Mixture of ensemble of bagged Markov trees

F. Schnitzler P. Geurts L. Wehenkel

fschnitzler@ulg.ac.be

University of Liège

20 September 2012

The goal of this research is to improve probabilistic reasoning in high-dimensional problems.

High-dimensional :

- high number of variables
- low number of samples

Great potential in many applications :

- Bioinformatics (tens of thousands of genes, hundreds of thousands proteins)
- Power networks (tens of thousands transmission nodes in Europe)

Two main problems :

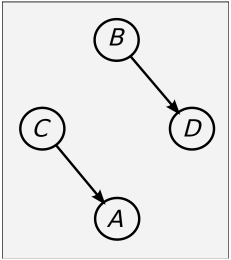
1. Few samples → high variance
2. Algorithmic complexity

→ Simple models must be used, e.g. Markov trees.

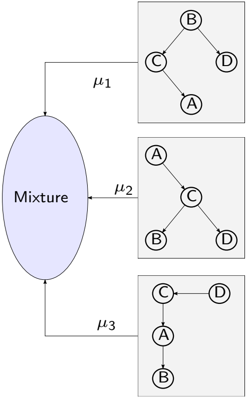
We try to use mixtures of Markov trees to improve a single Chow-Liu tree.

Mixtures or ensembles of trees build on the good properties of Markov trees.

A forest is a tree missing edges :



A mixture of trees is an ensemble method :

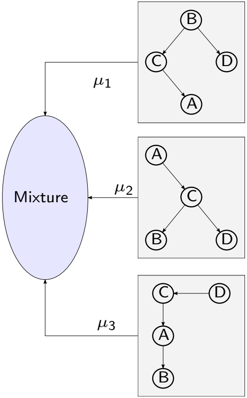


$$\mathbb{P}_{\hat{\mathcal{T}}}(X) = \sum_{i=1}^{m} \mu_i \mathbb{P}_{T_i}(X)$$

Mixtures or ensembles of trees build on the good properties of Markov trees.

Using several trees can improve the modelling accuracy while maintaining low complexity.

- inference is linear,
- learning : most algorithms are quadratic.

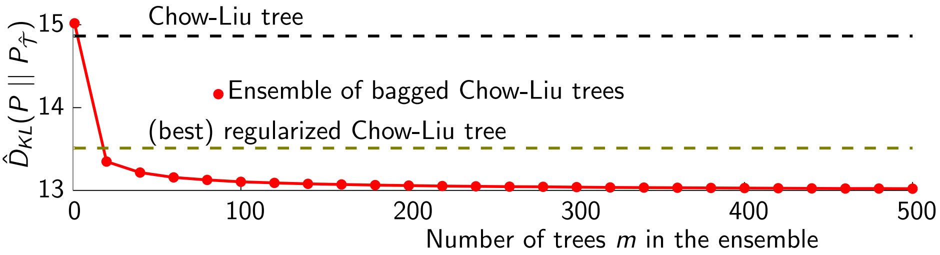


Some mixtures reduce the **variance** with respect to a single Chow-Liu tree.

Perturb and combine ensemble

- This approach can be viewed as an approximation of Bayesian learning in the space of Markov tree structures.
- The more trees, the better the reduction in variance of the ensemble.
- Can be constructed by a randomized [bagging, edge subsampling...] Chow-Liu algorithm applied several times.

Example : 200 variables and 200 samples

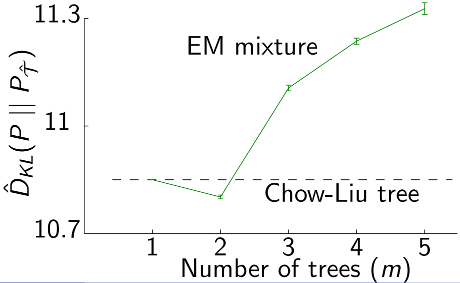


Some mixtures reduce the **bias** with respect to a single Chow-Liu tree.

Expectation-Maximization mixture

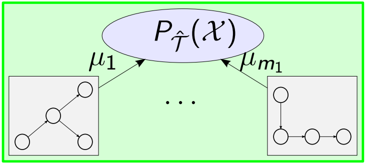
- Learning the mixture is viewed as a global optimization problem aiming at maximizing the data likelihood.
- There is a bias-variance trade-off associated to the number of terms.
- Partition of the learning set : each tree models a subset of observations.

Example : 200 variables and 2000 samples



We try to combine those two methods.

1. Build an EM mixture and associated soft partition $\{\mathbf{D}_k\}_{k=1}^{m_1}$.
2. Replace each tree T_k by an ensemble of bagged Chow-Liu trees based on $\mathbf{D}_k$.



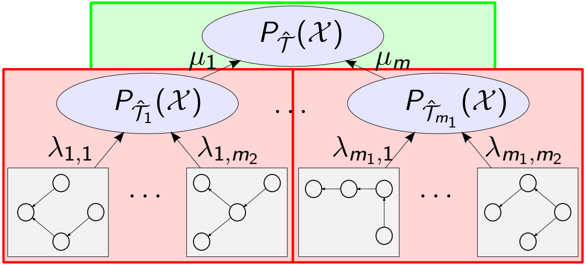
The EM algorithm iteratively :

- defines a soft partition of the data set into m_1 weighted learning samples $\mathbf{D}_k$, based on $\mathcal{T}$ and μ,
- optimizes μ based on the soft partition,
- optimizes each $T_k \in \mathcal{T}$ on $\mathbf{D}_k$ by the Chow-Liu algorithm,

until convergence.

We try to combine those two methods.

1. Build an EM mixture and associated soft partition $\{\mathbf{D}_k\}_{k=1}^{m_1}$.
2. Replace each tree T_k by an ensemble of bagged Chow-Liu trees based on $\mathbf{D}_k$.



$P_{\hat{\mathcal{T}}}(\mathcal{X})$
μ_1
μ_m
$P_{\hat{\mathcal{T}}_1}(\mathcal{X})$
...
$P_{\hat{\mathcal{T}}_{m_1}}(\mathcal{X})$
$\lambda_{1,1}$
λ_{1,m_2}
$\lambda_{m_1,1}$
λ_{m_1,m_2}
...
...

Two variants are tested for the second level ensemble. Each tree of the EM mixture is replaced by :

1. m_2 bagged Chow-Liu trees,
2. one Chow-Liu tree and $m_2 - 1$ bagged Chow-Liu trees.

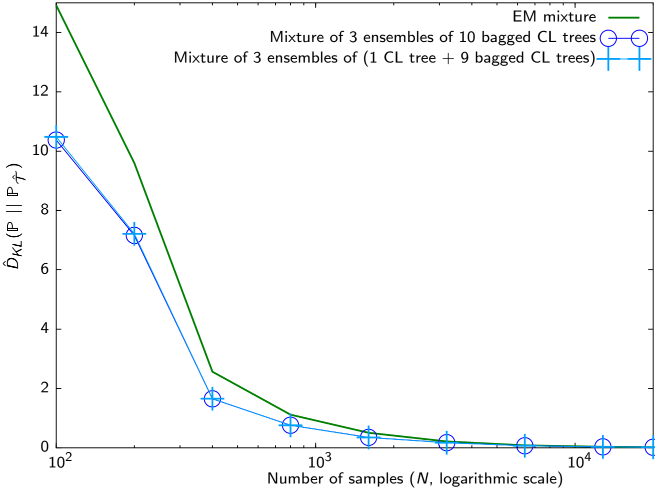
We developed 3 experimental settings.

Accuracy is estimated by the Kullback-Leibler divergence $\hat{D}_{KL}(\mathbb{P} \,||\, \mathbb{P}_{\hat{\mathcal{T}}})$ of the mixtures learned, averaged over several learning sets.

$\mathbb{P}$ can be either :

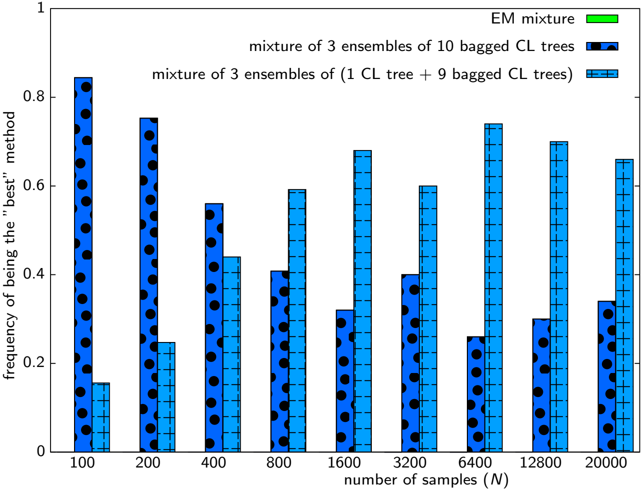
1. a synthetic mixture of three trees
2. synthetic Bayesian networks over binary variables
3. more realistic Bayesian networks (from the literature)

The combined methods achieves a higher accuracy than an EM mixture for learning a mixture of 3 Markov trees.



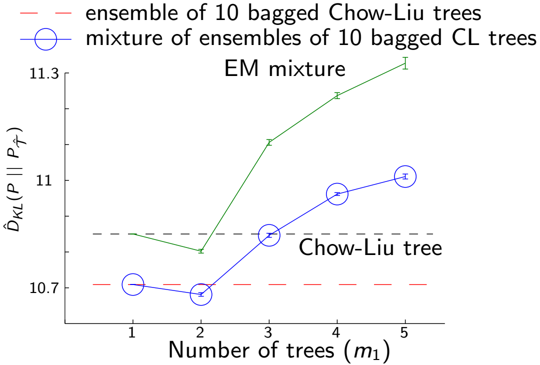
Averaged results on 1 uniformly weighted mixture of 3 randomly generated Markov trees (200 binary variables) × 1 learning set × many initializations of the EM algorithm.

Setting 1 : The accuracy of the EM mixture is always improved by replacing each tree by an ensemble.



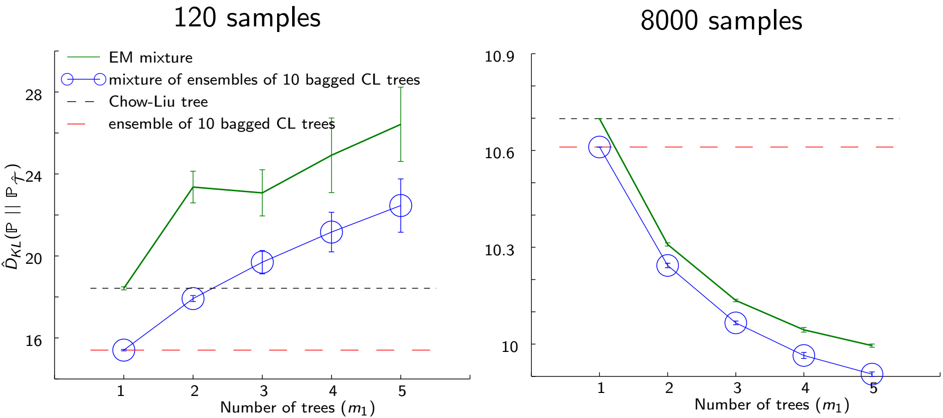
Relative number of runs where each method is the best, displayed by number of samples. MT-EM is always at 0.

Setting 2 : The mixture of ensembles of CL trees can lead to a better accuracy than both the EM mixture or an ensemble of bagged Chow-Liu trees.



Averaged results on 5 randomly generated Bayesian networks (200 binary variables) × 5 learning sets.

Setting 2 : The mixture of ensembles of CL trees is better than the EM mixture, no matter what the optimal m_1 is.



Averaged results on 5 randomly generated Bayesian networks (200 binary variables) × 5 learning sets.

Setting 3 : Larger 2nd level ensembles might be necessary when estimating more complex probability distributions.

Data set	$N = 200$		$N = 500$		$N = 2500$	
	MT-EM	+bag, bagCl	MT-EM	+bag, bagCl	MT-EM	+bag, bagCl
Child10	-	25	-	25	-	10
Pigs	21	4	-	25	-	10
Alarm10	3	22	-	25	-	10
Gene	25	-	-	25	-	10
Lung Cancer	25	-	8	17	-	10
Link	3	22	-	25	-	10
Insurance10	2	23	1	24	-	10
Munin	1	24	6	19	-	10
Hailfinder10	25	-	25	-	-	10
ALL	105	120	40	185	0	90

Table: Best methods on realistic data sets (by increasing complexity) for 5 learning sets × several initializations of the EM mixture, with $m_1 = 2$ and $m_2 = 10$. N is the number of learning samples.

Final words

We replaced each tree of a bias reducing mixture of Markov trees by an ensemble reducing the variance of a single tree.

Conclusions :

- This method
  - does not require the selection of a regularization parameter
  - avoids the use of MCMC exploration schemes.
- The resulting models are able to improve the accuracy of the base EM mixture.

Further work :

- Evaluate the accuracy against other variance reducing schemes for EM mixture (e.g. MCMC exploration).
- Apply it to high-dimensional problems... along with other methods ?

More realistic data sets (by C. Aliferis, A. Statnikov, I. Tsamardinos & al).

Name	p	$\|\mathcal{X}_i\|$	$\|E(\mathcal{G})\|$	$\|\theta\|$
Alarm10	370	2-4	570	5468
Child10	200	2-6	257	2323
Gene	801	3-5	977	8348
Hailfinder10	560	2-11	1017	97448
Insurance10	270	2-5	556	14613
Link	724	2-4	1125	14211
Lung Cancer	800	2-3	1476	8452
Munin	189	1-21	282	15622
Pigs	441	3-3	592	3675

TABLE: Distributions from the literature and their characteristics. p corresponds to the number of variables, $|\mathcal{X}_i|$ to the range of cardinalities of single variables, $|E(\mathcal{G})|$ to the number of edges and $|\theta|$ to the number of independent parameters in the original model.

Setting 3

Best methods on realistic data sets on 5 runs times 5 sets (resp. 10 runs times 1 set) of 200 or 500 (resp. 2500) samples, with $m_1 = 2$ and $m_2 = 10$.

Data set	p	$\|X_i\|$		$N = 200$			$N = 500$			$N = 2500$	
			MT-EM	-Bag	-BagCI	MT-EM	-Bag	BagCI	MT-EM	-Bag	-BagCI
Alarm10	370	2-4	3	5	17	-	6	19	-	8	2
Child10	200	2-6	-	1	24	-	4	21	-	4	6
Gene	801	3-5	25	-	-	-	9	16	-	8	2
Hailfinder10	560	2-11	25	-	-	25	-	-	-	1	9
Insurance10	270	2-5	2	1	22	1	9	15	-	7	3
Link	724	2-4	3	7	15	-	13	12	-	8	2
Lung Cancer	800	2-3	25	-	-	8	-	17	-	8	2
Munin	189	1-21	1	15	9	6	5	14	-	5	5
Pigs	441	3-3	21	1	3	-	16	9	-	9	1
ALL			105	30	90	40	62	123	0	58	32