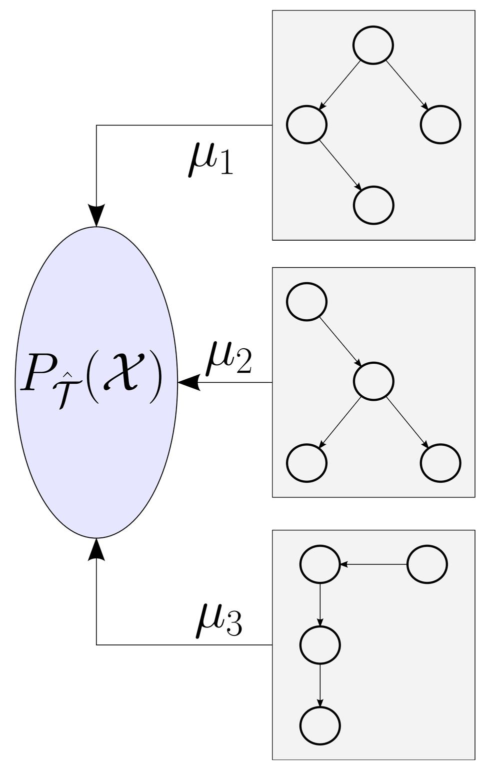
Mixtures of Bagged Markov Tree Ensembles

François Schnitzler, 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

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.

There are 2 approaches to improve over a single Chow-Liu tree:

Bias reduction, e.g. EM algorithm [1]

- Learning the mixture is viewed as a global optimization problem aiming at maximizing the data likelihood.
- There is a bias-variance trade-off associated with the number of terms.
- It leads to a partition of the learning set: each tree models a subset of observations.

Variance reduction, e.g. perturb and combine [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.

→ **Is it possible to reduce the variance of the EM mixture by combining both methods?**

Concept:

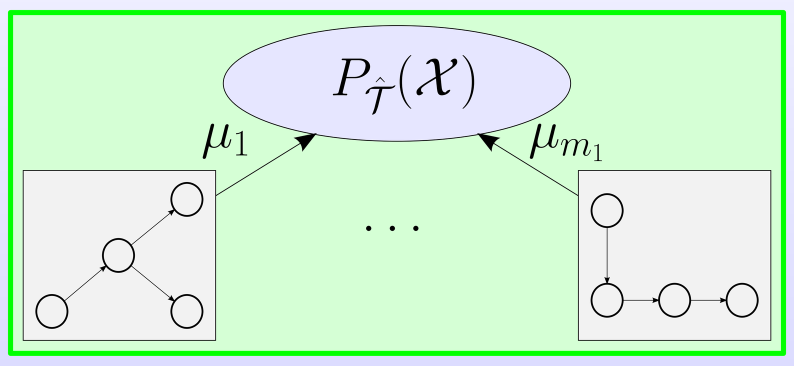
Building a mixture of ensemble of Markov trees:

$$\mathbb{P}_{\hat{\mathcal{T}}}(\mathcal{X}) = \sum_{k=1}^{m_1} \mu_k \mathbb{P}_{\hat{\mathcal{T}}_k}(\mathcal{X})$$

$$\mathbb{P}_{\hat{\mathcal{T}}_k}(\mathcal{X}) = \sum_{j=1}^{m_2} \lambda_{k,j} \mathbb{P}_{T_{k,j}}(\mathcal{X}) \ \ \forall k \in [1, m_1] \ .$$

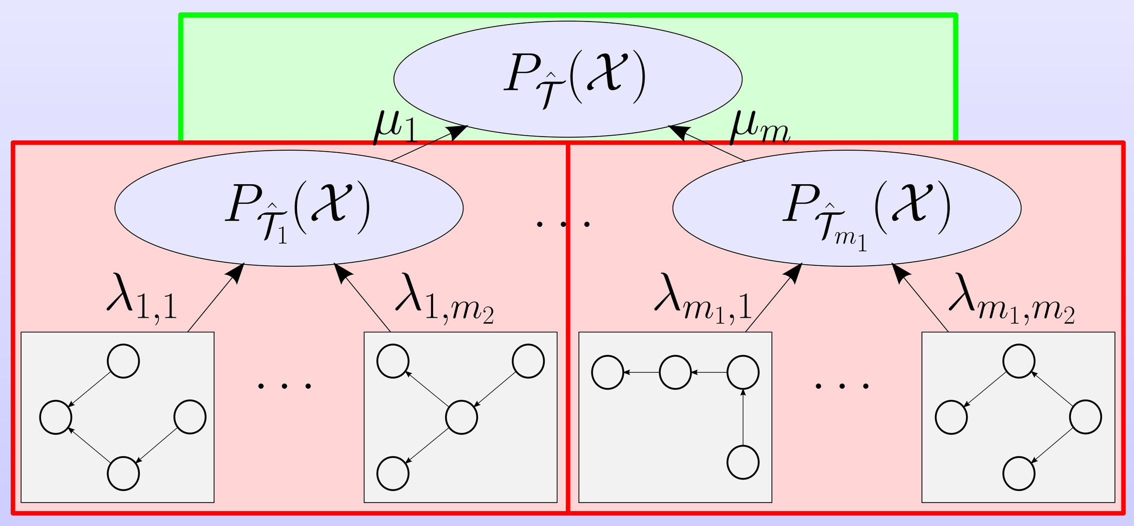
Algorithms tested:

1: build an EM mixture

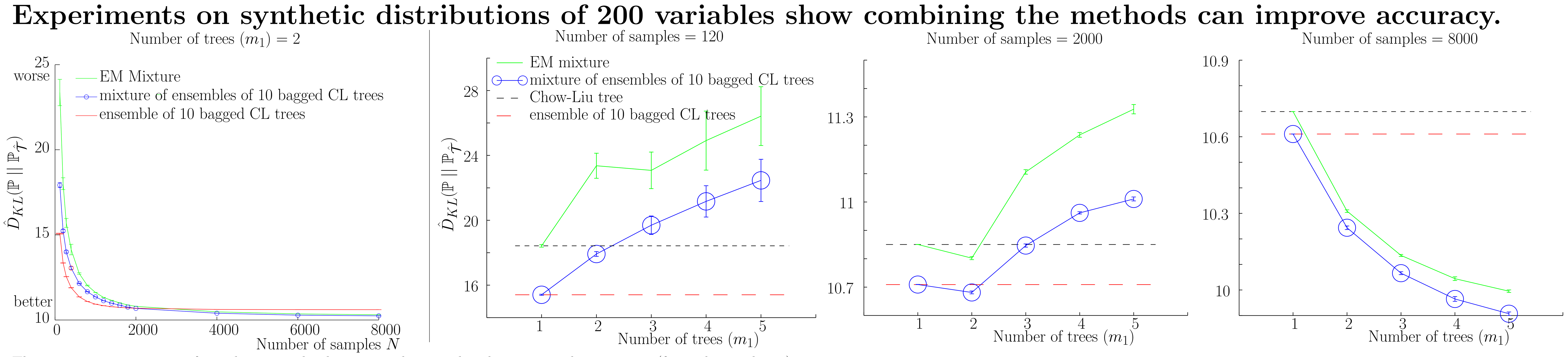


2: replace each tree by either:

- **a mixture of 10 bagged Chow-Liu trees,**
- **the original tree and 9 bagged Chow-Liu trees.**

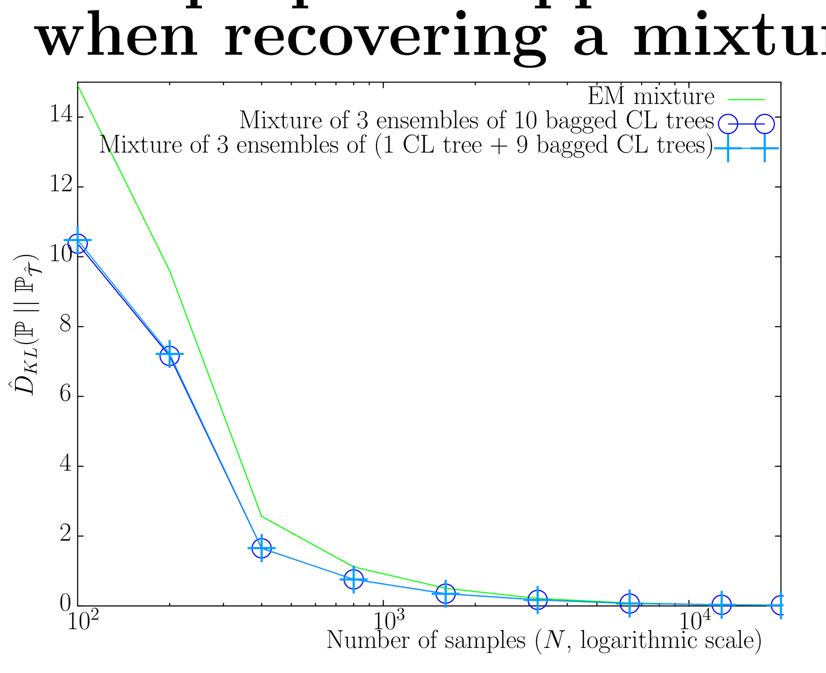


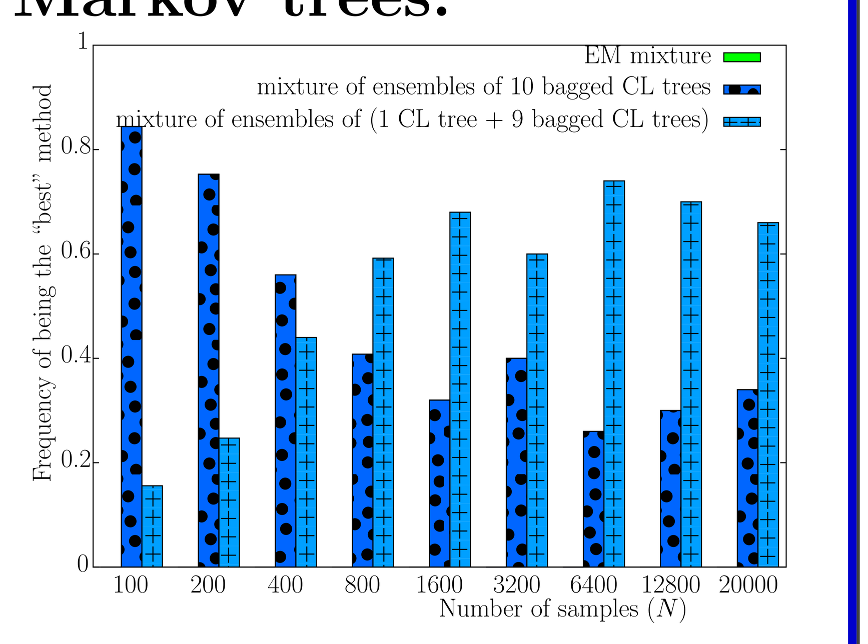
Experiments on synthetic distributions of 200 variables show combining the methods can improve accuracy.



These experiments were performed on 5 randomly generated target distributions × 5 learning sets (for each samplesize).

The proposed approach improves over an EM mixture when recovering a mixture of Markov trees:





- The original mixture has 3 Markov trees, 200 binary variables, and uniform weights.
- The number of trees in the target mixture is known.
- The accuracy of the EM mixture is always improved by replacing each tree by an ensemble.
- The first ensemble seems better for low samplesize, the second better when N increases.

Number of runs of the EM algorithm (with a different initial point, on the same learning set), by number of samples:

Number of samples	100	200	400	800	1600	6400	12800	20000
Number of runs	905	531	400	250	50	50	50	50

More terms might be necessary in the ensembles when estimating more complex probability distributions:

The following table displays the number of runs where each method was better than the two other methods, for different configurations:

Data set	p	$\lvert X_i \rvert$	$\lvert\theta\rvert$	$N = 200$			$N = 500$			$N = 2500$		
				EM	+Bag	+Bag&Cl	EM	+Bag	+Bag&Cl	EM	+Bag	+Bag&Cl
Child10	200	2-6	2323	-	1	24	-	4	21	-	4	6
Pigs	441	3-3	3675	21	1	3	-	16	9	-	9	1
Alarm10	370	2-4	5468	3	5	17	-	6	19	-	8	2
Gene	801	3-5	8348	25	-	-	-	9	16	-	8	2
Lung Cancer	800	2-3	8452	25	-	-	8	-	17	-	8	2
Link	724	2-4	14211	3	7	15	-	13	12	-	8	2
Insurance10	270	2-5	14613	2	1	22	1	9	15	-	7	3
Munin	189	1-21	15622	1	15	9	6	5	14	-	5	5
Hailfinder10	560	2-11	97448	25	-	-	25	-	-	-	1	9
ALL				105	30	90	40	62	123	0	58	32

- These experiments were performed on 5 learning sets. For 200 and 500 samples, 5 runs of the EM algorithm were performed per learning set. Only 2 runs were carried out at 2500 samples.
- When the number of samples is low, the EM mixture is not always improved by the addition of the ensembles.
- This may be due to a number of trees in the ensembles ($m_2 = 10$) too low for the problems. Indeed, increasing m_2 leads to better results.
- Is it possible to optimize m_2 for each ensemble?

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.