Apprentissage de structure à partir de données incomplètes et application à la classification

Olivier François, Philippe Leray
Francois.Olivier.C.H@gmail.com,
Philippe.Leray@insa-rouen.fr

Laboratoire d'Informatique, Traitement de l'Information, Systèmes.

3ème Journées Francophones sur les Réseaux bayésiens
organisées à valenciennes (Nord)

les 18 et 19 juin.

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes
LITIS

Plan

Problématique

Soit $\mathcal{X}$ un système complexe.
$\mathcal{X}$ est représenté par de nombreux attributs $\{X_i\}_{1 \leqslant i \leqslant n}$.

- Certains attributs sont observés **systématiquement**,
- d'autres sont observés **occasionnellement**,
  - état *critique* du système ?
  - mesure *couteuse* ?...
- et de nombreux autres ne sont **jamais** observés,
  - parce que leur *influence/pertinence est faible* ?
  - parce que l'on ne *pas connaissance* de leur intérêt ?...

Par Exemple : Pour une base de 2000 exemples sur 20 attributs,
20% des mesures sont `manquantes complètement au hasard`
$\Longrightarrow$ en moyenne *seulement* 23 cas complets (c-à-d $\%EI \simeq 99\%$)

Types de données manquantes

Notations :

$$\mathbf{D} =< \mathbf{O}, \mathbf{H} >= ((d_{il}))_{n \times m}$$

$\mathbf{R} = ((r_{il}))_{n \times m}$, une matrice où $r_{il} = 1$ si d_{il} est manquant, 0 sinon.
Θ, paramètres de la loi qui a généré $\mathbf{D}$,
μ, paramètres de la loi qui a généré $\mathbf{R}$.

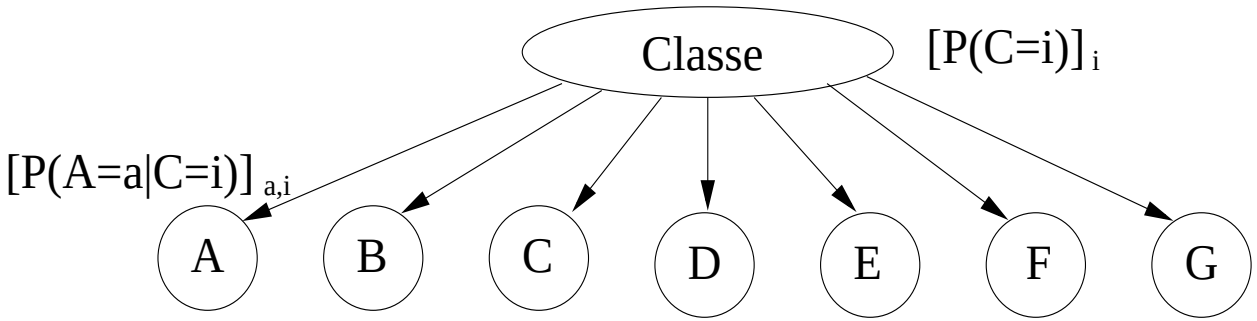
Données manquantes ? (Rubin,76)

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

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

Le réseau Naïf

Supposons que *la classe a une influence sur toutes les variables*, **mais** indépendamment

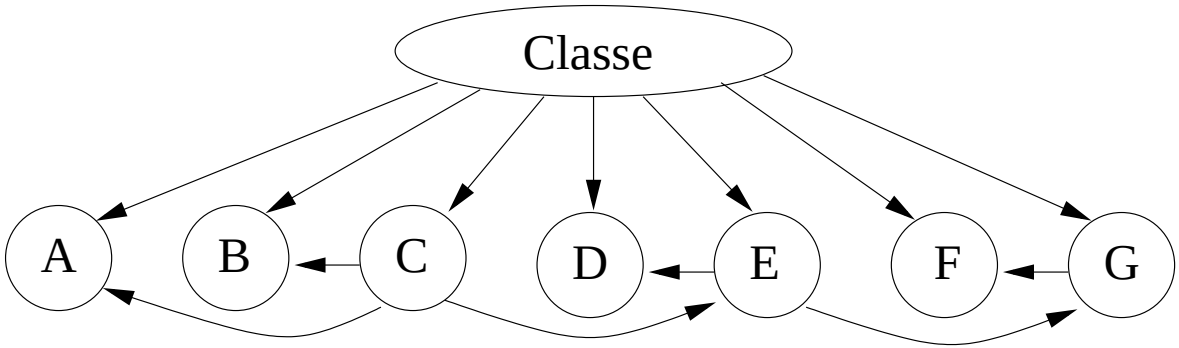


Ce qu'il est possible de faire :

- l'apprentissage des paramètres avec DI (par ex. avec EM),
- l'inférence avec des Données Incomplètes.

Le réseau Naïf

Supposons que *la classe a une influence sur toutes les variables*, **mais** *plus* indépendamment



Ce qu'il est *toujours* possible de faire :

- l'apprentissage des paramètres avec DI (par ex. avec EM),
- l'inférence avec des Données Incomplètes.

et
si l'on veut ajouter des dépendances automatiquement : ?

(peu de méthodes efficaces à partir de DI)

Plan

Calculer un score

Soit $\mathcal{S}(\mathcal{M}|\mathbf{D}_c)$, score pour un modèle $\mathcal{M}$ et des données complètes $\mathbf{D}_c$.
$\rightarrow$ approximation de $\mathcal{S}$ pour $\mathcal{M}$ et la base incomplète $\mathbf{D} =< \mathbf{O}, \mathbf{H} >$

$$Q^{\mathcal{S}}(\mathcal{M}|\mathbf{D}) = \mathbb{E}_{\mathbf{H}\sim\mathbb{P}(\mathbf{H}|\mathbf{O},\mu)}\left[\mathcal{S}(\mathcal{M}|\mathbf{O},\mathbf{H})\right]$$

Mais la loi $\mathbb{P}(\mathbf{H}|\mathbf{O},\mu)$ est inconnue.

Principe EM

Supposons que le modèle $\mathcal{M}^0$ a généré la base $\mathbf{D}$ alors

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

Ou la loi *a posteriori* $\mathbb{P}(\mathbf{H}|\mathbf{O},\mathcal{M}^0)$ est connue.

Structural-EM

- Choisir un modèle $\mathcal{M}^0$ ($\Rightarrow \mathbb{P}(\mathbf{H}|\mathbf{O}, \mathcal{M}^0)$)
- Trouver un modèle $\mathcal{M}^{i+1}$ qui maximise** un score $Q^{\mathcal{S}}(\mathcal{M} : \mathcal{M}^i|\mathbf{D})$
- Utiliser le nouveau modèle comme base pour l'itération suivante jusqu'à convergence.

**EM généralisé : augmente le score

AMS-EM : le nouveau modèle est choisi parmis les *voisins* du graphe courant (Friedman,97)

MWST-EM

AMS-EM : le nouveau modèle est choisi parmis les *voisins* du graphe courant (Friedman,97)$\longrightarrow$ nombreuses itérations

MWST-EM : nous trouvons le 'meilleur' modèle dans l'espace des arbres (François & Leray,05)$\longrightarrow$ peu d'itérations

Pour cela, on utilise un algorithme de type `Kruskal`
sur la matrice de score suivante :

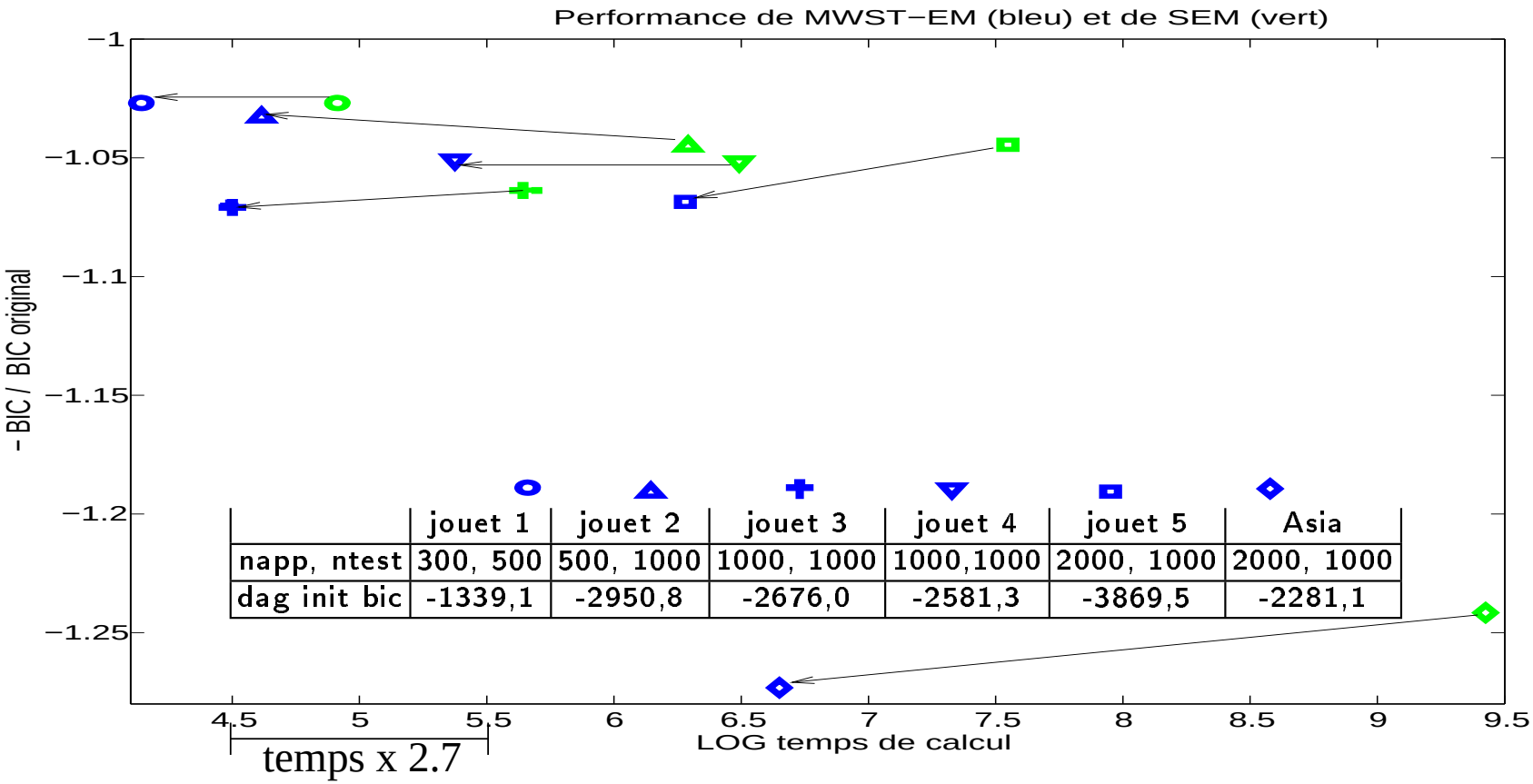
$$\left[M^Q_{ij}\right]_{i,j} = \left[Q^{bic}(X_i, P_i = \{X_j\}, \Theta_{Xi|X_j} : \mathcal{T}^*, \Theta^*) - Q^{bic}(X_i, P_i = \emptyset, \Theta_{Xi} : \mathcal{T}^*, \Theta^*)\right]$$

où $Q^{BIC}(\mathcal{G}, \Theta : \mathcal{G}^*, \Theta^*) = \sum_i Q^{bic}(X_i, P_i, \Theta_{Xi|P_i} : \mathcal{G}^*, \Theta^*)$ et

$$Q^{bic}(X_i, P_i, \Theta_{Xi|P_i} : \mathcal{G}^*, \Theta^*) = \sum_{X_i = x_k} \sum_{P_i = pa_j} N^*_{ijk} \log \theta_{ijk} - \frac{\log N}{2} \mathrm{Dim}(\Theta_{Xi|P_i})$$

avec $N^*_{ijk} = E_{\mathcal{G}^*,\Theta^*}[N_{ijk}] = N * P(X_i = x_k, P_i = pa_j | \mathcal{G}^*, \Theta^*)$.

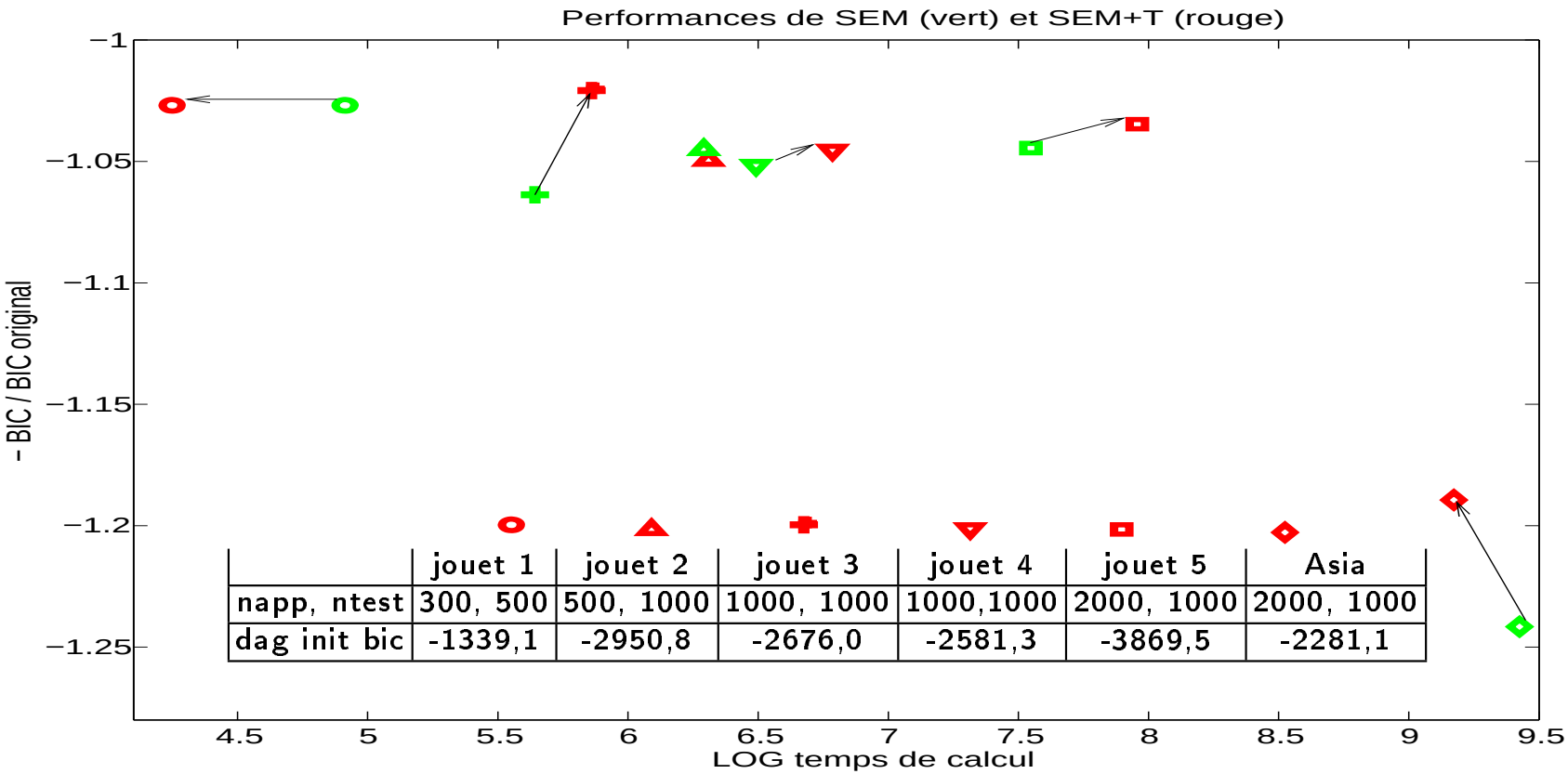
Résultats Préliminaires : données MCAR



	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 bic	-1339,1	-2950,8	-2676,0	-2581,3	-3869,5	-2281,1

- MWST-EM donne de 'bons' résultat.
  - L'arbre est proche de la solution
  - Le temps de calcul est faible

Résultats Préliminaires : données MCAR

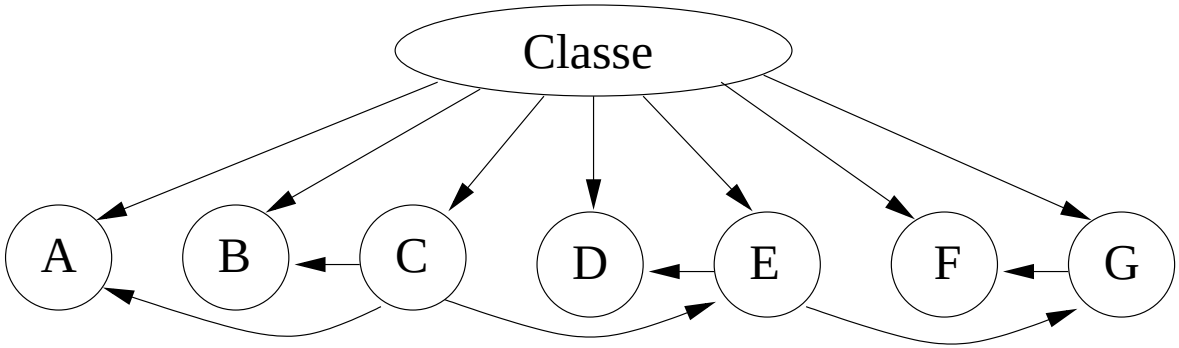


	jouet 1	jouet 2	jouet 3	jouet 4	jouet 5	Asia
napp, ntest	300, 500	500, 1000	1000, 1000	1000,1000	2000, 1000	2000, 1000
dag init bic	-1339,1	-2950,8	-2676,0	-2581,3	-3869,5	-2281,1

- SEM n'est pas *stable* (initialisation).
- Initialiser SEM avec un arbre optimal
  - stabilise la méthode,
  - *meilleurs résultats* et/ou mêmes résultats plus *rapidement*.

Plan

TAN-EM



Pour augmenter le réseau naïf par un arbre,

la classe fait toujours partie de l'ensemble des parents

et la matrice de score devient ($i, j \neq classe$) :

$$\left[M_{ij}^{Q}\right]_{i,j} = \Big[Q^{bic}(X_i, P_i = \{C, X_j\}, \Theta_{Xi|X_j} : \mathcal{T}^*, \Theta^*) - Q^{bic}(X_i, P_i = \{C\}, \Theta_{Xi} : \mathcal{T}^*, \Theta^*)\Big]$$

Résultats

	N	N app	N test	#C	%EI
Hepatitis	20	90	65	2	8.4
House	17	290	145	2	46.7
Horse	28	300	300	2	88.0
Thyroid	22	2800	972	2	29.9
Mushrooms	23	5416	2708	2	30.5

	NB-EM	MWST-EM	TAN-EM	AMS-EM	AMS-EM+T
Hepatitis	70.8%	73.8%	**75.4%**	66.1%	66.1%
	-1224 ; 29	**-1147** ; 90	**-1148** ; 88	-1211.5 ; 1213	-1207 ; 1478
House	89.7%	**93.8%**	92.4%	92.4%	**93.8%**
	-2203 ; 110	-2518 ; 157	**-2022** ; 180	-2524 ; 1732	-2195 ; 3327
Horse	75%	77.9%	**80.9%**	66.2%	66.2%
	-5589 ; 227	**-5199** ; 656	-5354 ; 582	-5348 ; 31807	-5318 ; 10054
Thyroid	95.3%	93.8%	**96.2%**	93.8%	93.8%
	-39348 ; 1305	-38881 ; 3173	**-38350** ; 3471	**-38303** ; 17197	-39749 ; 14482
Mushrooms	**92.8%**	74.7%	91.3%	74.9%	74.9%
	-97854 ; 2028	-108011 ; 6228	**-87556** ; 5987	-111484 ; 70494	-110828 ; 59795

Conclusions

La méthode TAN-EM permet d'obtenir :

- de bonnes performances en classification,
- de bonnes vraissemblances des modèles obtenus,
- un excellent rapport *performances/rapidité*
  (car basée sur le réseau naïf et MWST-EM **?**).

Néanmoins, cette méthode

- est limitée aux tâches de classification,
- augmente le NB *forcement* par un arbre.

Perspectives

Pour TAN-EM :

- adaptation à la classification non-supervisée,
- tests sur des données générées (MAR),
- tests en non-supervisé.

Pour MWST-EM :

- passer de l'espace des arbres à
  - l'espace des *forets* ($\Longrightarrow$ FAN-EM),
  - l'espace des *équivalents de Markov* ($\Longrightarrow$ GES-EM et BNAN-EM ?),
- remplacer les principes de EM par d'autres (Robust Bayesian Estimator...) $\Longrightarrow$ NMAR.

Merci pour votre attention.

- Questions ?
- Remarques ?
- Suggestions ?