Apprentissage de structure de réseaux bayésiens à partir de réseaux markoviens

Christophe Gonzales et *Nicolas Jouve*

{Christophe.Gonzales, Nicolas.Jouve}@lip6.fr

LIP6 – Université Paris 6

Introduction

Apprentissage de structure : étant donné un échantillon d'une distribution P, trouver un graphe représentant P « au mieux »

Deux approches :

- *Contraintes :* tests d'indépendance conditionnelle (*e.g.* χ^2)
- *Optimisation :* exploration heuristique de l'espace des modèles, dotés d'une mesure de qualité (*e.g.* BIC)

Introduction

Apprentissage de structure : étant donné un échantillon d'une distribution P, trouver un graphe représentant P « au mieux »

Deux approches :

- *Contraintes :* tests d'indépendance conditionnelle (*e.g.* χ^2)
- *Optimisation :* exploration heuristique de l'espace des modèles, dotés d'une mesure de qualité (*e.g.* BIC)

Introduction

Approche optimisation : quel espace utiliser ?

- *RB :* structures équivalentes forment des plateaux dans la fonction score
- *Classes d'équivalence* (graphes partiellement orientés) :
  - espace plus petit et mieux adapté à l'exploration
  - garantie d'optimalité (algo Greedy Equivalent Search)
- *RM :* contrepartie non orientée des RB
  - exponentiellement plus petit
  - voisinage plus simple

Motivation : peut-on tirer profit de l'espace des RM ?

Introduction

Approche optimisation : quel espace utiliser ?

- *RB :* structures équivalentes forment des plateaux dans la fonction score
- *Classes d'équivalence* (graphes partiellement orientés) :
  - espace plus petit et mieux adapté à l'exploration
  - garantie d'optimalité (algo Greedy Equivalent Search)
- *RM :* contrepartie non orientée des RB
  - exponentiellement plus petit
  - voisinage plus simple

Motivation : peut-on tirer profit de l'espace des RM ?

Introduction

Approche optimisation : quel espace utiliser ?

- *RB :* structures équivalentes forment des plateaux dans la fonction score
- *Classes d'équivalence* (graphes partiellement orientés) :
  - espace plus petit et mieux adapté à l'exploration
  - garantie d'optimalité (algo Greedy Equivalent Search)
- *RM :* contrepartie non orientée des RB
  - exponentiellement plus petit
  - voisinage plus simple

Motivation : peut-on tirer profit de l'espace des RM ?

Introduction

Approche optimisation : quel espace utiliser ?

- *RB :* structures équivalentes forment des plateaux dans la fonction score
- *Classes d'équivalence* (graphes partiellement orientés) :
  - espace plus petit et mieux adapté à l'exploration
  - garantie d'optimalité (algo Greedy Equivalent Search)
- *RM :* contrepartie non orientée des RB
  - exponentiellement plus petit
  - voisinage plus simple

Motivation : peut-on tirer profit de l'espace des RM ?

Plan de l'exposé

- Réseaux bayésiens et markoviens
- Notre méthode
- Apprentissage du RM
- Orientation
- Raffinement
- Conclusions et perspectives

Réseaux markoviens et bayésiens

Réseau	markovien	bayésien
Graphe	non orienté	orienté sans circuits
Critère	Séparation : $\mathbf{X} \perp_s \mathbf{Y} \mid \mathbf{Z}$ si toute chaîne entre **X** et **Y** possède un nœud dans **Z**	d-séparation : $\mathbf{X} \perp_{ds} \mathbf{Y} \mid \mathbf{Z}$si $\forall$ chaîne *ch* entre **X** et **Y** $\exists$ un nœud *S* de *ch* t.q. - si *S* est *convergent* sur *ch*, ni *S* ni aucun de ses descendants n'appartiennent à **Z**, - sinon, *S* appartient à **Z**.
Factoris.	si $P > 0$, $P(\mathcal{V}) = \prod_{\mathbf{C} \in \mathcal{C}} \psi(\mathbf{C})$, où $\mathcal{C}$ est l'ens. des cliques du graphe.	$P(\mathcal{V}) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$

Réseaux markoviens et bayésiens

Réseau	markovien	bayésien
Graphe	non orienté	orienté sans circuits
Critère	Séparation : $\mathbf{X} \perp_s \mathbf{Y} \mid \mathbf{Z}$ si toute chaîne entre **X** et **Y** possède un nœud dans **Z**	d-séparation : $\mathbf{X} \perp_{ds} \mathbf{Y} \mid \mathbf{Z}$si $\forall$ chaîne *ch* entre **X** et **Y** $\exists$ un nœud *S* de *ch* t.q. - si *S* est *convergent* sur *ch*, ni *S* ni aucun de ses descendants n'appartiennent à **Z**, - sinon, *S* appartient à **Z**.
Factoris.	si $P > 0$, $P(\mathcal{V}) = \prod_{\mathbf{C} \in \mathcal{C}} \psi(\mathbf{C})$, où $\mathcal{C}$ est l'ens. des cliques du graphe.	$P(\mathcal{V}) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$

Réseaux markoviens et bayésiens

Réseau	markovien	bayésien
Graphe	non orienté	orienté sans circuits
Critère	Séparation : $\mathbf{X} \perp_s \mathbf{Y} \mid \mathbf{Z}$ si toute chaîne entre **X** et **Y** possède un nœud dans **Z**	d-séparation : $\mathbf{X} \perp_{ds} \mathbf{Y} \mid \mathbf{Z}$ si $\forall$ chaîne *ch* entre **X** et **Y** $\exists$ un nœud *S* de *ch* t.q. - si *S* est *convergent* sur *ch*, ni *S* ni aucun de ses descendants n'appartiennent à **Z**, - sinon, *S* appartient à **Z**.
Factoris.	si $P > 0$, $P(\mathcal{V}) = \prod_{\mathbf{C} \in \mathcal{C}} \psi(\mathbf{C})$, où $\mathcal{C}$ est l'ens. des cliques du graphe.	$P(\mathcal{V}) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$

Réseaux markoviens et bayésiens

Réseau	markovien	bayésien
Graphe	non orienté	orienté sans circuits
Critère	Séparation : **X** $\perp_s$ **Y** | **Z** si toute chaîne entre **X** et **Y** possède un nœud dans **Z**	d-séparation : **X** $\perp_{ds}$ **Y** | **Z** si $\forall$ chaîne *ch* entre **X** et **Y** $\exists$ un nœud *S* de *ch* t.q. - si *S* est *convergent* sur *ch*, ni *S* ni aucun de ses descendants n'appartiennent à **Z**, - sinon, *S* appartient à **Z**.
Factoris.	si $P > 0$, $P(\mathcal{V}) = \prod_{\mathbf{C} \in \mathcal{C}} \psi(\mathbf{C})$, où $\mathcal{C}$ est l'ens. des cliques du graphe.	$P(\mathcal{V}) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$

Optimalité

- Ppté de Markov Globale : $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \Longrightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$
- Optimalité : $\mathcal{G}$ contenant P est optimale s'il n'existe pas $\mathcal{G}'$ contenant P tq (i) $\mathcal{G}$ « contient » $\mathcal{G}'$ et (ii) $\mathcal{G}$ et $\mathcal{G}'$ ne sont pas équivalentes
- *P DAG-isomorphe* $\Longrightarrow \exists \mathcal{B}^*$ RB optimal, unique aux équivalents près
- *P DAG-isomorphe* $\Longrightarrow P > 0 \Longrightarrow \exists! \mathcal{G}^*$ RM optimal, graphe moral de $\mathcal{B}^*$

Optimalité

- Ppté de Markov Globale : $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \Longrightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$
- Optimalité : $\mathcal{G}$ contenant P est optimale s'il n'existe pas $\mathcal{G}'$ contenant P tq (i) $\mathcal{G}$ « contient » $\mathcal{G}'$ et (ii) $\mathcal{G}$ et $\mathcal{G}'$ ne sont pas équivalentes
- *P DAG-isomorphe* $\Longrightarrow \exists \mathcal{B}^*$ RB optimal, unique aux équivalents près
- *P DAG-isomorphe* $\Longrightarrow P > 0 \Longrightarrow \exists ! \mathcal{G}^*$ RM optimal, graphe moral de $\mathcal{B}^*$

De Markov à Bayes : tout est dans la V-structure

- L'information liée à l'orientation réside dans les nœuds convergents, et plus précisément dans les *V-structures*
- Passer d'un modèle à l'autre :
  - RB = RM ssi le graphe est triangulé (NB : DAG sans VS est triangulé)
  - RB $\longrightarrow$ RM : moralisation
  - RM $\longrightarrow$ RB : démoralisation...

De Markov à Bayes : tout est dans la V-structure

- L'information liée à l'orientation réside dans les nœuds convergents, et plus précisément dans les *V-structures*
- Passer d'un modèle à l'autre :
  - RB = RM ssi le graphe est triangulé (NB : DAG sans VS est triangulé)
  - RB $\rightarrow$ RM : moralisation
  - RM $\rightarrow$ RB : démoralisation...

De Markov à Bayes : tout est dans la V-structure

- L'information liée à l'orientation réside dans les nœuds convergents, et plus précisément dans les *V-structures*
- Passer d'un modèle à l'autre :
  - RB = RM ssi le graphe est triangulé (NB : DAG sans VS est triangulé)
  - RB $\rightarrow$ RM : moralisation
  - RM $\rightarrow$ RB : démoralisation...

De Markov à Bayes : tout est dans la V-structure

- L'information liée à l'orientation réside dans les nœuds convergents, et plus précisément dans les *V-structures*
- Passer d'un modèle à l'autre :
  - RB = RM ssi le graphe est triangulé (NB : DAG sans VS est triangulé)
  - RB $\rightarrow$ RM : moralisation
  - RM $\rightarrow$ RB : démoralisation...

Plan de l'exposé

- Réseaux bayésiens et markoviens
- Notre méthode
- Apprentissage du RM
- Orientation
- Raffinement
- Conclusions et perspectives

Notre méthode

Hypothèses : P est DAG-isomorphe et les données sont en nombre suffisant

- construire le RM optimal $\mathcal{G}^*$
- l'orienter en un RB $\mathcal{B}_0$
- raffiner $\mathcal{B}_0$ dans l'espace des CE jusqu'à obtenir $\mathcal{B}^*$

Approche exacte (comme GES) malgré la NP-difficulté du problème mais...

- le pire cas ne semble pas fréquent
- on obtient des réseaux aux propriétés intéressantes en temps polynomial
- la phase de raffinement est *anytime*

Notre méthode

Hypothèses : P est DAG-isomorphe et les données sont en nombre suffisant

- construire le RM optimal $\mathcal{G}^*$
- l'orienter en un RB $\mathcal{B}_0$
- raffiner $\mathcal{B}_0$ dans l'espace des CE jusqu'à obtenir $\mathcal{B}^*$

Approche exacte (comme GES) malgré la NP-difficulté du problème mais...

- le pire cas ne semble pas fréquent
- on obtient des réseaux aux propriétés intéressantes en temps polynomial
- la phase de raffinement est *anytime*

Plan de l'exposé

- Réseaux bayésiens et markoviens
- Notre méthode
- Apprentissage du RM
- Orientation
- Raffinement
- Conclusions et perspectives

Exploration de l'espace des RM

Deux approches :

- *Optimisation :* impossible dans ce contexte car l'estimation du MV est très coûteuse dans le cas général
- $\Rightarrow$ *Contraintes*, mais les tests IC sont non-significatifs si l'ensemble conditionnant est trop grand (pour une quantité de données raisonnable)

Exploration de l'espace des RM

Deux approches :

- *Optimisation :* impossible dans ce contexte car l'estimation du MV est très coûteuse dans le cas général
- $\Rightarrow$ *Contraintes*, mais les tests IC sont non-significatifs si l'ensemble conditionnant est trop grand (pour une quantité de données raisonnable)

Apprentissage de $\mathcal{G}^*$

1. $\mathcal{G}$ graphe vide
2. *Phase d'ajouts :*
   TQ c'est possible,
   - Choisir $(X, Y) \notin \mathcal{G}$ tq $X \not\perp\!\!\!\perp Y \mid Sep_{\mathcal{G}}(X, Y)$
   - L'ajouter
3. *Phase de retraits :*
   $\forall (X, Y) \in \mathcal{G}$ tq $X \perp\!\!\!\perp Y \mid Sep_{\mathcal{G}}(X, Y)$, ôter (X, Y)

$\hookrightarrow$ On montre qu'on obtient $\mathcal{G}^*$, en temps polynomial

Apprentissage de $\mathcal{G}^*$

1. $\mathcal{G}$ graphe vide
2. *Phase d'ajouts :*
   TQ c'est possible,
   - Choisir $(X, Y) \notin \mathcal{G}$ tq $X \not\perp\!\!\!\perp Y \mid Sep_{\mathcal{G}}(X, Y)$
   - L'ajouter
3. *Phase de retraits :*
   $\forall (X, Y) \in \mathcal{G}$ tq $X \perp\!\!\!\perp Y \mid Sep_{\mathcal{G}}(X, Y)$, ôter (X, Y)

$\hookrightarrow$ On montre qu'on obtient $\mathcal{G}^*$, en temps polynomial

Comment éviter les tests non-significatifs ?

- Construction incrémentale plutôt qu'agrégation de couvertures de Markov
- Calculs de séparateurs (presque) minimaux par triangulation du graphe courant et collecte-diffusion dans l'arbre de jonction
- Choix heuristique de l'arête à ajouter : dépendance la plus grande, mesurée par l'écart normalisé au seuil du χ^2

Comment éviter les tests non-significatifs ?

- Construction incrémentale plutôt qu'agrégation de couvertures de Markov
- Calculs de séparateurs (presque) minimaux par triangulation du graphe courant et collecte-diffusion dans l'arbre de jonction
- Choix heuristique de l'arête à ajouter : dépendance la plus grande, mesurée par l'écart normalisé au seuil du χ^2

Comment éviter les tests non-significatifs ?

- Construction incrémentale plutôt qu'agrégation de couvertures de Markov
- Calculs de séparateurs (presque) minimaux par triangulation du graphe courant et collecte-diffusion dans l'arbre de jonction
- Choix heuristique de l'arête à ajouter : dépendance la plus grande, mesurée par l'écart normalisé au seuil du χ^2

Plan de l'exposé

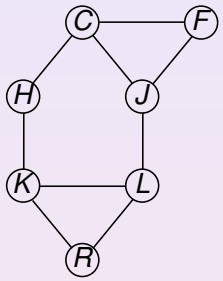
- Réseaux bayésiens et markoviens
- Notre méthode
- Apprentissage du RM
- Orientation
- Raffinement
- Conclusions et perspectives

Orientation

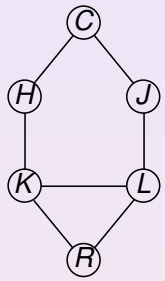
Boucle sur les nœuds :

- Choisir X n'appartenant qu'à une unique clique
- Orienter les arêtes adjacentes vers lui
- Tenter d'ôter des arêtes de moralisation entre ses voisins

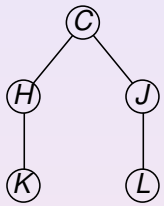
Orientation



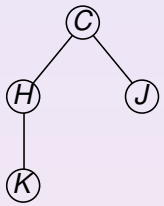
$\mathcal{G}^*$



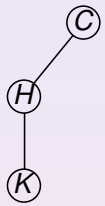
$\mathcal{G}_1$



$\mathcal{G}_2$



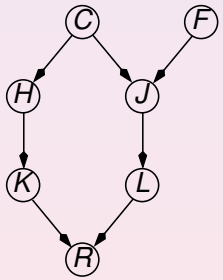
$\mathcal{G}_3$



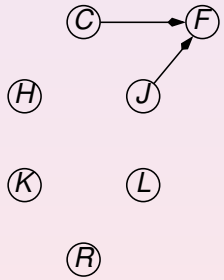
$\mathcal{G}_4$

$\mathcal{G}_5$

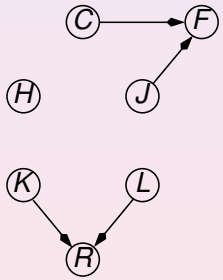
$\mathcal{G}_6$



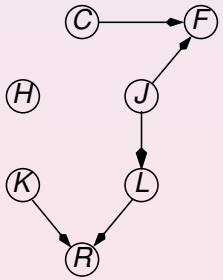
$\mathcal{B}^*$



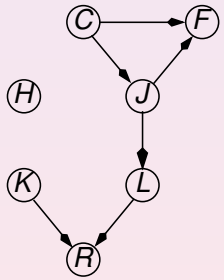
$\mathcal{B}_1$



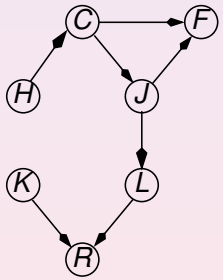
$\mathcal{B}_2$



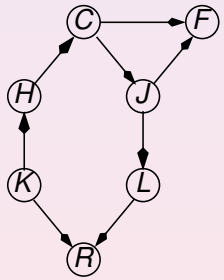
$\mathcal{B}_3$



$\mathcal{B}_4$



$\mathcal{B}_5$



$\mathcal{B}_6 = \mathcal{B}$

Propriétés

- A chaque pas, il existe bien un nœud n'appartenant qu'à une unique clique si l'algo est appliqué à $\mathcal{G}^*$
- Sinon, il faut trianguler localement
- On obtient un RB $\mathcal{B}_0$ contenant P, de graphe moral $\mathcal{G}^*$, en temps polynomial
- On a révélé des VS de $\mathcal{B}^*$ mais en général pas toutes

Propriétés

- A chaque pas, il existe bien un nœud n'appartenant qu'à une unique clique si l'algo est appliqué à $\mathcal{G}^*$
- Sinon, il faut trianguler localement
- On obtient un RB $\mathcal{B}_0$ contenant P, de graphe moral $\mathcal{G}^*$, en temps polynomial
- On a révélé des VS de $\mathcal{B}^*$ mais en général pas toutes

Propriétés

- A chaque pas, il existe bien un nœud n'appartenant qu'à une unique clique si l'algo est appliqué à $\mathcal{G}^*$
- Sinon, il faut trianguler localement
- On obtient un RB $\mathcal{B}_0$ contenant P, de graphe moral $\mathcal{G}^*$, en temps polynomial
- On a révélé des VS de $\mathcal{B}^*$ mais en général pas toutes

Plan de l'exposé

- Réseaux bayésiens et markoviens
- Notre méthode
- Apprentissage du RM
- Orientation
- Raffinement
- Conclusions et perspectives

Raffinement

Greedy Equivalent Search comprend deux phases (exponentielles) :

- Phase d'ajouts : construction d'un RB contenant P
- Phase de retraits : raffinement d'un RB contenant P jusqu'à $\mathcal{B}^*$

$\Longrightarrow$ on peut appliquer la phase 2 de GES à $\mathcal{B}_0$

Raffinement

Greedy Equivalent Search comprend deux phases (exponentielles) :

- Phase d'ajouts : construction d'un RB contenant P
- Phase de retraits : raffinement d'un RB contenant P jusqu'à $\mathcal{B}^*$

$\Longrightarrow$ on peut appliquer la phase 2 de GES à $\mathcal{B}_0$

Plan de l'exposé

- Réseaux bayésiens et markoviens
- Notre méthode
- Apprentissage du RM
- Orientation
- Raffinement
- Conclusions et perspectives

Conclusions

- on conserve la propriété d'optimalité de GES
- on remplace sa première phase exponentielle par une phase polynomiale
  - plus rapide !
  - on obtient de bons réseaux (optimaux pour l'inférence) en temps polynomial
- la phase de raffinement (exponentielle) est *anytime*

$\hookrightarrow$ *Principe :* exploiter d'abord et intégralement l'information accessible polynomialement, à savoir l'aspect non orienté du modèle

Conclusions

- on conserve la propriété d'optimalité de GES
- on remplace sa première phase exponentielle par une phase polynomiale
  - plus rapide !
  - on obtient de bons réseaux (optimaux pour l'inférence) en temps polynomial
- la phase de raffinement (exponentielle) est *anytime*

$\hookrightarrow$ *Principe :* exploiter d'abord et intégralement l'information accessible polynomialement, à savoir l'aspect non orienté du modèle

Conclusions

- on conserve la propriété d'optimalité de GES
- on remplace sa première phase exponentielle par une phase polynomiale
  - plus rapide !
  - on obtient de bons réseaux (optimaux pour l'inférence) en temps polynomial
- la phase de raffinement (exponentielle) est *anytime*

$\hookrightarrow$ *Principe :* exploiter d'abord et intégralement l'information accessible polynomialement, à savoir l'aspect non orienté du modèle

Conclusions

- on conserve la propriété d'optimalité de GES
- on remplace sa première phase exponentielle par une phase polynomiale
  - plus rapide !
  - on obtient de bons réseaux (optimaux pour l'inférence) en temps polynomial
- la phase de raffinement (exponentielle) est *anytime*

$\hookrightarrow$ *Principe :* exploiter d'abord et intégralement l'information accessible polynomialement, à savoir l'aspect non orienté du modèle

Perspectives

- Poursuivre les expérimentations
- Exploiter les propriétés de $\mathcal{B}_0$ pour optimiser la phase 2
- Etudier plus précisément la robustesse de la méthode à ses deux hypohtèses