Efficiently Approximating Markov Tree Bagging for High-Dimensional Density Estimation

François Schnitzler, Sourour Ammar, Philippe Leray, Pierre Geurts, Louis Wehenkel

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

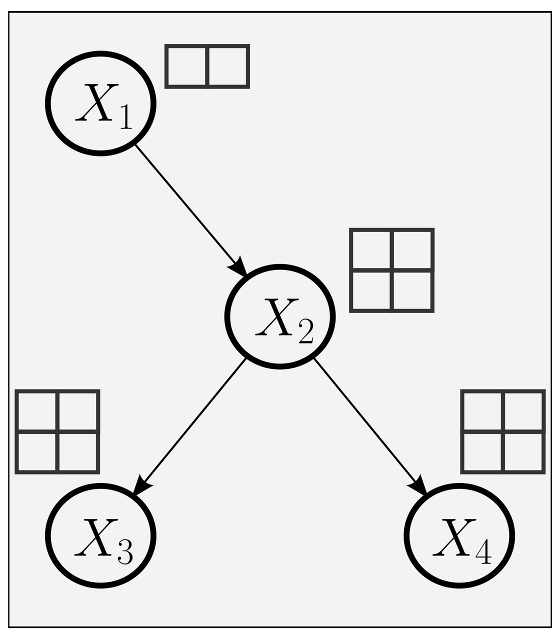
fschnitzler@ulg.ac.be,P.Geurts@ulg.ac.be,L.Wehenkel@ulg.ac.be

Ecole Polytechnique de l'Université de Nantes, Knowledge and Decision Team,
Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241, France.

sourour.ammar,philippe.leray@univ-nantes.fr

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 and outperform a single Markov tree maximizing the data likelihood. We show how learning **Mixtures of Bagged Markov Trees** can be accelerated using a by-product from computing a first tree so as to avoid considering poor candidate edges in the subsequently generated trees.

Markov tree T :

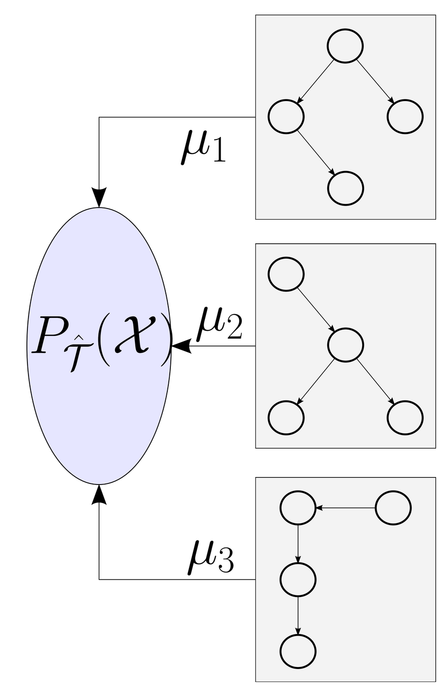


- A class of Bayesian Networks.
- No cycle, each variable has only one parent.
- Encodes a joint probability distribution over n variables $\mathcal{X}$:

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

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

Mixture of Markov trees [1]:



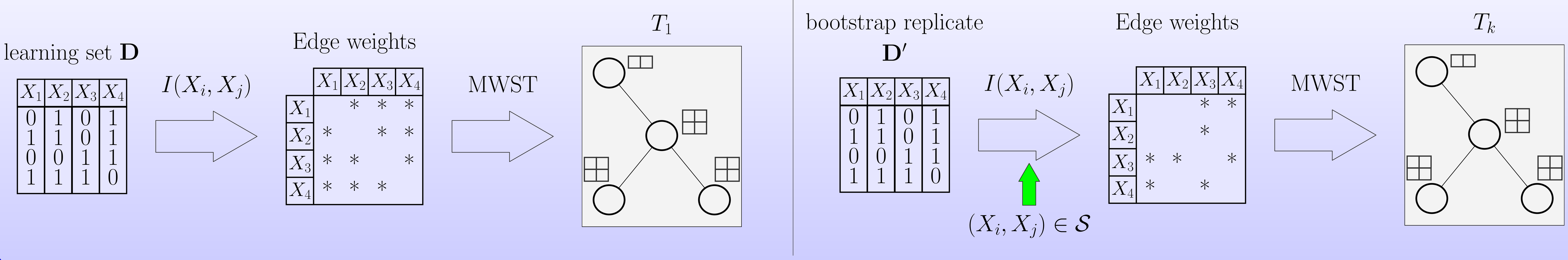
- Composed of a set $\hat{\mathcal{T}} = \{T_1, \ldots, T_m\}$ of m elementary Markov Tree densities and a set $\{\mu_k\}_{k=1}^{m}$ of weights.
- Convex combination of tree predictions :

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

Key points:

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

We approximate a mixture of bagged Markov trees by exploiting previous trees to select a good subset $\mathcal{S}_k$ of candidate edges for building the subsequent tree:



We developed two strategies to build $\mathcal{S}_k$:

Strategy a: inertial:

$\mathcal{S}_k$ depends on T_{k-1}.

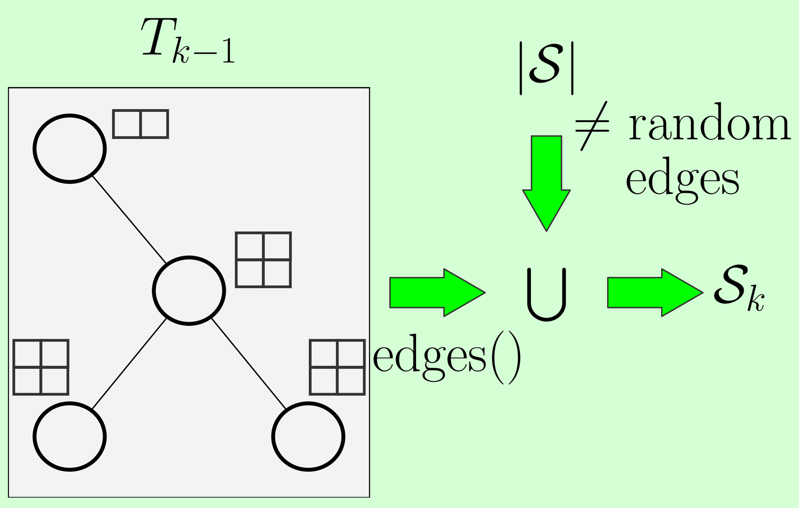
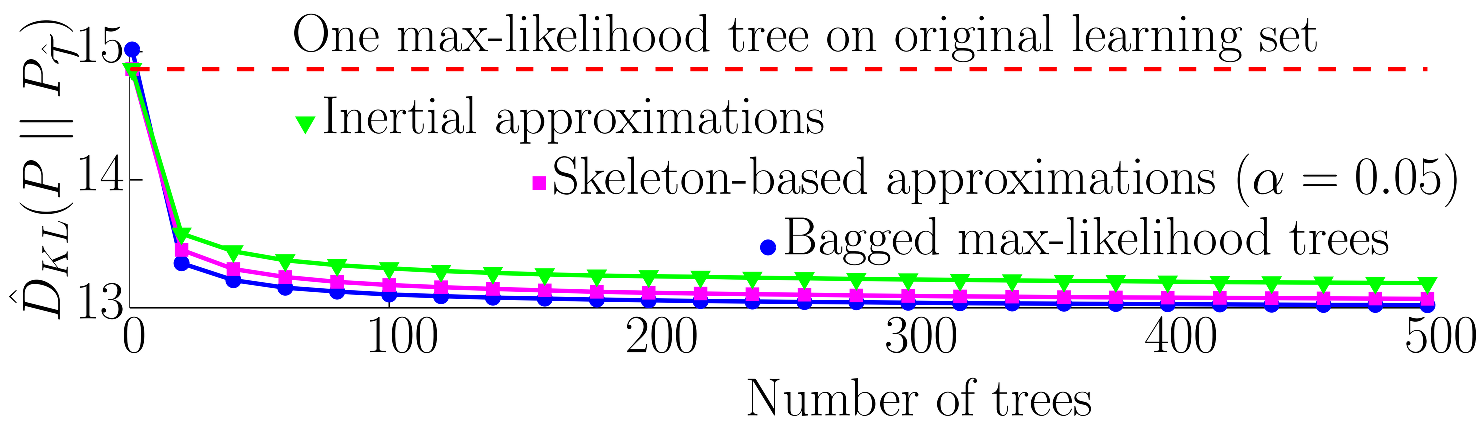


Illustration of the complexity/quality tradeoff :

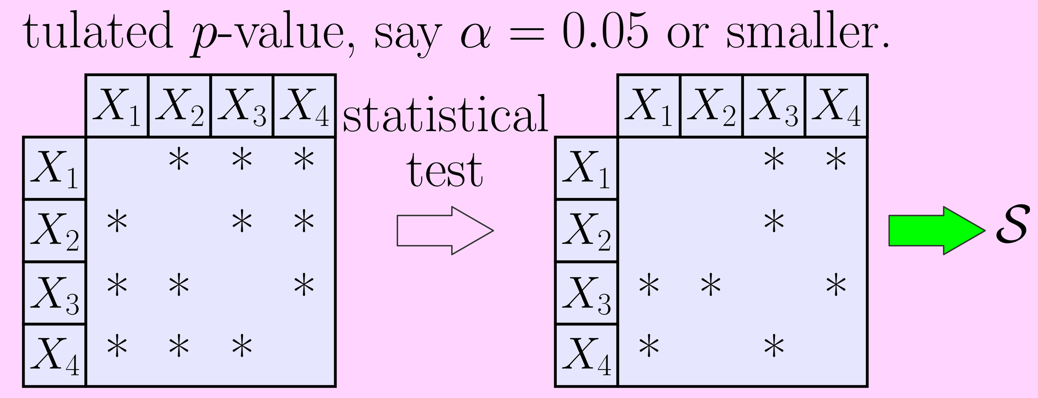
Synthetic data set, 200 variables 200 samples:



Run-time (max-likelihood tree : 1) for 500 trees : ▼ 45; ■ 21; ● 532

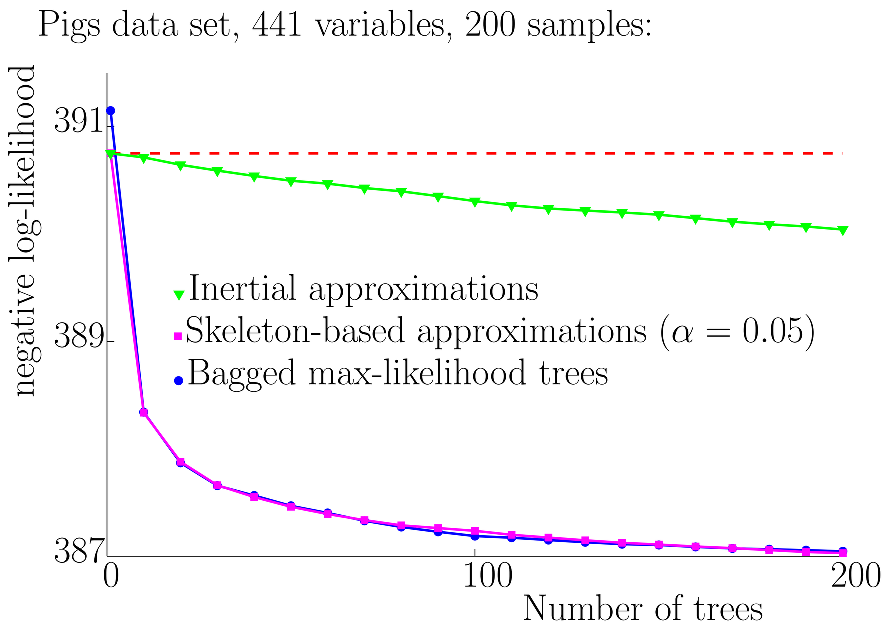
Strategy b: skeleton-based:

$\mathcal{S}_k = \mathcal{S}\ \forall k$ and is obtained by comparing $I_{\mathbf{D}}(X_i; X_j)$ to a threshold depending on a postulated p-value, say $\alpha = 0.05$ or smaller.

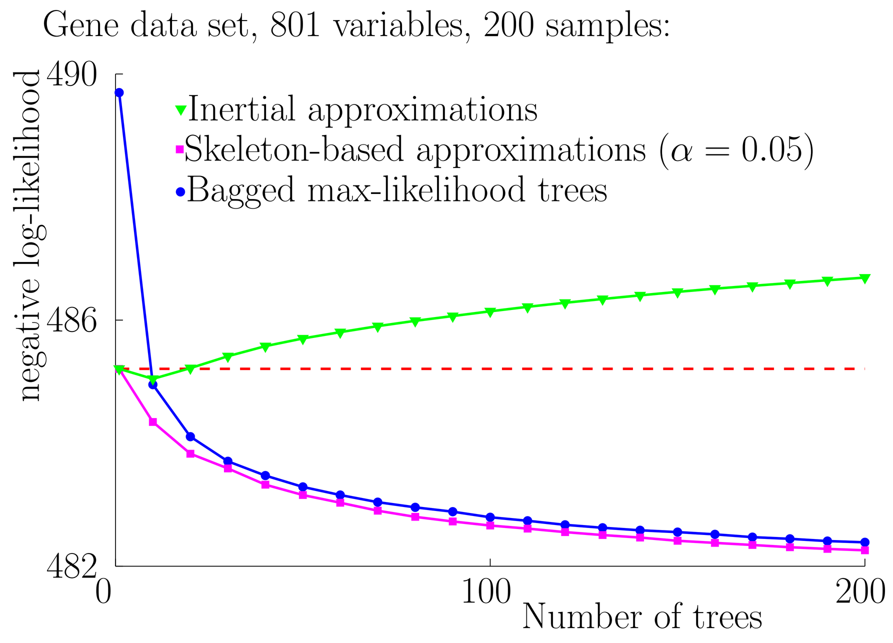


Evaluation on real data sets [2]:

Pigs data set, 441 variables, 200 samples:

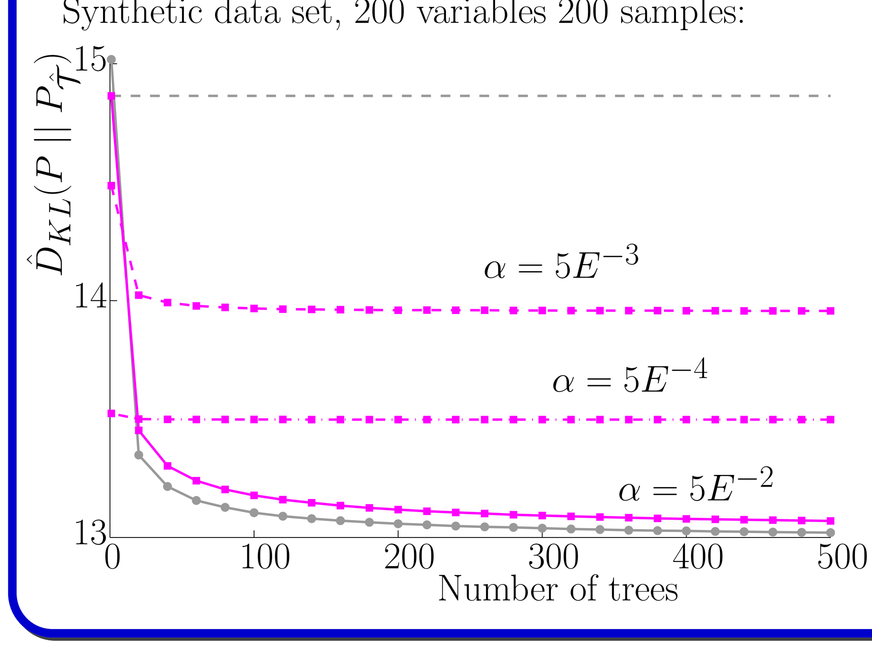


Gene data set, 801 variables, 200 samples:

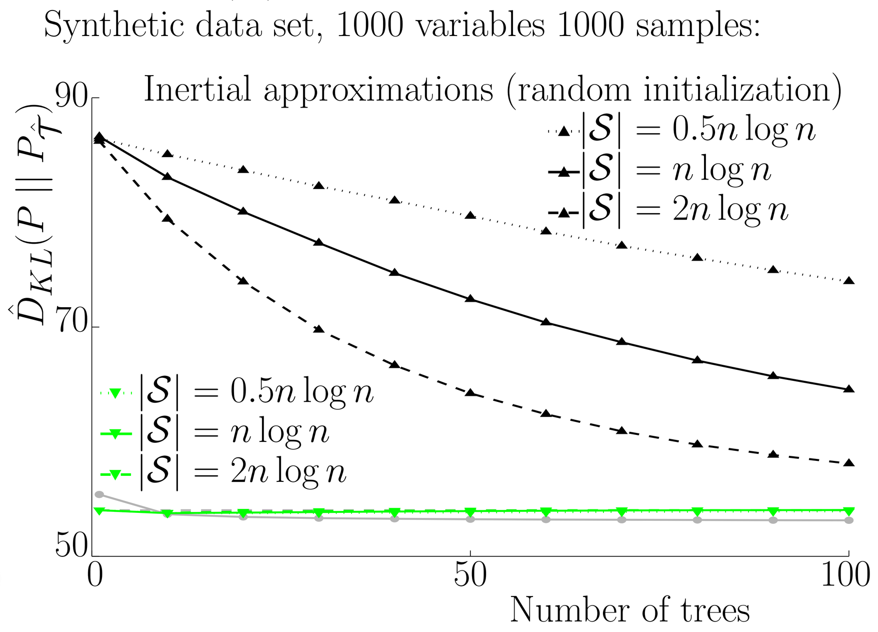


Effects of the parameters:

Influence of α in the skeleton-based approximation:
Synthetic data set, 200 variables 200 samples:



Influence of $|\mathcal{S}|$ in the inertial approximation:
Synthetic data set, 1000 variables 1000 samples:



References

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

[2] Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.: Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. JMLR 11, 171–234 (2010)

Acknowledgements

François Schnitzler is supported by a F.R.I.A. scholarship. Pierre Geurts is a research associate of the FNRS, Belgium. This work was also funded by the Biomagnet IUAP network of the Belgian Science Policy Office and the Pascal2 network of excellence of the EC. The scientific responsibility is the authors'.