Modèles graphiques probabilistes

Francis Bach
Centre de Morphologie Mathématique
Ecole des Mines de Paris

Mai 2006

Modèles graphiques probabilistes

Modèle graphique probabiliste = représentation graphique d'un ensemble de lois de probabilité multivariées

- Modularité
  - Gestion de la complexité
- Modèle probabiliste
  - flexibilité d'utilisation (inférence et apprentissage)
- Formalisme commun à de nombreux modèles/domaines
  - Transferts : théorie/applications, théorie/théorie

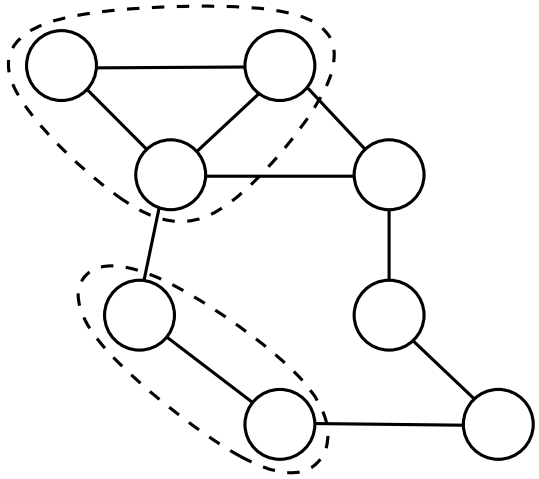
Modèles graphiques probabilistes
Plan de la présentation

- Définition
- Inférence
- Apprentissage
  - Paramètres
  - Structure
- Applications et perspectives

Modèle graphique

- Variables aléatoires: $X = (X_1, \ldots, X_n)$
- Modèle graphique = structure (graphe) + collection de lois locales
- Graphe
  - Sommets $\sim$ variables
  - Absence d'arêtes $\sim$ indépendances conditionnelles
- Deux grandes familles:
  - Graphe orienté
  - Graphe non orienté

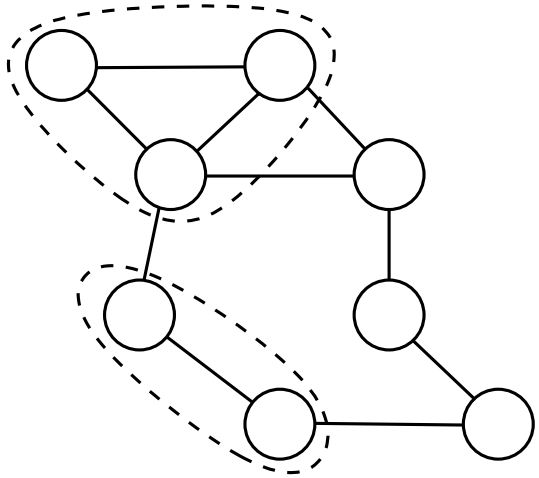
Modèle graphique non orienté (champ de Markov)



- Graphe, cliques
- (1) Loi factorisée: $p(x) = \frac{1}{Z} \prod_{j=1}^{p} \phi_{C_j}(x_{C_j})$

  Z est une constante de normalisation

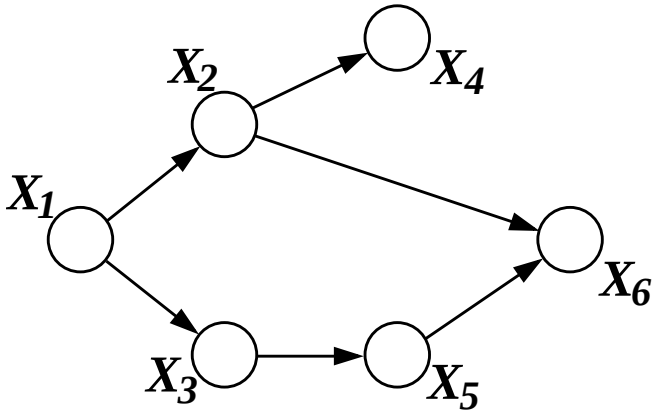
Modèle graphique non orienté (champ de Markov)



- Graphe, cliques
- (1) Loi factorisée: $p(x) = \frac{1}{Z} \prod_{j=1}^{p} \phi_{C_j}(x_{C_j})$

  Z est une constante de normalisation
- (2) Indépendances conditionnelles: sachant ses voisins, chaque variable est indépendante du reste du graphe (**séparation**)
- Propriétés équivalentes sous certaines conditions (Théorème de Hammersley-Cliford)

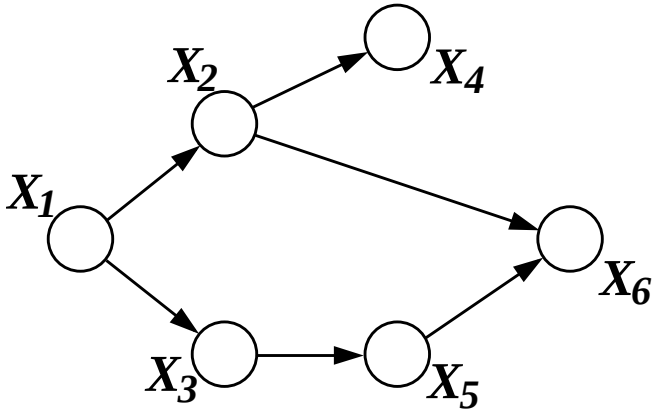
Modèle graphique orienté ("Réseaux Bayésiens")



- DAG (Directed acyclic graph): graphe orienté sans cycle
- (1) Loi factorisée: $p(x) = \prod_{i=1}^{n} p(x_i | x_{\text{parents}(i)})$

  (pas de contrainte de normalisation)

Modèle graphique orienté ("Réseaux Bayésiens")

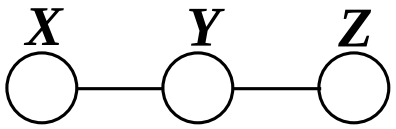


- DAG (Directed acyclic graph): graphe orienté sans cycle
- (1) Loi factorisée: $p(x) = \prod_{i=1}^{n} p(x_i | x_{\text{parents}(i)})$
  (pas de contrainte de normalisation)
- (2) Indépendance conditionnelles: chaque variable est indépendante de ses non-descendants sachant ses parents (**d-séparation**)

Modèles à trois noeuds

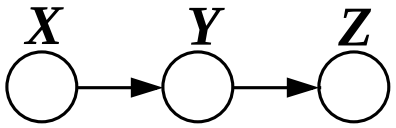
- Modèles triviaux

- Non orienté:



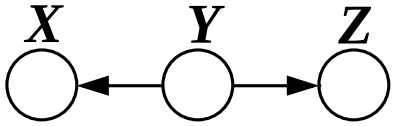
$$X \perp Z \mid Y$$

- Orienté 1 : Chaîne de Markov



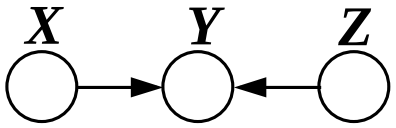
$$X \perp Z \mid Y$$

- Orienté 2 : variable latente commune



$$X \perp Z \mid Y$$

- Orienté 3 : "Explaining away"



$$X \perp Z$$

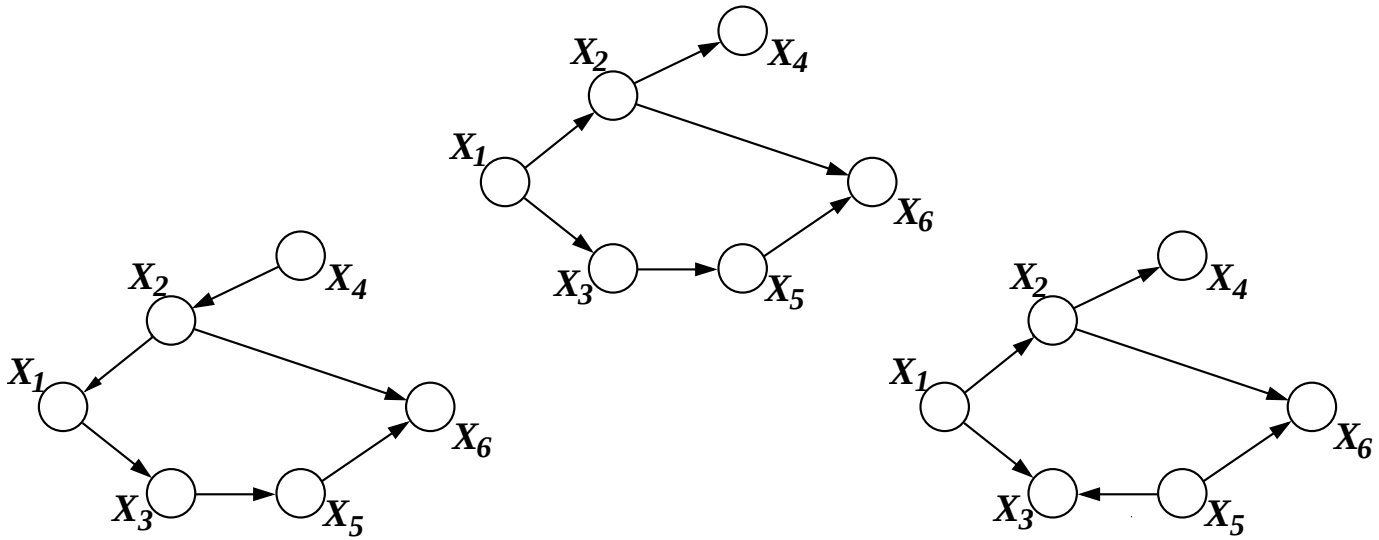
  – exemple: X, Z deux dés indépendants, Y leur somme

Markov-équivalence

- Deux graphes G_1, G_2 sont Markov-équivalents ssi ils définissent la même famille de lois de probabilité
- G_1, G_2 non orientés: Markov-équivalence $\Leftrightarrow$ égalité

Markov-équivalence

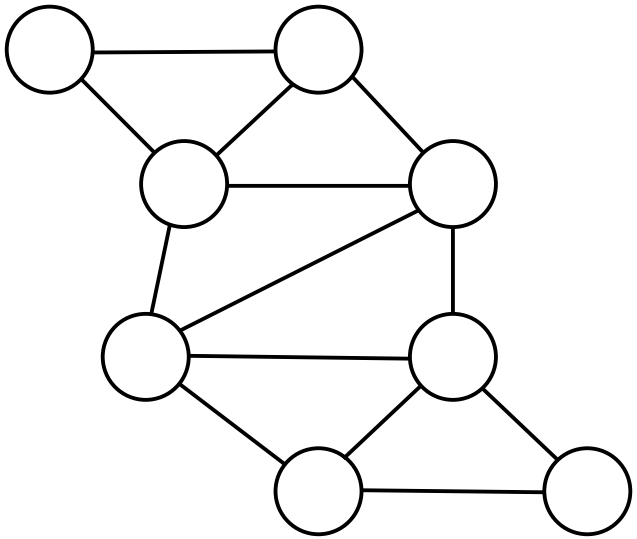
- Deux graphes G_1, G_2 sont Markov-équivalents ssi ils définissent la même famille de lois de probabilité
- G_1, G_2 non orientés: Markov-équivalence $\Leftrightarrow$ égalité
- G_1, G_2 orientés: Markov-équivalence $\Leftrightarrow$ G_1 et G_2 ont les mêmes "v-structures" et des graphes non orientés égaux



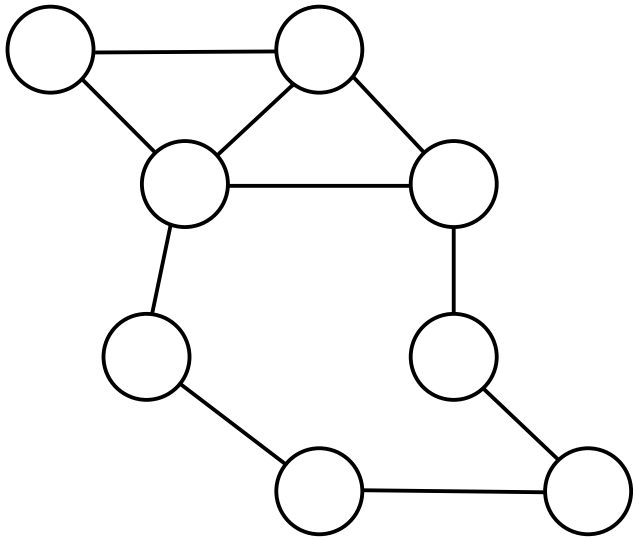
Graphes décomposables

- G_1 orienté, G_2 non orienté: Markov-équivalence possible pour les **graphes décomposables** (i.e., triangulés)
  - G non orienté est décomposable/triangulé ssi il n'existe pas de cycle d'ordre supérieur ou égal à 4 sans cordes.

Triangulé



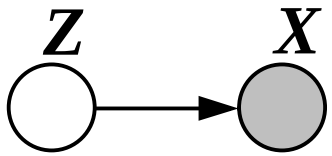
Non triangulé



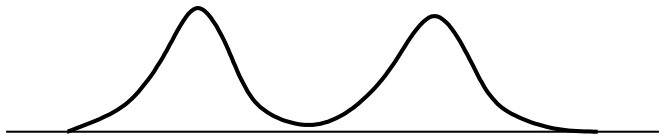
Modèles classiques formulés en modèles graphiques

- Modèles de mélange
- Analyse factorielle
- Modèle de Markov caché
- Filtre de Kalman
- Modèles pour inférence Bayésienne
- Champs de Markov

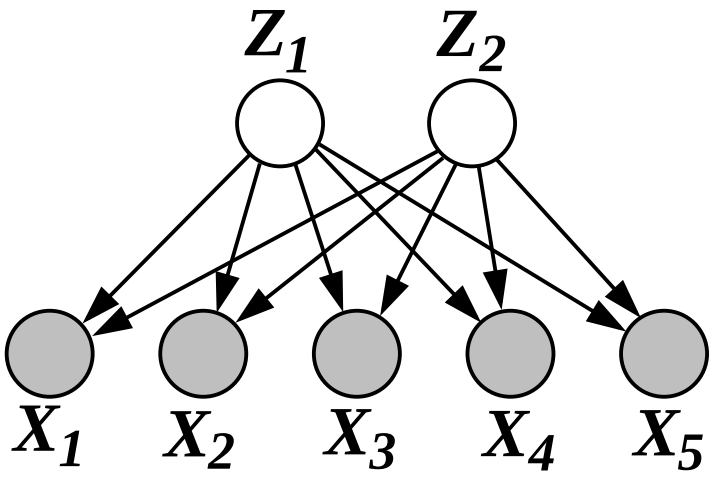
Modèles de mélange



- $p(x) = \sum_k p(z = k)p(x|z = k) = \sum_k \pi_k p_k(x)$
- NB: moyen simple pour modéliser des variables aléatoires non Gaussiennes/non standard:

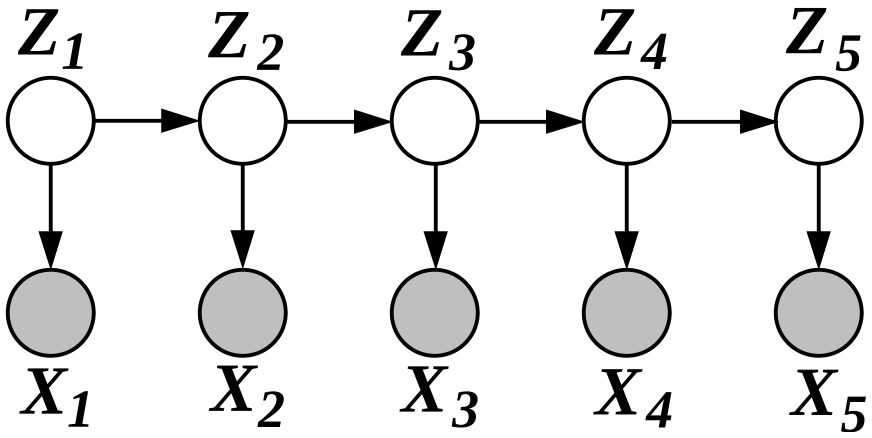


Analyse factorielle

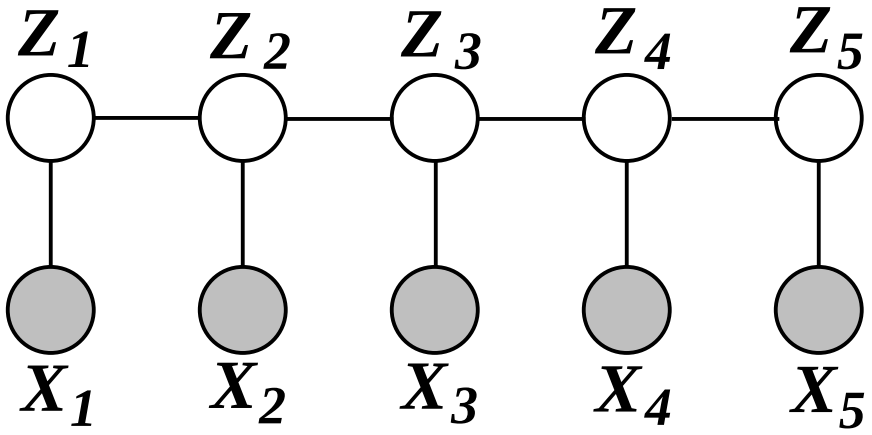


- $z_i \sim \mathcal{N}(0,1)$, $x_j|z \sim \mathcal{N}(\sum_j w_{ji} z_i, \sigma_j^2)$
- NB: si $\sigma_j^2 = \sigma^2$, $\forall j$, équivalent à l'analyse en composantes principales (Tipping & Bishop, 1999)

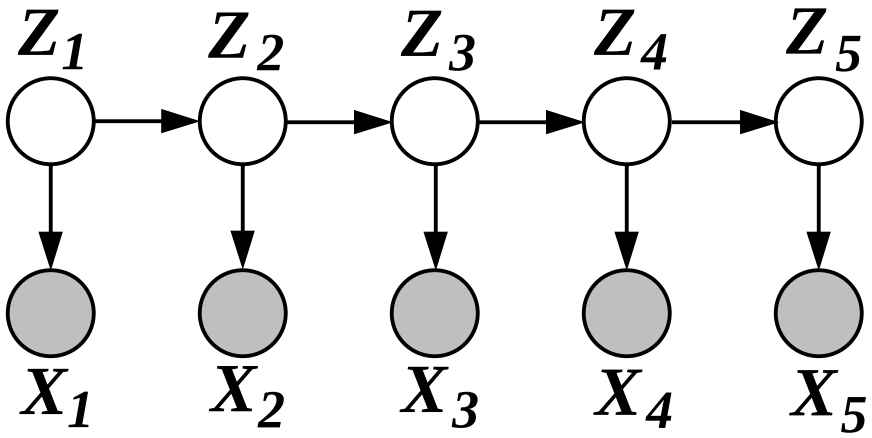
Modèles de Markov cachés



- Z_i discrets
- NB: équivalent à

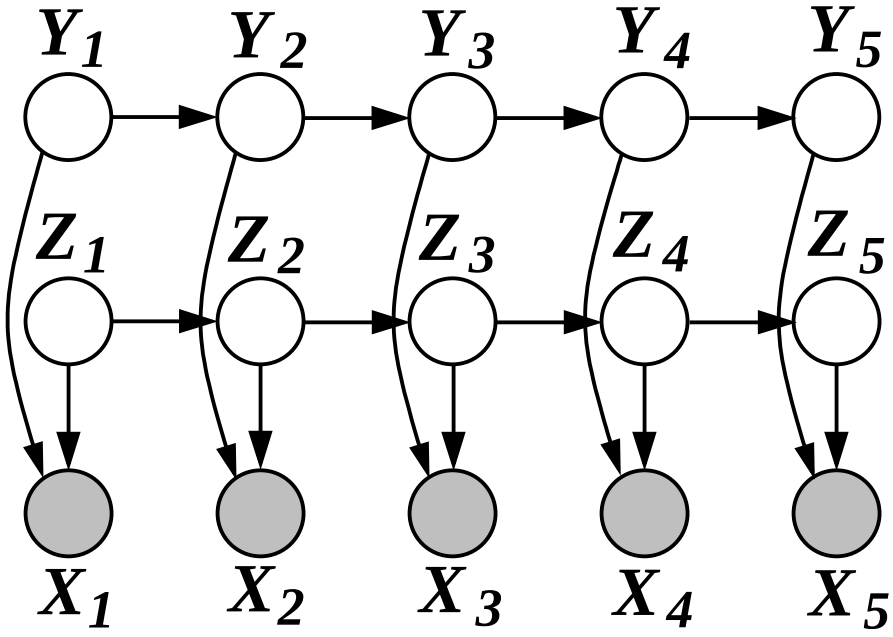


Filtre de Kalman



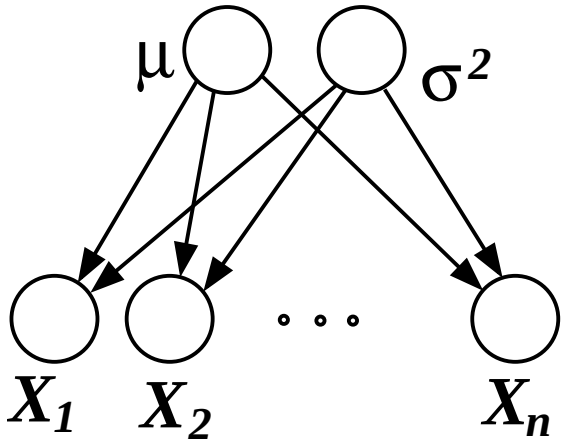
- Densités Gaussiennes

Modèles factoriels

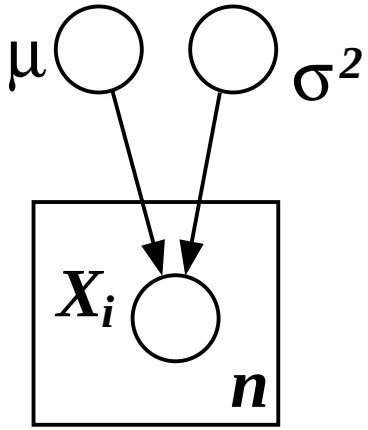


- Modèles de Markov cachés factoriels (Ghahramani & Jordan, 1997)
- Variables latentes discrètes et continues (switching Kalman filter)

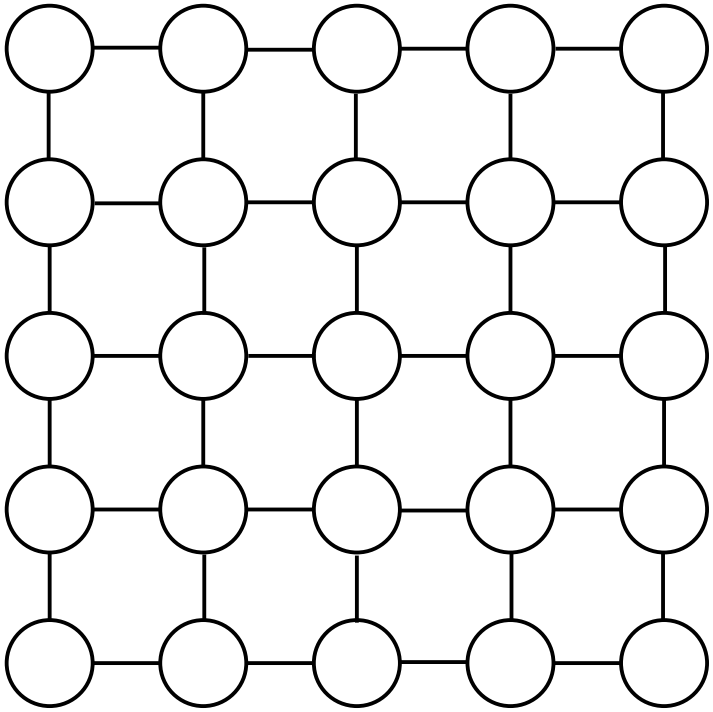
Modèle pour inférence Bayésienne



- Paramètres considérés comme variables aléatoires
- Exemple: $x_i|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma^2 \sim \Gamma(\alpha, \beta)$, $\mu \sim \mathcal{N}(\mu_0, t^2)$
- Notation "plate":



Champs de Markov



- Images/Pixels
- Physique statistique

Modèles graphiques probabilistes
Plan de la présentation

- Définition
- Inférence
- Apprentissage
  - Paramètres
  - Structure
- Applications et perspectives

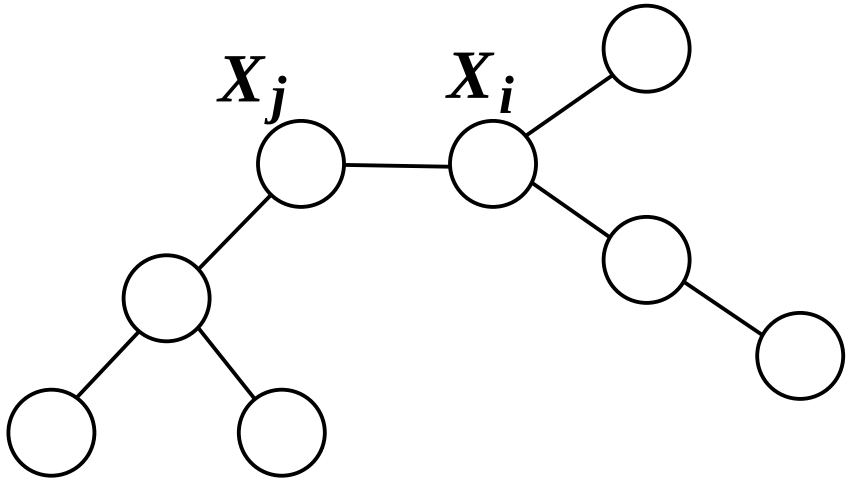
Inférence dans les modèles graphiques

- $\{1, \ldots, n\} = O \cup C$, observations et variables cachées
- Calculer $p(x_C|x_O)$ ou $\arg\max_{x_C} p(x_C|x_O)$
- Inférence naïve: $p(x_C|x_O) = p(x_C, x_O)/p(x_O)$
  et $p(x_O) = \sum_{x_C} p(x_C, x_O)$

Inférence dans les modèles graphiques

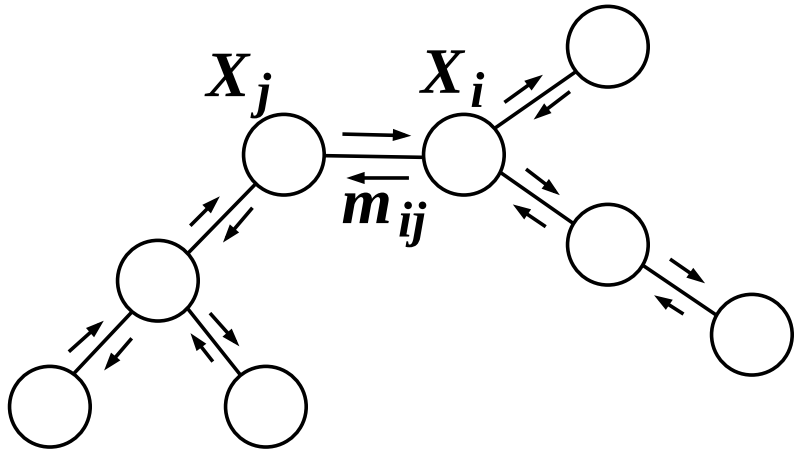
- $\{1, \ldots, n\} = O \cup C$, observations et variables cachées
- Calculer $p(x_C|x_O)$ ou $\arg\max_{x_C} p(x_C|x_O)$
- Inférence naïve: $p(x_C|x_O) = p(x_C, x_O)/p(x_O)$
  et $p(x_O) = \sum_{x_C} p(x_C, x_O)$
- Inférence exacte non naïve
  - utilise la factorisation des lois
  - cadre non orienté plus adapté
  - Tâche générique: calculer $Z = \sum_{x_1,\ldots,x_n} \prod_{j=1}^{p} \phi_{C_j}(x_{C_j})$ ou $\arg\max_{x_1,\ldots,x_n} \prod_{j=1}^{p} \phi_{C_j}(x_{C_j})$
- Inférence approchée

Inférence exacte dans les arbres



- Arbre (non orienté): graphe sans cycle
  - Potentiels: $\psi_i(x_i)$, $\psi_{ij}(x_i, x_j)$
  - $p(x) = \frac{1}{Z} \prod_i \psi_i(x_i) \prod_{i,j} \psi_{ij}(x_i, x_j)$
  - But: calculer $Z = \sum_{x_1,\ldots,x_n} \prod_i \psi_i(x_i) \prod_{i,j} \psi_{ij}(x_i, x_j)$

Inférence exacte dans les arbres



- But: calculer $Z = \sum_{x_1,\ldots,x_n} \prod_i \psi_i(x_i) \prod_{i,j} \psi_{ij}(x_i, x_j)$
- Propagation de $2n$ messages le long des arêtes

$$m_{ij}(x_j) = \sum_{x_i} \psi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i)\setminus\{j\}} m_{ki}(x_i)$$

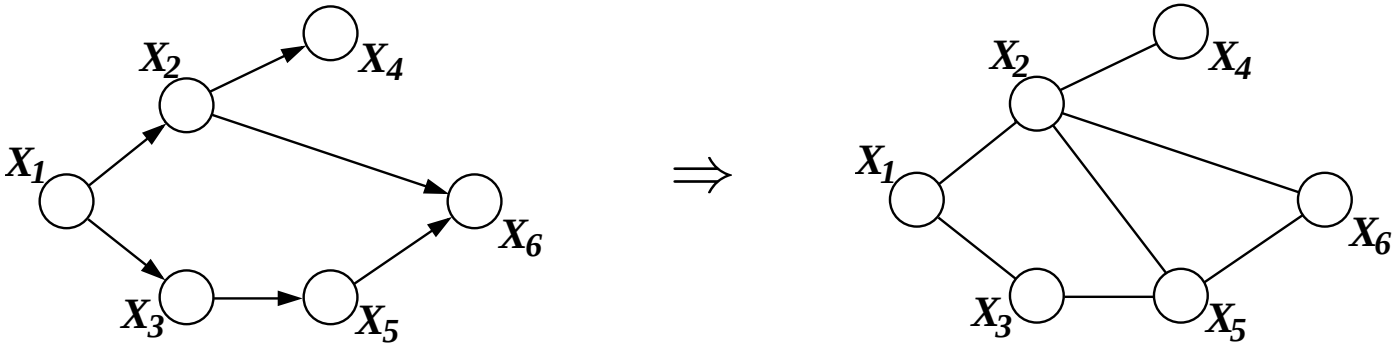
- Complexité linéaire
- NB: configuration la plus probable: remplacer $\sum$ par $\max$

Inférence exacte

(Lauritzen & Spiegelhalter,1988)

- Transformation des graphes en arbres non orientés

1. Moralisation des modèles graphiques orientés

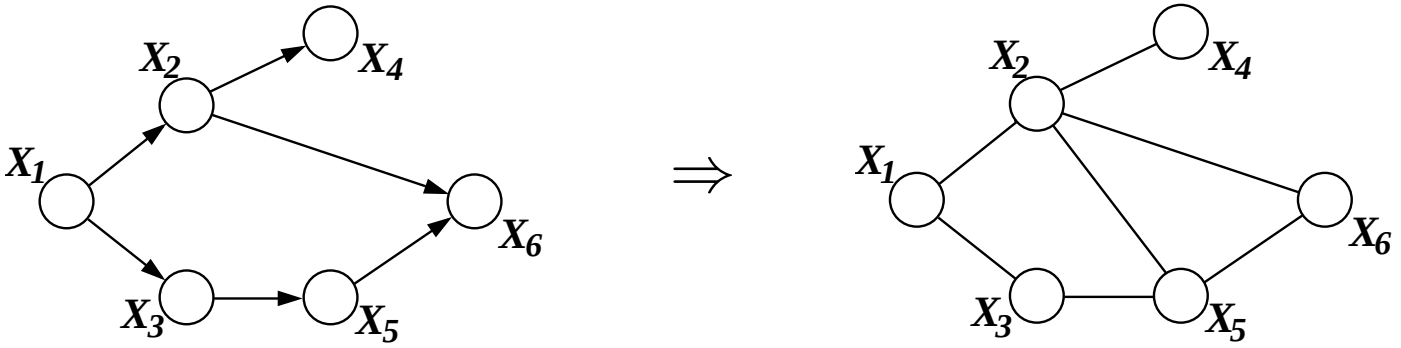


Inférence exacte

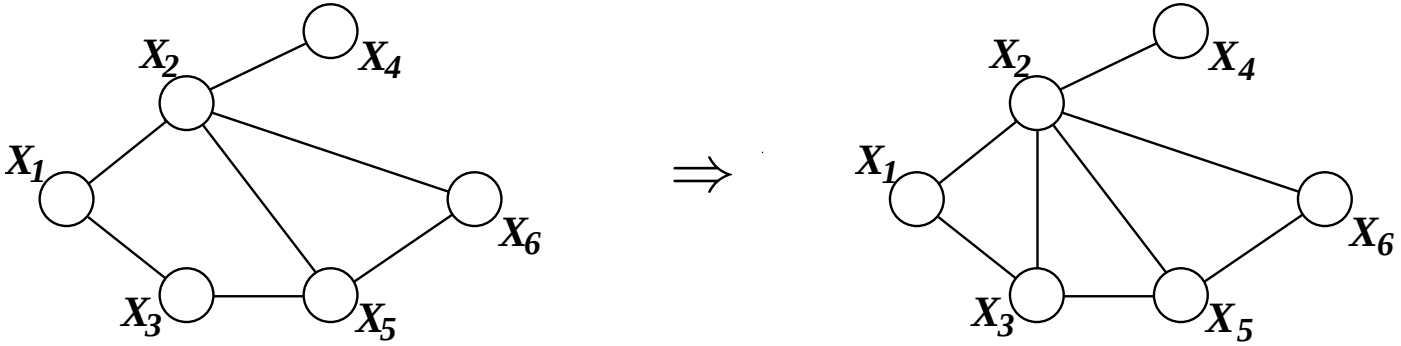
(Lauritzen & Spiegelhalter,1988)

- Transformation des graphes en arbres non orientés

1. Moralisation des modèles graphiques orientés



2. Triangulation du graphe non orienté

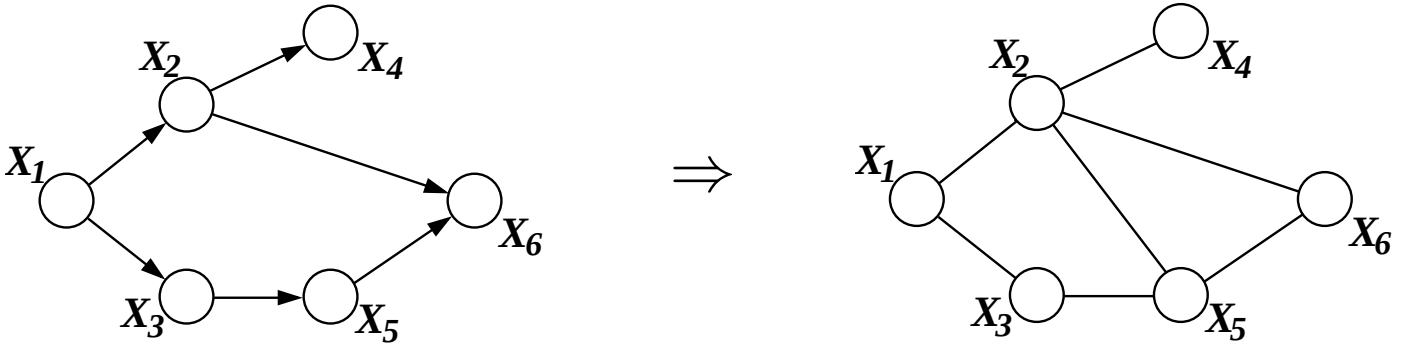


Inférence exacte

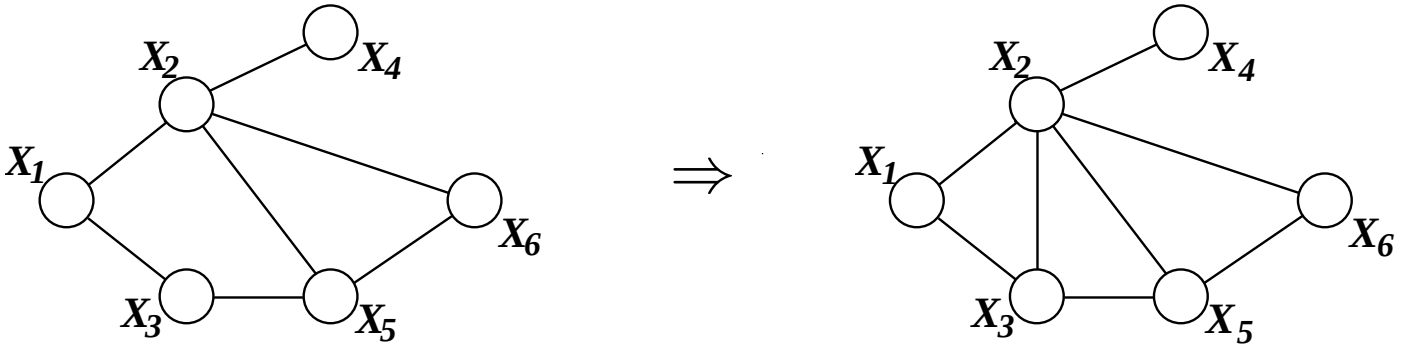
(Lauritzen & Spiegelhalter,1988)

- Transformation des graphes en arbres non orientés

1. Moralisation des modèles graphiques orientés



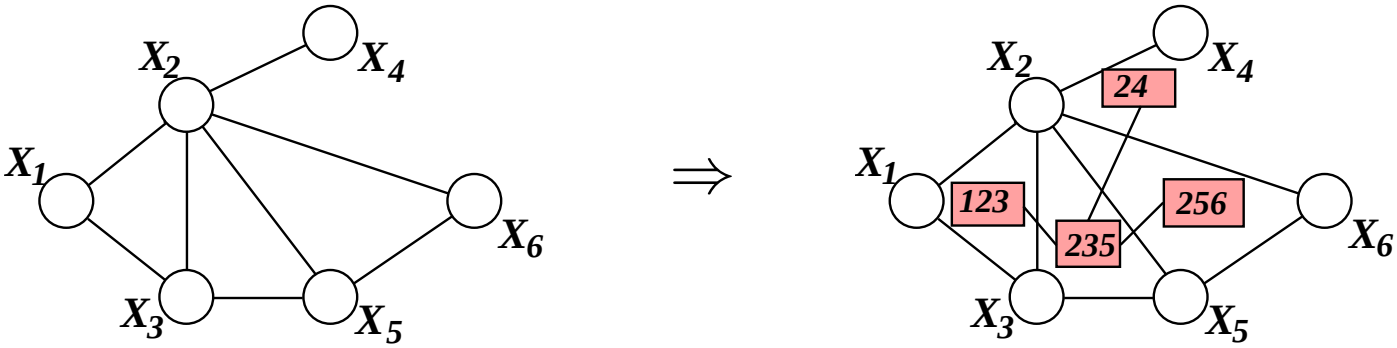
2. Triangulation du graphe non orienté



3. Construction d'un arbre de cliques

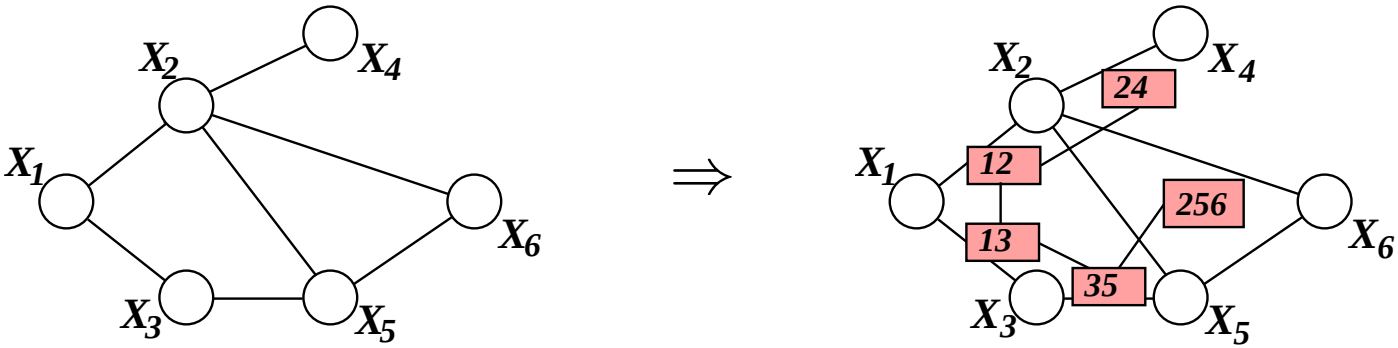
Construction d'un arbre de cliques

- Arbre de cliques



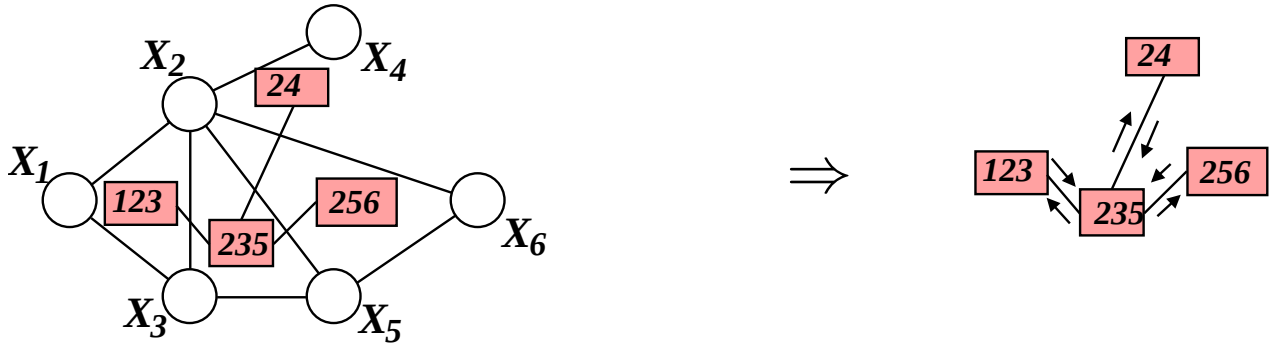
- Si le graphe est triangulé, arbre de clique = arbre de jonctions (propriété d'intersection courante)

- Sinon...



Propagation des messages

- Arbre de jonctions



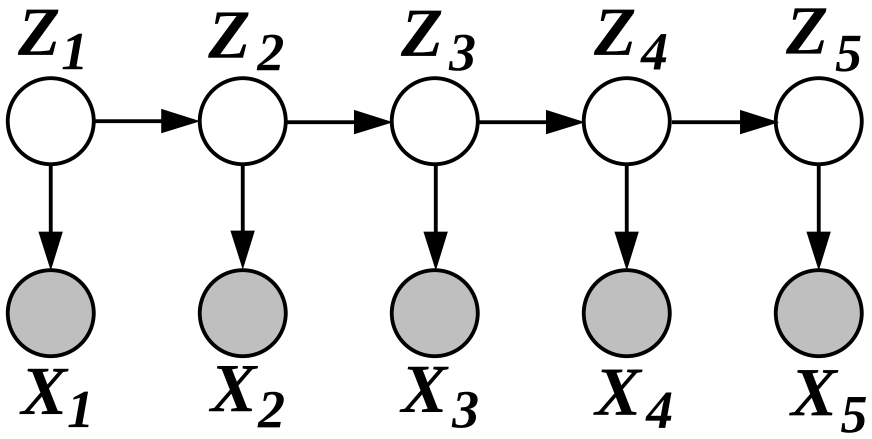
- Propagation de $2m$ messages (m = nombre de cliques), "MAX" ou "SUM"

- Cohérence locale $\Rightarrow$ cohérence globale !

- Complexité: exponentielle dans le cardinal de la plus grande clique

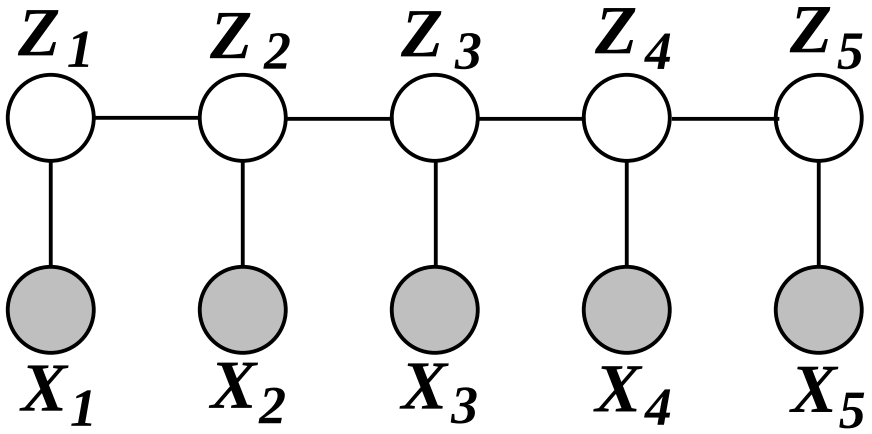
- Largeur arborescente (treewidth)
  = Cardinal de la plus grande clique d'un graphe triangulé optimal

Inférence
Modèle de Markov cachés



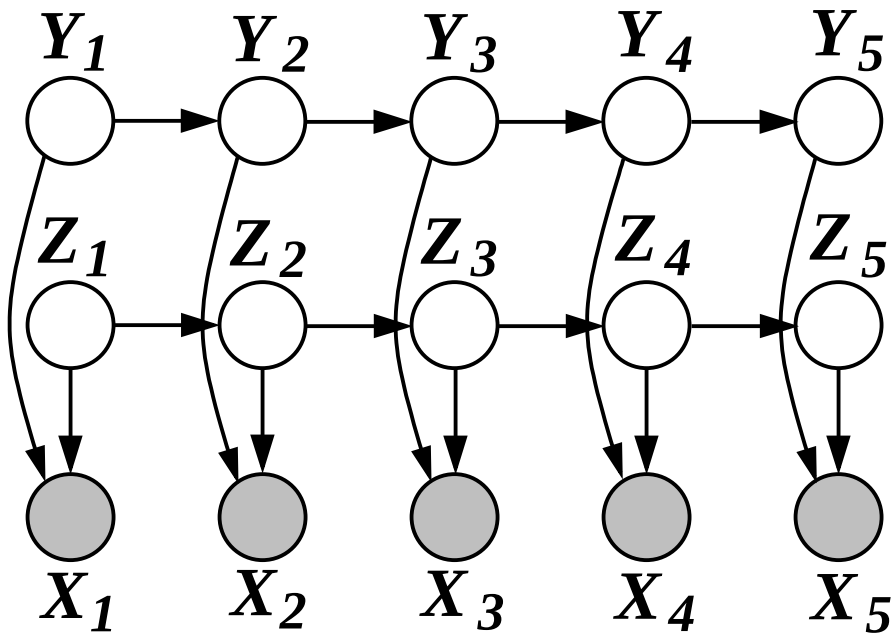
- Z_i discrets
- Max-propagation $=$ algorithme de Viterbi

Filtre de Kalman



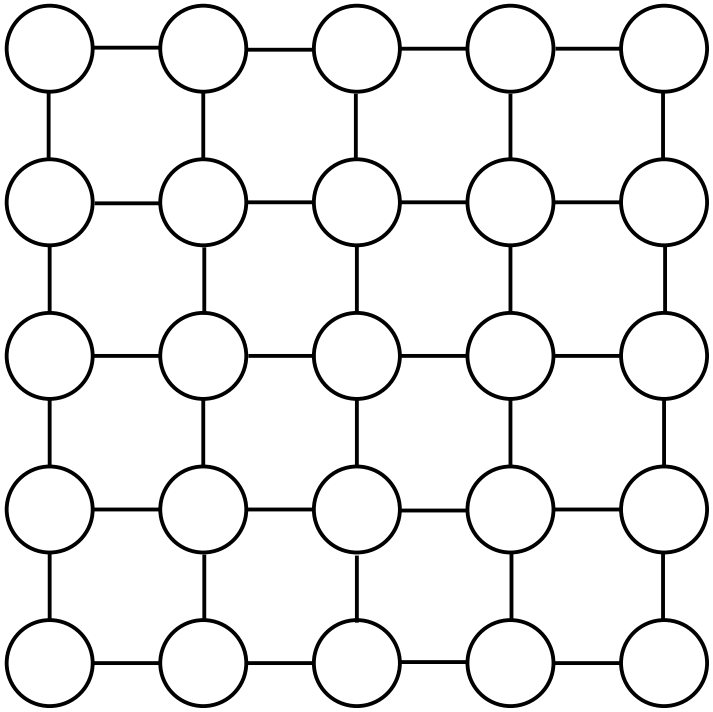
- Densités Gaussiennes
- Propagation de vecteurs moyennes et de matrice de covariances

Modèles factoriels



- Modèles de Markov cachés factoriels
- Si m chaînes de Markov latentes, complexité $O(2^{m+1})$ au lieu de $O(2^{2m})$ (cas binaire)
- Données hétérogènes ?

Champs de Markov



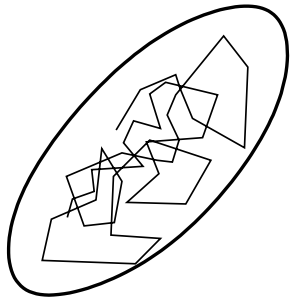
- Largeur arborescente non bornée

Problèmes de l'inférence exacte

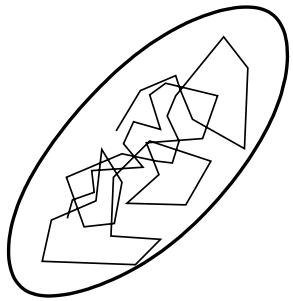
- (1) Grande largeur arborescente
- (2) Données hétérogènes
- Inférence approchée
  - Méthodes stochastiques (MCMC)
  - Méthodes variationnelles

Markov chain Monte Carlo (MCMC)

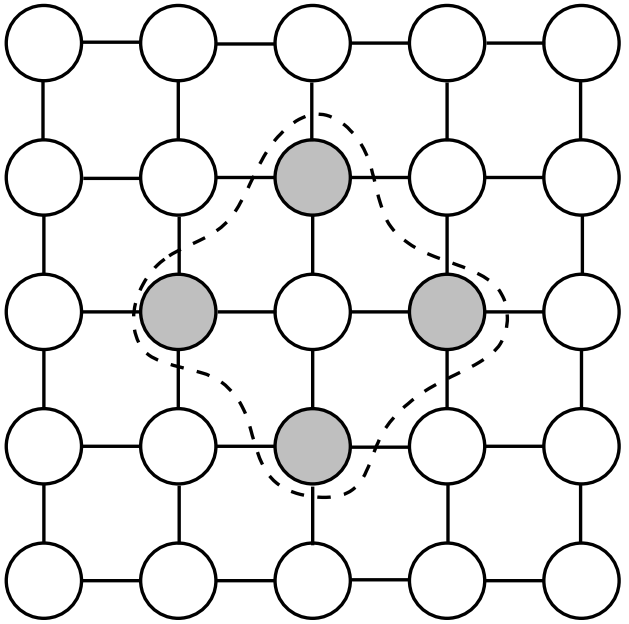
- Echantillonner $p(x_H C|x_O)$ à l'aide d'une chaîne de Markov dont la loi stationnaire est exactement $p(x_C|x_O)$



Markov chain Monte Carlo (MCMC)



- Echantillonner $p(x_H C|x_O)$ à l'aide d'une chaîne de Markov dont la loi stationnaire est exactement $p(x_C|x_O)$
- Construction simple de la loi de transition: échantillonage de Gibbs (Geman & Geman, 1984)
  - Itération: pour i de 1 à n, échantilloner $p(x_i|x_{\text{voisins}(i)})$
  - "burn-in"



Markov chain Monte Carlo (MCMC)

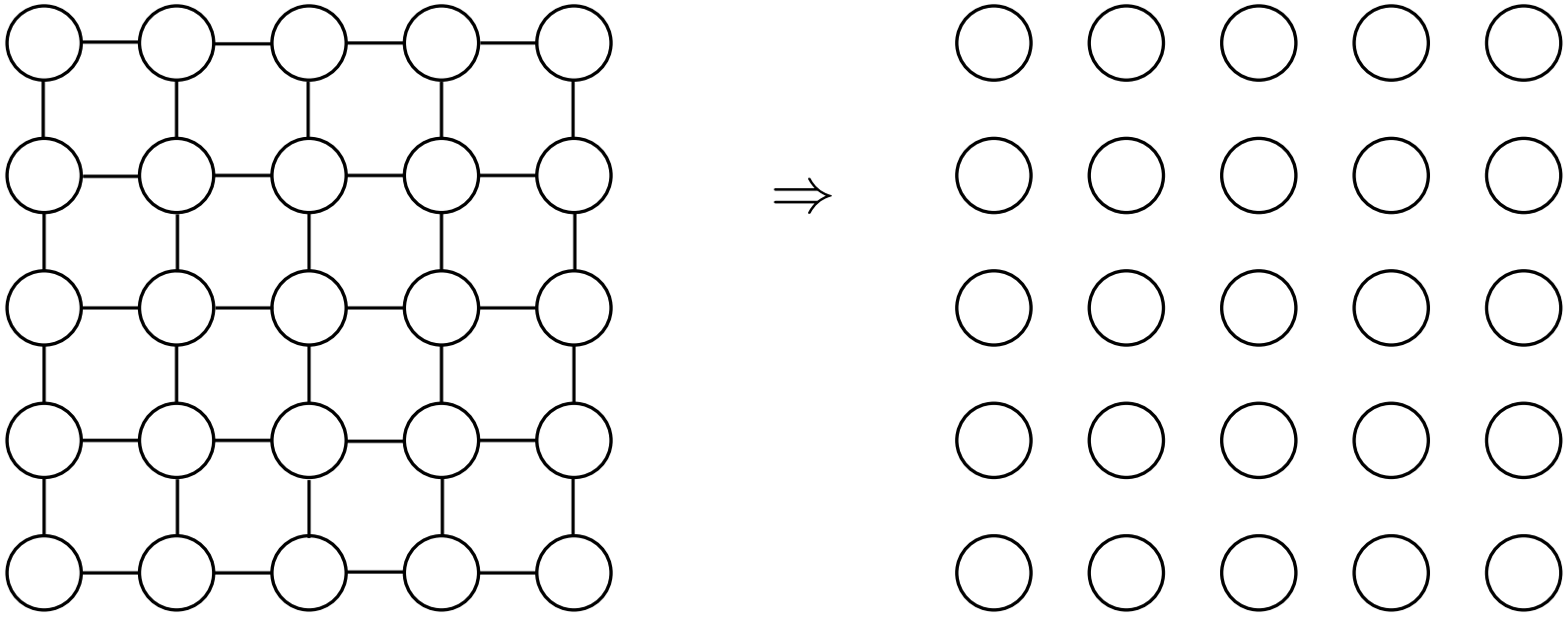
- Avantages
  - converge vers l'inférence exacte
  - peut toujours être utilisé
- Inconvénients
  - Non triviale à mettre en oeuvre correctement
  - Lenteur

Méthodes variationnelles d'inférence (Jordan & al., 1997)

- Approcher $p(x)$ par une loi $q(x, \lambda)$ plus simple
  - Optimization (efficace) du paramètre variationnel λ

Méthodes variationnelles d'inférence (Jordan & al., 1997)

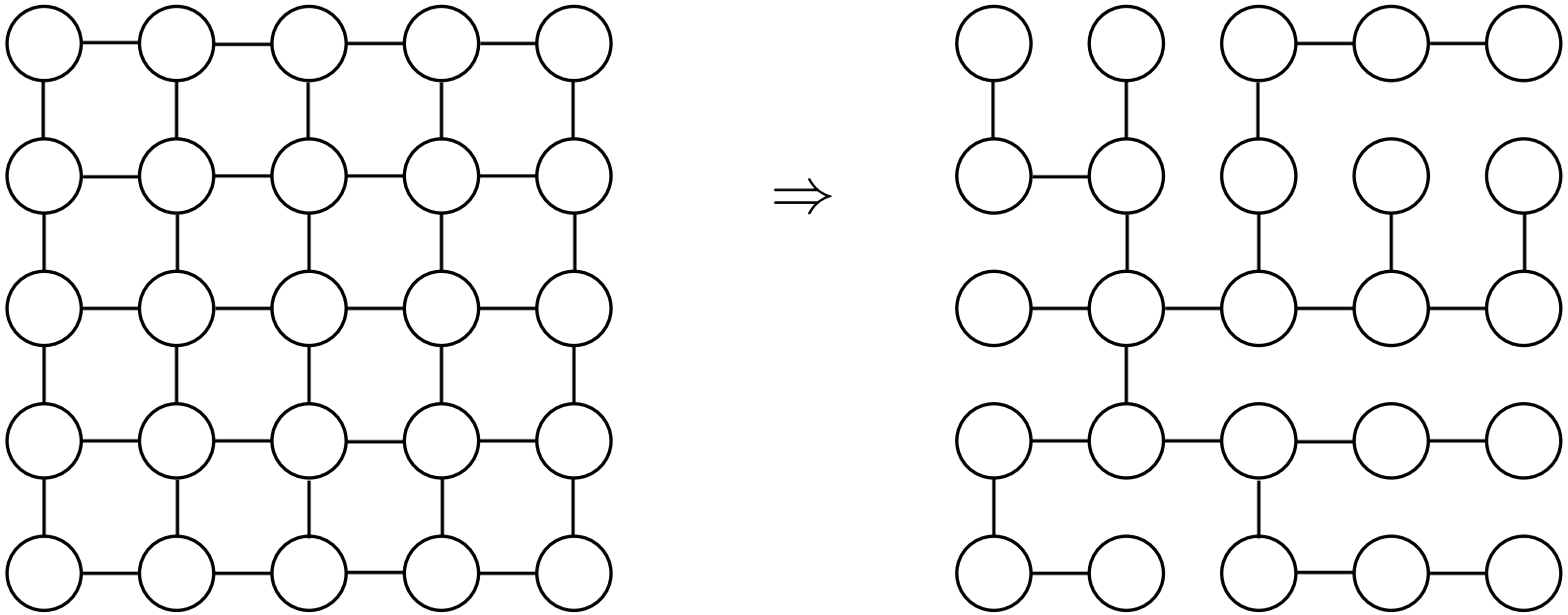
- Approcher $p(x)$ par une loi $q(x, \lambda)$ plus simple
  - Optimization (efficace) du paramètre variationnel λ
- Modèles graphiques: $q(x, \lambda)$ obtenue en enlevant des arêtes



Méthode du champ moyen (physique statistique)

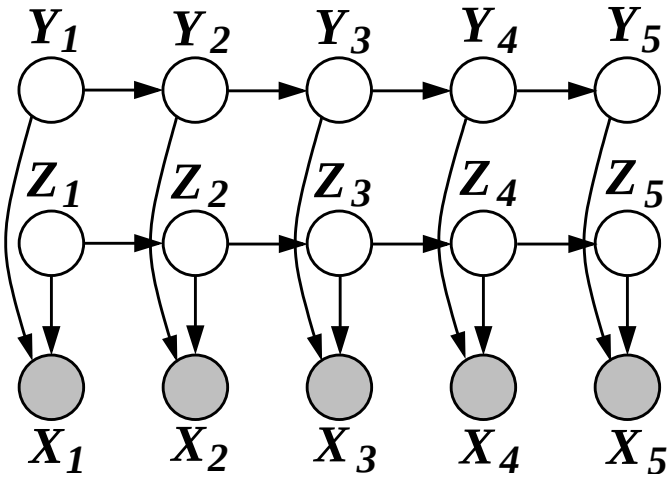
Méthodes variationnelles d'inférence (Jordan & al., 1997)

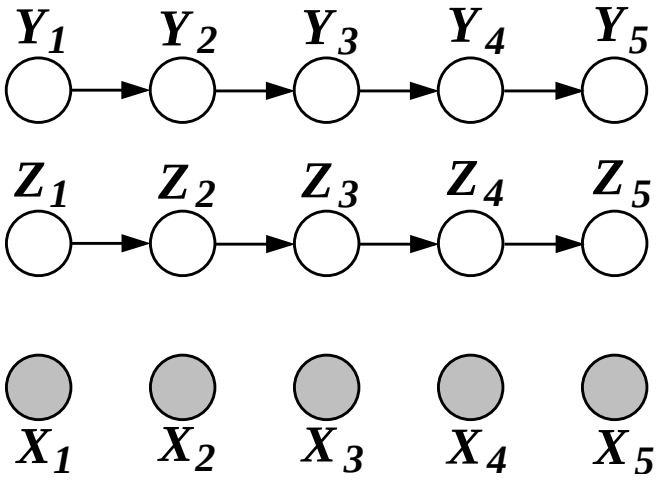
- Approcher $p(x)$ par une loi $q(x, \lambda)$ plus simple
  - Optimization (efficace) du paramètre variationnel λ
- Modèles graphiques: $q(x, \lambda)$ obtenue en enlevant des arêtes



Méthode structurée : utilisation de sous-structures simples

Méthodes variationnelles d'inférence
Modèle factoriel





Méthodes variationnelles d'inférence

- Avantages
  - Déterministe
  - Simple à mettre en oeuvre
- Inconvénients
  - ne converge pas vers l'inférence exacte

Modèles graphiques probabilistes
Plan de la présentation

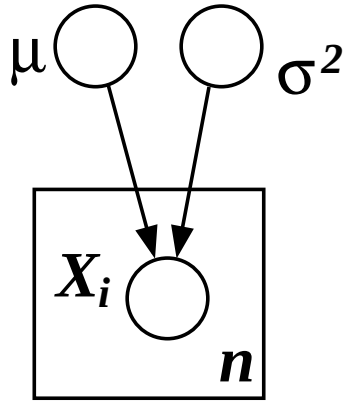
- Définition
- Inférence
- Apprentissage
  - Paramètres
  - Structure
- Applications et perspectives

Apprentissage des paramètres

- Nécessaire en pratique
- Hypothèse: structure fixe (graphe + paramétrisation des lois locales)
- Cadre fréquentiste: estimateur de maximum de vraisemblance bien adapté aux modèles graphiques
- Cadre Bayésien

Apprentissage des paramètres

- Cadre Bayésien: paramètres considérés comme variables aléatoires
  - Probabilités *a priori* $p(\theta)$, *a posteriori* $p(\theta|x)$
  - Formule de Bayes: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$
  - Apprentissage "=" inférence
  - "Vrais Bayésiens" : jamais d'estimateurs ponctuels
  - "Faux Bayésiens" : maximum a posteriori (MAP)



- Deux cadres asymptotiquement équivalents

“Réseaux Bayésiens Bayésiens” et “réseaux Bayésiens non Bayésiens” ?

Données complètes
Maximum de vraisemblance

- Modèles orientés: $p(x|\theta) = \prod_{i=1}^{n} p(x_i|x_{\mathsf{parents}(i)}, \theta_i)$
  - Découplage du maximum de vraisemblance ($\forall$ treewidth !)
  - Estimations locales
  - Exemples: modèles discrets / Gaussiens

Données complètes
Maximum de vraisemblance

- Modèles orientés: $p(x|\theta) = \prod_{i=1}^{n} p(x_i|x_{\text{parents}(i)}, \theta_i)$
  - Découplage du maximum de vraisemblance ($\forall$ treewidth !)
  - Estimations locales
  - Exemples: modèles discrets / Gaussiens

- Modèles non orientés: $p(x|\theta) = \frac{1}{Z(\theta_1, \ldots, \theta_p)} \prod_{j=1}^{p} \phi_{C_j}(x_{C_j}, \theta_j)$
  - Pas de découplage du maximum de vraisemblance
  - Algorithmes itératifs (IPF, IS). Cf. Della Pietra et. al (1997), Jirousek, 1995.

Données incomplètes
Algorithme Expectation-Maximisation (EM)

- Modèle $p(x, z|\theta)$
  z non observée (données manquantes ou modèles de mélange)

- log vraisemblance: $\log p(x|\theta) = \log \sum_z p(x, z|\theta)$

- Algorithme itératif
  - E-step: calculer $J(\theta) = E_{p(z|x,\theta)} \log p(x, z|\theta)$
  - M-step: maximiser $J(\theta)$ par rapport à θ

- Propriétés de convergence
  - Croissance de la vraisemblance à chaque itération
  - Minimas locaux

Apprentissage de la structure

- Deux visions des modèles graphiques $\longrightarrow$ deux types de méthodes
  - Apprentissage de la structure par tests d'indépendances conditionnelles (IC/PC: Pearl, 2000, Spirtes & al, 1993)
    - Problème de complexité
    - Problème de cohérence/fiabilité

Apprentissage de la structure

- Deux visions des modèles graphiques $\longrightarrow$ deux types de méthodes
  - Apprentissage de la structure par tests d'indépendances conditionnelles (IC/PC: Pearl, 2000, Spirtes & al, 1993)
    - Problème de complexité
    - Problème de cohérence/fiabilité
  - Apprentissage de la structure par des méthodes statistiques de sélection de modèles
    - Utilisation de la loi factorisée
    - Calcul d'un score pour chaque structure
    - Fouille dans l'espace des graphes **orientés**

Apprentissage de la structure
Scores pour données complètes - Modèle orienté

- Cadre Bayésien
  - Calcul de la probabilité marginale du modèle
  - Formule analytique dans certains cas (e.g., BDe pour discrets/Gaussiens, Heckerman & al, 1995)

- Cadre fréquentiste

Apprentissage de la structure
Scores pour données complètes - Modèle orienté

- Cadre fréquentiste
  - Le maximum de vraisemblance ne permet pas de faire de la sélection de modèles

Apprentissage de la structure
Scores pour données complètes - Modèle orienté

- Cadre fréquentiste
  - Le maximum de vraisemblance ne permet pas de faire de la sélection de modèles
  - Pénaliser les modèles plus complexes: AIC - BIC/MDL

$$J(G) = \sum_{i=1}^{n} I(x_i, x_{\pi_i(G)}) + \text{cste} \times \sum_{i=1}^{n} \#(i, \pi_i(G))$$

I information mutuelle calculée à l'aide des distributions empiriques (multinomiales, Gaussiennes)

Apprentissage de la structure
Scores pour données complètes - Modèle orienté

$$J(G) = \sum_{i=1}^{n} I(x_i, x_{\pi_i(G)}) + \text{cste} \times \sum_{i=1}^{n} \#(i, \pi_i(G))$$

- Optimisation du score
  - Arbres
    - Problème d'arbre couvrant de poids maximal
    - Algorithme de Chow-Liu (1968!)
  - Cas général: optimization gloutonne
  - Problèmes de Markov-équivalence

Apprentissage de la structure
Données incomplètes

- Structural EM (Friedman, 1988)
  - Alterner entre apprentissage des paramêtres et apprentissage de la structure
- Problèmes de minimas locaux

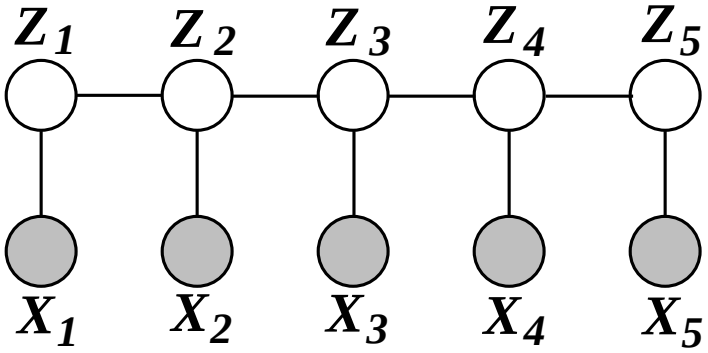
Modèles graphiques probabilistes
Plan de la présentation

- Définition
- Inférence
- Apprentissage
  - Paramètres
  - Structure
- Applications et perspectives

Perspectives

- Applications
  - Bio, image, texte, son, etc...
- Thèmes récents/intéressants
  - Inférence, inférence, inférence ...
  - Modèles graphiques discriminants
  - Modèles causaux
  - Méthodes non-paramétriques
  - Filtres particulaires

Modèles graphiques discriminants
Conditional random fields (CRF)



- Situation courante:
  - x toujours observé, but=prédire z
  - Apprentissage avec données complètes: $\max_\theta p(x, z|\theta)$
  - Utilisation du modèle: $\arg\max_z p(x, z|\theta) = \arg\max_z p(z|x, \theta)$

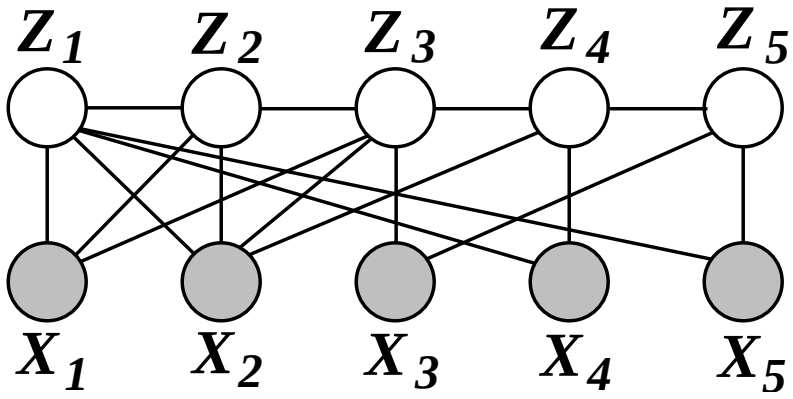
- Modèles graphiques discriminants:
  1. Apprentissage **discriminant**
  2. Interactions **longue portée** possibles

Modèles graphiques discriminants

- Apprentissage **discriminant**
  - $\max_\theta p(z|x,\theta)$ au lieu de $\max_\theta p(z,x|\theta)$
  - cf. régression logistique vs. analyse linéaire discriminante
  - Complexité numérique supérieure
  - Modèle "correct/incorrect"

Modèles graphiques discriminants

- Interactions **longue portée**



- Biologie, traitement du texte (Lafferty & al, 2001)

Causalité

- La corrélation n'implique pas la causalité
- En général, le sens d'une arête dans un modèle orienté ne correspond pas à une relation de causalité

Causalité

- La corrélation n'implique pas la causalité
- En général, le sens d'une arête dans un modèle orienté ne correspond pas à une relation de causalité
- Sans hypothèses supplémentaires, des relations de causalité ne peuvent pas être inférée sans interventions

Causalité

- La corrélation n'implique pas la causalité
- En général, le sens d'une arête dans un modèle orienté ne correspond pas à une relation de causalité
- Sans hypothèses supplémentaires, des relations de causalité ne peuvent pas être inférée sans interventions
- Ceci dit...
  - Modèles graphiques causaux (Pearl, 2000, Spirtes & al, 1993)

Modèles causaux

- Modèles orientés
- Chaque $p(x_i|x_{\pi_i})$ représente un processus stochastique autonome
  - Permet les interventions
- Inférence causale
- Apprentissage de relations de causalité
  - Rappel: La corrélation n'implique pas la causalité
  - Hypothèse de sufficience causale
  - Apprentissage sans expérimentation
  - Apprentissage avec expérimentation

Méthodes non-paramétriques

- Hypothèses paramétriques classiques
  - Données binaires: loi de Bernoulli
  - Données catégoriques: loi multinomiale
  - Données continues: loi Gaussienne

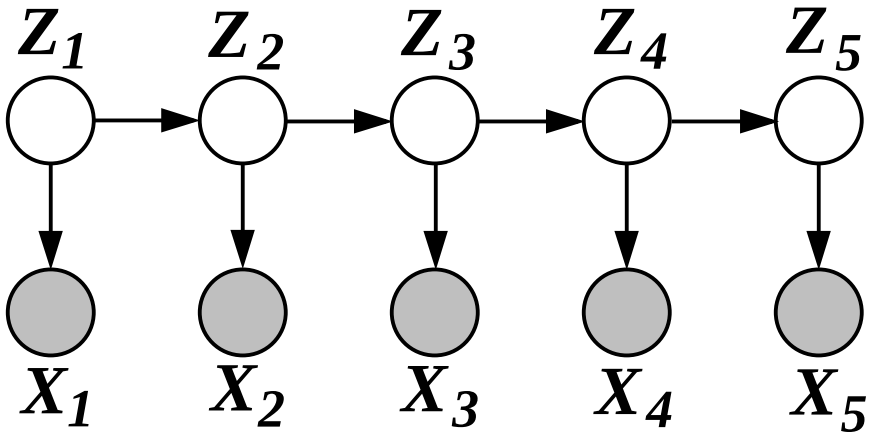
Méthodes non-paramétriques

- Hypothèses paramétriques classiques
  - Données binaires: loi de Bernoulli
  - Données catégoriques: loi multinomiale
  - Données continues: loi Gaussienne
  - Familles exponentielles (Poisson, Gamma, Beta) pour $p(x)$
  - Modèles linéaires généralisés (GLIM) pour $p(y|x)$

Méthodes non-paramétriques

- Hypothèses paramétriques classiques
  - Données binaires: loi de Bernoulli
  - Données catégoriques: loi multinomiale
  - Données continues: loi Gaussienne
  - Familles exponentielles (Poisson, Gamma, Beta) pour $p(x)$
  - Modèles linéaires généralisés (GLIM) pour $p(y|x)$
- Non-paramétrique: permet de s'affranchir d'hypothèses restrictives sur les lois locales
  - par obligation, ignorance, ou paresse...
  - sans complexité numérique additionnelle majeure
- Cf. tutoriel de Michael Jordan (UC Berkeley), NIPS 2005

Filtres particulaires pour modèles graphiques dynamiques



- Données, dynamique, et/ou observations complexes (i.e., non Gaussiennes)
- Faible largeur arborescente, mais propagation délicate
- Méthodes particulaires: mise à jour efficace d'un ensemble de particules qui suivent la loi $p(z_1, \ldots, z_t | x_1, \ldots, x_t)$ pour $t = 1, \ldots$.
- cf. Murphy (2002), Doucet & al. (2001)

Conclusion

- Modèle graphique probabiliste = outil flexible de modélisation
- Echanges théorie/applications

Références

- M. Tipping and C. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society B, 61, 611-622, 1999
- Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. Mach. Learn., 29(2-3):245–273, 1997.
- S. L. Lauritzen and D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, J. Royal Stat. Society, B 50 (1988), no. 2, 157–223.
- S. L. Lauritzen. Graphical Models. Clarendon Press, Oxford, 1996.
- F. V. Jensen. "Bayesian Networks and Decision Graphs". Springer. 2001.
- J. Pearl. "Causality". Cambridge. 2000.
- R. G. Cowell, A. P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter. "Probabilistic Networks and Expert Systems". Springer-Verlag. 1999.
- M. I. Jordan (ed). "Learning in Graphical Models". MIT Press. 1998.
- S. Roweis & Z. Ghahramani, 1999. A Unifying Review of Linear Gaussian Models, Neural Computation 11(2) (1999) pp.305-345
- R. McEliece and S. M. Aji, 2000. The Generalized Distributive Law, IEEE Trans. Inform. Theory, vol. 46, no. 2 (March 2000), pp. 325–343.
- Kschischang, B. Frey and H. Loeliger, 2001. Factor graphs and the sum product algorithm, IEEE Transactions on Information Theory, February, 2001.

- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, 1997. "An introduction to variational methods for graphical models."
- D. MacKay, 1998. "An introduction to Monte Carlo methods".
- Spirtes, P., Glymour, C. and Scheines, R. (1993) Causation, Prediction, and Search. Springer-Verlag, NY.
- D. Heckerman, D. Geiger, D. M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20 (1995) 197–243,
- C. Chow and C. Liu. 1968. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 14(3):462–467.
- N. Friedman. The Bayesian structural EM algorithm. In Proc. UAI, 1998.
- J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. ICML, 2001.
- Kevin Murphy. Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, UC Berkeley, Computer Science Division, July 2002.
- Doucet, A., de Freitas, J.F.G. and Gordon, N.J. (2001). Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag.
- S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans. PAMI, 6:721–741, 1984.
- S. Della Pietra, V. J. Della Pietra, and J. D. Lafferty. Inducing features of random fields. IEEE Trans. PAMI, 19(4):380–393, 1997.
- Jirousek, R. & Preucil, S. (1995). On the effective implementation of the iterative proportional fitting procedure, Computational Statistics & Data Analysis. 19: 177-189.