A Reparameterization of Mixtures of Truncated Basis Functions and its Applications

Antonio Salmerón antonio.salmeron@ual.es
Department of Mathematics and
Centre for the Development and Transfer of Mathematical Research to Industry (CDTIME)
University of Almería (Spain)

Helge Langseth helge.langseth@ntnu.no
Department of Computer Science
Norwegian University of Science and Technology, Trondheim (Norway)

Andrés Masegosa arma@cs.aau.dk
Department of Computer Science
Aalborg University, Copenhagen (Denmark)

Thomas D. Nielsen tdn@cs.aau.dk
Department of Computer Science
Aalborg University, Aalborg (Denmark)

Abstract

Mixtures of truncated basis functions (MoTBFs) are a popular tool within the context of hybrid Bayesian networks, mainly because they are compatible with efficient probabilistic inference schemes. However, their standard parameterization allows the presence of negative mixture weights as well as non-normalized mixture terms, which prevents them from benefiting from existing likelihood-based mixture estimation methods like the EM algorithm. Furthermore, the standard parameterization does not facilitate the definition of a Bayesian framework ideally allowing conjugate analysis. In this paper we show how MoTBFs can be reparameterized applying a strategy already used in the literature for Gaussian mixture models with negative terms. We exemplify how the new parameterization is compatible with the EM algorithm and conjugate analysis.

Keywords: Mixtures of truncated basis functions ; hybrid Bayesian networks ; parameter learning ; EM algorithm.

1. Introduction

Mixtures of truncated basis functions (MoTBFs) (Langseth et al., 2012a) provide a flexible framework for handling hybrid Bayesian networks, i.e., Bayesian networks where discrete and continuous variables coexist. MoTBFs generalize two other models previously proposed within the context of hybrid Bayesian networks, namely the so-called mixtures of truncated exponentials (MTEs) (Moral et al., 2001) and mixtures of polynomials (MoPs) (Shenoy and West, 2011; López-Cruz et al., 2012).

MoTBFs admit hybrid Bayesian networks with no structural restrictions on the relations between the continuous and discrete variables, unlike conditional Gaussian (CG) models (Lauritzen, 1992), where discrete variables are not allowed to have continuous parents. Furthermore, MoTBFs are closed under addition, multiplication, and integration, which facilitates the use of exact probabilistic inference methods like the Shenoy-Shafer

architecture (Shenoy and Shafer, 1990) or the variable elimination algorithm (Zhang and Poole, 1996).

Methods for learning MoTBFs from data have previously been studied and cover algorithms for learning both marginal (Langseth et al., 2012a) and conditional MoTBF densities (Langseth et al., 2009, 2014; Pérez-Bernabé et al., 2015). However, there are still some limitations in relation to learning. For instance, given that MoTBFs can include negative terms which are themselves not proper densities (they are not normalized), it is not possible to use popular mixture learning approaches based on optimizing the likelihood function, remarkably the EM algorithm. Also, MoTBFs lack of a Bayesian formulation allowing the definition of prior distributions on the parameters that could be updated when new data arrived.

The presence of negative mixture terms has been successfully addressed in the case of Gaussian mixture models by Zhang and Zhang (2005), who propose a reparameterization of the mixture density compatible with an iterative EM algorithm. However, such parameterization is not directly applicable to MoTBFs, because the latter contain mixture terms that are not proper densities. In this paper, we define a similar reparameterization for MoTBFs and illustrate that it can serve as a basis for developing an EM-inspired parameter estimation algorithm, as well as a Bayesian formulation of the learning problem.

2. Preliminaries

The *mixture of truncated basis functions* framework (Langseth et al., 2012b) is based on the abstract notion of real-valued *basis function*, which includes both polynomial and exponential functions as special cases. It is formally defined as follows.

Definition 1 *Let $\mathbf{X}$ be a mixed n-dimensional random vector. Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of $\mathbf{X}$, respectively, with $c + d = n$. Let $\Psi = \{\psi_i(\cdot)\}_{i=0}^{\infty}$ with $\psi_i : \mathbb{R} \to \mathbb{R}$ define a collection of real basis functions. We say that a function $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ is a* mixture of truncated basis functions *potential to level k wrt. Ψ if one of the following two conditions holds:*

- *f can be written as*

$$f(\mathbf{x}) = f(\mathbf{y}, \mathbf{z}) = \sum_{i=0}^{k} \prod_{j=1}^{c} \theta_{i,\mathbf{y}}^{(j)} \psi_i\left(z_j\right), \tag{1}$$

  where $\theta_{i,\mathbf{y}}^{(j)}$ are real numbers.

- *There is a partition $\Omega_{\mathbf{X}}^1, \ldots, \Omega_{\mathbf{X}}^m$ of $\Omega_{\mathbf{X}}$ for which the domain of the continuous variables, $\Omega_{\mathbf{Z}}$, is divided into hyper-cubes and such that f is defined as*

$$f(\mathbf{x}) = f_\ell(\mathbf{x}) \quad \text{if} \quad \mathbf{x} \in \Omega_{\mathbf{X}}^\ell,$$

  where each f_ℓ, $\ell = 1, \ldots, m$ can be written in the form of Eq. (1)*.*

An MoTBF potential is a *density* if $\sum_{\mathbf{y}\in\Omega_{\mathbf{Y}}}\int_{\Omega_{\mathbf{z}}} f(\mathbf{y},\mathbf{z})d\mathbf{z} = 1$. Similarly, an MoTBF $f(\mathbf{y},\mathbf{z})$ is a *conditional density* for $\mathbf{Z}' \subseteq \mathbf{Z}$ and $\mathbf{Y}' \subseteq \mathbf{Y}$ given $\mathbf{Z} \setminus \mathbf{Z}'$ and $\mathbf{Y} \setminus \mathbf{Y}'$ if $\sum_{\mathbf{y}'\in\Omega_{\mathbf{Y}'}}\int_{\Omega_{\mathbf{z}'}} f(\mathbf{y}',\mathbf{y}'',\mathbf{z}',\mathbf{z}'')d\mathbf{z}' = 1$, for all $\mathbf{z}'' \in \Omega_{\mathbf{Z}\setminus\mathbf{Z}'}$ and $\mathbf{y}'' \in \Omega_{\mathbf{Y}\setminus\mathbf{Y}'}$. Following Langseth et al. (2012b) we furthermore assume that the influence a set of continuous parent variables $\mathbf{Z}$ have on their child variable X is encoded only through the partitioning of $\Omega_{\mathbf{Z}}$ into hyper-cubes, and not directly in the functional form of $f(x|\mathbf{z})$ inside the hyper-cube $\Omega^j_{\mathbf{Z}}$. That is, for a partitioning $\mathcal{P} = \{\Omega^1_{\boldsymbol{Z}}, \ldots, \Omega^m_{\boldsymbol{Z}}\}$ of $\Omega_{\boldsymbol{Z}}$, the conditional MoTBF is defined for $\mathbf{z} \in \Omega^j_{\boldsymbol{Z}}$, $1 \leq j \leq m$, as

$$f_k^{(j)}(x|\mathbf{z} \in \Omega^j_{\boldsymbol{Z}}) = \sum_{i=0}^{k} \theta_i^{(j)}\, \psi_i^{(j)}(x). \tag{2}$$

In the remainder of this paper we shall assume that a conditional MoTBF density includes only a single 'head' variable, i.e., $|\mathbf{Z}' \cup \mathbf{Y}'| = 1$. Under this assumption, a hybrid Bayesian network can be fully specified just using univariate MoTBF densities. In nodes with no parents, a univariate MoTBF density would be specified. In nodes with parents, the conditional density is specified by giving a univariate MoTBF for each partition of the parents, as in Eq. (2). Thus, in this paper we will focus on univariate MoTBFs.

A remarkable particular case is obtained if we instantiate the basis functions as polynomials (i.e., $\psi_i(x) = x^i$, for $i = 0, \ldots, k$), in which case the MoTBF model reduces to an MoP (mixture of polynomials) model (Shenoy and West, 2011). Similarly, if we let $\Psi = \{1, e^{-x}, e^{x}, e^{-2x}, e^{2x}, \ldots\}$, the MoTBF model implements an MTE (mixture of truncated exponentials) model (Moral et al., 2001).

Typically, a univariate MoTBF for a variable X does not rely on a partitioning of Ω_X. Furthermore, in this work we will assume that all MoTBFs are defined on the unit interval. The next proposition shows the conditions under which an MoTBFs can be translated to the unit interval without loss of information and keeping the same set of basis functions, regardless of the domain in which it is initially defined.

Proposition 2 *Let X be a continuous random variable with univariate MoTBF density*

$$f_X(x) = \sum_{i=0}^{k} \theta_i \psi_i(x), \quad l < x < r,$$

and assume there exists a real function $h(l,r)$ such that for all the basis functions it holds that

$$\psi_i((r-l)x + l) = h(l,r)\psi_i(x), \quad i = 0, \ldots, k. \tag{3}$$

Then, the random variable $Y = \dfrac{X-l}{r-l}$ has a univariate MoTBF density defined on the unit interval as

$$f_Y(x) = \sum_{i=0}^{k} \theta_i (r-l) h(l,r) \psi_i(x), \quad 0 < x < 1. \tag{4}$$

Proof Since X takes values on (l,r), it is clear that $Y \in (0,1)$. Now, let us denote by F_X and F_Y the cummulative distribution functions of X and Y, respectively. Then,

$$F_Y(x) = P(Y \leq x) = P\left(\frac{X-l}{r-l} \leq x\right) = P(X \leq (r-l)x + l) = F_X((r-l)x + l). \tag{5}$$

On the other hand, if we let η_i be a primitive function of ψ_i, $i = 0, \ldots, k$, we can write

$$\begin{aligned} F_X(x) = P(X \leq x) = \int_l^x f_X(t)dt = \int_l^x \left(\sum_{i=0}^k \theta_i \psi_i(t) \right) dt = \sum_{i=0}^k \theta_i \int_l^x \psi_i(t)dt \\ = \sum_{i=0}^k \theta_i [\eta_i(t)]_{t=l}^{t=x} = \sum_{i=0}^k \theta_i (\eta_i(x) - \eta_i(l)), \quad l \leq x \leq r. \end{aligned} \tag{6}$$

Therefore, it follows from Eqs. (5) and (6) that

$$F_Y(x) = F_X((r-l)x + l) = \sum_{i=0}^k \theta_i (\eta_i((r-l)x + l) - \eta_i(l)), \quad 0 \leq x \leq 1,$$

and thus

$$\begin{aligned} f_Y(x) = \frac{\partial}{\partial x} F_Y(x) = \sum_{i=0}^k \theta_i (r-l) \eta_i'((r-l)x + l) \\ = \sum_{i=0}^k \theta_i (r-l) \psi_i((r-l)x + l) = \sum_{i=0}^k \theta_i (r-l) h(l,r) \psi_i(x), \quad 0 \leq x \leq 1. \end{aligned} \tag{7}$$

■

It can be easily checked that for both MTEs and MoPs, it is possible to find the corresponding function $h(l, r)$.

3. A Reparameterization of MoTBFs

From the point of view of parameter estimation, mixtures with negative components are difficult to handle, because they do not admit a straightforward application of the EM algorithm, since the weights of the terms in the mixture can no longer be regarded as probability values. There has been, however, a successful attempt to adapt the EM algorithm to Gaussian mixture models with negative components, based on re-parameterizing the original mixture density (Zhang and Zhang, 2005). This process assumes that the mixture components are *proper densities*, meaning that each component is non-negative and integrates to one and that each weight is non-negative. This is not the case for MoTBFs, for which the basis functions ψ_i, $i = 0, \ldots, k$ are not necessarily densities (they are not normalized and may not be positive over the full domain) and the weights θ_i, $i = 0, \ldots, k$ may violate the positivity constraint.

In this section we adapt the reparameterization scheme introduced by Zhang and Zhang (2005) to MoTBFs. The next theorem is the key result towards such reparameterization.

Theorem 3 *Let $f(x) = \sum_{i=0}^k \theta_i \psi_i(x)$ be an MoTBF density on the unit interval. Then, there exist a real number $a > 0$ and two proper mixture densities $p^+(x)$ and $p^-(x)$ on the unit interval such that f can be written as*

$$f(x) = (1 + a)p^+(x) - ap^-(x), \quad 0 < x < 1. \tag{8}$$

Proof For each $i = 0, \ldots, k$, consider two positive real numbers θ_i^+ and θ_i^- such that $\theta_i = \theta_i^+ - \theta_i^-$. Now, we define

$$f^+(x) = \sum_{i=0}^{k} \theta_i^+ \psi_i(x)$$

and

$$f^-(x) = \sum_{i=0}^{k} \theta_i^- \psi_i(x),$$

so that $f(x)$ can be written as $f(x) = f^+(x) - f^-(x)$. Now let $c_i = \int_0^1 \psi_i(x)dx, \quad i = 0, \ldots, k$. Then,

$$f^+(x) = \sum_{i=0}^{k} \theta_i^+ c_i \frac{1}{c_i} \psi_i(x)$$

and

$$f^-(x) = \sum_{i=0}^{k} \theta_i^- c_i \frac{1}{c_i} \psi_i(x).$$

By denoting $a_i^+ = \theta_i^+ c_i$ and $a_i^- = \theta_i^- c_i$, we get

$$f^+(x) = \sum_{i=0}^{k} a_i^+ \frac{1}{c_i} \psi_i(x)$$

and

$$f^-(x) = \sum_{i=0}^{k} a_i^- \frac{1}{c_i} \psi_i(x).$$

Note that

$$\begin{aligned} 1 &= \int_0^1 f(x)dx = \int_0^1 (f^+(x) - f^-(x))dx \\ &= \sum_{i=0}^{k} \left(a_i^+ \int_0^1 \frac{1}{c_i} \psi_i(x)dx - a_i^- \int_0^1 \frac{1}{c_i} \psi_i(x)dx \right) = \sum_{i=0}^{k} (a_i^+ - a_i^-). \end{aligned}$$

If we denote $a = \sum_{i=0}^{k} a_i^-$, then $1 = \sum_{i=0}^{k} a_i^+ - a \Rightarrow \sum_{i=0}^{k} a_i^+ = (1 + a)$.

Let $\beta_i^+ = \dfrac{a_i^+}{1+a}$ and $\beta_i^- = \dfrac{a_i^-}{a}$, $i = 0, \ldots, k$. Then,

$$\sum_{i=0}^{k} \beta_i^+ = \sum_{i=0}^{k} \beta_i^- = 1. \tag{9}$$

If we define

$$\begin{aligned} p^+(x) &= \sum_{i=0}^{k} \beta_i^+ \frac{1}{c_i} \psi_i(x), \\ p^-(x) &= \sum_{i=0}^{k} \beta_i^- \frac{1}{c_i} \psi_i(x), \end{aligned}$$

then it holds that $f^+(x) = (1+a)p^+(x)$ and $f^-(x) = ap^-(x)$. Hence, any MoTBF f can be written as

$$f(x) = (1+a)p^+(x) - ap^-(x),$$

where $a > 0$ and $p^+(x)$ and $p^-(x)$ are proper mixture densities, since the weights sum to 1 (see Eq. (9)) and $\frac{1}{c_i}\psi_i(x), i = 0, \ldots, k$ are proper densities (they are non negative and their integral on the unit interval is equal to one). ■

According to Theorem 3, instead of the initial parameters $\theta_i, i = 0, \ldots, k$, the new parameters after the reparameterization are $a, \beta_i^+, \beta_i^-, c_i$, $i = 0, \ldots, k$. However, notice that c_i is defined as $c_i = \int_0^1 \psi_i(x)dx$. If the basis functions ψ_i are polynomials, it therefore holds that

$$\int_0^1 x^i dx = \frac{1}{i+1} \Rightarrow c_i = i+1.$$

On the other hand, if the basis functions are exponentials, we find that

$$\int_0^1 e^{ix} dx = \frac{e^i - 1}{i} \Rightarrow c_i = \frac{i}{e^i - 1}.$$

Therefore, for the particular cases of MTEs and MoPs, the c_i are not really free parameters, so the only parameters in the model are a, β_i^+, β_i^-, $i = 0, \ldots, k$.

The reparameterization in Theorem 3 has interesting potential applications both from the point of view of modelling and parameter estimation. In the next sections we will discuss both potential applications.

4. Iterative Parameter Estimation

The reparameterization in Theorem 3 paves the way to the definition of an iterative parameter estimation algorithm, inspired on the EM, which is aimed at optimizing the likelihood function. We will restrict the discussion in this section to polynomial basis functions, i.e. MoP densities, but similar arguments can be developed for MTEs.

The iterative procedure consists of the following steps:

1. Start with a random initialization of a, β_i^+, β_i^-, $i = 0, \ldots, k$.
2. Iterate until convergence (i.e. until the likelihood is no longer increased):
   (a) **Estimation of p^+:** Estimate β_i^+ with a and β_i^- fixed.
   (b) **Estimation of p^-:** Estimate β_i^- with a and β_i^+ fixed.
   (c) **Estimation of a:** Estimate a with β_i^- and β_i^+ fixed.

Regarding the estimation of the parameters of p^+, it is important to take into account that p^+ is a proper mixture, and therefore we can (potentially) estimate the weights (parameters) using the EM algorithm. However, in order to do that we need a sample drawn from p^+, which is not directly available (we are assuming that our data come from the original distribution, whose density is $f(x)$ rather than $p^+(x)$). But note that

$$f(x) \quad = \quad (1+a)p^+(x) - ap^-(x) \Rightarrow p^+(x) = \frac{1}{1+a}f(x) + \frac{a}{1+a}p^-(x).$$

Therefore, since we have shown that $p^+(x)$ is infact a mixture of $f(x)$ and $p^-(x)$, we can obtain a sample from p^+ using the composition method (Rubinstein, 1981):

- Take the original sample (coming from f)
- Sample p^- (whose parameters are known in this point) and add the items to the original sample, so that the fraction of items from the original sample is $1/(1+a) \times 100\%$ in the final sample.

Once we have the sample from $p^+(x)$, we can apply the following updating rules in order to obtain the EM-estimates of the parameters β_i^+. Note that the only parameters to estimate are the mixture weights, since the polynomial term itself does not contain any free parameter. Therefore, the updating rule is just the standard updating rule for the mixture weights (Dempster et al., 1977). Let Z_j be a hidden discrete random variable with $Z_j = i$ meaning that the j-th item in the sample comes from the i-th component. The updating rule from iteration t to iteration $t+1$ (assuming polynomial basis functions) for the parameters of p^+ is

$$T_{i,j}^{(t)} = P(Z_j = i | X_j = x_j, \beta_i^{+(t)}) = \frac{\beta_i^{+(t)}(i+1)x_j^i}{\sum_{l=0}^k \beta_l^{+(t)}(l+1)x_j^l},$$

$$\beta_i^{+(t+1)} = \frac{\sum_{j=1}^n T_{i,j}^{(t)}}{\sum_{j=1}^n \sum_{l=0}^k T_{l,j}^{(t)}} = \frac{1}{n}\sum_{j=1}^n T_{i,j}^{(t)},$$

where x_j denotes the j-th item in the sample drawn from p^+ and n is the sample size.

The parameters of p^- can be estimated in a similar way, as long as we have a sample from p^-, and taking into account the updated parameters of p^+. To obtain the sample from p^-, we notice that

$$p^-(x) = \frac{1+a}{a}p^+(x) - \frac{1}{a}f(x).$$

We can sample from p^- using the instantiation of the acceptance-rejection method for mixtures with negative terms defined by Bignami and De Matteis (1971):

- Sample an x from p^+.
- Choose $u \sim \mathcal{U}(0,1)$.
- If $u \leq \dfrac{\frac{1+a}{a}p^+(x) - \frac{1}{a}f(x)}{\frac{1+a}{a}p^+(x)}$, accept x as an item drawn from p^-.

These steps are repeated until the desired sample size is obtained. Notice that checking the acceptance condition above requires the evaluation of $f(x)$, which is unknown. We propose to use the empirical density or a kernel estimation of it instead.

Finally, assuming p^+ and p^- fixed, we can apply again the EM algorithm taking into account that

$$p^+(x) = \frac{1}{1+a} f(x) + \frac{a}{1+a} p^-(x).$$

Therefore, we can draw a sample from p^+ and use the EM to estimate the weights in the mixture above, i.e. $\frac{1}{1+a}$ and $\frac{a}{1+a}$. In fact, we only need to estimate one of them as both of them can be obtained by subtracting the other to 1. Once the weight is estimated, we can finally obtain a.

In order to formulate the updating rules, we define $Y_j = 0$ if the j-th element of the sample comes from f (first term in the mixture) and $Y_j = 1$ if it comes from p^- (second term in the mixture). We also define

$$s^{(t)} = \frac{1}{1+a^{(t)}}$$

and $Y_{i,j}^{(t)} = P(Y_j = i | X_j = x_j, s^{(t)}), \quad i = 0, 1$. The updating rule for a is given by

$$\begin{aligned} Y_{0,j}^{(t)} &= \frac{s^{(t)} f(x_j)}{s^{(t)} f(x_j) + (1 - s^{(t)}) p^-(x_j)}, \\ s^{(t+1)} &= \frac{\sum_{j=1}^n Y_{0,j}^{(t)}}{\sum_{j=1}^n (Y_{0,j}^{(t)} Y_{1,j}^{(t)})} = \frac{1}{n} \sum_{j=1}^n Y_{0,j}^{(t)}, \\ a^{(t+1)} &= \frac{1}{s^{(t+1)}} - 1 = \frac{n}{\sum_{j=1}^n Y_{0,j}^{(t)}} - 1. \end{aligned}$$

A problem of the updating rule above is that it does not guarantee that the resulting estimation of f will be a valid density. In fact, it may happen that the resulting density becomes negative in some parts of the unit interval. In order to avoid that, a restriction on a must be imposed. Note that

$$f(x) = (1+a)p^+(x) - ap^-(x) \geq 0 \Rightarrow p^+(x) + a(p^+(x) - p^-(x)) \geq 0.$$

If $p^+(x) - p^-(x) < 0$ then, in order to guarantee non-negativity, the value of a must be such that

$$a \leq \frac{p^+(x)}{p^-(x) - p^+(x)} \quad \forall x \in (0, 1). \tag{10}$$

4.1 Examples

As a proof of concept of the iterative procedure described above, we have conducted four simple experiments. In the first two experiments we used samples drawn from MoPs with degree 4 (example 1) and 6 (example 2). Then we used the iterative procedure to estimate a MoP with the same number of parameters. In the other two experiments, we used samples drawn from Beta distributions, more precisely Be(0.5, 0.5) (example 3) and Be(5, 5) (example 4). In all the experiments, the sample size was set to 500 and the number of iterations of the procedure to 1000. The initial parameters were initialized in order to start from a uniform distribution on the unit interval. In each example, we also learnt a MoP with the

same number of parameters, but using the least squares algorithm introduced by Langseth et al. (2014), using the implementation available in the MoTBFs `R` package (Pérez-Bernabé et al., 2020).

The estimated densities are displayed in Fig. 1. As a proof of concept, the examples seem to indicate that the iterative algorithm is able to find solutions with fairly high log-likelihood (see Table 1) even though the results are not as good as the ones obtained by the least squares method implemented in the MoTBFs `R` package. In any case, the goal of this experiment was to show that the reparameterization in Theorem 3 is potentially useful for developing an EM-inspired parameter estimation algorithm for MoTBFs, rather than developing such an algorithm.

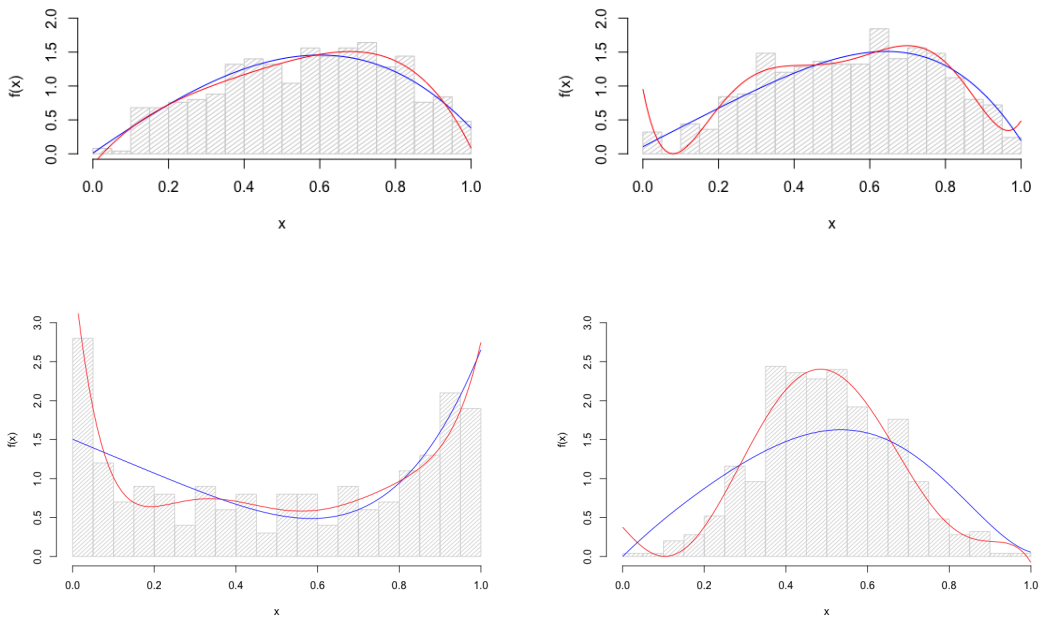

Figure 1: Histograms of the samples used in the examples, with the estimated MoPs overlaid. The upper row corresponds to examples 1 (left) and 2 (right), and the lower row to examples 3 (left) and 4 (right). The blue lines represent the MoPs estimated with the iterative EM-inspired procedure described in Section 4, and the red lines represent the MoPs estimated by least squares.

	Example 1	Example 2	Example 3	Example 4
Least squares	54.21318	74.55039	27.5904	198.0074
EM-inspired	58.91136	62.57779	18.65739	155.9439

Table 1: Log-likelihood of the models learnt in the examples.

5. A Bayesian Approach to Parameter Estimation

The estimation procedure outlined in Section 4 is based on optimizing the likelihood. However, adopting the parameterization in Theorem 3 it is also possible to define a Bayesian estimation scheme, where prior distributions on the parameters are specified and the estimation procedure consists of computing their posterior distribution given the data. From a practical point of view, the advantage of Bayesian inference with respect to a scheme based on the EM algorithm is that the former is usually much faster, specially if the prior distribution and the likelihood are *conjugate*, in which case the posterior distribution given the data can be typically computed in closed form.

Using the standard parameterization of MoTBFs in Eq. (1), it is difficult to define a Bayesian scheme where the parameters are considered as random variables. The reason is that the parameters can take any real value, while after the reparameterization, they take values on the unit interval. Furthermore, as we are considering all the MoTBFs to be defined on the unit interval, it is guaranteed that the posterior will also be defined in the same support, thus yielding always valid values for the parameters.

As an example of Bayesian formulation for MoTBFs, consider a random variable X with a degree 1 MoP as likelihood, given by

$$f(x|\beta_1^+) = (1+a)p^+(x) - ap^-(x), \quad 0 < x < 1, \tag{11}$$

with $p^+(x) = \beta_0^+ + \beta_1^+ x$ and $p^-(x) = \beta_0^- + \beta_1^- x$. Let us assume that the values of parameters a, β_0^- and β_1^- are known, which means that we only have one unknown parameter, say β_1^+, because the other one, β_0^+ is a function of β_1^+.

Assume we set a prior on β_1^+ which is also a MoP of degree 1, with initial parameters α_0^+ and α_1^+, i.e.

$$\pi(\beta_1^+) = \alpha_0^+ + \alpha_1^+ \beta_1^+, \quad 0 < \beta_1^+ < 1. \tag{12}$$

The posterior on β_1^+ given a data point x is

$$\begin{aligned}
\pi(\beta_1^+|x) &\propto f(x|\beta_1^+)\pi(\beta_1^+) = ((1+a)(\beta_0^+ + \beta_1^+ x) - ap^-(x))(\alpha_0^+ + \alpha_1^+\beta_1^+) \\
&= (\beta_0^+ + \beta_1^+ x + a\beta_0^+ + a\beta_1^+ x - ap^-(x))(\alpha_0^+ + \alpha_1^+\beta_1^+) \\
&= \beta_0^+\alpha_0^+ + \beta_0^+\alpha_1^+\beta_1^+ + \beta_1^+\alpha_0^+ x + \alpha_1^+{\beta_1^+}^2 x \\
&\quad + a\beta_0^+\alpha_0^+ + a\beta_0^+\alpha_1^+\beta_1^+ + a\alpha_0^+\beta_1^+ + a\alpha_1^+{\beta_1^+}^2 x \\
&\quad - ap^-(x)\alpha_0^+ - ap^-(x)\alpha_1^+\beta_1^+ \\
&= \beta_0^+\alpha_0^+ + a\beta_0^+\alpha_0^+ - ap^-(x)\alpha_0^+ \\
&\quad + \beta_0^+\alpha_1^+\beta_1^+ + x\alpha_0^+\beta_1^+ + a\beta_0^+\alpha_1^+\beta_1^+ + a\alpha_0^+\beta_1^+ - ap^-(x)\alpha_1^+\beta_1^+ \\
&\quad + \alpha_1^+{\beta_1^+}^2 + a\alpha_1^+ x{\beta_1^+}^2 \\
&= \theta_0^+ + \theta_1^+\beta_1^+ + \theta_2^+{\beta_1^+}^2,
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
\theta_0^+ &= \beta_0^+\alpha_0^+ + a\beta_0^+\alpha_0^+ - ap^-(x)\alpha_0^+, \\
\theta_1^+ &= \beta_0^+\alpha_1^+ + x\alpha_0^+ + a\beta_0^+\alpha_1^+ + a\alpha_0^+ - ap^-(x)\alpha_1^+, \\
\theta_2^+ &= \alpha_1^+ + a\alpha_1^+ x.
\end{aligned} \tag{14}$$

Therefore, the posterior on β_1^+ is again (up to a normalization constant) a MoP, in this case of second degree, whose parameters, θ_0^+, θ_1^+ and θ_2^+, can be obtained in closed form. However, note that the number of parameters is increased by one with respect to the prior distribution. Therefore, if we use the obtained posterior as new prior when new data arrives, the complexity of the posterior grows linearly with the amount of data.

6. Conclusions

We have introduced a reparameterization of MoTBFs in Theorem 3 that potentially paves the way to methodological advances in hybrid Bayesian networks. First, we illustrated how it facilitates the definition of an incremental EM algorithm (Neal and Hinton, 1998) for estimating the new parameters. The examples reported in this paper just served as proof of concept, and there is still work to do before it is shaped as a competitive parameter estimation algorithm for MoTBFs. For instance, in the examples run in this work, we have tested the restriction in Eq. (10) only in some points in the domain. A way to include such restrictions in the updating equations of the EM would likely improve the performance. Also, a way to avoid terms with very little weight, possibly by including some regularization term, is a promising way to improve our proposal. Note that the parameterization in Theorem 3 can also be formulated for multivariate MoTBFs, which is also another possibility to extend the work in this paper.

The application of the reparameterization to the definition of a Bayesian framework for parameter estimation also looks promising. However, the complexity issue that arises from the increase in the number of parameters of the posterior as new data arrives, is indeed challenging. One way to face it could be to carry out approximations when the complexity is too high. Similarly to the so-called *semi parametric Bayesian networks* (Atienza et al., 2022), where some conditionals are represented as Gaussians while some other are represented as kernel densities, we could consider fixed-complexity MoTBFs and seek the best approximation to the posterior within a given number of parameters when the complexity grows.

Acknowledgments

This research is part of Project PID2019-106758GB-C32 funded by MCIN/AEI/10.13039/501100011033, FEDER "Una manera de hacer Europa" funds, and was also partially funded by Junta de Andalucía grant P20-00091 and UAL-FEDER grant UAL2020-FQM-B1961.

References

D. Atienza, C. Bielza, and P. Larrañaga. Semiparametric Bayesian networks. *Information Sciences*, 584:564 – 582, 2022.

A. Bignami and A. De Matteis. A note on sampling from combinations of distributions. *IMA Journal of Applied Mathematics*, 8:80–81, 1971.

A. Dempster, N. Larid, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Seties B*, 39:1 – 38, 1977.

H. Langseth, T. Nielsen, R. Rumí, and A. Salmerón. Maximum likelihood learning of conditional MTE distributions. *ECSQARU 2009. Lecture Notes in Computer Science*, 5590:240–251, 2009.

H. Langseth, T. Nielsen, R. Rumí, and A. Salmerón. Inference in hybrid Bayesian networks with mixtures of truncated basis functions. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models (PGM'2012)*, pages 171–178, 2012a.

H. Langseth, T. Nielsen, R. Rumí, and A. Salmerón. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning*, 53(2):212–227, 2012b.

H. Langseth, T. Nielsen, I. Pérez-Bernabé, and A. Salmerón. Learning mixtures of truncated basis functions from data. *International Journal of Approximate Reasoning*, 55:940–956, 2014.

S. L. Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87:1098–1108, 1992.

P. L. López-Cruz, C. Bielza, and P. Larrañaga. Learning mixtures of polynomials from data using B-spline interpolation. In A. Cano, M. Gómez-Olmedo, and T. D. Nielsen, editors, *Proceedings of the 6th European Workshop on Probabilistic Graphical Models (PGM'12)*, pages 211–218, 2012.

S. Moral, R. Rumí, and A. Salmerón. Mixtures of truncated exponentials in hybrid Bayesian networks. In *ECSQARU'01. Lecture Notes in Artificial Intelligence*, volume 2143, pages 135–143, 2001.

R. M. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. Springer Netherlands, Dordrecht, 1998.

I. Pérez-Bernabé, A. Salmerón, and H. Langseth. Learning conditional distributions using mixtures of truncated basis functions. *ECSQARU'2015. Lecture Notes in Artificial Intelligence*, 9161:397–406, 2015.

I. Pérez-Bernabé, A. Maldonado, T. Nielsen, and A. Salmerón. MoTBFs: An R package for learning hybrid Bayesian networks using mixtures of truncated basis functions. *The R Journal*, 12:342–358, 2020.

R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley (New York), 1981.

P. P. Shenoy and G. Shafer. Axioms for probability and belief function propagation. In R. D. Shachter, T. S. Levitt, J. F. Lemmer, and L. N. Kanal, editors, *Uncertainty in Artificial Intelligence 4*, pages 169–198. North Holland, Amsterdam, 1990.

P. P. Shenoy and J. C. West. Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, 52:641–657, 2011.

B. Zhang and C. Zhang. Finite mixture models with negative components. In *MLDM 2005. Lecture Notes in Artificial Intelligence*, volume 3587, pages 31–41, 2005.

N. Zhang and D. Poole. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.