_n^2 \sigma_{k_n}^2} \right). \quad (16)$$

3 Estimating Gaussian Dependence-Tree Density

Let us recall the practical situation when the true probability density $P^*(\boldsymbol{x})$ is unknown and the dependence-tree approximation has to be constructed from data. Unlike the binary case (cf. Sec. 3.1), we have to assume parametric models of two-dimensional marginals $f(x_n, x_{k_n})$ in order to make the resulting dependence-tree model practically applicable. A natural choice here is to use two-dimensional Gaussian densities:

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

$$\Sigma_{nk} = \begin{pmatrix} \sigma_n^2 & \sigma_{nk} \\ \sigma_{nk} & \sigma_k^2 \end{pmatrix}, \quad n, k \in \mathcal{N}, \quad (17)$$

which imply the univariate marginals

$$f(x_n|\mu_n, \sigma_n) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left\{ -\frac{(x_n - \mu_n)^2}{2\sigma_n^2} \right\}, \quad n \in \mathcal{N}.$$

Considering the Gaussian dependence-tree density function with the structural parameters $\boldsymbol{\alpha}$, vector of means $\boldsymbol{\mu}$ and the covariance matrices $\boldsymbol{\Sigma}$:

$$\boldsymbol{\alpha} = (k_2, \ldots, k_N), \quad \boldsymbol{\mu} = \{\mu_1, \ldots, \mu_N\}, \quad \boldsymbol{\Sigma} = \{\Sigma_{nk_n}, n = 2, \ldots, N\}$$

we can write

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

$$= f(x_1|\mu_1, \sigma_1) \prod_{n=2}^{N} \frac{f(x_n, x_{k_n}|\mu_n, \mu_{k_n}, \Sigma_{nk_n})}{f(x_{k_n}|\mu_{k_n}, \sigma_{k_n})} \quad (18)$$

and the corresponding log-likelihood function is given by

$$L(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log P(\boldsymbol{x}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} [\log f(x_1|\mu_1, \sigma_1) -$$

$$- \sum_{n=2}^{N} \log f(x_{k_n}|\mu_{k_n}, \sigma_{k_n}) \Big] + \sum_{n=2}^{N} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log f(x_n, x_{k_n}|\mu_n, \mu_{k_n}, \Sigma_{nk_n}). \tag{19}$$

For any fixed dependence structure $\boldsymbol{\alpha}$ we can concentrate on estimating two-dimensional Gaussian densities which imply all univariate marginals and conditional densities. Using maximum-likelihood estimates of the underlying parameters, we obtain

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

and, making substitution in the formula (19), we can write

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

$$= \sum_{n=1}^{N} \frac{1}{2} \left[1 + \log(2\pi\hat{\sigma}_n^2)\right] + \sum_{n=2}^{N} \frac{1}{2} \log \left(1 - \frac{\hat{\sigma}_{nk_n}^2}{\hat{\sigma}_n^2 \hat{\sigma}_{k_n}^2}\right). \tag{20}$$

In the last equation only the second term is structure dependent and therefore the optimal dependence structure $\boldsymbol{\alpha}^*$ is defined by

$$\boldsymbol{\alpha}^* = \arg\max_{\alpha} \Big\{ \sum_{n=2}^{N} -\frac{1}{2} \log \left(1 - \frac{\hat{\sigma}_{nk_n}^2}{\hat{\sigma}_n^2 \hat{\sigma}_{k_n}^2}\right) \Big\}. \tag{21}$$

Note that the edge-weight of the underlying spanning tree is the same (cf. (16)) as in the deterministic approximation problem based on Gaussian densities.

Recall that, assuming Gaussian marginals $p(x_n, x_k)$ in the tree-dependence density function (10), we restrict the approximation power of the optimal tree dependence model $P(\boldsymbol{x}|\boldsymbol{\alpha}^*, \boldsymbol{\theta}^*)$ by the underlying global Gaussian hypothesis. The only advantage of the simplifying tree-dependence approximation (18) is then its applicability to high-dimensional spaces since the involved two-dimensional marginals can be well estimated even from limited data sets and avoid the risk of ill-conditioned high-dimensional matrices.

4 Mixtures of Dependence Trees

The product mixture model can be generalized by using mixtures of dependence trees (cf. [6], [18], [14]). We recall that the dependence tree distribution can explicitly describe statistical dependencies between pairs of variables at the level of individual components and therefore the approximation potential of the resulting mixture model may considerably increase. On the other hand, marginal distributions of the dependence-tree mixtures are not trivially available anymore and we lose some of the excellent properties of product

mixtures, as mentioned in the Introduction. Nevertheless, in some cases such properties may be unnecessary, while the increased complexity of components could become essential.

The tree-dependence model (3) can be easily generalized to mixtures (cf. [6], [18]) because the optimization of tree dependence structure proposed by Chow and Liu [3] is compatible with the EM algorithm. In particular, we assume the following concept of Gaussian dependence-tree mixtures

$$P(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) = \tag{22}$$

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

with the weight vector $\boldsymbol{w}=(w_1,w_2,\ldots,w_M)$, the structural parameters $\{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\ldots,\boldsymbol{\alpha}_M\}$ and the component parameters

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

$$\boldsymbol{\Sigma}=\{\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\ldots,\boldsymbol{\Sigma}_M\},\quad \boldsymbol{\Sigma}_m=\{\Sigma_{nk_n}^{(m)},n=2,\ldots,N\}.$$

Equation (22) can be equivalently rewritten in the form

$$P(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) = \tag{23}$$

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

Note that the approximation potential of the dependence-tree mixture (23) is no longer limited by the underlying Gaussian assumption.

To optimize the mixture of dependence-tree densities (23) we have to maximize the log-likelihood function

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

By using the EM algorithm, we reduce the mixture estimation problem to iterative maximization of the following weighted likelihood function (cf. Sec. 2):

$$Q_m(\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) = \sum_{x\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|}\log F(\boldsymbol{x}|\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) \tag{24}$$

with the conditional weights $q(m|\boldsymbol{x})$ and the corresponding component weights $w_m^{'}$:

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)}{P(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma})},\ w_m^{'} = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} q(m|\boldsymbol{x}),\quad (w_m^{'}|\mathcal{S}| = \sum_{x\in\mathcal{S}} q(m|\boldsymbol{x})). \tag{25}$$

Considering the formula (23), we can write

$$Q_m(\boldsymbol{\alpha}_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \sum_{x\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|} \log f(x_1|\mu_1^{(m)}, \sigma_1^{(m)}) +$$

$$+ \sum_{n=2}^{N} \sum_{x\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|} \log \frac{f(x_n, x_{k_n}|\mu_n^{(m)}, \mu_{k_n}^{(m)}, \Sigma_{nk_n}^{(m)})}{f(x_{k_n}|\mu_{k_n}^{(m)}, \sigma_{k_n}^{(m)})}. \tag{26}$$

Again, for any fixed dependence structure $\boldsymbol{\alpha}_m$, we may confine ourselves to estimating the Gaussian parameters $\mu_n^{(m)}, \sigma_n^{(m)}, \sigma_{nk}^{(m)}$ by using the weighted analogy of m.-l. estimates:

$$\mu_n^{'(m)} = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|} x_n, \quad (\sigma_n^{'(m)})^2 = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|} (x_n - \mu_n^{'(m)})^2, \tag{27}$$

$$\sigma_{nk}^{'(m)} = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \frac{q(m|\boldsymbol{x})}{w_m^{'}|\mathcal{S}|} (x_n - \mu_n^{'(m)})(x_k - \mu_k^{'(m)}), \quad n, k \in \mathcal{N}.$$

Making substitutions (27) into (26) in analogy with (20), we obtain

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

In the last equation only the second term is structure-dependent and therefore the optimal dependence tree $\boldsymbol{\alpha}_m^{'}$ is for each component defined by Eq.

$$\boldsymbol{\alpha}_m^{'} = \arg\max_{\alpha} \left\{ \sum_{n=2}^{N} -\frac{1}{2} \log \left(1 - \frac{(\sigma_{nk_n}^{'(m)})^2}{(\sigma_n^{'(m)}\sigma_{k_n}^{'(m)})^2}\right)\right\}. \tag{28}$$

The resulting EM algorithm for estimating Gaussian dependence-tree mixtures can thus be summarized by Eqs. (25), (27) and (28), (cf. [6], Eqs. (5.9)-(5.12)).

5 Preprocessing of Screening Mammograms

In order to illustrate the application possibilities of Gaussian dependence-tree mixtures we recompute our recent results on evaluation of screening mammograms simply by using dependence trees instead of product components. In the papers [12], [8], [11] we proposed preprocessing of screening mammograms by means of local statistical models with the aim of facilitating diagnostic evaluation.

The idea of the method is to emphasize diagnostically important details as "unusual" locations of high "novelty". First we estimate local statistical properties of gray levels in a suitably chosen search window in terms of a joint probability density. In particular, we use the data set produced by scanning a mammogram with a search window in order to estimate the density function in the form of a Gaussian mixture with product components. At the second stage we compute the value of the estimated mixture density at each position of the search window and display the corresponding log-likelihood value $\log P(\boldsymbol{x})$ as a gray level at the central pixel of the window. Thus the resulting "log-likelihood

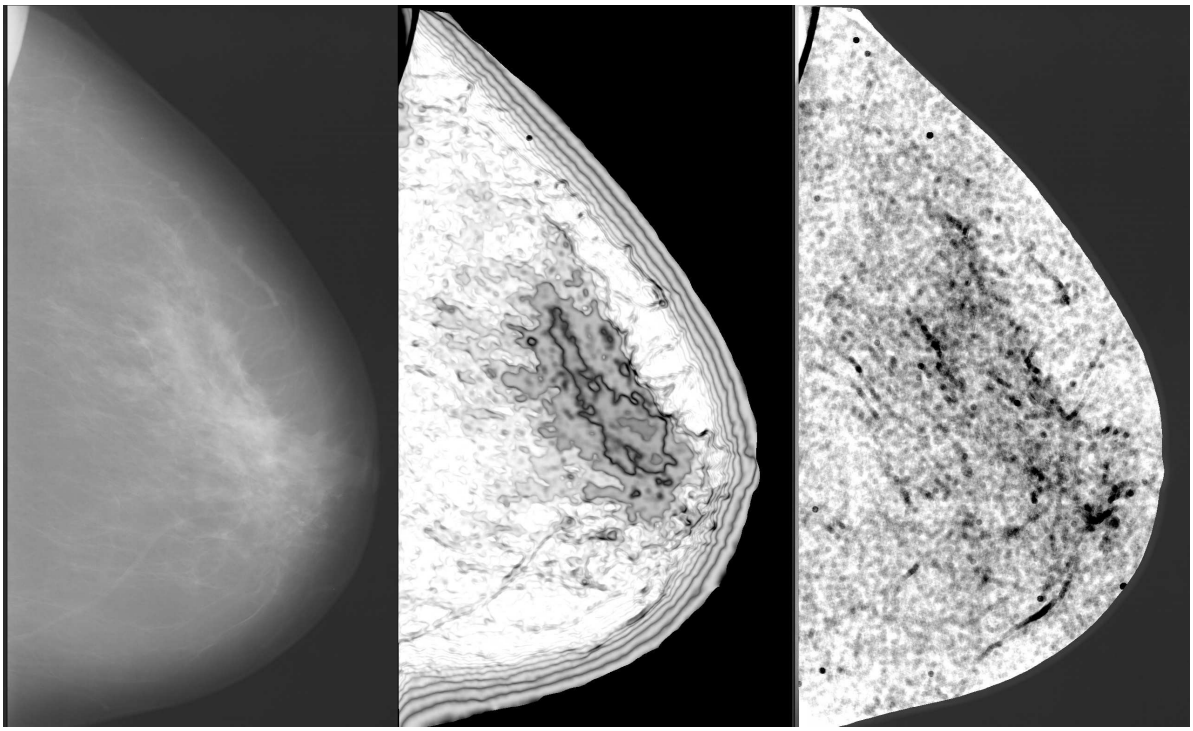

Figure 1: Comparison of the log-likelihood images for the digital mammogram C-0002-1 (left cranio-caudal part) from the DDSM database. From left to right: original mammogram, the log-likelihood image based on product mixture and the log-likelihood image based on mixture of dependence trees.

image" maps the unusual, atypical parts of the mammogram as dark regions and, in this way, the locations suspected of containing malignant lesions should be emphasized.

In the computational experiments we used mammograms from the DDSM database (cf. [13]) subsampled to the pixel size of about 0.1 mm (cf. Fig.1, left-hand part). The size of the search window was 13×13 pixels with trimmed corners. The resulting dimension of the local data vectors $\boldsymbol{x} \in \mathcal{S}$ was N=145 (=169-4×6). In all experiments we used the mixture model of 36 Gaussian components. A typical example of the log-likelihood image is shown in Fig. 1 (central part). Note that the dark regions are partly emphasized by contour lines and even small, barely visible micro-calcifications appear as dark spots (cf. [12] for detailed explanation).

The right-hand part of Fig.1 shows an analogous log-likelihood image obtained by means of a dependence-tree mixture. In particular, we used the same data to estimate the mixture of Gaussian dependence-tree densities (22). Not surprisingly, we obtained a much higher value of the maximized log-likelihood criterion using the same number of components (M=36, L=-249.1 versus L=-441.5), with the corresponding much better approximation of the underlying multivariate density. Nevertheless, the resulting log-likelihood image (cf. the right-hand part of Fig.1) tends to break down to small nearly point-like disconnected regions. The identification of micro-calcifications seems to contrast even more, but the contour lines vanish almost completely.

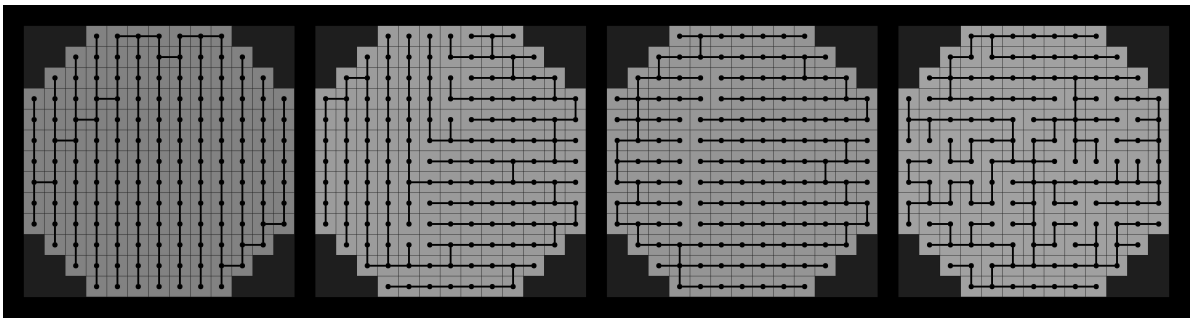

Figure 2: Gaussian dependence-tree mixture for local statistical model of screening mammogram - examples of estimated component means μ_{mn} in the search-window arrangement. The superimposed dependence-tree structures reflect statistical relations between the window fields.

It can be concluded that a simple product mixture model provides highly specific log-likelihood images characterized by connected dark regions, which are partly emphasized by contour lines (cf. [12] for detailed discussion). The typical features of the images can be explained by a certain "topological" continuity of product mixtures. Recall that a one-pixel shift of the search window generally yields a completely different data vector (despite the great overlap), because the shared gray-levels are assigned to different variables. Consequently, the likelihood values of neighboring window positions may generally differ, even by many orders. However, the differences are partly suppressed in the case of product components because the means $\mu_n^{(m)}$ are almost uniform for any given component. Thus the shift of the search window by one pixel actually changes only the order of product terms which are nearly the same and therefore, unlike dependence-tree components, the resulting product does not change very much.

Fig.2 illustrates the fact that the component means (in window arrangement) are almost uniform. The superimposed dependence-tree structures reflect statistical relations between the window fields. In view of the smooth background the most informative pairs of variables are nearly always neighboring.

6 Conclusion

It is intuitively clear that we can increase the approximation potential of density mixtures by using dependence-tree components instead of product densities. However, it appears that the increased approximation power of dependence trees is more relevant in case of a small number of multidimensional components. A large number of components makes the properties of density mixtures more similar to non-parametric Parzen estimates which are often optimized by a single smoothing parameter and very simple kernel functions. Accordingly, in the early paper [6] we observed in a numerical example, that the high approximation accuracy of dependence-tree mixture is simply achievable by increasing the number product components.

Acknowledgement

This work has been supported by the Czech Science Foundation Projects No. 14-02652S and 14-10911S.

References

[1] O. Boruvka, "On a minimal problem", *Transaction of the Moravian Society for Natural Sciences* (in czech), No. 3, 1926.

[2] B. Behsaz and M. Rahmati. "Estimation of Probability Density Function by Dependence Tree Methods for Pattern Recognition Systems." *Tech. Rep. U. Alberta*, 1, 2006.

[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.

[4] 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.

[5] 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*

[6] J. Grim, "On structural approximating multivariate discrete probability distributions", *Kybernetika*, Vol. 20, No. 1, pp. 1-17, 1984. *http://dml.cz/dmlcz/125676*

[7] J. Grim, "Multivariate statistical pattern recognition with nonreduced dimensionality", *Kybernetika*, Vol. 22, No. 2, pp. 142-157, 1986. *http://dml.cz/dmlcz/125022*

[8] 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

[9] J. Grim, M. Haindl, "Texture Modelling by Discrete Distribution Mixtures", *Computational Statistics and Data Analysis*, Vol. 41, No. 3-4 pp. 603-615, 2003.

[10] J. Grim and J. Hora, "Computational Properties of Probabilistic Neural Networks", *Artificial Neural Networks - ICANN 2010 Part II*, Springer: Berlin, LNCS 5164, pp. 52-61, 2010.

[11] J. Grim and G.L. Lee, "Evaluation of Screening Mammograms by Local Structural Mixture Models", In *Stochastic and Physical Monitoring Systems SPSM 2012*, Prague: Czech Technical University, pp. 51-61, 2012.

[12] J. Grim, P. Somol, M. Haindl and J. Daneš, "Computer-Aided Evaluation of Screening Mammograms Based on Local Texture Models," *IEEE Trans. on Image Processing*, Vol. 18, No. 4, pp. 765-773, 2009.

[13] M. Heath, K. W. Bowyer, and D. Kopans et al., "Current State of the Digital Database for Screening Mammography," *Digital Mammography*, Kluwer Academic Publishers, pp. 457460, 1998.

[14] 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.

[15] 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.

[16] J.B Kruskal, "On the shortest spanning sub-tree of a graph", *Proc. Amer. Math. Soc.*, No. 7, pp. 48-50, 1956.

[17] S. Kullback and R.A. Leibler, "On Information and Sufficiency", *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79-86, 1951.

[18] M. Meila and M.I. Jordan, "Learning with mixtures of trees", *Journal of Machine Learning Research*, Vol. 1, No. 9, pp. 1-48, 2001.

[19] E. Parzen, "On estimation of a probability density function and its mode," *Annals of Mathematical Statistics*, Vol. 33., pp. 1065-1076, 1962.

[20] R.C. Prim, "Shortest connection networks and some generalizations", *Bell System Tech. J.*, Vol. 36 , pp. 1389-1401, 1957.

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

[22] I. Vajda, *Theory of statistical inference and information*, Kluwer Academic Publishers (Dordrecht and Boston), 1989.