Quand devons-nous utiliser l'algorithme EM pour effectuer un apprentissage de structure de RB ?

Olivier François

Quand devons-nous utiliser l'algorithme EM pour effectuer un apprentissage de structure de RB ?

Illustration à l'aide de l'exemple ASIA.

Olivier François[1]

Construction Management and Engineering, URS building, Room 3n38, Whiteknights, PO Box 219, Reading, RG6 6AW, UK.

francois.olivier.c.h@gmail.com, http://ofrancois.tuxfamily.org

RÉSUMÉ. De nombreuses méthodes d'apprentissage de réseaux bayésiens sont apparues durant ces deux dernières décénies. Néanmoins, peu de techniques proposent actuellement d'effectuer un tel travail lorsque la base d'exemples est incomplète. Nous avons adapté certains algorithmes gloutons à cette tâche en utilisant soit l'algorithme Expectation-Maximisation*, soit la méthode* Pairwise Deletion*. La méthode EM s'avère être la plus communément utilisée pour traiter les bases incomplètes. Cependant, à travers une étude expérimentale basée sur l'exemple bien connu d'*ASIA*, nous allons montrer que l'utilisation de cet algorithme ne doit pas être favorisée dans toutes circonstances.*

ABSTRACT. Many Bayesian Network Structure Learning algorithms have been proposed during the twentieth last years but few technics deal with missing data. We have adapted classical gready algorithms like MWST, GS and GES to deal with incomplete datasets in two ways. First, we use these technics together with the EM algorithm as it is the main common used technics to deal with missing data. Then, we also have upgraded those gready algorithms with the pairwise data deletion estimation which is very computationaly efficient when learning could be done localy. Through an experimental study, we show that the EM algorithm does not have to be used in any circumstances.

MOTS-CLÉS : Apprentissage de structure, données incomplètes, EM.

KEYWORDS: Structure learning, missing data, pairwise data deletion.

1. J'aimerai tout particulièrement remercier Philippe Leray pour les nombreuses discussions que nous avons eues et qui ont guidé ces travaux

1. Introduction

La détermination d'un réseau bayésien $\mathcal{B} = (\mathcal{G}, \theta)$ nécessite la définition d'un graphe acyclique dirigé (DAG) $\mathcal{G}$ dont les sommets représentent un ensemble de variables aléatoires $X = \{X_1, \cdots, X_n\}$ (la structure), et de matrices de probabilités conditionnelles du nœud i connaissant l'état de ses parents $Pa(X_i)$ dans $\mathcal{G}$, $\theta_i = [\mathbb{P}(X_i / X_{Pa(X_i)})]$ (les paramètres).

De nombreuses méthodes d'apprentissage de structure de réseaux bayésiens ont vu le jour ces dernières années. Alors qu'il est possible de faire de l'apprentissage de paramètres de réseaux bayésiens à partir de données incomplètes et que l'inférence dans les réseaux bayésiens est possible même lorsque peu d'attributs sont observés (Jensen, 1996; Pearl, 1998; Naïm *et al.*, 2004), les algorithmes d'apprentissage de structure avec des données incomplètes restent rares.

Il est possible de différencier trois types de données manquantes selon le mécanisme qui les a générées. Le premier type représente les données manquantes au hasard (MAR, *missing at random*). Dans ce cas, la probabilité qu'une variable ne soit pas mesurée ne dépend que de l'état de certaines autres variables observées. Lorsque cette probabilité ne dépend plus des variables observées, les données manquantes sont dites MCAR (*missing completely at random*). Par contre lorsque la probabilité qu'une variable soit manquante dépend à la fois de l'état de certaines autres variables observées mais également de phénomènes extérieurs, les données sont dites NMAR (Rubin, 1976).

Lorsque les données sont incomplètes, il est possible de déterminer les paramètres et la structure du réseau bayésien à partir des entrées complètes de la base. Comme les données manquantes sont supposées l'être aléatoirement, nous construisons ainsi un estimateur sans biais. Néanmoins, dans l'exemple d'une base de 2000 cas sur 20 attributs, avec une probabilité de 20% qu'une mesure soit manquante, nous ne disposerons en moyenne que de 23 cas complets. Les autres données à notre disposition ne sont donc pas négligeables et il serait donc préférable de faire l'apprentissage en utilisant toute l'information à laquelle nous avons accès.

Un avantage des réseaux bayésiens est qu'il suffit que seules les variables X_i et $Pa(X_i)$ soient observées pour estimer la table de probabilité conditionnelle correspondante. Dans ce cas, il est alors possible d'utiliser tous les exemples (même incomplets) où ces variables sont observées (dans l'exemple précédent, en supposant que X_i possède trois parents, nous aurions 819 exemples en moyenne pour estimer les paramètres correspondants).

À la manière de (Friedman, 1997), nous proposons d'utiliser un algorithme EM pour l'apprentissage de structure de réseaux bayésien. Cette

méthode sera appliquée à l'algorithme MWST (Chow *et al.*, 1968; Heckerman *et al.*, 1994; François, 2006), à une méthode d'apprentissage gloutonne dans l'espace des graphes dirigés sans cycle (Cooper *et al.*, 1992; Friedman, 1998), ainsi qu'à la technique GES proposée par (Chickering, 2002a). Nous proposons également d'utiliser seulement les cas disponibles pour évaluer localement les paramètres d'un nœud conditionnellement à l'état de ses parents dans le graphe. Cette méthode que nous nomerons ACA pour *available cases analysis* est connu de la communauté statisticienne sous le nom de *pairwise data deletion* et est bien adaptée dans le cas ou les estimations doivent être faites localement, comme cela est le cas pour l'apprentissage d'un réseau bayésien.

D'autre méthodes ont étés proposées, certaines utilisant un échantillonneur de Gibbs pour effectuer de l'imputation de données (Myers *et al.*, 1999), certaines effectuent l'apprentissage de sous-structures locales en utilisant une méthodes d'estimation de probabilités robustes aux données incomplètes (Sebastiani *et al.*, 2001) ou d'autres qui utilisent des tests d'indépendances conditionnelles avec des bases incomplètes (Dash *et al.*, 2003).

Dans la section suivante, Nous présentons les méthodes d'estimation utilisées lorsque la base d'exemples est incomplète, puis nous présenterons les différentes techniques d'apprentissage comparées dans ce papier. Nous concluerons à la suite d'une étude expérimentale basés sur l'exemple ASIA.

2. Généralités sur l'estimation de probabilités à partir de bases d'exemples incomplètes

Pour estimer des probabilités en présence de données manquantes, il faut recourir

– a) soit à l'effacement des entrées incomplètes de la base d'exemples,

– b) soit à la substitution des valeurs manquantes par des valeurs admissibles,

– c) soit à une estimation par maximun de vraisemblance.

Les techniques les plus populaires pour traiter les bases d'exemples incomplètes sont alors les suivantes.

a) Méthodes d'effacement d'entrées

Étude des cas complets (CCA) : Cette méthode est appelée *listwise data deletion* dans la littérature consiste en le retrait des entrèes incomplètes de la base. Si les données manquantes sont aléatoirement distribuées (*i.e.* base d'exemples MCAR), ce retrait permet de conserver une base telle que l'estimateur de maximum de vraisemblance sera sans biais, sinon

elle introduira un biais. Lorsqu'il y a de nombreuses données manquantes, cette approche n'est pas satisfaisante car le nombre de cas complets peut alors être très faible ou nul.

Étude des cas disponibles (ACA) : Lorsque nous effectuons un test sur un nombre limité de variables, il est possible de ne considérer que les cas où 'ces' variables sont complètement observées. Cette méthode est appelés *pairwise data deletion* dans la littérature. Comme pour l'étude des cas complets, cette approche introduit un biais lorsque les données manquantes ne vérifient pas les hypothèses MCAR.

b) Méthodes d'imputation

Substitution par la moyenne : Remplacer les valeurs manquantes par la moyenne (ou le mode) de la variable correspondante calculée sur les cas complets de cette variable est une approche qui possède l'avantage de travailler avec la base entière. Néanmoins les principaux désavantages sont, d'une part que ce type de substitution fait artificiellement décroître les variations de score et plus il y a des données manquantes plus cette variation sera faible, et d'autre part, la substitution par les valeurs moyennes peut considérablement changer les corrélations entre les variables en introduisant un biais certain.

Imputation par regression : Les données sont remplies avec une fonction de regression et suppose que les valeurs sont alors correctement prédites. Cette méthode est préférable au simple remplacement par la moyenne, le biais introduit dépend alors de la qualité de la fonction de regression utilisée. Néanmoins, une telle fonction peut être très difficile à construire en pratique.

Multiple imputation : L'imputation multiple utilise différentes valeurs appropriées à partir des données brutes pour compléter la base de données existante. Typiquement, cinq à dix bases de données sont créées de cette façon. Ces matrices de données sont ensuite analysées comme si elles étaient des bases de données complètes. Les différents résultats sont alors combinés dans un unique modèle récapitulatif simple. Des exemples typiques d'implémentation de cette technique sont les méthodes de *Bootstrap* (Rubin, 1981) et de *Boosting* (Freund, 1995).

c) Méthodes itératives

Robust bayesian estimator : Tout au long de son estimation, cette méthode converve les valeurs minimum et maximum des probabilités à évaluer qui évolue itérativement. Après convergence, il est possible d'estimer les probabilités recherchés en 'choisissant' des valeurs dans les intervals finaux de manière cohérente (Ramoni *et al.*, 2000).

***Expectation-Maximisation* (EM)** : Il s'agit d'une méthode itérative qui tente de maximiser la vraisemblance de la probabilité cible en deux

étapes. La première *−expectation−* consiste en l'évaluation de la valeur moyenne sur les exemples complets. Puis dans l'étape *−maximisation−* la valeur manquante est remplacée par la valeur maximisant la vraisemblance (Dempster *et al.*, 1977). Cette méthode possède le désavantage d'être gourmande en temps de calcul.

La méthode EM est une des rares méthodes qui est prouvée n'introduisant pas un biais lorsque les données manquantes ne satisfont pas l'hypothèse MCAR et qui n'utilise pas de modèles intermédiaires difficile à construire utiles comme dans les méthodes d'imputation de données Ces deux raisons font qu'elle est aujourd'hui largement utilisée.

Par la suite, nous allons étudier l'apport de l'utilisation de EM par rapport à la méthode *pairwise data deletion* (notée ACA dans les figures) au cours de l'apprentissage de structure de réseaux bayésiens à l'aide méthodes gloutonnes.

3. Les méthodes mises en œuvre

3.1. *Score d'un réseau bayésien et données incomplètes*

Soient $X = \{X_1, \cdots, X_n\}$ un ensemble des variables aléatoires et $\mathbf{D}_c$ une base de m tirages de X indépendants et identiquement distribués. Supposons par ailleurs que seule une version incomplète $\mathbf{D}$ de la base $\mathbf{D}_c$ soit disponible, celle-ci peut alors se décomposer en

$$\mathbf{D} = [[\mathbf{X}_i^l]]_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant l \leqslant m}} = [\mathbf{O}, \mathbf{H}]$$

où $\mathbf{O}$ est l'ensemble des variables $\mathbf{X}_i^l$ mesurées et $\mathbf{H}$ l'ensemble des variables $\mathbf{X}_i^l$ manquantes.

Le score bayésien est défini par $BD(\mathcal{G}, \mathbf{D}) = \mathbb{P}(\mathcal{G}, \mathbf{D}) = \mathbb{P}(\mathcal{G})\mathbb{P}(\mathbf{D}|\mathcal{G})$. Pour une base d'exemples complète $\mathbf{D}_c$ et en supposant que toutes les variables sont discrètes, (Cooper *et al.*, 1992) ont donné le résultat suivant.

$$\mathbb{P}(\mathbf{D}_c|\mathcal{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{((\sum_{k=1}^{r_i} N_{ijk}) + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \qquad [1]$$

où N_{ijk} est le nombre d'instances où $\{X_i = x_{i,k}$ et $Pa(X_i) = pa_{i,j}\}$ et r_i le nombre d'états pouvant être pris par l'ensemble des parents de X_i.

Or, en présence de données incomplètes, nous avons

$$\mathbb{P}(\mathbf{D}|\mathcal{G}) = \sum_{\mathbf{H}} \int_{\Theta} \mathbb{P}(\mathbf{O}, \mathbf{H}|\mathcal{G}, \Theta)\mathbb{P}(\Theta|\mathcal{G})d\Theta \qquad [2]$$

L'équation 2 peut alors être évaluée par application multiple de l'équation 1 pour toutes les complétions possibles des variables manquantes $\mathbf{H}$. La

complexité d'un tel calcul est alors exponentielle en fonction du nombre de valeurs manquantes dans la base d'exemples. En pratique, ceci n'est pas utilisable, il faut donc utiliser une méthode d'approximation pour $\mathbb{P}(\mathbf{D}|\mathcal{G})$.

3.1.1. *Évaluation de l'équation 2 avec EM*

Soit $S(\mathcal{B}|\mathbf{D}_c)$ une fonction de score quelconque pour un modèle $\mathcal{B} = (\mathcal{G}, \Theta)$ et pour une base complète $\mathbf{D}_c$. Le score S peut être le score bayésien ou tout autre score (pour plus d'information sur les fonctions de score, voir (Naïm *et al.*, 2004)). Il est possible d'estimer le score de ce modèle avec des données incomplètes en calculant

$$Q^S(\mathcal{B}|\mathbf{D}) = \mathbb{E}_{\mathbf{H}\sim\mathbb{P}(\mathbf{H})}\big[S(\mathcal{B}|\mathbf{O},\mathbf{H})\big] \quad [3]$$

Malheureusement, nous n'avons pas accès à la loi $\mathbb{P}(\mathbf{H})$. Il va donc falloir approcher cette loi à partir d'un modèle supposé générateur de $\mathbf{D}$. Soit $\mathcal{B}^0$ un tel modèle, alors il est possible d'écrire

$$Q^S(\mathcal{B}:\mathcal{B}^0|\mathbf{D}) = \mathbb{E}_{\mathbf{H}\sim\mathbb{P}(\mathbf{H}|\mathcal{B}^0)}\big[S(\mathcal{B}|\mathbf{O},\mathbf{H})\big] = \sum_{\mathbf{H}} S(\mathcal{B}|\mathbf{O},\mathbf{H})\mathbb{P}(\mathbf{H}|\mathcal{B}^0) \quad [4]$$

La loi $\mathbb{P}(\mathbf{H}|\mathcal{B}^0)$ est codée par le réseau bayésien $\mathcal{B}^0$ et est obtenue par inférence. Cette méthode permet donc, à partir d'une fonction de score $S(\mathcal{B}|\mathbf{D}_c)$ quelconque, de créer une fonction de score $Q^S(\mathcal{B}:\mathcal{B}^0|\mathbf{D})$ qui donne un résultat (approché) sur des bases d'exemples incomplètes.

Ce score à la particularité de conserver la propriété de *décomposabilité*: si $S(\mathcal{B}|\mathbf{D}_c)=\sum_{i=1}^{n} s(X_i, Pa(X_i)|\mathbf{D}_c)$ alors $Q^S(\mathcal{B}:\mathcal{B}^0|\mathbf{D})=\sum_{i=1}^{n} q^s(X_i, Pa(X_i):\mathcal{B}^0|\mathbf{D})$.
Par exemple, pour le score BIC cela donne:

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

3.1.2. *Évaluation de l'équation 2 avec ACA*

Pour approcher l'équation 2 en utilisant que les exemples disponibles de la base va être beaucoup plus simple. Au lieu d'avoir à faire une méthode itérative comme dans le cas de l'algorithme EM pour évaluer la probabilité $p_{ijk} = \mathbb{P}(X_i = x_{i,k}\, et\, Pa(X_i) = pa_{i,j})$, puis en déduire $N_{ijk} = p_{ijk} \times N$, nous allons évaluer la probabilité p_{ijk} sur une sous-base d'exemples contenant seulement les variables $X_i \cup Pa(X_i)$ de laquelle nous n'allons extraire que les exemples complets.

3.2. *Recherche de l'arbre couvrant de score maximum*

3.2.1. *MWST-EM*

(Chow *et al.*, 1968) ont proposé une méthode d'identification de distributions de probabilité à l'aide de structure de dépendance arborescente qui a été reprise dans le cadre des réseaux bayésiens par (Heckerman *et al.*, 1994) et que nous avons généralisée au cadre des bases d'exemples incomplètes (François *et al.*, 2006).

Cette méthode est inspirée de la méthode de (Friedman, 1997), sauf qu'elle utilise un algorithme EM dans l'espace des structures arborescentes au lieu d'un algorithme EM généralisé, c'est-à-dire qui n'optimise que dans un voisinage de la solution courante au lieu de l'espace dans sa globalité.

Elle utilise un algorithme EM pour évaluer la matrice de score suivante

$$M_{ij}^{k} = \begin{cases} \mathbb{E}_{\mathbf{H}\sim\mathbb{P}(\mathbf{H}|\mathcal{B}^{k-1})}\Big[s(X_i, X_j|\mathbf{O},\mathbf{H}) - s(X_i,\emptyset|\mathbf{O},\mathbf{H})\Big] & \text{si } i \neq j \\ 0 & \text{sinon} \end{cases} \quad [5]$$

L'algorithme de Prim (Prim, 1957), ou l'algorithme de Kruskal (Jarník, 1930), permettent ensuite d'obtenir, grâce à cette matrice, l'arbre qui maximise $Q(\mathcal{G},.:\mathcal{G}^{k-1},\Theta^{k,l})$.

3.2.2. *MWST-ACA*

Une version de cet algorithme construisant la matrice de score à l'aide de la méthode d'estimation ACA est également mise en œuvre. Cet méthode possède l'avantage d'être la seule méthode 'directe' d'apprentissage de structure de réseau bayésien à partir de données incomplètes *à ma connaissance* jusqu'à présent (François, 2006).

3.3. *Recherche dans l'espace des graphes dirigés sans cycle*

3.3.1. *GS-EM*

(Friedman, 1997) a été un des premiers à proposer une méthode déterministe efficace pour faire de la recherche de structure à partir de données incomplètes. Le principe général de la méthode nomée AMS-EM par (Friedman, 1997) est rappelé dans l'algorithme 1.

3.3.2. *GS-ACA*

Une implémentation de cette méthode inspirée de (Heckerman *et al.*, 1994; Cooper *et al.*, 1992) utilisant l'estimation à l'aide de la méthode *pairwise data delation*, va également être testée.

ALGORITHME 1 Algorithme EM pour l'apprentissage de structure

$k = 0$
Choisir le graphe $\mathcal{G}^k$ et les paramètres $\Theta^{k,0}$ aléatoirement ou en utilisant une heuristique
Tant que l'on a pas convergence et $k \leqslant k_{max}$ **Faire**
  $l = 0$
  Tant que l'on a pas convergence et $l \leqslant l_{max}$ **Faire**
    $l = l + 1$
    $\Theta^{k,l} = \underset{\Theta}{\text{argmax}}\ Q(\mathcal{G}^n, \Theta : \mathcal{G}^k, \Theta^{k,l-1})$
  Fin Tant que
  $k = k + 1$
  $\mathcal{G}^k = \underset{\mathcal{G}}{\text{argmax}}\ Q(\mathcal{G}, . : \mathcal{G}^{k-1}, \Theta^{k,l})$
  $\Theta^{k,0} = \underset{\Theta}{\text{argmax}}\ Q(\mathcal{G}^k, \Theta : \mathcal{G}^{k-1}, \Theta^{k-1,l})$
Fin Tant que

3.4. *Recherche dans l'espace des représentants des classes d'équivalence de Markov*

3.4.1. *GES-EM*

Des travaux récents (Chickering, 2002a; Castelo *et al.*, 2002; Auvray *et al.*, 2002) montrent qu'il peut être plus profitable de travailler dans l'espace des CPDAG (représentant des classes d'équivalence de Markov) plutôt que dans l'espace des DAG. Cet espace possède moins de plateaux pour la fonction de score, car de nombreux DAG avec des scores égaux sont représentés par un seul CPDAG.

Une méthode de recherche gloutonne dans l'espace des équivalents de Markov nommée GES pour *Greedy Equivalent Search* a été introduite par (Meek, 1997). Elle consiste en deux étapes itératives. La première étape construit un graphe pas à pas en parcourant l'espace des équivalents de Markov (en fait, l'espace des *graphes essentiels*, CPDAG, représentants des classes d'équivalence de Markov). la seconde étape consiste en le retrait des arcs superflus éventuels. L'optimalité de cet algorithme a été prouvée et dépend de la *conjecture de Meek* démontrée depuis par (Kočka *et al.*, 2001; Chickering, 2002b).

L'algorithme GES-EM reprend les principes de la méthode proposée par (Friedman, 1997) en utilisant cette fois un voisinage qui est construit grâce à cette notion d'équivalence de Markov.

3.4.2. *GES-ACA*

Une version utilisant l'estimation à l'aide des *cas disponibles* du critère de score pris en considération ici, le score BIC pour toutes les implémentations de ces algorithmes, est également mise en œuvre.

4. Expérimentations

4.1. *Protocole expérimental*

Nous avons utilisé Matlab, et plus précisément la Bayes Net Toolbox (Murphy, 2004) qui fournit déjà certaines méthodes présentées ci-dessous. Le code des fonctions mises en œuvre dans ces expérimentations est mis à disposition par l'intermédiaire du *Structure Learning Package* décrit dans (Leray *et al.*, 2004). Le site propose une introduction et des tutoriels en français. Nous avons utiliser le réseau bayésien ASIA (Jensen, 1996) représenté à droite du tableau 1 ainsi que la méthode proposée par (François *et al.*, 2007; François, 2006) pour générer des bases d'exemples de différentes tailles et avec différents taux de données manquantes. Les valeurs retenues sont données dans le tableau.

taux\tailles	56	361	983	2000
0%	√	√	√	√
5%	√	√	√	
15%	√	√	√	
30%	√	√	√	
50%		√	√	

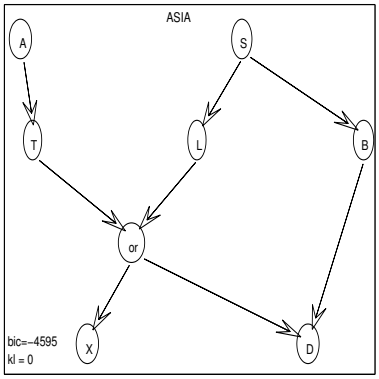

Tableau 1. *À gauche, les valeurs retenue pour les tailles et les taux de données manquantes des bases d'exemples, la base de taille* 2000 *sert pour donner les scores BIC en test. À droite la structure du réseau utilisé.*

4.2. *Résultats*

Les résultats des expérimentations sont donnés sous forme graphique dans la table 2 lorsque la taille de la base augmente ou le taux de données manquantes diminu. Le score BIC, calculé sur une base de test de 2000 points, et la divergence de Kullback-Leiber, donnée par la formule suivante entre le réseau bayésien ASIA et le réseau bayésien obtenu, sont donnés en moyenne pour cinq apprentissage des paramètres Θ^1 différents.

$$KL(\mathscr{B}^1||\mathscr{B}^2) = \sum_{i=1}^{n} \sum_{j^1=1}^{q_i^1} \sum_{k=1}^{r_i} \theta_{ijk}^1 \log \left(\frac{\theta_{ijk}^1}{\mathbb{P}(X_i = k | Pa^1(X_i) = j^1, \mathscr{G}^2, \Theta^2)} \right)$$

où $Pa^1(X_i)$ représente l'ensemble des parents de la variable X_i dans la structure de $\mathscr{B}^1 = (\mathscr{G}^1, \Theta^1)$ et où $\theta_{ijk}^1 = \dfrac{N_{ijk}}{N}$ est un paramètre de $\mathscr{B}^1$.

La table 2 illustre les performances qui peuvent être obtenues par les différents algorithmes d'apprentissage testés utilisant soit ACA soit EM

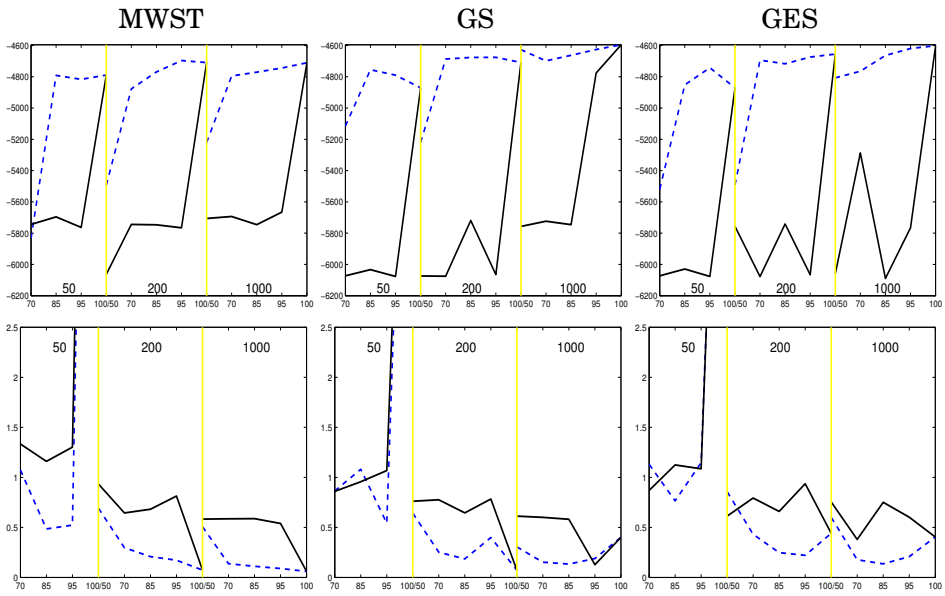

Tableau 2. *En pointillé la méthode ACA a été utilisée et en trait plein l'algorithme EM. Sur la première ligne, les scores BIC obtenus pour les différentes tailles (50, 200, 1000) et taux de données complètes (50, 70, 85, 95 et 100%). Sur la seconde ligne, la moyenne des divergences de Kullback-Leiber obtenues pour cinq apprentissages des paramètres.*

comme méthodes d'estimation de probabilités. De la gauche vers la droite, le nombre de données en apprentissage augmente[1].

Premièrement, nous pouvons remarquer que l'estimation à l'aide de ACA permet d'obtenir de meilleurs scores BIC ainsi que de meilleures divergences que lorsque l'algorithme EM est utilisé. Par ailleurs, les temps de calcul moyens avec ACA sont nettement inférieur de 80 fois pour MWST à 300 fois pour GES sur ces exemples.

5. Conclusions

Nous avons étudié les différences de performances de méthodes d'apprentissage de structure gloutonnes à partir de bases d'exemples incomplètes. Sur l'exemple considéré, il semble que l'utilisation de ACA soit à privilégier sur celle de EM. Néanmoins, nous devons remarquer que les structures obtenues à l'aide de ACA ont tendance à contenir beaucoup

1. L'artefact pour la base de 50 exemples complets est dû à un sur-apprentissage.

plus d'arcs que celles obtenues avec EM. Ce nombre important d'arc permet d'obtenir des modèles plus représentatifs (d'où de meilleurs scores) mais plus difficilement interprétables. Ceci qui peut être un tort lorsque que l'objectif est d'effectuer de l'extraction d'information ou un avantage si l'objectif est d'effectuer une modélisation à des fins de simulation ou de classification.

6. Bibliographie

Auvray V., Wehenkel L., « On the Construction of the Inclusion Boundary Neighbourhood for Markov Equivalence Classes of bayesian Network Structures », *in* A. Darwiche, N. Friedman (eds), *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, Morgan Kaufmann Publishers, S.F., Cal., p. 26-35, 2002.

Castelo R., Kocka T., Towards an inclusion driven learning of bayesian networks, Technical Report n° UU-CS-2002-05, Institute of information and computing sciences, University of Utrecht, 2002.

Chickering D., « Learning equivalence classes of bayesian-network structures », *Journal of machine learning research*, vol. 2, p. 445-498, 2002a.

Chickering D., « Optimal Structure Identification with Greedy Search », *Journal of Machine Learning Research*, vol. 3, p. 507-554, November, 2002b.

Chow C., Liu C., « Approximating discrete probability distributions with dependence trees », *IEEE Transactions on Information Theory*, vol. 14, n° 3, p. 462-467, 1968.

Cooper G., Hersovits E., « A bayesian Method for the Induction of Probabilistic Networks from Data », *Maching Learning*, vol. 9, p. 309-347, 1992.

Dash D., Druzdzel M., « Robust Independence Testing for Constraint-Based Learning of Causal Structure », , Proceedings of The Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03), pp 167-174, 2003.

Dempster A., Laird N., Rubin D., « Maximum Likelihood from Incomplete Data Via the EM Algorithm », *Journal of the Royal Statistical Society*, vol. B 39, p. 1-38, 1977.

François O., De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes, PhD thesis, Institut National des Sciences Appliquées de Rouen (INSA), `http://ofrancois.tuxfamily.org/these.html`, 2006.

François O., Leray P., « Learning the Tree Augmented Naive Bayes Classifier from incomplete datasets », *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM'06), ISBN: 80-86742-14-8*, Prague, Czech Republic, p. 91-98, september, 2006.

François O., Leray P., « Incomplete Datasets Generation using the Bayesian Networks Formalism », *Proceedings of the International Joint Conferences on Neural Networks (IJCNN 2007)*, Orlando, Florida, 2007.

Freund Y., « Boosting a week learning algorithm by majority », *Information and Computation*, 1995.

Friedman N., « Learning belief networks in the presence of missing values and hidden variables », *Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, p. 125-133, 1997.

Friedman N., « The bayesian Structural EM Algorithm », *in* G. F. Cooper, S. Moral (eds), *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, Morgan Kaufmann, San Francisco, p. 129-138, July, 1998.

Heckerman D., Geiger D., Chickering M., « Learning Bayesian networks: The combination of knowledge and statistical data », *in* R. L. de Mantaras, D. Poole (eds), *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, CA, USA, p. 293-301, July, 1994.

Jarník V., « O jistem problemu minimalnim », *raca Moravske Prirodovedecke Spolecnosti (in Czech)*, vol. 6, p. 57-63, 1930.

Jensen F., *An introduction to Bayesian Networks*, Taylor and Francis, London, United Kingdom, 1996.

Kočka T., Bouckaert R., Studenỳ M., « On characterization inclusion of bayesian Networks », *in* J. Breese, D. Koller (eds), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence.*, Morgan Kaufmann, p. 261-268, 2001.

Leray P., François O., BNT Structure Learning Package: Documentation and Experiments, Technical Report n° 2004/PhLOF, Laboratoire PSI, INSA de Rouen, 2004. `http://bnt.insa-rouen.fr`.

Meek C., Graphical Models: Selecting causal and statistical models, PhD thesis, Carnegie Mellon University, 1997.

Murphy K., « Bayes Net Toolbox v5 for MATLAB », , Cambridge, MA: MIT Computer Science and Artificial Intelligence Laboratory, 2004. `http://bnt.sourceforge.net`.

Myers J., Laskey K., Lewitt T., « Learning bayesian Network from Incomplete Data with Stochastic Search Algorithms », *the Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI99)*, 1999.

Naïm P., Wuillemin P.-H., Leray P., Pourret O., Becker A., *Réseaux bayésiens*, Eyrolles, ISBN : 2-212-11137-1, 2004.

Pearl J., « Graphical Models for Probabilistic and Causal Reasoning », *in* D. M. Gabbay, P. Smets (eds), *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Volume 1: Quantified Representation of Uncertainty and Imprecision*, Kluwer Academic Publishers, Dordrecht, p. 367-389, 1998.

Prim R., « Shortest connection networks and some generalisations », *Bell System Technical Journal*, vol. 36, p. 1389?1401, 1957.

Ramoni M., Sebastiani P., « Robust Learning with Missing Data », *Machine Learning*, vol. 45, p. 147-170, 2000.

Rubin D., « Inference and missing data », *Biometrika*, vol. 63, p. 581-592, 1976.

Rubin D., « The bayesian boostrap », *Ann. Statistics*, vol. 9, p. 130-134, 1981.

Sebastiani P., Ramoni M., « Bayesian Selection of Decomposable Models With Incomplete Data », , Journal of the American Statistical Association, 96, No. 456, pp 1375-1386, 2001.