Inférence de dépendance conditionnelles et réseaux de gènes

Christophe Giraud

CMAP, Ecole Polytechnique

Palaiseau, 26 août 2009

Sujets de recherche en statistique

- **Sélection de modèle:** parmi plusieurs modèles probabilistes en concurrence, le(s)quel(s) reflète(nt) au mieux les données? (complexité des modèles / collection de modèles / nb d'observations / risque)

- **Modèles graphiques:** représentation des relations de dépendances conditionnelles entres différentes variables aléatoires.

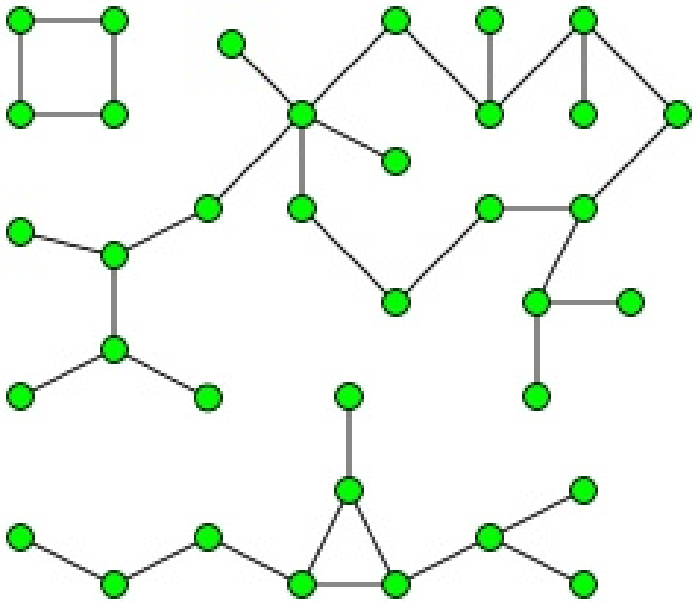
Content

Modèles graphiques

Modèles graphiques

Modèles non orientés

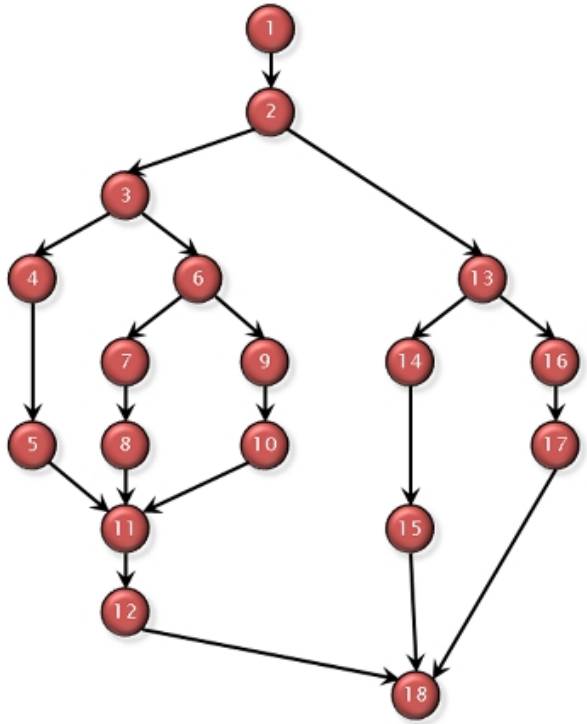
Markov random Fields / Gibbs Fields



X_i indépendant de $\{X_j : j \not\sim i\}$ sachant $\{X_j : j \sim i\}$

Modèles orientés

DAG models / Bayesian networks



X_i indpt de $\{X_j : i \not\to \ldots \not\to j\}$ sachant $\{X_j : j \to i\}$

Attention à l'interprétation des modèles orientés!
Non unicité du graphe minimal $\Longrightarrow$ prudence

Ex: $X_{i+1} = \alpha X_i + \varepsilon_i$ avec ε_i indépendant de $X_1, \ldots, X_{i-1}$.
Modèle graphique minimal:

$$1 \rightarrow 2 \rightarrow \ldots \rightarrow p$$

mais aussi

$$1 \leftarrow 2 \leftarrow \ldots \leftarrow p$$

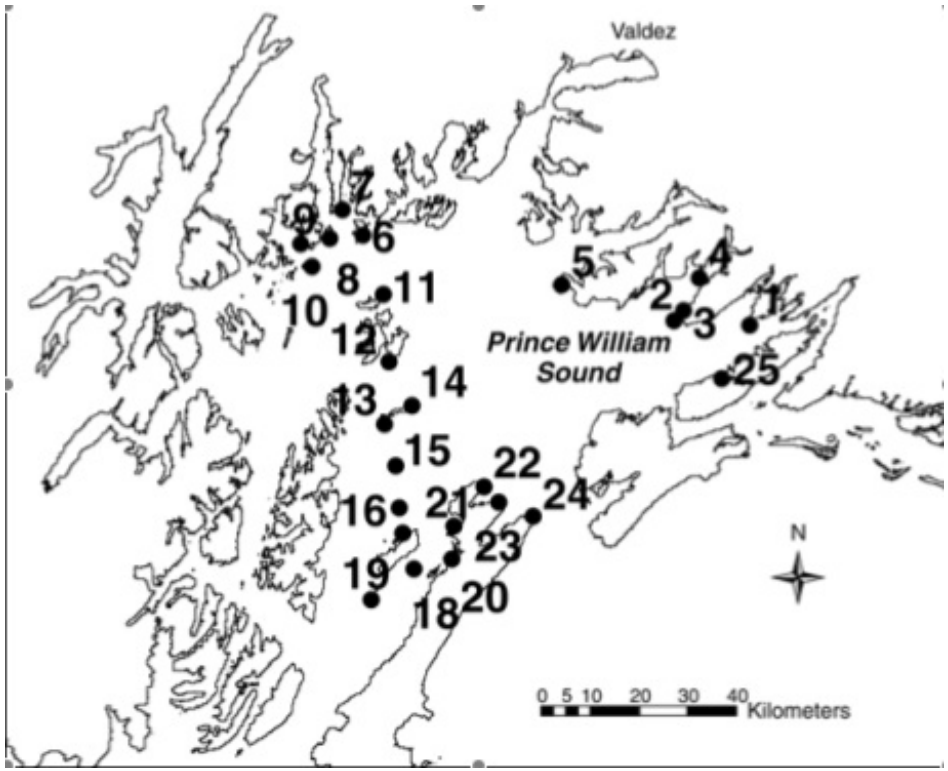
Modèle non orienté: unicité du graphe minimal qui représente les dépendances conditionnelles.

Applications

- Modélisation des réseaux de gènes / métaboliques / etc
- Ecologie: modèles hiérarchiques pour analyser les données (ad hoc ou basés sur des modèles)
- Génétique: modélisation de réseaux phylogéniques ou généalogiques prenant en compte recombinaisons et transferts horizontaux...
- Signal: computer vision / reconnaissance vocale
- physique statistique...

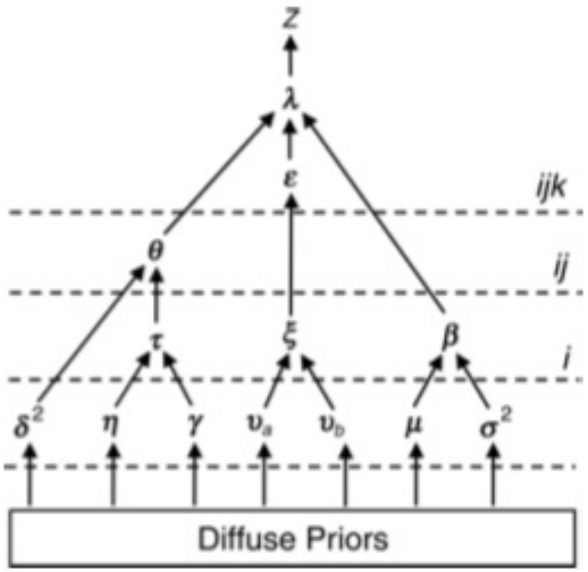
Modèles hiérarchiques: phoques en baie Prince William

D'après J. Ver Hoef et K. Frost *A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska* (Environmental and Ecological Statistics 2003)



Relevés sur 25 sites pendant 10 ans après Exxon-Valdez

modèle purement ad-hoc



comptage:

$$\text{loi}(Z_{i,t,k}|\lambda_{i,t,k}) = \text{Poisson}(\lambda_{i,t,k})$$

$$\log(\lambda_{i,t,k}) = \theta_{i,t} + f(\beta_i, X_{i,t,k}) + \varepsilon_{i,t,k}$$

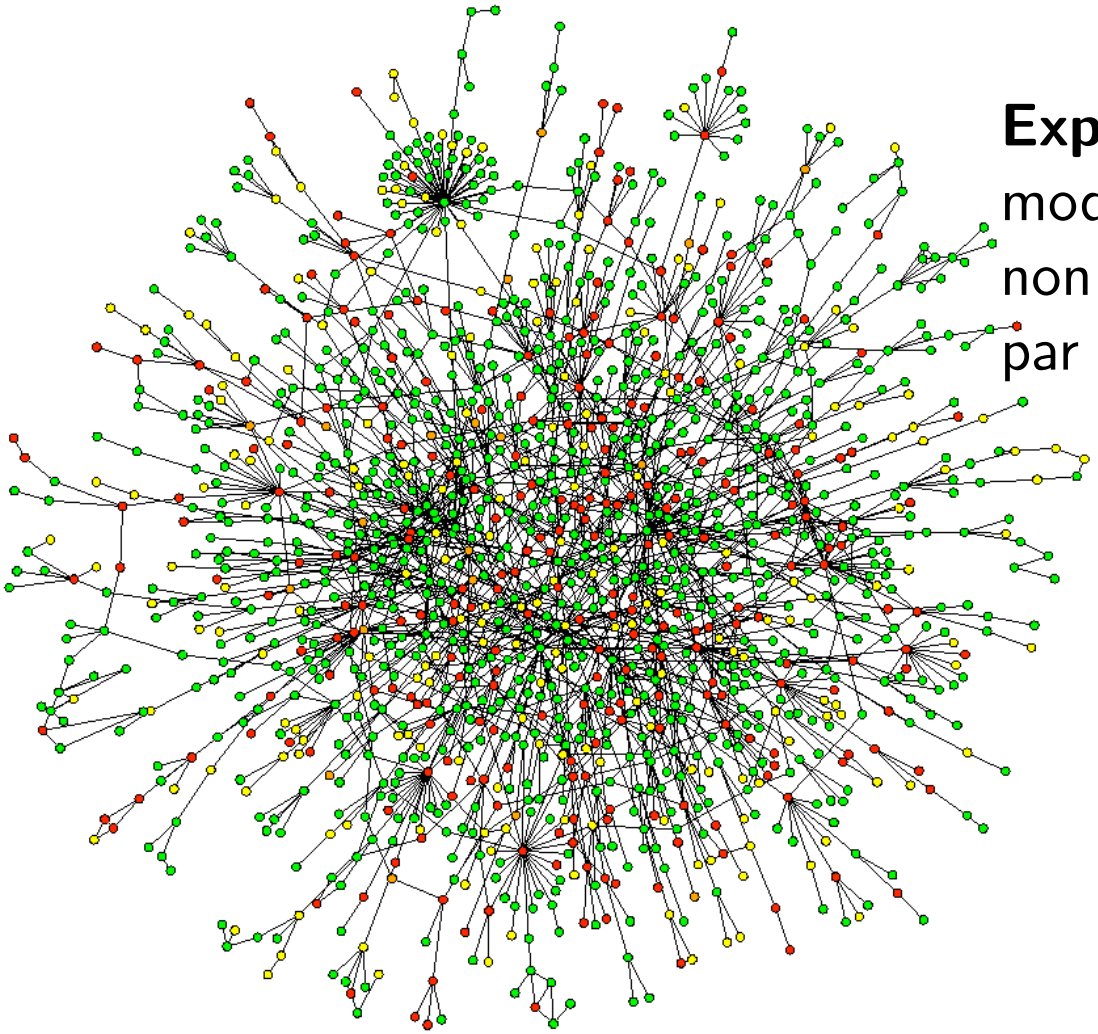
$$\text{loi}(\theta_{i,t}|\tau, \delta) = \text{Gauss}(\tau_{0,i} + t \times \tau_{1,i}, \delta^2)$$

etc

tendance:

$$\alpha = \frac{\sum_i e^{-\tau_{0,i}} \tau_{1,i}}{\sum_i e^{-\tau_{0,i}}}$$

Réseaux de gènes



Expression des gènes: modèle graphique Gaussien non orienté de graphe donné par le réseau

Quelques problèmes statistiques

- **Inférence avec observations partielles:**
  - graphe connu / observations partielles (noeuds cachés)
    Ex: phoques, analyse du génome par HMM.
  - estimer les paramètres / la valeur des noeuds cachés sachant la valeur des noeuds observés.
  - méthodes: EM, MCMC
  - challenge: grande dimension

- **Apprentissage du graphe:**
  - graphe inconnu / observations complètes (en général)
    Ex: réseaux de gènes.
  - estimer la structure et les lois conditionnelles.
  - méthode: sélection par minimisation (parfois stochastique) de critères pénalisés.
  - challenges:
    - critères optimaux
    - cas où nb de répétitions $<<$ nb de noeuds.

Inférence de réseaux de gènes

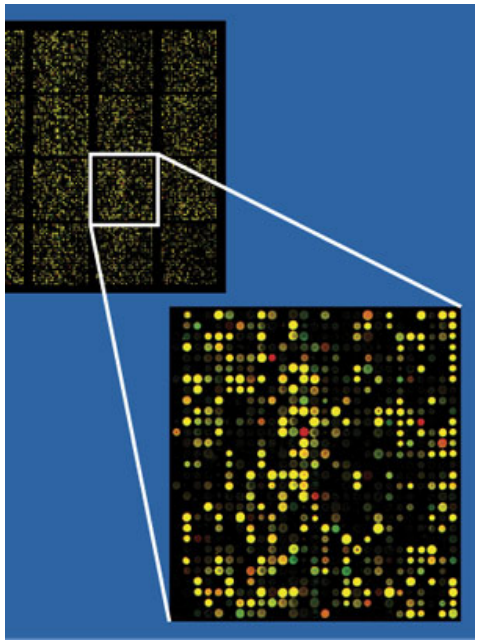
Démarche

Pour un organisme donné, inférer sur une base statistique un réseau de gène à partir de données transcriptomiques.

Démarche exploratoire:

- proposer des pistes d'investigations concernant les interactions entre gènes,
- suggérer des fonctions possibles pour des gènes orphelins.

Données et modélisation



- **Données** données transcriptomiques obtenues par microarray
- **Modélisation** les niveaux d'expression sont modelisés à l'aide d'un modèle graphique Gaussien de graphe **g** inconnu.

objectif statistique: estimer à partir des données le graphe **g** du GGM.

Difficultés

Difficulté principale: $n \ll p$

- $p \approx$ quelques 100 à quelques 1000 genes
- $n \approx$ quelques 10

Nouveaux algorithmes: par seuillage ou régularisation

Multiple testing	**Convex minimization**
- Drton & Perlman (2004)	- Meinshausen & Bühlmann (2006)
- Schäfer & Strimmer (2005)	- Huang *et al.* (2006)
- Wille & Bühlmann (2006)	- Yuan & Lin (2007)
- Verzelen & Villers (2007)	- Banerjee *et al.* (2007)
- Bühlmann & Kalisch (2008)	- Friedman *et al.* (2007)
...	...

GGMselect

collaboration avec S. Huet (INRA Jouy-en-Josas)
et N. Verzelen (Univ. d'Orsay)

R package:

Input: données + information a priori (si disponible)

Output: graphe estimé

Cahier des charges:

- procédure justifiée théoriquement dans un cadre non-asymptotique
- procédure performante en pratique

Principe de la procédure:

- Construction de diverses familles de graphes à partir des données (+ éventuellement information a priori)
- sélection d'un graphe parmi les familles construites par minimisation d'un critère pénalisé.