Estimation de densité par ensembles aléatoires de poly-arbres

Sourour Ammar, Philippe Leray, Louis Wehenkel

Estimation de densité par ensembles aléatoires de poly-arbres

Sourour Ammar[1 2] — Philippe Leray[1] — Louis Wehenkel[3]

[1] *Laboratoire d'Informatique de Nantes Atlantique (LINA) UMR 6241*
École Polytechnique de l'Université de Nantes, France
philippe.leray@univ-nantes.fr

[2] *Laboratoire d'Informatique, Traitement de l'Information et des Systèmes (LITIS) EA 4108 - Institut National des Sciences Appliquées de Rouen, France*
sourour.ammar@etu.univ-nantes.fr

[3] *Département d'Electricité, d'Electronique et Informatique & Groupe Interdisciplinaire de Génoprotéomique Appliquée- Université de Liège, Belgique*
l.wehenkel@ulg.ac.be

RÉSUMÉ. *La notion de mélange de modèles simples aléatoires est de plus en plus utilisée et avec succès dans la littérature de l'apprentissage supervisé ces dernières années. Parmi les avantages de ces méthodes, citons l'amélioration du passage à l'échelle des algorithmes d'apprentissage grâce à leur aspect aléatoire et l'amélioration de l'exactitude de la prédiction des modèles induits grâce à une flexibilité plus élevée en ce qui concerne le compromis biais/variance. Dans le présent travail, nous proposons d'explorer cette idée dans le contexte de l'estimation de la densité. Nous proposons une nouvelle famille de méthodes d'apprentissage non-supervisé à base de mélange de grands ensembles aléatoires de poly-arbres. La caractéristique spécifique de ces méthodes est leur passage à l'échelle, aussi bien en terme de nombre de variables que de données à traiter. Cette étude, exploratoire, compare empiriquement ces méthodes sur un ensemble de problèmes de test discrets de taille et de complexité croissantes et ouvre de nombreuses perspectives auxquelles nous prévoyons de nous intéresser.*

ABSTRACT. *Ensembles of weakly fitted randomized models have been studied intensively and used successfully in the supervised learning literature during the last two decades. Among the advantages of these methods, let us quote the improved scalability of the learning algorithm thanks to its randomization and the improved predictive accuracy the induced models thanks to the higher flexibility in terms of bias/variance trade-off.*
In the present work we propose to explore this idea in the context of density estimation. We propose a new family of unsupervised learning methods of mixtures of large ensembles of ran-

domly generated poly-trees. The specific feature of these methods is their scalability to very large numbers of variables and training instances. We explore these methods empirically on a set of discrete test problems of growing size. We finally discuss possible extensions which we plan to study.

MOTS-CLÉS : *estimation de densité, génération aléatoire de poly-arbres, mélange de modèles.*

KEYWORDS: *density estimation, polytree random generation, mixture of models.*

1. Introduction

L'apprentissage des réseaux bayésiens à partir de données s'avère un problème de pointe depuis une dizaine d'années (Leray, 2006; Naïm *et al.*, 2007). La motivation de ce domaine est de retrouver, à partir d'un ensemble de données d'apprentissage, un modèle qui représente le mieux le modèle sous-jacent et qui pourrait ensuite être utilisé dans le cadre de l'aide à la décision. Ce modèle peut aussi être utilisé dans le cadre de la simulation, car il est sensé être générateur des données observées.

Dans ce travail, nous nous intéressons de manière générale au problème de l'estimation de la densité, i.e. comment modéliser le mieux possible la distribution des données sous-jacentes, et ce avec un modèle pouvant supporter facilement le passage à l'échelle.

Les méthodes classiques d'apprentissage de la structure d'un réseau bayésien permettent de déterminer un modèle optimal (souvent local) au sens d'un score ou en tant que bon représentant des indépendances conditionnelles découvertes dans les données. Pourtant, dans le cas par exemple où l'ensemble de données de test est relativement petit par rapport à la taille du problème, plusieurs solutions potentielles existent. Les méthodes standards d'apprentissage de structure ne choisiront qu'un seul modèle parmi cet ensemble, alors que l'exploitation de tous les modèles possibles permettrait d'améliorer l'estimation de la densité de probabilité.

Nous nous intéressons donc dans ce travail à définir un modèle issu en théorie d'un mélange de toutes les structures possibles pour un problème donné. Autrement dit, nous nous intéressons à estimer :

$$p(X|D) = \sum_m p(m|D)\, p(X|m, D) \qquad [1]$$

où $p(X|D)$ représente la loi de probabilité des variables X du problème sachant les données D.

L'espace des structures des réseaux bayésiens est de taille super-exponentielle en fonction du nombre de variables du problème. Le calcul de cette somme doit donc se faire sur un espace plus restreint grâce à plusieurs approximations proposées dans (Madigan *et al.*, 1994; Madigan *et al.*, 1995). Certains travaux comme ceux de (Madigan *et al.*, 1995) utilisent une méthode d'échantillonnage comme les MCMC pour la génération de modèles m possibles pour l'estimation de l'équation 1. (Friedman *et al.*, 2000) propose une approche du même type, en échantillonnant non pas directement les modèles, mais les ordres sur les variables, et en cherchant les modèles optimaux pour chaque ordre.

Dans cet article, nous proposons d'utiliser une méthode de mélange de modèles simples. Nous utilisons pour cela deux approximations pour le calcul de la somme de l'équation 1 en restreignant le calcul à un ensemble de modèles m simples (poly-arbres).

Nous commencerons donc par aborder le problème général de mélange de modèles dans la section 2 de cet article. Ensuite, le section 3 décrira notre approche à base de mélange de poly-arbres aléatoires. L'algorithme de génération aléatoire de poly-arbres est alors détaillé dans la section 4. Une première évaluation de cette approche sur des problèmes de taille et de complexité croissante est décrite dans la section 5. Ces expériences sont les prémisses d'une étude plus poussée qui est actuellement en cours. La section 6 nous permettra donc de tirer les premières conclusions de ces travaux et surtout d'en détailler les prochaines étapes.

2. Notion de mélange de modèles

Soit un ensemble de données D pour un problème dont le domaine consiste en un ensemble de variables $X = (X_1, ..., X_n)$. La probabilité d'un évènement (observation x de X) sachant cet ensemble de données D est donnée par l'espérance sur tous les modèles possibles m pour l'ensemble des variables ainsi que leurs paramètres θ_m (Chickering *et al.*, 1997) :

$$p(x|D) = \sum_m p(m|D)\ p(x|m, D) \qquad [2]$$

$p(x|D, m)$ étant l'intégrale sur toutes les valeurs possibles des paramètres θ_m du modèle m :

$$p(x|m, D) = \int p(x|\theta_m, m)\ p(\theta_m|m, D) d\theta_m \qquad [3]$$

Ainsi $p(x|D)$ s'écrit :

$$p(x|D) = \sum_m p(m|D) \int p(x|\theta_m, m)\ p(\theta_m|m, D) d\theta_m \qquad [4]$$

avec $p(x|\theta_m, m)$ la vraisemblance de l'observation x pour le modèle m muni des paramètres θ_m.

Notons que dans cette équation, l'estimation de la densité de probabilité de X ne se fait pas par l'apprentissage d'un seul modèle à partir des données mais par une somme pondérée sur tous les modèles possibles, pour tous les paramètres possibles. Toutefois, dans la pratique, il est empiriquement peu intéressant de faire un tel calcul.

Des approximations peuvent être faites pour réduire le domaine de calcul. (Chickering *et al.*, 1997) montre qu'en appliquant une approximation de Laplace, l'intégrale de la vraisemblance de l'équation 3 sur les valeurs des paramètres θ_m possibles pour un modèle m peut être approchée par la vraisemblance en une seule valeur de $\theta_m = \tilde{\theta}_m$ qui maximise $p(\theta_m|m, D)$. $\tilde{\theta}_m$ est dite la configuration *maximum à posteriori* (MAP) de θ_m sachant D.

La deuxième approximation à faire est celle relative à la sommation sur tous les modèles m possibles. (Robinson, 1977) a montré que le nombre de structures

possibles de graphes en fonction du nombre n de variables est super-exponentiel. (Friedman *et al.*, 2000) a évoqué l'intérêt de réduire ce nombre de modèles en restreignant l'espace de calcul à un ensemble G de graphes et a discuté les différentes propositions de la littérature pour la détermination de cet ensemble. Cette approche est connue sous le nom de *sélection de modèles*.

En appliquant ces deux approximations, l'équation 4 s'écrit alors :

$$p(x|D) = \sum_{m \in G} p(m|D) p(x|\tilde{\theta}_m, m) \quad [5]$$

Notons ici que les approches classiques d'apprentissage de structure essaient de simplifier une fois de plus cette équation, en ne conservant que le modèle $m = \tilde{m}$ maximisant $p(m|D)$:

$$p(x|D) = p(x|\tilde{\theta}_{\tilde{m}}, \tilde{m}) \quad [6]$$

Cette approche de mélange de modèles est aussi à relier à une autre technique d'apprentissage, le boosting, proposé initialement dans le cadre de la classification supervisée par (Freund *et al.*, 1995) et appliqué aux réseaux bayésiens dans le cas initial de la classification (Choudhury *et al.*, 2002) et pour l'estimation de densité (Rosset *et al.*, 2002). Dans le cas du boosting, le mélange de modèles est obtenu de manière itérative, chaque modèle étant appris sur des données où les modèles précédents n'étaient pas pertinents, et les coefficients de mélange sont déterminés au cours de l'apprentissage.

3. Mélanges aléatoires de poly-arbres

Dans ce travail, nous proposons de restreindre l'ensemble des modèles G de l'équation 5 à l'ensemble des poly-arbres, et plus exactement à un ensemble de grande taille de poly-arbres générés aléatoirement.

Cette restriction de G à l'espace des poly-arbres est guidé par des considérations de passage à l'échelle. Pouvoir construire et manipuler rapidement des modèles de grande taille avec un grand nombre de données nécessite l'utilisation de modèles simples. Ainsi, en supposant le modèle déjà construit, l'algorithme d'inférence classique du *Message Passing* proposé par Pearl (Pearl, 1986) ne marche que pour des arbres et des poly-arbres. L'inférence dans des modèles plus complexes nécessite l'utilisation d'algorithmes d'inférence plus complexes comme *Junction Tree* (Jensen *et al.*, 1990) ou des algorithmes approchés stochastiques ou variationnels (Jordan *et al.*, 1998).

L'espace des arbres ou celui des poly-arbres sont donc les meilleurs espaces dans lesquels nous pourrons faire des calculs d'inférence très rapidement. Entre ces deux espaces, nous choisissons celui des poly-arbres, plus riche, puisqu'il permet de représenter des indépendances conditionnelles impossibles à représenter par des arbres (grâce aux V-structures du type $A \rightarrow C \leftarrow B$).

Notre second choix concerne l'utilisation d'un mélange aléatoire de poly-arbres. Ce choix a été guidé par plusieurs considérations. Tout d'abord, se restreindre à la meilleure structure dans cet espace n'est pas une solution adéquate. La recherche d'un arbre optimal, i.e. utiliser l'équation 6 avec G l'espace des arbres, proposée initialement par (Chow *et al.*, 1968) a donné lieu à un algorithme d'apprentissage de structure MWST (Maximum Weight Spanning Tree) très rapide mais trop contraint (pas de cycles, ni de V-structures) (Francois *et al.*, 2003). Malheureusement les travaux de (Dasgupta, 1999) montrent que la recherche d'un poly-arbre optimal est une alternative trop complexe pour être intéressante en ce qui concerne le passage à l'échelle.

(Meila-Predoviciu, 1999) utilise alors un modèle de mélange, mais dans l'espace des arbres, en choisissant de déterminer les structures optimales de ces arbres et les coefficients de mélange durant l'apprentissage (en utilisant conjointement MWST pour trouver les structures et l'algorithme EM pour obtenir les paramètres de mélange).

Ainsi, nous proposons, de notre côté, de travailler dans l'espace des poly-arbres, mais de ne pas chercher à chaque fois le meilleur poly-arbre possible, en considérant plutôt un mélange de modèles tirés au hasard dans l'espace des poly-arbres, en nous inspirant des méthodes d'échantillonnage de modèles utilisées dans (Madigan *et al.*, 1995; Friedman *et al.*, 2000).

Pour cela, nous allons générer aléatoirement un ensemble de modèles en construisant une séquence $P = (P_1, P_2, ..., P_M)$ de M poly-arbres.

L'estimation de la densité par ce modèle de mélange sera calculée par :

$$p(x|D) = \sum_{i=1}^{M} p(P_i|D) p(x|\tilde{\theta}_{P_i}, P_i) \qquad [7]$$

Dans cette première étude, nous considérerons constants les coefficients de mélange $p(P_i|D) = 1/M$, ce qui nous donne finalement la formule :

$$p(x|D) = \frac{1}{M} \sum_{i=1}^{M} p(x|\tilde{\theta}_{P_i}, P_i) \qquad [8]$$

où $\tilde{\theta}_{P_i}$ sont les paramètres obtenus par maximum a posteriori pour le poly-arbre P_i.

4. Génération aléatoire d'arbres

La mise en œuvre de l'équation 8 nécessite la génération aléatoire de poly-arbres. De plus, nous proposons dans la section 5 une série d'expériences pour essayer d'approcher par ce mélange de poly-arbres des modèles de complexité croissante en commençant par les chaînes ou les arbres. Pour tout cela, il va nous falloir générer aléatoirement différentes structures : chaînes, arbres, poly-arbres. Nous avons choisi d'utiliser les algorithmes proposés dans (Quiroz, 1989). Ces algorithmes se basent sur les

propriétés de codage de Prûfer pour les structures d'arbres étiquetées. Ces algorithmes nous permettent de démontrer que la génération des structures se fait selon une loi uniforme dans l'espace des arbres. En effet, Prûfer a établit une bijection entre l'ensemble des arbres complètement étiquetés à n noeuds et l'ensemble des listes, dites de *Prûfer*, de $(n-2)$ nombres entiers dans $\{1, 2, ..., n\}$ (répétition autorisée).

La probabilité d'une liste dans cet ensemble est $p = 1/n^{n-2}$, ce qui confirme le théorème de Cayley disant que le nombre d'arbres complètement étiquetés à n noeuds est n^{n-2}.

Pour construire un arbre non dirigé à n noeuds, les étapes sont les suivantes :

1) Générer une liste a de $(n-2)$ entiers dans $\{1, 2, ..., n\}$. Soit $a = (a_1, a_2, ..., a_{n-2})$, et $b = (1, 2, ..., n)$.

2) Chercher b_1 le plus petit entier dans b non dans a. Joindre a_1 à b_1 pour former la première arête de l'arbre. Retirer a_1 de a et b_1 de b.

3) Répéter l'étape 2 jusqu'à vider la liste a.

4) Joindre les deux entiers restant dans b pour former la dernière arête de l'arbre.

L'orientation des arêtes de l'arbre étiqueté ainsi obtenu peut se faire en utilisant une exploration en profondeur classique à partir d'un nœud racine tiré au hasard parmi les n noeuds du graphe. L'ordre de visite des arêtes du graphe nous permet alors de transformer l'arbre non orienté en un arbre orienté.

La génération aléatoire d'un poly-arbre ne peut pas se faire aussi facilement. L'orientation des arêtes doit être décidée d'une façon aléatoire selon une loi de probabilité uniforme $(p = 1/2)$ de manière à pouvoir obtenir des V-structures non présentes dans les arbres orientés.

La génération aléatoire d'une chaîne à n noeuds peut se faire de deux manières : soit en tirant au hasard un ordre sur les entiers de 1 à n, le premier entier correspondant à la racine de la chaîne, soit en utilisant le codage de Prûfer. En effet. la liste de Prûfer correspondant à une chaîne est tout simplement une liste ne présentant pas de répétitions. Il suffit donc de générer une telle liste, de construire la chaîne avec l'algorithme présenté ci-dessus, et d'orienter la chaîne comme dans le cas d'un arbre.

5. Résultats expérimentaux

5.1. *Protocole*

Afin de tester notre approche, nous envisageons un ensemble de scénarii d'expériences. L'objectif de tous ces scénarii est d'évaluer la qualité de l'estimation d'un mélange de poly-arbres (équation 8) pour des problèmes plus complexes.

Pour évaluer l'intérêt de notre approche dans des situations les plus variées possibles, nous proposons de générer des réseaux bayésiens de structures et paramètres variés qui nous permettront de générer des données qui seront ensuite utilisées pour

l'apprentissage du mélange de modèles. Cette procédure nous permet ainsi de contrôler la complexité des problèmes à modéliser en choisissant respectivement des chaînes, des arbres, des poly-arbres puis des graphes orientés sans circuit quelconques.

La qualité de l'estimation de notre mélange sera estimée par la mesure de divergence KL (Kullback *et al.*, 1951) entre ce modèle estimé et le modèle théorique avec lequel les données sont réellement générées.

Dans ce travail, nous ne considérons que le cas d'ensembles de données complètes pour un ensemble de variables discrètes. Les paramètres des poly-arbres sont estimés par maximum a posteriori avec des a priori de Dirichlet uniformes.

Ainsi, pour un modèle théorique initial (structure et paramètres), une expérience consiste en :

1) la génération avec ce modèle d'un ensemble de données d'apprentissage D.

2) la génération aléatoire d'un ensemble de poly-arbres dont les paramètres respectifs sont déterminés à partir des données D

3) le calcul de la divergence KL entre ce mélange de poly-arbres et le modèle théorique
La taille de l'ensemble de poly-arbres augmentant progressivement de 10 à 1000 pour évaluer l'importance du nombre M de poly-arbres dans la qualité de l'approximation.

Nous choisissons de répéter 10 fois cette expérience, i.e. utiliser 10 bases d'apprentissage D différentes pour un même modèle théorique initial afin que les performances mesurées ne dépendent pas d'un seul jeu de données qui pourrait être un cas particulier. La qualité de cette expérience (un modèle théorique, 10 bases d'apprentissage) est alors calculée par la moyenne des divergences KL obtenues pour chaque base d'apprentissage.

Afin de prendre en compte ensuite la variabilité des modèles théoriques initiaux, nous décidons de répéter cette série de 10 expériences plusieurs fois, i.e. pour 10 structures théoriques générées aléatoirement selon le principe décrit dans 4. La divergence KL moyenne sur cette série de 10x10 expériences nous donnera alors la qualité de l'estimation de notre modèle pour une classe de modèles théoriques fixés (chaînes, arbres, poly-arbres, ...).

De plus, pour vérifier l'hypothèse faite dans la section 3 préférant les mélanges de poly-arbres aux mélanges d'arbres, nous décidons de réaliser les mêmes séries d'expériences pour des mélanges aléatoires d'arbres.

5.2. *Résultats*

Nous avons utilisé dans nos expériences un modèle théorique à 8 variables booléennes. La taille des bases de données générées pour l'apprentissage est de 2000. Nos différents algorithmes de génération de modèles ont été implémentés en C++ à l'aide

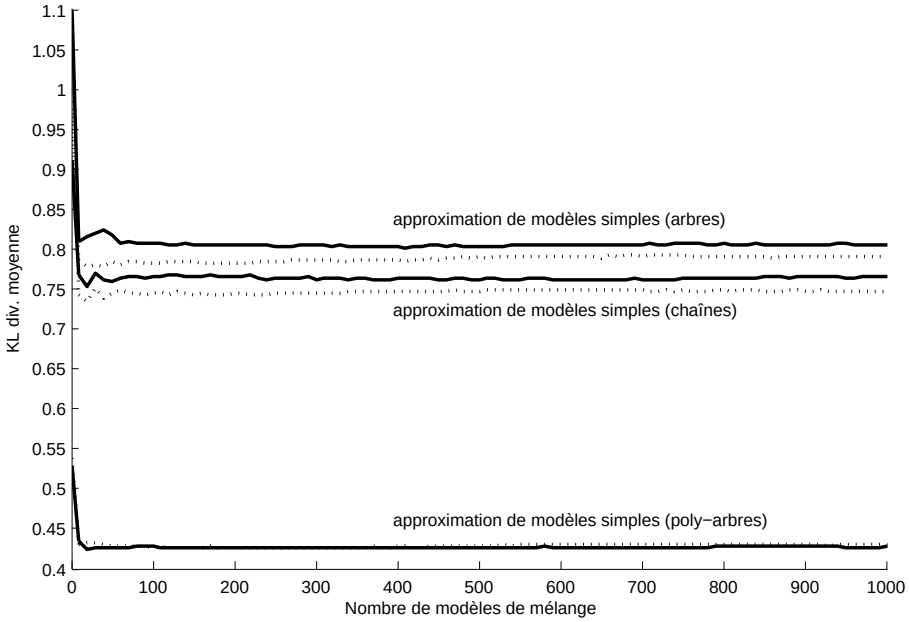

Figure 1. *Approximation de modèles simples (chaînes, arbres et poly-arbres) par des mélanges d'arbres (trait pointillé) et de poly-arbres (trait plein)*

de la librairie Boost disponible sur `http ://www.boost.org/` et des API fournies par la plateforme ProBT© disponible sur `http ://bayesian-programming.org`.

La figure 1 trace la divergence KL moyenne sur 10x10 expériences, en fonction du nombre de modèles dans le mélange, pour les trois familles de modèles théoriques simples (chaînes, arbres et poly-arbres) et pour deux types de mélange (mélange d'arbres en pointillé, mélange de poly-arbres en trait plein).

Nous voyons sur ces premières expériences que les mélanges d'arbres sont légèrement meilleurs que les mélanges de poly-arbres lorsque le modèle théorique à approcher est très simple (chaîne ou arbre). Par contre, les deux semblent se valoir lorsque le modèle théorique est un poly-arbre !

Nous voyons aussi qu'un mélange de taille 1 est visiblement moins bon qu'un mélange de plusieurs modèles. L'utilisation d'un mélange de plusieurs modèles améliore alors la qualité de l'estimation.

Cette même figure nous indique que le nombre de modèles dans le mélange ne semble plus influer sur la qualité de l'estimation à partir de $M = 150$ modèles environ.

6. Conclusions et perspectives

Nous avons présenté ici les tous premiers résultats d'une étude sur l'estimation de densité par mélange aléatoire de poly-arbres.

Après avoir rappelé quelques éléments théoriques à propos de la notion de mélanges de modèles, nous avons détaillé notre approche et présenté les algorithmes de génération de modèles simples (chaînes, arbres et poly-arbres) utilisés dans la phase expérimentale.

Nous avons ensuite expliqué le protocole expérimental que nous nous proposons de suivre, et décrit les premiers résultats de ces expériences. Ces résultats sont assez surprenants puisque pour les modèles théoriques les "moins" simples de ces expériences (poly-arbres), les mélanges d'arbres donnent d'aussi bons résultats que les mélanges de poly-arbres.

Puisque ce travail est encore dans sa phase préliminaire, il nous reste de nombreuses perspectives à creuser. Nous comptons tout d'abord répéter les mêmes expériences sur davantage de modèles (100x100 expériences au lieu de 10x10) pour obtenir des performances moyennes plus représentatives.

Nous poursuivrons ensuite en reprenant les mêmes expériences avec des modèles théoriques plus complexes (graphes orientés sans circuit) pour observer si un mélange d'arbres donne encore des résultats du même niveau qu'un mélange de poly-arbres.

Notre objectif est d'étudier si ces mélanges de modèles peuvent remplacer les approches d'apprentissage de structure classiques. Pour cela, nous comparerons nos résultats à ceux obtenus par des algorithmes d'apprentissage comme l'arbre de recouvrement maximal, la recherche gloutonne dans l'espace des graphes orientés sans circuit ou dans l'espace des représentants des classes d'équivalence de Markov, aussi bien en terme de qualité d'estimation que de temps d'apprentissage et d'inférence.

Pour tout cela, nous comptons aussi nous intéresser à l'influence du nombre de données utilisées sur la qualité de l'estimation, et à l'influence du passage à l'échelle en augmentant significativement le nombre de variables considérées.

Pour finir, nous sommes réfléchissons actuellement à une méthode automatique d'estimation des paramètres de mélange au lieu de les considérer constants, méthode devant aussi être capable de résister au passage à l'échelle.

7. Bibliographie

Chickering D. M., Heckerman D., « Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables », *Machine Learning*, vol. 29, n° 2-3, p. 181-212, 1997.

Choudhury T., Rehg J., Pavlovic V., Pentland A., « Boosting and Structure Learning in Dynamic Bayesian Networks for Audio-Visual Speaker Detection », *Proceedings of ICPR 2002*, 2002.

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

Dasgupta S., « Learning polytrees », *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, Morgan Kaufmann, San Francisco, CA, p. 134-14, 1999.

Francois O., Leray P., « Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens », *Proceedings of RJCIA 2003, plateforme AFIA 2003*, Laval, France, p. 167-180, 2003.

Freund Y., Schapire R. E., « A decision-theoretic generalization of on-line learning and an application to boosting », *European Conference on Computational Learning Theory*, p. 23-37, 1995.

Friedman N., Koller D., « Being Bayesian about Network Structure », *in* C. Boutilier, M. Goldszmidt (eds), *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-00)*, Morgan Kaufmann Publishers, SF, CA, p. 201-210, June 30– July 3, 2000.

Jensen F. V., Lauritzen S. L., Olesen K. G., « Bayesian Updating in Causal Probabilistic Networks by Local Computations », *Computational Statistics Quaterly*, vol. 4, p. 269-282, 1990.

Jordan M. I., Ghahramani Z., Jaakkola T. S., Saul L., « An Introduction to Variational Methods for Graphical Models », *in* M. I. Jordan (ed.), *Learning in Graphical Models*, Kluwer Academic Publishers, Boston, 1998.

Kullback S., Leibler R. A., « On Information and Sufficiency », *Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79-86, 1951.

Leray P., « Réseaux Bayésiens : Apprentissage et Diagnostic de Systemes Complexes », , Habilitation à Diriger les Recherches, Université de Rouen, France, novembre, 2006.

Madigan D., Raftery A. E., « Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window », *The American Statistical Association*, vol. 89, p. 1535-1546, 1994.

Madigan D., York J., « Bayesian graphical models for discrete data », *International Statistical Review*, vol. 63, p. 215-232, 1995.

Meila-Predoviciu M., Learning with Mixtures of Trees, PhD thesis, MIT, 1999.

Naïm P., Wuillemin P.-H., Leray P., Pourret O., Becker A., *Réseaux bayésiens*, 3 edn, Eyrolles, Paris, 2007.

Pearl J., « Fusion, Propagation, and Structuring in Belief Networks », *Artificial Intelligence*, vol. 29, p. 241-288, 1986.

Quiroz A., « Fast random generation of binary, t-ary and other types of trees », *Journal of Classification*, vol. 6, n° 1, p. 223-231, December, 1989. available at http ://ideas.repec.org/a/spr/jclass/v6y1989i1p223-231.html.

Robinson R. W., « Counting unlabeled acyclic digraphs », *in* C. H. C. Little (ed.), *Combinatorial Mathematics V*, vol. 622 of *Lecture Notes in Mathematics*, Springer, Berlin, p. 28-43, 1977.

Rosset S., Segal E., « Boosting Density Estimation », *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada, p. 267-281, 2002.