Two-level Mixtures of Markov Trees

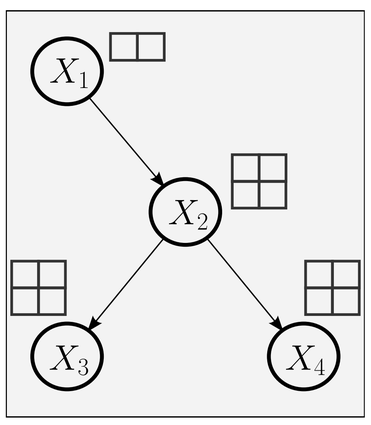
Franois Schnitzler, Louis Wehenkel

Systems and Modeling & GIGA research, University of Liège, Belgium

fschnitzler@ulg.ac.be,L.Wehenkel@ulg.ac.be

Bayesian Networks efficiently encode a probability distribution on a large set of variables but their **poor scaling** in terms of the number of variables may make them unfit to tackle learning and inference problems of increasing size. **Mixtures of Markov trees** however scale well by design. We investigate whether the two approaches to **learn such mixtures from data** (maximum likelihood and variance reduction) can be combined together by building a **two level Mixture** of Markov Trees. Our experiments on synthetic data show the interest of this model.

Markov Tree T :

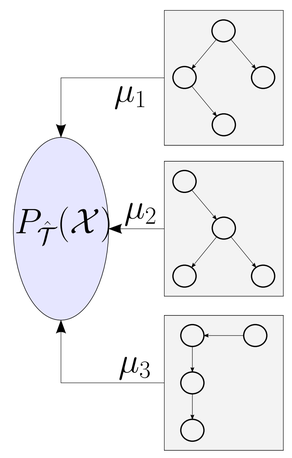


- A class of Bayesian Network.
- No cycle, each variable has only one parent.
- Encodes a joint probability distribution over n variables $\mathcal{X}$:

$$P_T(\mathcal{X}) = \prod_{i=1}^{n} P(X_i | Pa_{\mathcal{G}}(X_i)) \ .$$

- Learning from a data set is $\mathcal{O}(n^2 \log(n))$ (Chow-Liu algorithm).
- Inference is $\mathcal{O}(n)$.

Mixtures of Markov trees :



- Composed of a set $\hat{\mathcal{T}} = \{T_1, \ldots, T_m\}$ of m elementary Markov Tree densities and a set $\{\mu_i\}_{i=1}^m$ of weights.
- Convex combination of tree predictions :

$$P_{\hat{\mathcal{T}}}(\mathcal{X}) = \sum_{k=1}^{m} \mu_k P_{T_k}(\mathcal{X}) \ .$$

Key points:

- Trees → efficient algorithms.
- Mixture → improved modeling power.

There are 2 approaches to learn such mixtures from data :

Maximum-likelihood [1]

- Learning the mixture is viewed as a global optimization problem aiming at maximizing the data likelihood.
- Can be solved (for a fixed m) using the EM algorithm by iteratively :
  - optimizing μ based on $\mathcal{T}$,
  - optimizing a soft partition the data set based on $\mathcal{T}$ and μ,
  - optimizing each $T \in \mathcal{T}$ on a part by the Chow-Liu algorithm.

Variance reduction [2]

- This approach can be viewed as an approximation of Bayesian learning in the space of Markov Tree structures.
- A sequence of trees is generated by a randomized Chow-Liu algorithm :
  - pure random structure,
  - edge subsampling,
  - bootstrapping...

We attempt to combine both.

Motivation:

- Variance reduction methods are good on low samples sets.
- Maximum-likelihood methods partition the data set.

→ It seems natural to combine them.

Concept:

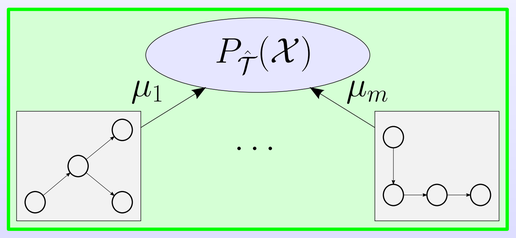
Replacing in the ML mixture the Chow-Liu algorithm by a variance reduction mixture learning algorithm:

$$P_{\hat{\mathcal{T}}}(\mathcal{X}) = \sum_{k=1}^{m} \mu_k P_{\hat{\mathcal{T}}_k}(\mathcal{X}) \ ,$$

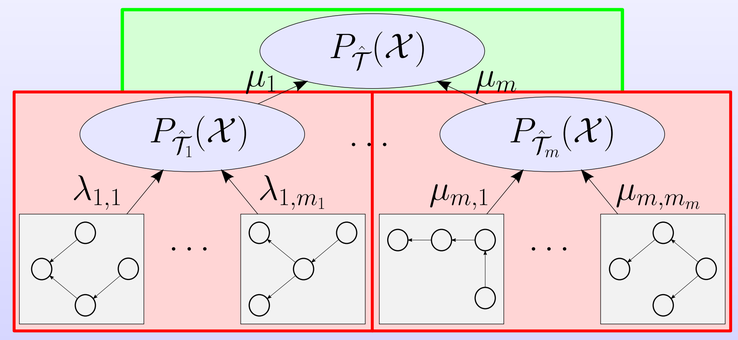
$$P_{\hat{\mathcal{T}}_k}(\mathcal{X}) = \sum_{j=1}^{m_k} \lambda_{k,j} P_{T_{k,j}}(\mathcal{X}) \ \ \forall k \ .$$

Algorithm tested:

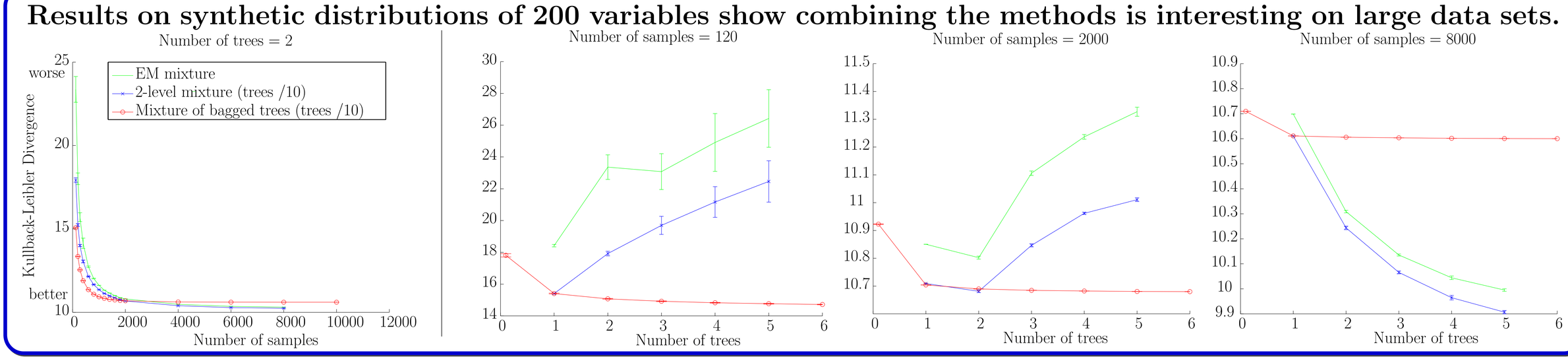
1 : build an EM mixture



2 : replace each tree by a mixture of 10 bagged trees



Results on synthetic distributions of 200 variables show combining the methods is interesting on large data sets.



References

[1] Meila, M., Jordan, M.: Learning with mixtures of trees. JMLR 1, 1–48 (2001)

[2] Ammar, S., Leray, P., Schnitzler, F., Wehenkel, L.: Sub-quadratic Markov tree mixture learning based on randomizations of the Chow-Liu algorithm. In: The Fifth European Workshop on Probabilistic Graphical Models. pp. 17–24 (2010)

Acknowledgement

This work was funded by the Belgian Fund for Research in Industry and Agriculture (FRIA), the Biomagnet IUAP network of the Belgian Science Policy Office and the Pascal2 network of excellence of the EC. It is not under those organisms scientific responsibility.