Learning Bayesian Networks from incomplete datasets

Philippe Leray and Olivier François
Ahmad Faour[1,2], Iyad Zaarour[1,2]

2nd Workshop on Applying Graphical Models
January 24-25, 2005

[1]INSA de Rouen - PSI Laboratory fre cnrs 2645, Rouen, France
[2]Lebanese university, Beyrouth, Lebanon

The PSI Lab.

PSI for `Perseption, Systems, Information`

3 thematics :

- DiCO: knowledge extraction from documents (electronic or manuscript) and *e*-learning
- PerSe: vision and sensor team (medical image analysis, pedestrian detection)
- **TAC**: contextual learning team (Machine Learning, Classification)

The TAC is member of the PASCAL european network of excellence (Pattern Analysis, Statistical Modelling and Computational Learning)

- SVM team
- **BN** team
- others : multi-modal modelisation, neural networks...

BN team collaboration and application of BN

A. **European Aeronautic Defence and Space compagny** (past)
⇒ Detection of illicit activities on Internet
B.Grilheres, S.Brunessaux, Ph.Leray : *Combining classifiers for harmful document filtering*, RIAO'2004, `(BN as fusion classifiers)`

B. **Lebanese university** and **PsyCO Lab.** (past)
⇒ *Clustering and bayesian network approaches for discovering handwriting strategies of primary school children*, International Journal of Pattern Recognition and Artificial Intelligence → I.Zaarour

C. **Lebanese university** and **'Rectorat' of Upper Normandy**
⇒ Informatic network intrusion detection → A.Faour

BN team collaboration and application field of BN

- **Vrije Universiteit Brussel** (new collaboration)
  ⇒ Hybrid score-based and constraint-based learning methods from incomplete data → S. Meganck

- **Thales Air Defence** and **CORIA Lab.** (future ?)
  ⇒ Nonlinear thermal models (Regional research project)

- **IMdR-SdF** and **INERIS** (future ?)
  ⇒ Industrial Risks Control (PhD proposal)

My prefered applications...

- Medical diagnosis support
- Robotics

...but as this time my work is purely academic.

Plan

Problems

if $\mathcal{S}$ is a complex system.
$\mathcal{S}$ is represented by the attributes $\{X_i\}_{1 \leqslant i \leqslant n}$.

- Some attributes are systemticaly observed,
- Some other are occasionally observed,
  - critical states of the system ?
  - expensive measurement ?...
- many others are never observed,
  - because their influence/pertinence is weak ?
  - because their interest is not known ?...

Goal

To find how the $\neq$ attributes interact ?

$\longrightarrow$ *Using a probability law*

- good means of modelling chaotic or complex systems (medicine, weather prediction...)
- a same configuration of the observed attributes leads to differents state.

Favour of such a modeling :

- essential statistics for the state prediction,
- confidence measurement in the prediction (its conditional likelyhood).

Types of missing data

If $R = (r_{li})_{m \times n}$ is the matrix where $r_{li} = 1$ if $d_{li} =$ 'missing' and 0 if not and $\mathbf{D} =< \mathbf{O}, \mathbf{H} >= (d_{li})_{m \times n}$.

$$\mathbb{P}(\mathbf{O}, \mathbf{H}, R|\Theta, \mu) = \mathbb{P}(\mathbf{O}, \mathbf{H}|\Theta) \times \mathbb{P}(R|\mathbf{O}, \mathbf{H}, \mu)$$

Where $\Theta \longrightarrow$ generate the dataset $\mathbf{D}$
and $\mu \longrightarrow$ generate the missing data $\mathbf{H}$

3 types :

- MCAR : $\mathbb{P}(R|\mathbf{O}, \mathbf{H}, \mu) = \mathbb{P}(R|\mu)$
- MAR : $\mathbb{P}(R|\mathbf{O}, \mathbf{H}, \mu) = \mathbb{P}(R|\mathbf{O}, \mu)$
- NMAR : $\mathbb{P}(R|\mathbf{O}, \mathbf{H}, \mu)$

Learning with complete data

Which technique ?

- Causality search
- Score-based search
- Mixed methods

Which space ?

- exhaustive search : impossible / space size.

  number of possible structures from n nodes (Robinson 77)

  $$NS(n) = n^{2^{\mathcal{O}(n)}} \qquad NS(5) = 29281 \qquad NS(10) \simeq 4.2 \times 10^{18}$$

- Trees
- DAG respecting a nodes ordering
- DAG (heuristics)
- Markov equivalent representatives (heuristics)

Only one network ?

State of the art

Two complementary approaches :

- with statistical tests
  - PC [Spirtes 93],
  - IC [Pearl 93],
  - BNPC [Cheng 02]
  - QFCI [Badea 04]...
- score based ($S(\mathcal{G}) = LL + P = \sum_{i=1}^{n} S(X_i|Pa(X_i)) \rightarrow$ local)
  - MWST [Heckerman 94],
  - K2 [Cooper 92],
  - GS,
  - GES [Chickering 02],
  - Algo. G. pour trouver un ordre [Larrañaga 96],
  - Algo. G. [Wong 03],
  - Fourmis [De Campos 02],
  - MCMC pour trouver un ordre [Friedman 00],
  - MCMC [Murphy 01]...

Mixed methods

- BENEDICT [De Campos 01]
- [Dash & Druzdzel 99]...

Learning with incomplete data

RB can perform inference and parameters learning from incomplete datasets (EM, MCMC, multiple imputation...).

What happens for the structure learning ?

- AMS-EM : gready search in the dag space with bic score (Friedman 97),
- BS-EM : gready search in the dag space with *bayesian* score (Friedman 98),
- Algo. G. and MCMC (Myers 99),
- RBE (Ramoni & Sebastiani 01),
- Hybrid-IT (Dash & Druzdzel 03).

Scoring a BN with incomplete data

If $S(\mathcal{M}|\mathbf{D}_c)$ is a score for $\mathcal{M}$ and for a complete dataset $\mathbf{D}_c$.
$\rightarrow$ approximating the score for $\mathcal{M}$ and an incomplete dataset $\mathbf{D}$

$$Q^S(\mathcal{M}|\mathbf{D}) = \mathbb{E}_{\mathbf{H}\sim\mathbb{P}(\mathbf{H})}(S(\mathcal{M}|\mathbf{O},\mathbf{H})) \quad (1)$$

But, the law $\mathbb{P}(\mathbf{H})$ is unknown.

If $\mathcal{M}^0$ is a supposed generative model of $\mathbf{D}$. Then :

$$Q^S(\mathcal{M}|\mathbf{D}) \approx Q^S(\mathcal{M}:\mathcal{M}^0|\mathbf{D}) = \mathbb{E}_{\mathbf{H}\sim\mathbb{P}(\mathbf{H}|\mathcal{M}^0)}(S(\mathcal{M}|\mathbf{O},\mathbf{H}))$$

$$Q^S(\mathcal{M}:\mathcal{M}^0|\mathbf{D}) = \sum_{\mathbf{H}} S(\mathcal{M}|\mathbf{O},\mathbf{H})\mathbb{P}(\mathbf{H}|\mathcal{M}^0)) \quad (2)$$

Now, $\mathbb{P}(\mathbf{H}|\mathcal{M}^0)$ is known.

Generic EM for structure learning

- Choose a model $\Rightarrow \mathbb{P}(\mathbf{H}|\mathcal{M}^0)$
- Find 'a' model which improve $Q^S(\mathcal{M} : \mathcal{M}^0|\mathbf{D})$
- Use this new model for the next itération
  until 'convergence'.

SEM : 'the' better model is choosen in the *neigborhood* of the curent network.
MWST-EM : Take 'the' best model in the *tree space*.

Motivation

Previously seen [RFIA04]

1. MWST : good speed / perf. ratio than GS
2. GS+MWST : makes GS more stable
3. AMS-EM : GS+EM

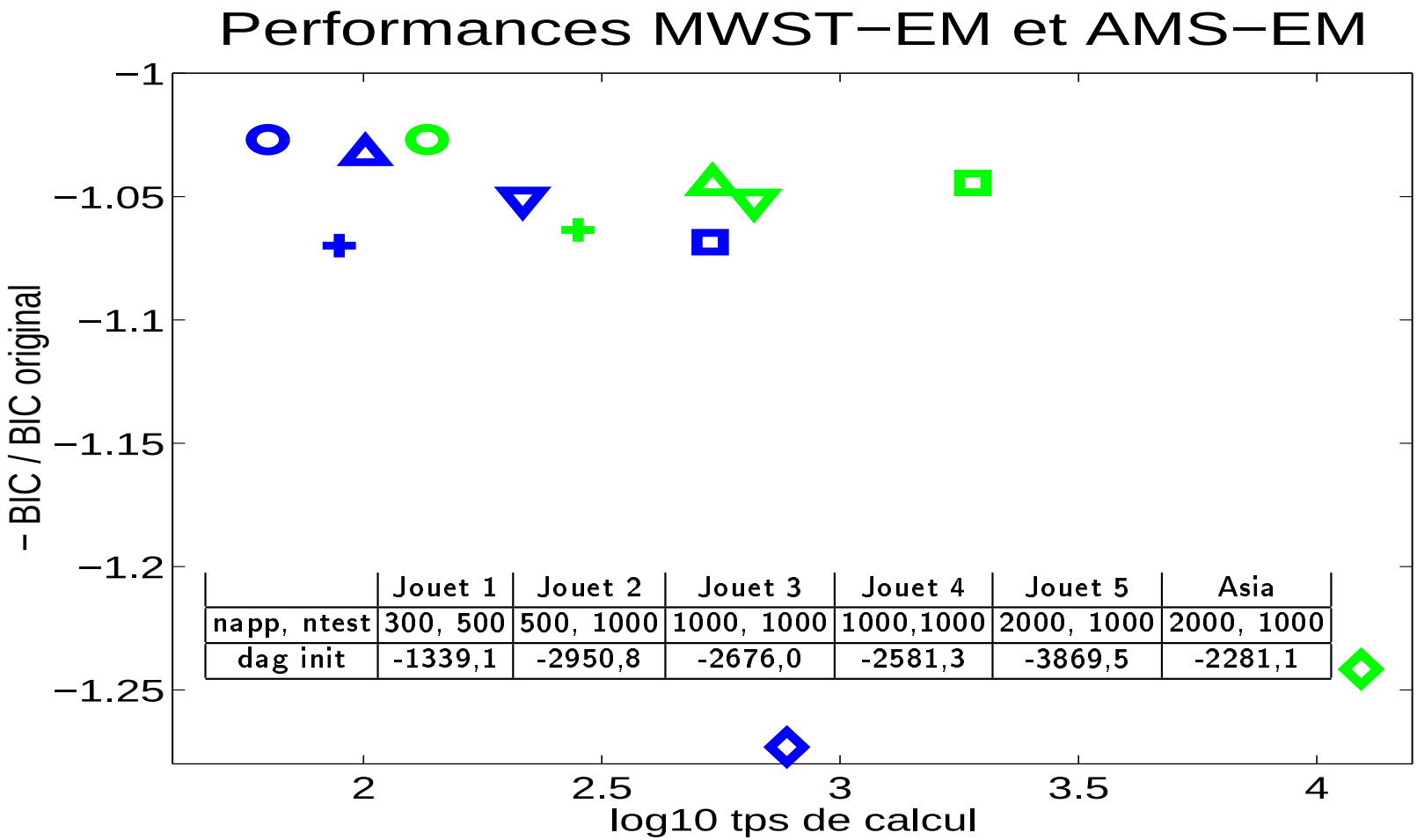
Goal : $\boxed{\text{MWST-EM} = \text{MWST} + \text{EM}}$

$\longrightarrow$ adaptation of the score evaluation :

$$q_i^{BIC}(X_i, Pa(X_i) : \mathcal{B}^0 | \mathbf{D}) = \sum_{X_i} \sum_{Pa(X_i)} \log(\widehat{\theta}_{X_i|Pa(X_i)}) \mathbb{E}_{\mathbf{H} \sim \mathbb{P}(\mathbf{H}|\mathcal{B}^0)}(N_{X_i, Pa(X_i)}) - \frac{1}{2} Dim(X_i, Pa(X_i)) \log N$$

A. MWST-EM : speed / perf. ratio than AMS-EM ?
B. MWST-EM to initialise de AMS-EM ?

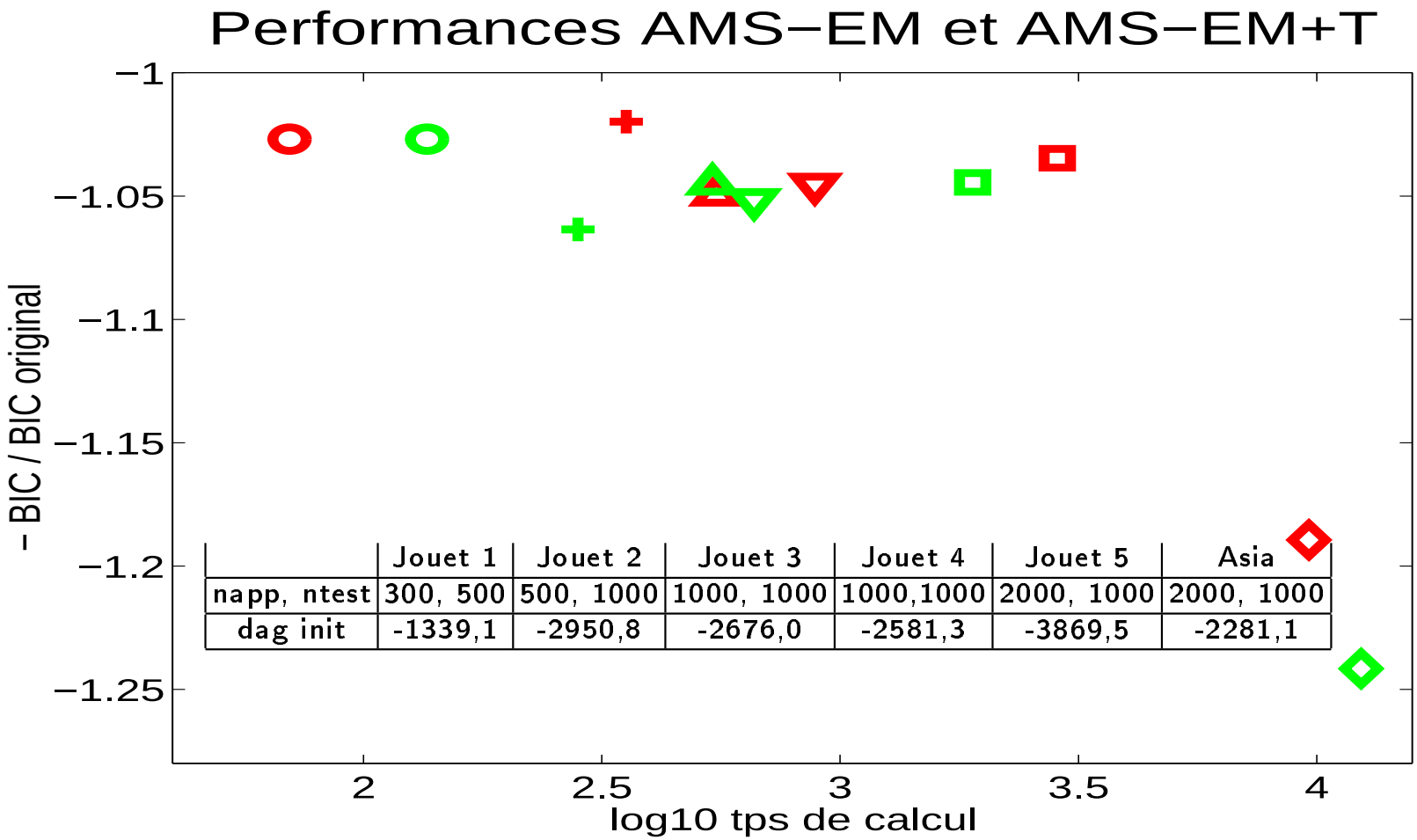
Préliminaries Results: MCAR data



	Jouet 1	Jouet 2	Jouet 3	Jouet 4	Jouet 5	Asia
napp, ntest	300, 500	500, 1000	1000, 1000	1000,1000	2000, 1000	2000, 1000
dag init	-1339,1	-2950,8	-2676,0	-2581,3	-3869,5	-2281,1

- MWST-EM speedly gives a 'good' result.

Préliminaries Results: MCAR data



	Jouet 1	Jouet 2	Jouet 3	Jouet 4	Jouet 5	Asia
napp, ntest	300, 500	500, 1000	1000, 1000	1000,1000	2000, 1000	2000, 1000
dag init	-1339,1	-2950,8	-2676,0	-2581,3	-3869,5	-2281,1

- AMS-EM is a 'rare' efficient method but *not very stable*.
- Initialising AMS-EM with a tree allow to stabilise the method, to improve performances and/or to reduce the computing time.

Short-term prospects

In progress :

- Testing these methods on MAR data.
- Testing these methods on classification problems.

Change of space :

- AMS-EM : space of DAG
- MWST-EM : part of the Markov equivalent representatives space
- $\Longrightarrow$ *SEEM = GES + EM*
- $\rightarrow$ *SEEM+T ?*

Long-term prospects

classical BN :

- Incoporate operators to detect hidden variables
- ... and to estimate their sizes.

dynamic BN :

- Adapt this work to the slices learning,
- " to the interslices structure learning.

Thank you for your attention.