Efficiently Approximating Markov Tree Bagging for High-Dimensional Density Estimation

F. Schnitzler[1] S. Ammar[2] P. Leray[2] P. Geurts[1] L. Wehenkel[1]

fschnitzler@ulg.ac.be

[1]University of Liège

[2]University of Nantes

6 September 2011

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

Great potential in many applications :

- Bioinformatics (21 000 genes, 1 000 000 proteins)
- Power networks (10 000 transmission nodes in Europe)

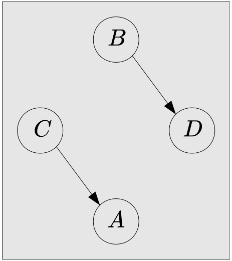
Two main problems :

- Few samples
- Algorithmic complexity

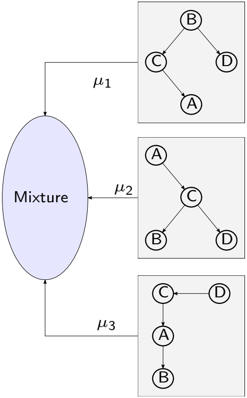
→ Simple models must be used

Mixtures 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 :



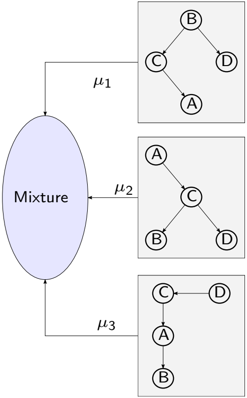
$$\mathbb{P}_{\hat{\mathcal{T}}}(X) = \sum_{i=1}^{m} \mu_i \mathbb{P}_{T_i}(X)$$

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

- Several models → large modeling power
- Simple models → low complexity :
  - inference is linear,
  - learning : most algorithms are quadratic.

There are two types of mixtures :

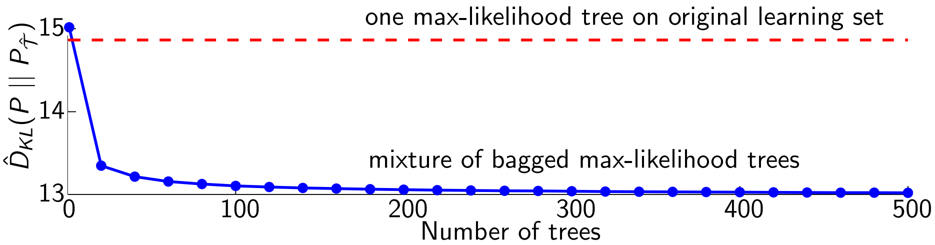
- Maximum likelihood
- Variance reduction



Bagging is a good variance reduction method.

- average over m max-likelihood trees learnt from m bootstrap replicates
  - $\rightarrow$ typically exhibits a lower variance
  - $\rightarrow$ reduction in overfitting
- A bootstrap replicate $\mathbf{D}'$ of a sample set $\mathbf{D}$ is the same size as $\mathbf{D}$ and is drawn with replacement from $\mathbf{D}'$.
- Each additional term improves the mixture.

Example : 200 variables and 200 samples



We developed approximation strategies to accelerate it.

Complexity : $\mathcal{O}(mn^2 \log n)$

- Our goal : speeding up learning without sacrificing accuracy.
- Motivation : We need many terms : it keeps improving.
- Bottleneck : number of candidate edges for each tree.

$$T_i(\mathbf{D}') = \arg\max_T \sum_{(X,Y)\in\mathcal{E}(T)} I_{\mathbf{D}'}(X;Y) \ ,$$

Replicate

A	B	C	D
0	1	0	1
1	1	0	1
0	0	1	1
1	1	1	0

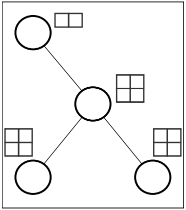
$I(X,Y)$

Edge weights

	A	B	C	D
A		*	*	*
B	*		*	*
C	*	*		*
D	*	*	*	

MWST

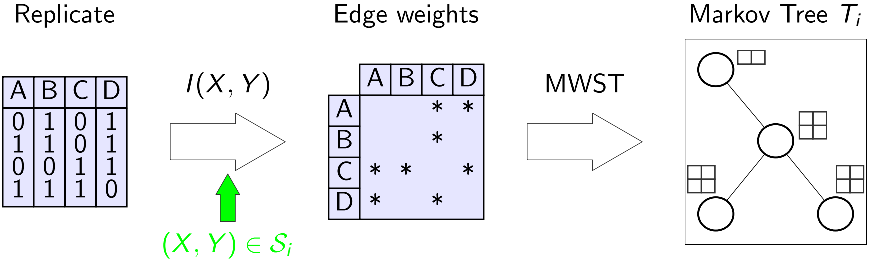
Markov Tree T_i



Key idea of approximation strategies

- Ideas :
  - start with a max-likelihood tree on the original data set
  - exploit previous trees to select a good subset $\mathcal{S}_i$ of candidate edges.

    $\rightarrow$ trees are not independent
- We developed two methods for selecting $\mathcal{S}_i$ of fixed size $|\mathcal{S}|$:
  - Complexity : $\mathcal{O}(mn^2 \log n) \rightarrow \mathcal{O}(n^2 \log n + m|\mathcal{S}| \log |\mathcal{S}|)$
  - Run time : one order of magnitude faster



Replicate

A	B	C	D
0	1	0	1
1	1	0	1
0	0	1	1
1	1	1	0

$I(X, Y)$

$(X, Y) \in \mathcal{S}_i$

Edge weights

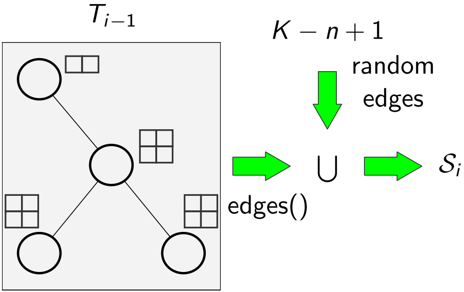
	A	B	C	D
A			*	*
B			*	
C	*	*		*
D	*		*	

MWST

Markov Tree T_i

1 : In the inertial approach, $\mathcal{S}_i$ is based on the previous tree T_{i-1}.

- $|\mathcal{S}_i| = \text{K}$ is a parameter.
- $\forall i \geqslant 2$, $\mathcal{S}_i$ is composed of
  - $n-1$ edges of T_{i-1},
  - $K-n+1$ other randomly sampled edges.
- Explores the set of all Markov Trees defined on the variables.



T_{i-1}

$K-n+1$ random edges

edges()

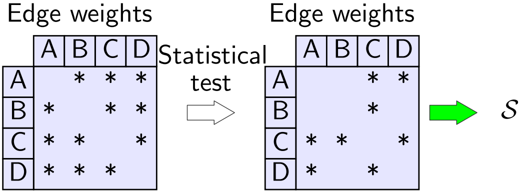
$\bigcup$

$\mathcal{S}_i$

2 : In the skeleton-based approach, all $\mathcal{S}_i$ are equal and based on the first tree.

- Edges with weak weights are
  - not likelily to be part of a tree (even if weights are perturbed),
  - probably not meaningful (noise or not direct relation).

  → We can ignore them in the search.
- $\mathcal{S}$ contains only edges whose associated weight is high.
- Explores the subset of trees (or forests) spanning $\mathcal{S}$.



Edge weights

	A	B	C	D
A		*	*	*
B	*		*	*
C	*	*		*
D	*	*	*	

Statistical test

Edge weights

	A	B	C	D
A			*	*
B			*	
C	*	*		*
D	*		*	

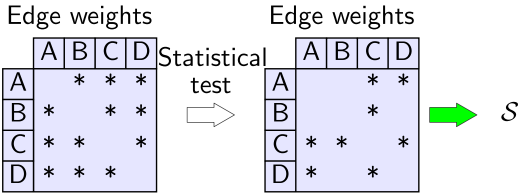
$\mathcal{S}$

Edges are tested for independence before inclusion in $\mathcal{S}$.

- Related to regularization :

$$T^{\lambda}_{CL}(\mathbf{D}) = \arg\max_T \sum_{(X,Y)\in\mathcal{E}(T)} I_{\mathbf{D}}(X;Y) - \lambda|T|$$

- Comparing $I_{\mathbf{D}}(X;Y)$ (χ-square distributed under independence) to a threshold depending on a postulated p-value, say $\alpha = 0.05$ or smaller.
- $\mathcal{S}$ contains the pairs of variables whose mutual information (on the original data set) is above the threshold.
- Mutual information values are a by-product of the computation of the first tree.



Edge weights

	A	B	C	D
A		*	*	*
B	*		*	*
C	*	*		*
D	*	*	*	

Statistical test

Edge weights

	A	B	C	D
A			*	*
B			*	
C	*	*		*
D	*		*	

$\mathcal{S}$

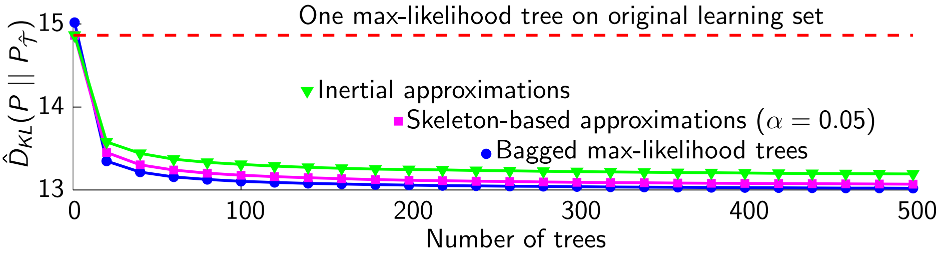
We evaluated our algorithms on synthetic and more realistic data sets.

Synthetic bayesian networks over binary variables :

- for each X_i
  - draw the number of parents in $[0, max(5, i-1)]$
  - randomly selecting these parents in $\{X_1, ..., X_{i-1}\}$.
- 200 and 1000 variables ; 200, 600 and 1000 observations.
- Validation by Monte-Carlo estimation of the Kullback-Leibler divergence (50 000 observations).

The two approaches are working well.

200 samples, 200 variables :

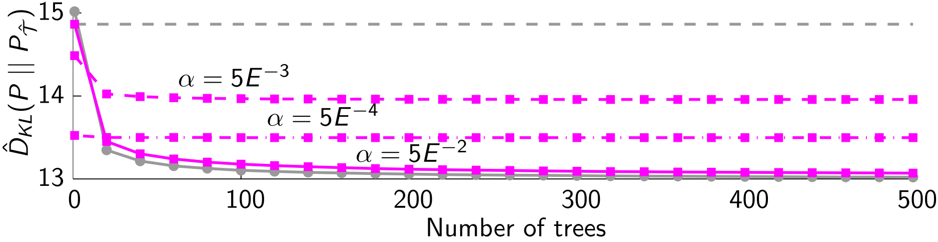


Relative run-time for mixtures of 500 trees (one max-likelihood tree : 1) :

- Bagged max-likelihood trees : 532
- Inertial approximations : 45
- Skeleton-based approximations : 21

Influence of the parameter α in the Skeleton-based approximation :

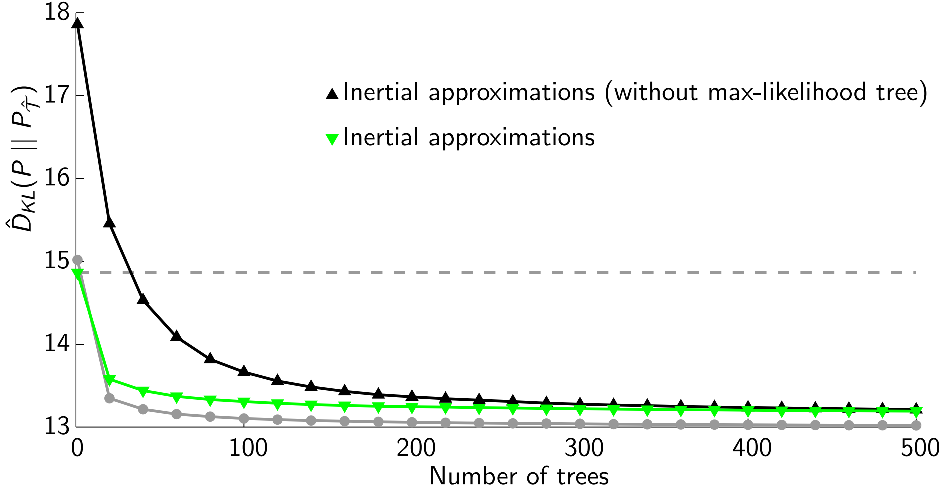
200 samples, 200 variables :



- The lower α, the faster the convergence.
- Regularization improves the first tree, but averaging over more diverse trees leads to better approximations.

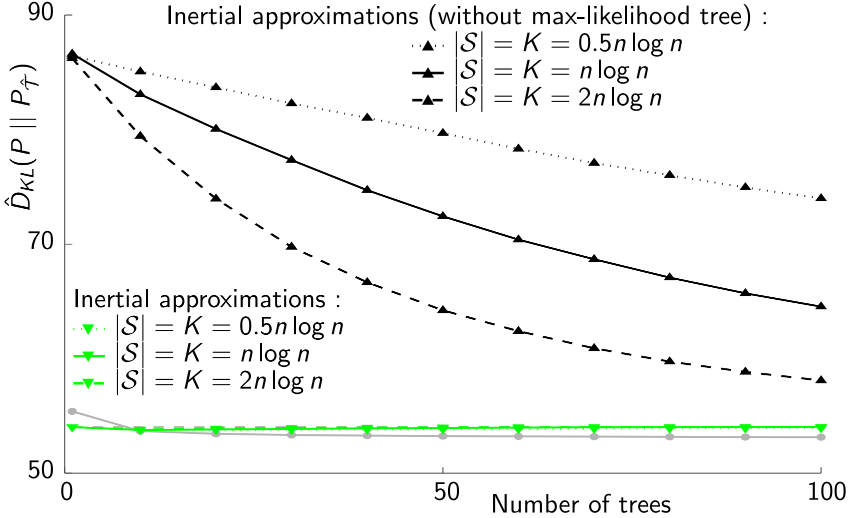
Starting by the max-likelihood tree is necessary in the inertial method.

200 samples, 200 variables :



Starting by the max-likelihood tree is necessary in the inertial method.

1000 samples, 1000 variables :



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

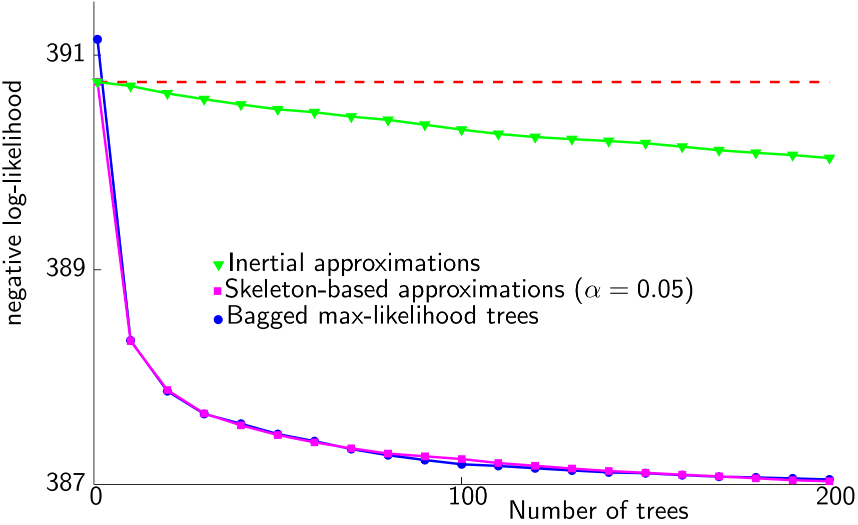
- 9 models ranging from 200 to 801 variables ; 200 and 500 samples :
  - 4 classical networks extended by tiling (Child10, Insurance10, Alarm10, Hailfinder10)
  - 2 data sets ressimulated from gene expression data (Gene, Lung Cancer)
  - 3 expert systems (Munin, Link,Pigs)
- validation by negative log-likelihood of an independent set of 5000 observations

Summary :

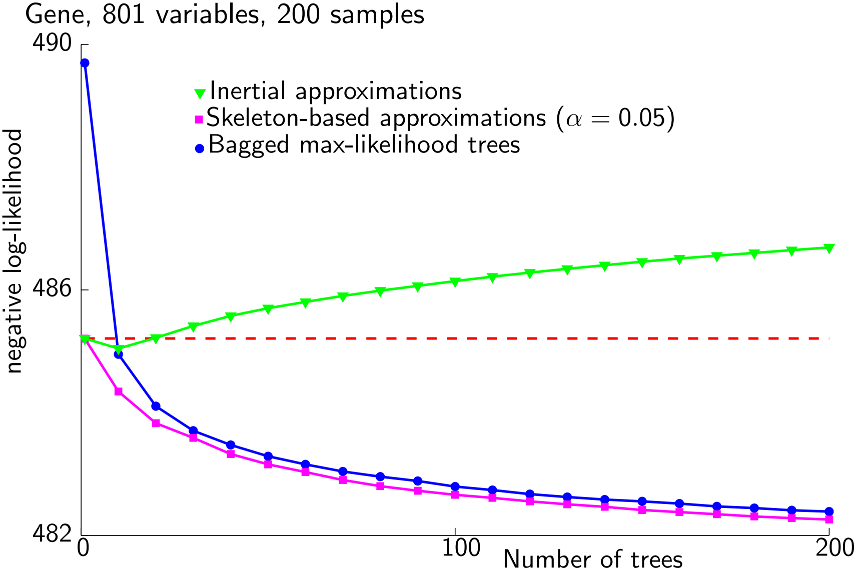
- Both approximations methods are working well : 2 instances
- Only the skeleton approach is working well : 8 instances
- 8 instances where we cannot conclude.

Both approximations are better than a maximal-likelihood tree in two experimental cases.

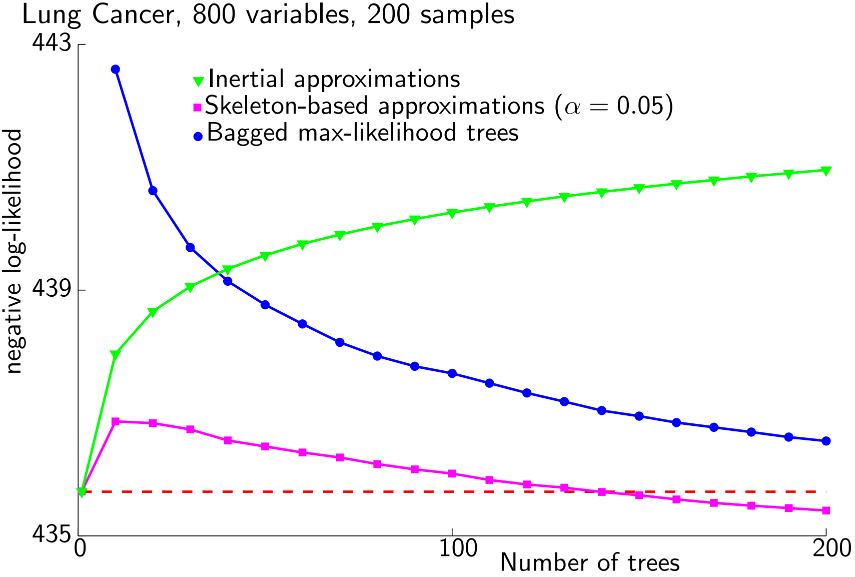
Pigs, 441 variables, 200 samples



In most cases only the skeleton-based approximation is good.



In one case the skeleton approach first degrades the maximum-likelihood tree before slowly improving.



Conclusions

- We propose two algorithms for learning mixtures of Markov trees designed to approach the quality of approximation of mixtures of bagged Chow-Liu trees at a lower computational cost.
- They exploit the computation of the previous or first tree of the mixture in order to test fewer edges in the subsequent trees.
- Searching only significant edges (as assessed on the original data set) is the most robust approach.

TABLE: Impact of the parameter α on the number of edges, averaged on 5 densities times 6 data sets for $n = 1000$ variables and $p = 200$ samples

Edges	Numbers (% of the total) for $\alpha =$			
	$1E^{-1}$	$5E^{-2}$	$5E^{-3}$	$5E^{-4}$
in T_1	998	997.9	993.2	626.8
in $\mathcal{S}$	52278(10.5%)	26821(5.36%)	3311(0.66%)	683 (0.13%)