# Spatial Triplet Markov Trees for Auxiliary Variational Inference in Spatial Bayes Networks

Hugo Gangloff[1, 2], Jean-Baptiste Courbot[3], Emmanuel Monfrini[4], and Christophe Collet[1]

[1] ICube, Université de Strasbourg - CNRS UMR 7357, Illkirch, France
(E-mail: hugogangloff@unistra.fr, c.collet@unistra.fr)
[2] GEPROVAS, Strasbourg, France
[3] IRIMAS UR 7499 Université de Haute-Alsace, Mulhouse, France
(E-mail: jean-baptiste.courbot@uha.fr)
[4] SAMOVAR - CNRS UMR 5157, Évry, France
(E-mail: emmanuel.monfrini@telecom-sudparis.eu)

**Abstract.** In this article, we develop a Triplet Markov Tree model with auxiliary random variables for an approximate inference in an intractable probabilistic model. It is based on recent advances on probabilistic modeling and varational inference with auxiliary random variables. The new Triplet Markov Tree model performs better than the classical Mean-Field variational inference and than a tree-structured variational inference. Our study provides insights and motivations for the developping work around models involving auxiliary random variables.
**Keywords:** Variational Inference, Auxiliary Random Variables, Triplet Markov Models.

## 1 Introduction

### 1.1 Position of the problem

Modeling the relations between random variables of probabilistic models requires to find a compromise between tractability and richness of the correlations.

Markov tree-structured probabilistic models are therefore a popular choice because they offer the ability to introduce correlations while preserving tractability, in particular, the determistic retrieval of the marginals [14]. Highly structured trees, known as dyadic- or quadtree-Markov Trees (DMT) [6] [8] [11] [19] (Figure 1), have been used for the probabilistic treatment of mono-dimensional or bi-dimensional data series.

The modelization with Markov process has been generalized by using auxiliary variables in [4] [5] [12]. In these models, a carefully chosen auxiliary process enables the introduction of richer correlations while, preserving the Markov property and the model tractability.

Based on the successful structured and hierarchical dyadic Markov trees, we present the Spatial Bayes Network (SBN) model which introduces, with

$6^{th}$ *SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain*



respect to the DMT model, additional correlations between neighboring nodes (Figure 2), which might be an appealing feature for the practitioner. However, as soon as those additional correlations are established, the network contains many loops, inference becomes intractable and it must be approximated.

### 1.2 Variational Inference

Variational inference (VI) [9] [14] [3] [18] is an approach to perform inference in a complex probability distribution using a simpler one, called the *variational distribution*. The inference problem is then recast as an optimization problem. There exists several designs of VI. The Mean-Field (MF) VI is the most popular approach to VI: it consists in using a variational distribution with a fully factorized, unstructured, form [9]. While computation is easy in this context, the result of the optimization problem can fail to reflect key correlations between random variables, because of the lack of structure. Therefore, over the years, *structured* VI has been developped. It consists in using a structured variational distribution, in which computation is still tractable, in order to better approximate the correlations in the original intricate distribution. Structured VI can lead to dramatic increase in performances thanks to the improved modelling of the correlations [2] [7] [15]. Generally speaking, VI leads to a non-convex optimization problem. However, adding more structure makes some local optima disappear [18], hence the enhanced inference results. A recent trend in structured VI is to propose variational distribution using auxiliary variables, which leads to further improvements in the inference [1] [16] [17].

### 1.3 Outline of the paper

The main goal of this article is to develop a Spatial Triplet Markov Tree (STMT) model of [5] to exhibit the richness of the correlations it induces. We do so by using the STMT distribution in a VI procedure with auxiliary variables to approximate the intractable SBN distribution.

We first present the definitions of the probabilistic models mentioned in the introduction: DMT, SBN and STMT. We then develop the variational inference framework required to perform approximate inference in the intractable SBN. To the best of our knowledge, structured VI has never been studied with neither DMT nor STMT as the variational distribution. Finally we numerically show on an example the interest of STMT and auxiliary variable modeling to enhance the modelization.

## 2 Model definitions

### 2.1 Dyadic Markov Tree (DMT)

Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ be a random process which can be real or discrete-valued. We will refer to the realizations of the random variable with the notation $p(\boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x})$.

We define a tree graph over a set of nodes $\mathcal{S} = \{\mathcal{S}^0, \ldots, \mathcal{S}^{n-1}\}$, where each $\mathcal{S}^k$ is a resolution. $\mathcal{S}^0$ has a unique node $r$, the root node. Let $\bar{\mathcal{S}} = \mathcal{S} \setminus \mathcal{S}^0$, then $\forall s \in \bar{\mathcal{S}}$, a parent node of $s$ is denoted $s^-$. In *dyadic* trees each node has 2 sons. Note that Markov trees have no cycles; each node has exactly one parent.

A DMT (illustrated in Figure 1) has the following joint distribution [11]:

$$p(\boldsymbol{x}) = p(x_r) \prod_{s \in \bar{\mathcal{S}}} p(x_s | x_{s^-}). \tag{1}$$

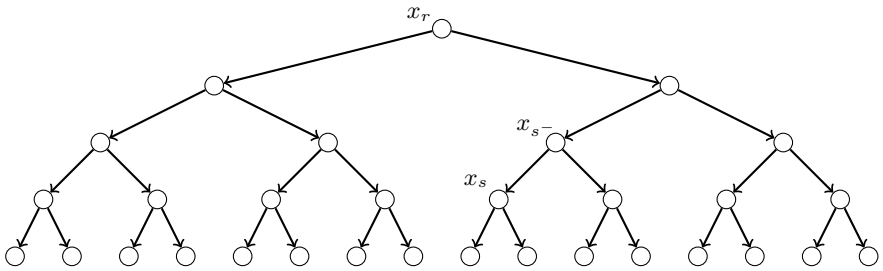


Fig. 1: Graphical model corresponding to a DMT. As an illustration, a node $x_s$, its father $x_{s^-}$ and the root node $x_r$ have been annotated.

In all tree structured graphs inference can be carried out exactly by a message passing technique [14]. In DMT, message passing is often stated in the form of the the upward-downward approach as given in [6].

### 2.2 Spatial Bayes Network (SBN)

We introduce a new model which we call Spatial Bayes Network (SBN). The model is a hierarchical Bayes network which contains cycles, in order to better capture local correlations between random variables.

In order to distinguish between the two sons of a father node $s^-$ in a dyadic tree: let $s_L$ and $s_R$ be respectively the *left* and *right* son of $s^-$. We also define $s^{\leftarrow}$ (resp. $s^{\rightarrow}$) as the left (resp. right) neighbouring node of $s$. Let $v \colon \bar{\mathcal{S}} \to \bar{\mathcal{S}}$ be a mapping from a node to a neighbouring node of its father, such that:

$$v \colon s \mapsto \begin{cases} (s^-)^{\leftarrow} \text{ if } s \text{ is a } \textit{left} \text{ node,} \\ (s^-)\rightarrow \text{ if } s \text{ is a } \textit{right} \text{ node,} \end{cases} \tag{2}$$

The SBN model (illustrated in Figure 2) has the following distribution:

$$p(\boldsymbol{x}) = p(x_{s_r}) \prod_{s \in \bar{\mathcal{S}}} p(x_s | x_{s^-}, x_{v(s)}). \tag{3}$$

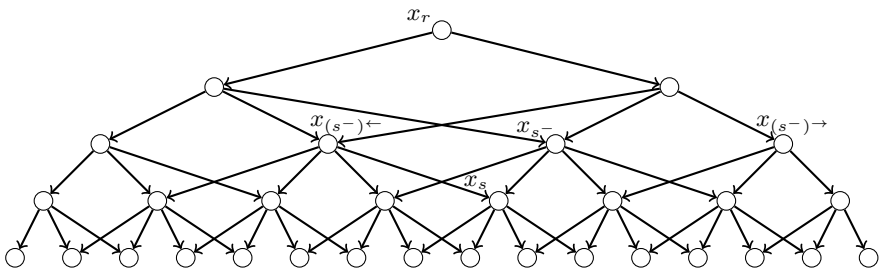


Fig. 2: Graphical model corresponding to a dyadic SBN. As an illustration, a node $x_s$, its father $x_{s^-}$, the left neighbour of its father $x_{(s^-)\leftarrow}$, the right neighbour of its father $x_{(s^-)\rightarrow}$ and the root node $x_r$ have been annotated.

### 2.3 Spatial Triplet Markov Tree (STMT)

Along with the previously introduced process $\boldsymbol{X}$, let $\boldsymbol{V} = (V_1, \ldots, V_N)$ be a (real- or discrete-valued) random process which is an *auxiliary* process. Let $\boldsymbol{T}$ be a process such that $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{V})$ and having the joint distribution:

$$p(\boldsymbol{t}) = p(\boldsymbol{t}_r) \prod_{s \in \bar{\mathcal{S}}} p(\boldsymbol{t}_s | \boldsymbol{t}_{s-}). \tag{4}$$

We now describe the hypothesis and the transitions to build a dyadic *Spatial* Triplet Markov Tree model as introduced in [5] (illustrated in Figure 4)[1]. A node at site $s$ is a 3-tuple $(X_s, V_s^{\leftarrow}, V_s^{\rightarrow})$; we then assume that we have the factorization, $\forall s \in \bar{\mathcal{S}}$:

$$\begin{aligned} p(\boldsymbol{t}_s | \boldsymbol{t}_{s-}) &= p(x_s, \boldsymbol{v}_s | x_{s^-}, \boldsymbol{v}_{s^-}), \\ &= p(x_s | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\leftarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\rightarrow} | x_{s^-}, \boldsymbol{v}_{s^-}). \end{aligned} \tag{5}$$

To further define the transition laws of each of the variable we define the notion of *inner* and *outer* variables for the $V$ variables. We define that, within *Left* (resp. *Right*) sons, $V_{s^L}^{\leftarrow}$ (resp. $V_{s^R}^{\rightarrow}$) is an outer variable and $V_{s^L}^{\rightarrow}$ (resp. $V_{s^L}^{\leftarrow}$) is an inner variable. Figure 3 illustrates these concepts for a particular node.

We now detail Equation 5, with special care on the variable type (*left*, *right*, *inner* or *outer*). For $X_s$ sons:

$$\begin{cases} p(x_s^L | x_{s^-}, \boldsymbol{v}_{s^-}) & = p(x_s^L | x_{s^-}, v_{s^-}^{\leftarrow}) \\ p(x_s^R | x_{s^-}, \boldsymbol{v}_{s^-}) & = p(x_s^R | x_{s^-}, v_{s^-}^{\rightarrow}), \end{cases} \tag{6}$$

for inner $V_s$ sons:

$$\begin{cases} p(v_{s^L}^{\rightarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) & = p(v_{s^L}^{\rightarrow} | x_{s^-}, v_{s^-}^{\rightarrow}) \\ p(v_{s^R}^{\leftarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) & = p(v_{s^R}^{\leftarrow} | x_{s^-}, v_{s^-}^{\leftarrow}), \end{cases} \tag{7}$$

[1] We keep refering to our model as *Spatial Triplet* as introduced in [5] because of its origin and the close definition: the original model is linked with image processing hence *Spatial* and it uses a triplet Markov tree. While in this article we only have two processes, this does not change the construction because the missing *observed* process only plays a marginal role in the original model.

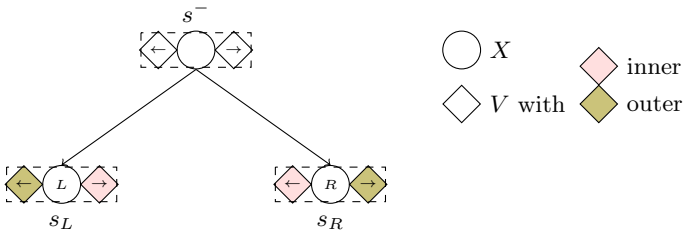


Fig. 3: Details of the STMT construction: a father node $s^-$ linked to its 2 sons $(s^L, s^R)$. The directionality of the $V_s$ is specified as well as their type (inner or outer).

and for outer $V_s$ sons:

$$\begin{cases} p(v^{\leftarrow}_{s^L}|x_{s^-}, \boldsymbol{v}_{s^-}) & = p(v^{\leftarrow}_{s^L}|v^{\leftarrow}_{s^-}, x_{s^-}) \\ p(v^{\rightarrow}_{s^R}|x_{s^-}, \boldsymbol{v}_{s^-}) & = p(v^{\rightarrow}_{s^R}|v^{\rightarrow}_{s^-}, x_{s^-}). \end{cases} \tag{8}$$

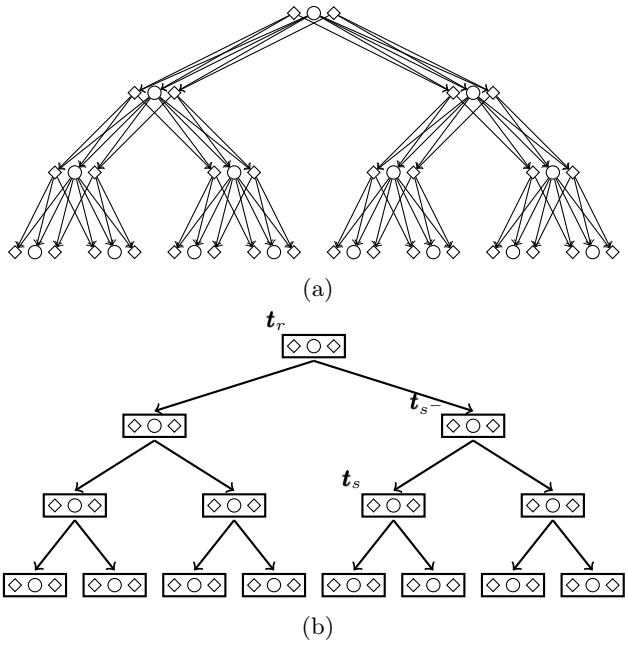


Fig. 4: Graphical model corresponding to a dyadic STMT. (a) depicts all the links, (b) gives a condensed view highlighting the preserved Markov tree strucuture. On (b), a node $\boldsymbol{t}_s$, its father $\boldsymbol{t}_{s^-}$ and the root node $\boldsymbol{t}_r$ have been annotated.

**Remark:** The conditioning of the $V_s$ variables can be seen as being the same conditioning as the $X$ variable it spatially "refers" to. The Triplet tree

framework then provides a way to simulate $X$ variables conditionally to the realizations of $V$ variables which behave similarly to the neighboring $X$ variables, hence the closeness with SBN we emphasize in this article.

A generalized version of the upward-downward algorithm of [6] enables us to compute deterministically the marginals in the STMT model [5].

## 3 Variational inference in SBNs

Within the Variational Inference (VI) framework [14], we aim at finding a distribution $q(\boldsymbol{x})$ which approximates well $p(\boldsymbol{x})$. We commonly seek to maximize the opposite Kullback-Leibler divergence $\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x}))$. By definition:

$$-\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x})) = \mathbb{E}_q[\log p(\boldsymbol{x})] - \mathbb{E}_q[\log q(\boldsymbol{x})]. \tag{9}$$

In this article, $p(\boldsymbol{x})$ is the joint distribution of a SBN (Section 2.2):

$$p(\boldsymbol{x}) = \prod_{s \in \mathcal{S}} p(x_s | x_{s^-}, x_{v(s^-)}) \tag{10}$$

In the context of structured variational inference, we will reparametrize and work with another representation of Equation 10. Indeed, we need to use a concise notation embedding the notion of clusters of variables for each of the terms of the product. The notations are adapted from [2]. We can write:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_{d_s} p_{d_s} \tag{11}$$

where $d_s$ represents the cluster of variables $(x_s, x_{s^-}, x_{v(s^-)})$ (note that these clusters overlap).

In the following we develop the necesary material to approximate Equation 10 with DMT and STMT, which benefit from their tree structure and deterministic marginalization. Therefore the approximations obtained by each of the models of interest will be caused by the enhanced correlations that some models have with respect to others.

Algorithmically, the steps of the VI procedure for a variational distribution $q$ can be summarized:

1. Initialize all the factors of $q$.
2. $\forall j \in \mathcal{S}$, update $q_{c_j}$ with the expression that minimizes the Kullback-Leibler divergence.
3. Check convergence and repeat step 2 if needed.

The convergence can be assessed, *e.g.*, by monitoring the values of the cost function or monitoring stationarity in the estimated variational parameters.

### 3.1 Structured variational inference with DMTs

The variational distribution $r$ is here defined with the structure of a Markov Tree (DMT) (Section 2.1):

$$r(\boldsymbol{x}) = \prod_{s \in \mathcal{S}} r(x_s | x_{s^-}) = \prod_{c_s} r_{c_s}, \tag{12}$$

where $c_s$ represents the cluster of variables $(x_s, x_{s^-})$. Using a MT structure as a posterior to approximate the more complex posterior $p$ is a compromise between additional structureand tractability. Indeed, once the posterior transitions will be learnt through the VI optimization, the posterior marginals are easily computable with the upwards-downwards algorithm [6].

The quantity to maximize (Equation 9) becomes in this section:

$$-\mathbb{KL}(r||p) = \sum_{\boldsymbol{x}} \prod_{c_s} r_{c_s} \left( \log p(\boldsymbol{x}) - \sum_{c_k} \log r_{c_k} \right). \tag{13}$$

Let us now isolate one of the factors, $r_{c_j}$ of the variational distribution $r$. This leads to splitting sums and product according to the different clusters $c_s$. It follows that:

$$-\mathbb{KL}(r||p) = \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x} \setminus \{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \Big( \log p(\boldsymbol{x}) - \log r_{c_j} - \sum_{c_k \neq c_j} \log r_{c_k} \Big). \tag{14}$$

We want to maximize this quantity with the constraint $\prod_{c_s} r_{c_s} = 1$. Therefore, we introduce Lagrangian multipliers and consider the functional derivative. Setting this derivative to 0, we have the expression of the variational factor $r^*_{c_j}$:

$$\log r^*_{c_j} = \mathbb{E}_{\prod_{c_i \neq c_j} r_{c_i}} \left[ \sum_{d_s} \log p_{d_s} - \sum_{c_k \neq c_j} \log r_{c_k} \right] + \text{const.}, \tag{15}$$

where const. denotes the group of constant terms with respect to $c_j$. This expression can be simplified because additional terms in the three sums do not depend on $c_j$ (and these terms can merge with the const. term). Let $\mathcal{D}_j$ be a set whose elements are clusters of variables $d_j$ containing $x_j$, $\mathcal{B}_j$ a set of clusters of variables $b_j$ also containing $x_j$ but with the condition $b_j \neq c_j$. With such reparametrization, it follows that Equation 15 leads to:

$$r^*_{c_j} = \frac{1}{Z_j} \exp \left( \mathbb{E}_{\prod_{c_i \neq c_j} r_{c_i}} \left[ \sum_{d_j \in \mathcal{D}_j} \log p_{d_j} - \sum_{b_j \in \mathcal{B}_j} \log r_{b_j} \right] \right), \tag{16}$$

with $Z_j$ a normalization constant.

**Remark:** The expectation appearing in Equation 16 requires to sample from the joint law $\boldsymbol{x} \setminus \{x_j, x_{j^-}\}$ given $x_j$ and $x_{j-}$. Such a sampling can be done when $r$ is an MT because of its straightforward sparse decomposition. Note also that, in fact, we only need to sample the variables in each cluster (upon which the expectation is computed, *i.e.* $d_j$ or $b_j$). Then $\boldsymbol{X}$ never needs to be fully sampled.

### 3.2 Auxiliary variable variational inference with STMTs

We now develop VI over STMT using auxiliary variables as initiated in [1]. Let $\boldsymbol{V}$ be the auxiliary process which will be used to augment both $p$ and $t$. The new quantity to maximize is:

$$\begin{aligned} -\mathbb{KL}(t(\boldsymbol{x},\boldsymbol{v})||p(\boldsymbol{x},\boldsymbol{v})) &= \mathbb{E}_t[\log p(\boldsymbol{x},\boldsymbol{v})] - \mathbb{E}_t[\log t(\boldsymbol{x},\boldsymbol{v})], \\ &= \mathbb{E}_t[\log p(\boldsymbol{x})] + \mathbb{E}_t[\log p(\boldsymbol{v}|\boldsymbol{x})] - \mathbb{E}_t[\log t(\boldsymbol{x},\boldsymbol{v})]. \end{aligned} \tag{17}$$

Following [10], we have that the Kullback-Leibler divergence with auxiliary variables is lower-bounded by the Kullback-Leibler divergence without auxiliary variables:

$$\mathbb{KL}(t(\boldsymbol{x},\boldsymbol{v})||p(\boldsymbol{x},\boldsymbol{v})) \geq \mathbb{KL}(t(\boldsymbol{x})||p(\boldsymbol{x})). \tag{18}$$

So the cost function with auxiliary variables is at best equal to the cost function without auxiliary variables.

However, the use of auxiliary variables offers much more flexibility in the modelization and enables the modeling of richer correlations between the variables of interest.

We focus now on STMT as the variational distribution to approximate the SBN. The auxiliary variables have been introduced to better reflect the correlations in SBNs, while keeping the Markov-tree property. Hence, at the end of the VI procedure, exact marginal computation can be done, again with a generalized upward-downward algorithm.

Let $t$ be the STMT distribution, we have from Section 2.3:

$$\begin{aligned} t(\boldsymbol{x},\boldsymbol{v}) &= \prod_{s\in\mathcal{S}} t(x_s, \boldsymbol{v}_s | x_{s^-}, \boldsymbol{v}_{s^-}), \\ &= \prod_{s\in\mathcal{S}} t(x_s|x_{s^-}, v_{n(s^-)}) t(v_s^{\leftarrow}|x_s, v_{n'(s^-)}) t(v_s^{\rightarrow}|x_s, v_{n''(s^-)}), \\ &= \prod_{c_s} t_{c_s} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s}, \end{aligned} \tag{19}$$

with $c_s = (x_s, x_{s^-}, v_{n(s^-)})$, $c'_s = (v_s^{\leftarrow}|x_s, v_{n'(s^-)})$ and $c''_s = (v_s^{\rightarrow}|x_s, v_{n''(s^-)})$.

Then we want to maximize:

$$\begin{aligned} -\mathbb{KL}(t||p) = \sum_{\boldsymbol{x},\boldsymbol{v}} \prod_{c_s} t_{c_s} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \Bigg( \log p(\boldsymbol{x}) - \\ \Big( \sum_{c_k} \log t_{c_k} + \sum_{c'_k} \log t_{c'_k} + \sum_{c''_k} \log t_{c''_k} \Big) \Bigg). \end{aligned} \tag{20}$$

Obtaining the update equations for each variational transition follows the same steps as for DMT-VI (which are omitted here for brevity). We have $\forall t_{c_j}$:

$$\begin{aligned}\log t^*_{c_j} =& \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s}} \Bigg[ \log p(\boldsymbol{x}) - \\ & \sum_{c_k \neq c_j} \log t_{c_k} - \sum_{c'_k} \log t_{c'_k} - \sum_{c''_k} \log t_{c''_k} \Bigg] + \text{const.},\end{aligned} \tag{21}$$

where the last term regroups constant terms with respect to $c_j$.

Update equations for $t_{c'_j}$ and $t_{c''_j}$ are similar but the term $\log p(\boldsymbol{x})$ is replaced by $\log p(\boldsymbol{v}|\boldsymbol{x})$ in both cases.

Now STMT VI is still a relatively sparse and highly structure network, hence, the updates equations can be simplified as in Equation 16. We need to carefully select the subsets of the variables involved in the expectation while the other join the constant term that is unimportant.

## 4 Experiments and Results

### 4.1 Experimental set-up

We now consider variational inference on the small SBN network given in Figure 5a. Similar experiments have been conducted in the same context, to evaluate a VI approximation, for example in [13]. Due to its small size the SBN of Figure 5a represents a slightly modified probability distribution from that the SBN used up to now. Indeed, we needed to treat in a specific fashion the root node to induce SBN-like correlations on a 3-layered network only. It follows that $p$ has the following distribution:

$$\begin{aligned}p(a, a^{\leftarrow}, a^{\rightarrow}, b, c, d, e, f, g) =& p(a)p(a^{\leftarrow})p(a^{\rightarrow})p(b|a, a^{\leftarrow})p(c|a, a^{\rightarrow}) \\ & p(d|b)p(e|b, c)p(f|c, b)p(g|c).\end{aligned} \tag{22}$$

We are interested in retrieving the marginals in the SBN using VI. . We successively consider 3 VI techniques: MF VI (Figure 6a)[1], MT VI (Figure 6b) and STMT VI (with auxiliary nodes) (Figure 6c).

The developments of the previous section can be straightforwardly used to conduct VI over DMT and STMT. Note that for the STMT VI, we also need to provide the SBN with auxiliary nodes [1]. We need to keep the property that $p(\boldsymbol{x}, \boldsymbol{v}) = p(\boldsymbol{x})p(\boldsymbol{v}|\boldsymbol{x})$, where $\boldsymbol{x} = \{a, a^{\leftarrow}, a^{\rightarrow}, b, c, d, e, f, g\}$ and $\boldsymbol{v} = \{b^{\leftarrow}, b^{\rightarrow}, c^{\leftarrow}, c^{\rightarrow}, d^{\leftarrow}, d^{\rightarrow}, e^{\leftarrow}, e^{\rightarrow}, f^{\leftarrow}, f^{\rightarrow}, g^{\leftarrow}, g^{\rightarrow}\}$; in order to ensure that $p(\boldsymbol{x})$ (Equation 22) is the same between the three VI procedures. In STMT VI, $p$ is then:

$$\begin{aligned}p(\boldsymbol{x}, \boldsymbol{v}) =& p(\boldsymbol{x})p(b^{\leftarrow}|b, a^{\rightarrow})p(b^{\rightarrow}|a, a^{\rightarrow})p(c^{\leftarrow}|a, a^{\leftarrow})p(c^{\rightarrow}|a, a^{\leftarrow}) \\ & p(d^{\leftarrow}|c^{\rightarrow}, b^{\leftarrow})p(d^{\rightarrow}|b^{\leftarrow}, c^{\rightarrow})p(e^{\leftarrow}|b^{\rightarrow}, c^{\leftarrow})p(e^{\rightarrow}|c^{\leftarrow}, b^{\rightarrow}) \\ & p(g^{\leftarrow}|c^{\rightarrow}, b^{\leftarrow})p(g^{\rightarrow}|b^{\leftarrow}, c^{\rightarrow})p(f^{\rightarrow}|b^{\rightarrow}, c^{\leftarrow})p(f^{\leftarrow}|c^{\leftarrow}, b^{\rightarrow}),\end{aligned} \tag{23}$$

[1]This approach is the most popular and the associated equations have not been developped in this article for brevity. Many resources cover the topic, *e.g.* [14].

with

$$
\begin{aligned}
p(b|a, a^{\leftarrow}) &= p(b^{\leftarrow}|a, a^{\leftarrow}) = p(c^{\leftarrow}|a, a^{\leftarrow}),\\
p(c|a, a^{\rightarrow}) &= p(b^{\rightarrow}|a, a^{\rightarrow}) = p(c^{\rightarrow}|a, a^{\rightarrow}),\\
p(d|b) &= p(d^{\leftarrow}|b^{\leftarrow}) = p(e^{\leftarrow}|b^{\leftarrow}),\\
p(e|b, c) &= p(d^{\rightarrow}|b^{\leftarrow}, b^{\rightarrow}) = p(f^{\leftarrow}|c^{\leftarrow}, b^{\rightarrow}),\\
p(f|c, b) &= p(g^{\leftarrow}|c^{\rightarrow}, c^{\leftarrow}) = p(e^{\rightarrow}|b^{\rightarrow}, c^{\leftarrow}),\\
p(g|c) &= p(f^{\rightarrow}|c^{\rightarrow}) = p(g^{\rightarrow}|c^{\rightarrow}).
\end{aligned}
$$

The model $p$ with auxiliary nodes is described in Figure 5b.

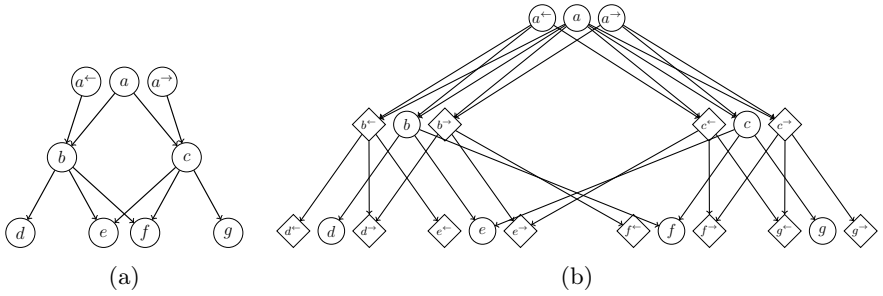

Fig. 5: (a) Adapted SBN, (b) Adapted SBN with auxiliary nodes.

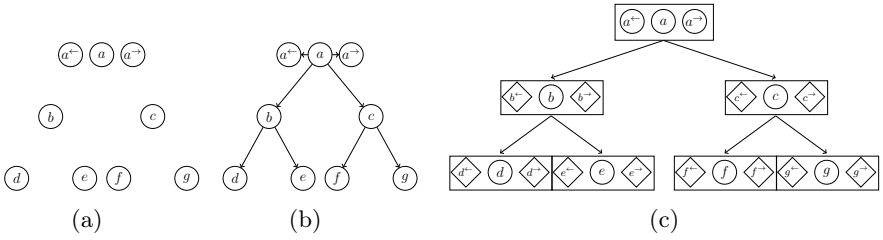

Fig. 6: Variational distributions for the VI procedure. (a) MF VI, (b) DMT VI and (c) STMT VI.

The random variables are chosen with values in $\{0, 1\}$. Our goal is to estimate the *true* marginals $p(x = 1), \forall x \in \boldsymbol{x}$ of the SBN of Figure 5a. In this synthetic example those true marginals are given as well as the transitions in $p$ (we do not cover the parameter estimation problem): the transitions of Equation 22 which also totally define Equation 23 are taken randomly.

### 4.2 Results

We define a marginal *error* for a variational distribution $q$ and a random variable $x$ by $e_q(x) = p(x = 1) - q(x = 1)$. These errors are computed and stored

over 1000 different SBNs $p$ whose transitions are randomly chosen. Figure 7 depicts the values of the Kullback-Leibler divergence of the three VI procedures. We see a rapid convergence which is related to the small size of the considered SBN. We then choose to stop the VI process after 30 iterations. The value for MF VI and DMT VI are comparable and are shown on the same graph. The minimization is better in the case of DMT VI, the structured VI; this is reflected in Figure 8 which illustrates the goodness of the estimated marginals.

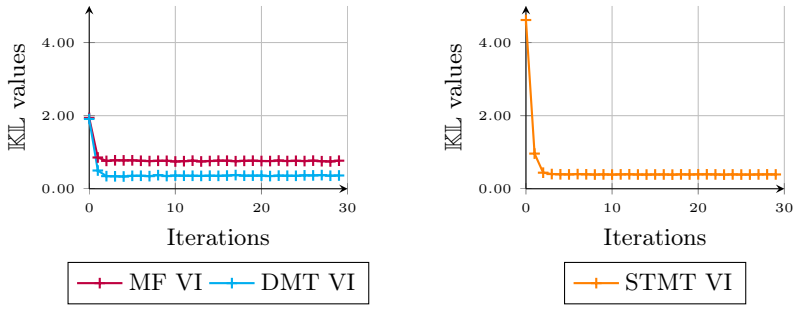


Fig. 7: The values of the cost function (Kullback-Leibler divergence values) of the optimization problem for the three VI procedures. Note that the STMT VI cost function integrates auxiliary variables and is not comparable with the others, hence it is plotted in another graph.

The analysis of Figure 8 illustrates the interest of the STMT structure to approximate the marginals of SBN: this VI procedure exhibits in all cases the smallest error with respect to the true marginals. We also observe that progressively adding structure increases the quality of the approximation since MF VI performs worse than DMT VI which performs in turn worse than STMT VI. Note that the left/right symmetry of the SBN can be found in the behaviour of the error rates: $d$ is similar to $g$, $e$ is similar to $f$, and so on. Moreover, we notice that the DMT VI performs worse for random variables $b$ and $c$. This can be explained by the fact that those variables are central in the SBN and are more correlated with others variables. We note also that the errors computed at nodes $b$ and $c$ in the STMT VI remain stable.

## 5 Conclusion

In this article we explored the potential of adding auxiliary random variables in order to develop more complex and rich correlations in probabilistic models. We showed that the triplet Markov framework enables to handle the additional auxiliary variables while preserving tractability of computations and analytical solutions.

We developped the STMT model and illustrated its potential as a variational distribution to approximate a much more complex Bayesian network: the SBN model. STMT performed better than the classical MF variational inference but also than the DMT variational inference methods.

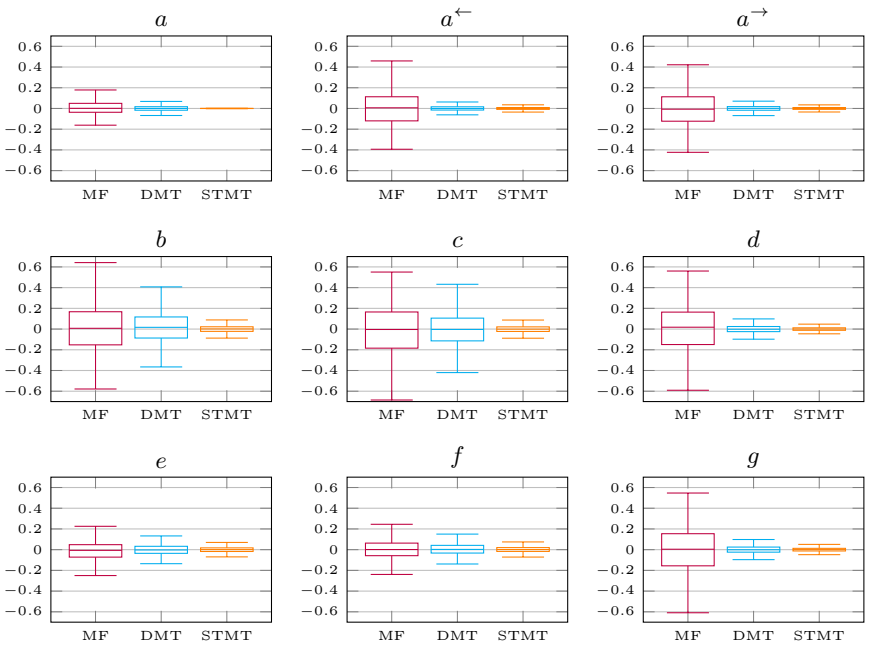


Fig. 8: Boxplots to illustrate the error with respect to the true marginal, for each node, for the three VI procedures. The serie of errors spans over 1000 different experiments (different $p$ transitions randomly chosen).

Further work might consider studying theoretical results to quantify and qualify the relation of the SBN and STMT models. We might also extend the developed algorithms from dyadic trees to quadtrees [11] [5] in order to treat bi-dimensional data such as images.

## References


[1] F. V. Agakov and D. Barber. "An auxiliary variational method". In: *International Conference on Neural Information Processing*. Springer. 2004, pp. 561–566.

[2] C. Bishop and J. Winn. "Structured variational distributions in VIBES". In: (2003).

[3] D. M. Blei et al. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[4] J.-B. Courbot et al. "Oriented triplet Markov fields". In: *Pattern Recognition Letters* 103 (2018), pp. 16–22.

[5] J.-B. Courbot et al. "Triplet Markov Trees for Image Segmentation". In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. 2018, pp. 233–237.

[6] J.-B. Durand et al. “Computational methods for hidden Markov tree models-An application to wavelet trees”. In: *IEEE Transactions on Signal Processing* 52.9 (2004), pp. 2551–2560.

[7] Z. Ghahramani and M. I. Jordan. “Factorial hidden Markov models”. In: *Advances in Neural Information Processing Systems*. 1996, pp. 472–478.

[8] H. Hanzouli-Ben Salah et al. “A framework based on hidden Markov trees for multimodal PET/CT image co-segmentation”. In: *Medical physics* 44.11 (2017), pp. 5835–5848.

[9] M. I. Jordan et al. “An introduction to variational methods for graphical models”. In: *Machine learning* 37.2 (1999), pp. 183–233.

[10] D. P. Kingma. “Variational inference & deep learning: A new synthesis”. PhD thesis. 2017.

[11] J.-M. Laferté et al. “Discrete Markov image modeling and inference on the quadtree”. In: *IEEE Transactions on image processing* 9.3 (2000), pp. 390–404.

[12] P. Lanchantin et al. “Unsupervised segmentation of randomly switching data hidden with non-Gaussian correlated noise”. In: *Signal Processing* 91.2 (2011), pp. 163–175.

[13] S. L. Lauritzen and D. J. Spiegelhalter. “Local computations with probabilities on graphical structures and their application to expert systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 50.2 (1988), pp. 157–194.

[14] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] V. Olariu et al. “Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages”. In: *Bioinformatics* 25.21 (2009), pp. 2824–2830.

[16] R. Ranganath et al. “Hierarchical variational models”. In: *International Conference on Machine Learning*. 2016, pp. 324–333.

[17] T. Salimans, D. A. Knowles, et al. “Fixed-form variational posterior approximation through stochastic linear regression”. In: *Bayesian Analysis* 8.4 (2013), pp. 837–882.

[18] C. Zhang et al. “Advances in variational inference”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.

[19] P. Zwiernik. *Semialgebraic statistics and latent tree models*. Chapman and Hall/CRC, 2015.