Graphical models
Magistère, Rennes 2018

Valérie Monbet

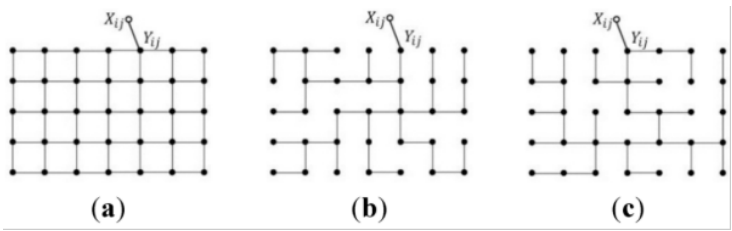
IRMAR, Université de Rennes 1

Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

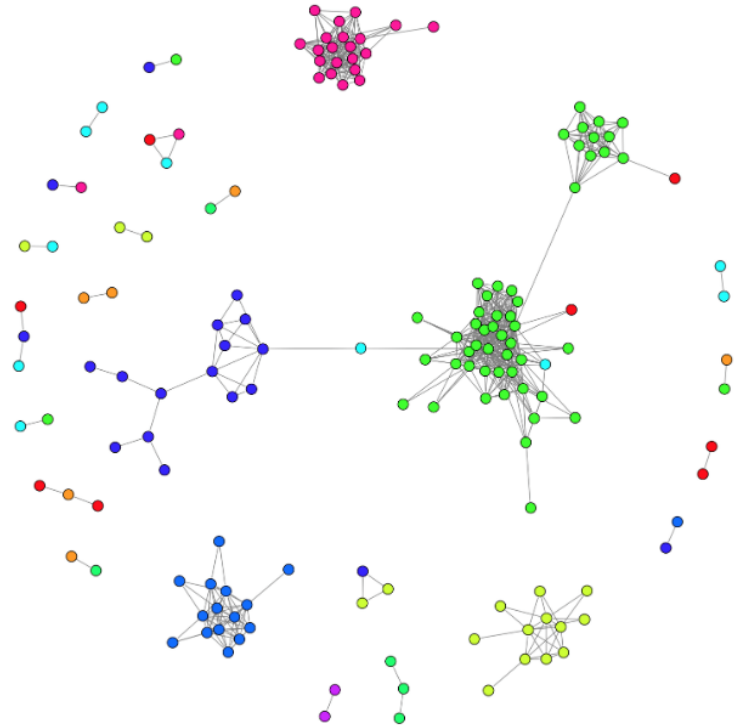
Introduction

- Graphical models became recently common tools in statistic and learning to describe dependence structures in (high dimensional) complex data such as images, social network data, gene expression, etc.
- Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's.
- Data can be structured (time series, images) or not (social network, gene network, ...)



Structured data

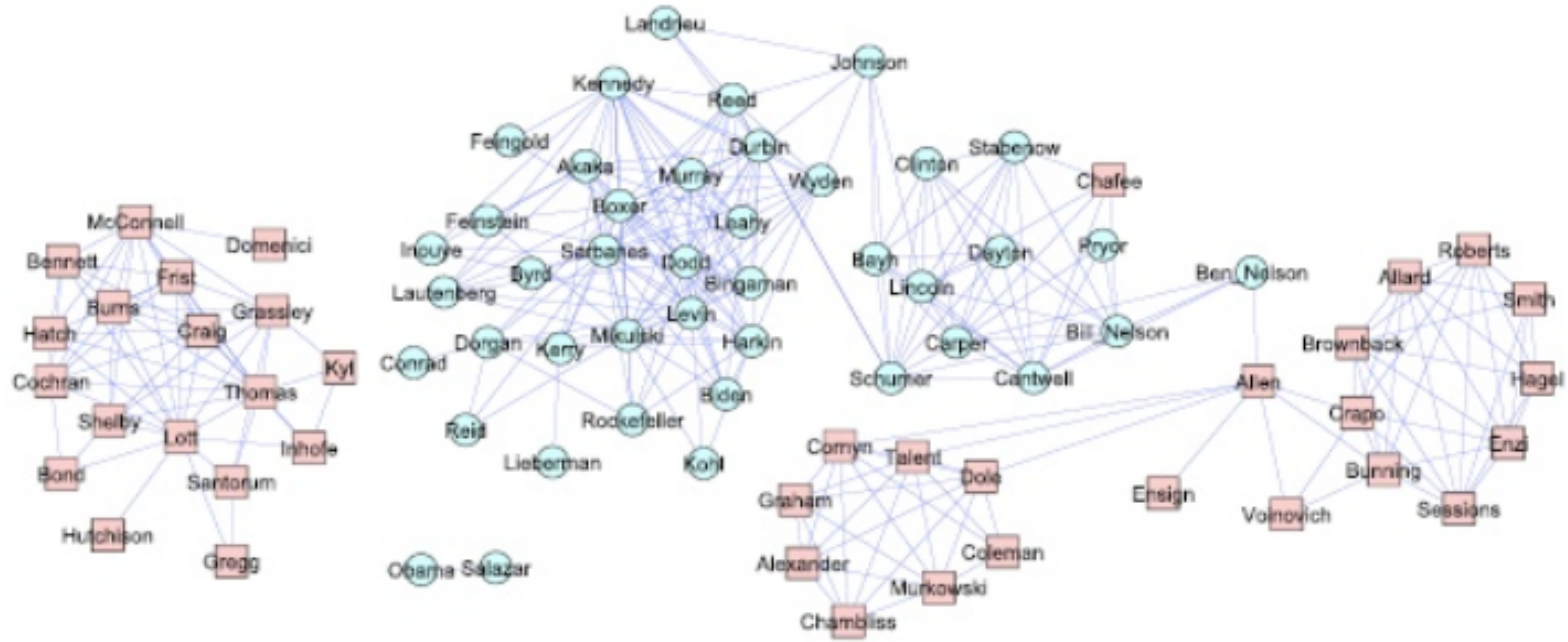
Gene network



Stock data

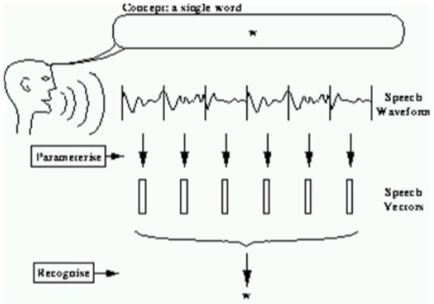
Introduction

- Graph representing votes for the American senators (2004-2006). 100 variables corresponding to 100 senators democrates (bullet) and republicans (square). 542 observations.
- In the graph, most democrats have demecrates as neighbors.
- Exception : Chafee. It is coherent with the mediatic positions of Chasbee at that time.

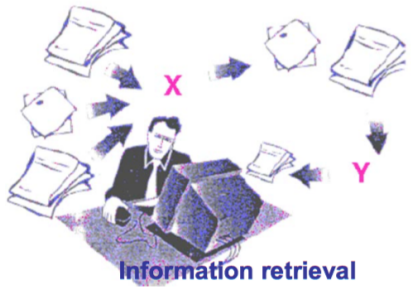


Introduction

Reasoning under uncertainty!

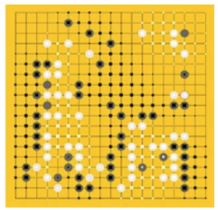


Speech recognition

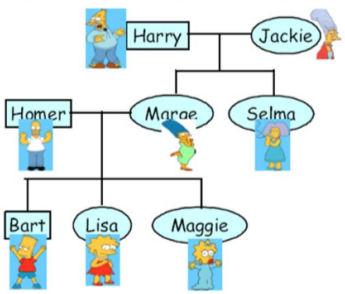


Information retrieval

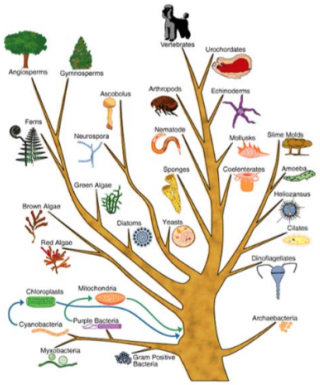
Computer vision



Games

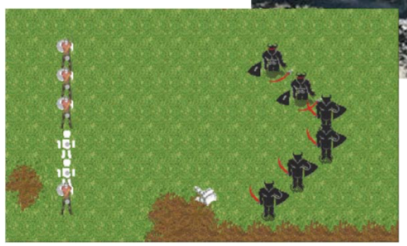


Pedigree



Evolution

Robotic control



Planning

Introduction

- Graphical models are used for
  - better understanding conditional dependencies,
  - segmentation (image),
  - pattern recognition,
  - signal denoising,
  - $\cdots$.
- Data associated with graphical models

$$X_1^{(i)}, \cdots, X_p^{(i)}, \ i = 1, \cdots, n$$

  (one variable per node of the graph).
- Graphical models are based on probabilistic models such as the multivariate Gaussian distribution or multivariate multinomial distribution.
- A graphical model can be interpreted as a multivariate distribution under constraints given by a network structure.
- Applications of GM : computer vision, natural language processing, decision making under uncertainty, computational biology, genetics and medical diagnosis/prognosis, etc.

Basic problems in graphical models

- Representation : what is the joint probability distribution on multiple variables ?

$$P(X_1, \cdots, X_p)$$

  - How many state configuration in total ? — ex for multinomial data 2^p
  - Are they all needed to be represented ?
  - Do we have any scientific/medical insight ?

- Learning : estimation of the parameters of the model
  - Maximal-likelihood estimation ? With how many data ?
  - Are there other estimation principles ?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities ?

- Inference : If not all variables are observable, how to compute the conditional distribution of latent variables given observations ?

Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

Representing multivariate distribution

- Representation : what is the joint probability distribution on multiple variables ?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

  - How many state configuration in total ? — 2^8
  - Are they need to be represented ?
  - Do they get any scientific/medical insight ?

- Factored representation : the chain-rule

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)\cdots$$
$$P(X_7|X_1, X_2, X_3, X_4, X_5, X_6)P(X_8|X_1, X_2, X_3, X_4, X_5, X_6, X_7)$$

  - This factorization is true for any distribution and any variable ordering
  - Do we save any parameterization cost ?

- If X_i's are independent : $P(X_i|.) = P(X_i)$ and

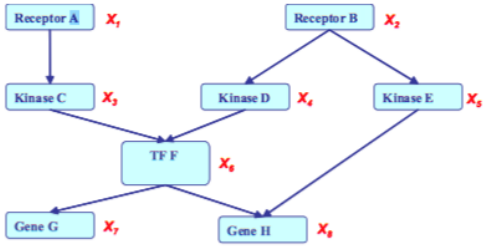
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \prod_i P(X_i)$$

  - What we gain ?
  - What we lose ?

Two types of graphical models

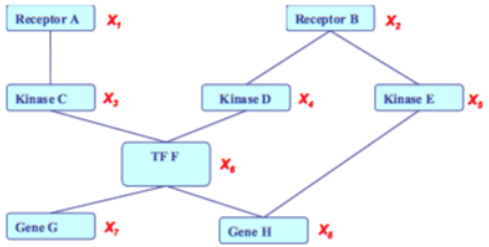
- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model) :

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2) P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6)$$



- Undirected edges simply give (partial) correlations between variables (Markov Random Field or Undirected Graphical model) :

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \frac{1}{Z} \exp\{E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2) + E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\}$$



where Z is a normalization constante and E represent *"energy" functions.*

Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
   - I. Quantitative Specification
   - II. Independence properties
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

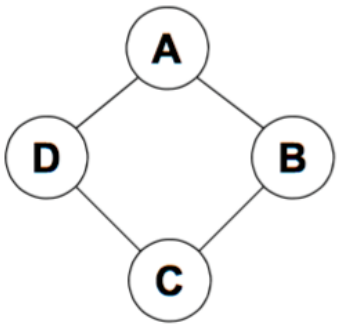
Canonical example

- Let us consider a family of distributions such that

$$A \perp C | \{B, D\} \text{ and } B \perp D | \{A, C\}$$

  There is no causality relationship between the variables.

- It can be represented by a simple graph. The graph helps to interpret the model.



MRF

- Person **A** *and person* **B** *tend to agree about things*
- Person **A** *and person* **D** *tend to disagree about things*
- Person **C** *tends to believe things are true, is more likely to be swayed by person* **D** *than by person* **B**, *and has no direct contact with person* **A**
- Person **B** *and person* **D** *have no direct contact*

- Factorization of the distribution

$$P(A, B, C, D) = P(A|B, D)P(C|B, D)P(B, D) = P(B|A, C)P(D|A, C)P(A, C)$$

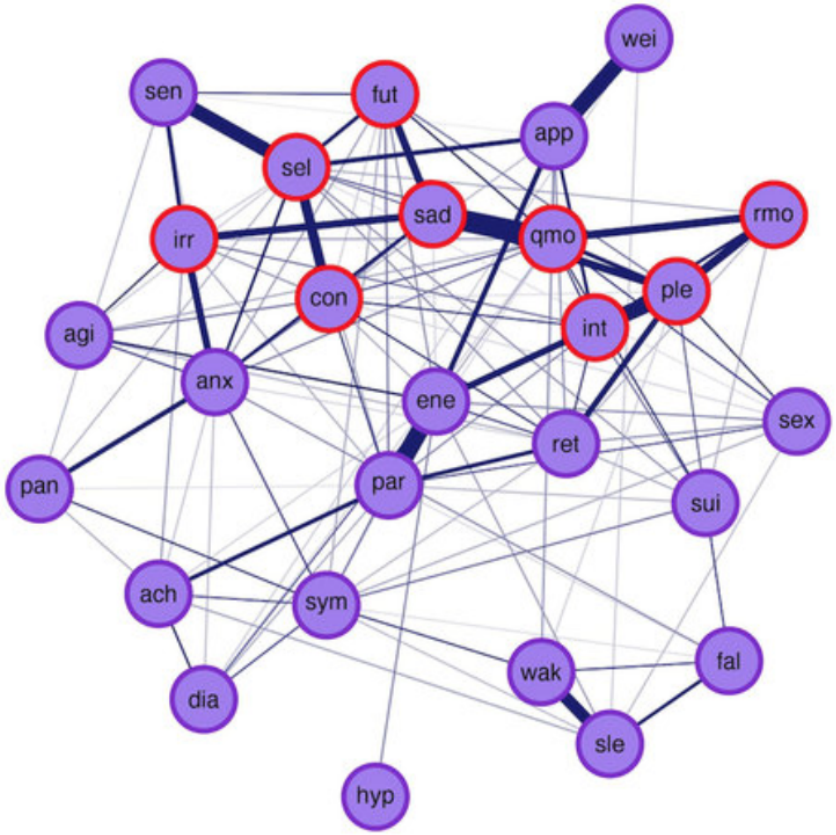
Canonical example - Exercice

- Let us consider a family of distributions such that

$$A \perp B|\{B, D, E\} \text{ and } E \perp A, C, D|\{B\} \text{ and } D \perp C|\{A, B, E\}$$

  There is no causality relationship between the variables.
- Plot the associated graph.
- Write the factorization of the distribution

Example 1



ach: aches and pains
agi: psychomotor agitation
anx: feeling anxious
app: change of appetite
con: concentration problems
dia: diarrhea/constipation
ene: energy level
fal: falling asleep
fut: view of myself
hyp: hypersomnia
int: general interest
irr: feeling irritable
pan: panic/phobic symptoms
par: leaden paralysis
ple: capacity for pleasure (not sex)
qmo: quality of mood
ret: psychomotor retardation
rmo: respons of mood
sad: feeling sad
sel: view of oneself
sen: interpersonal sensitivity
sex: interest in sex
sle: sleep during the night
sui: suicidal thoughts
sym: other bodily symptoms
wak: waking up too early
wei: change of weight

The resulting network structure of a group of healthy controls and people with a current or history of depressive disorder ($N = 1108$). Cognitive symptoms are displayed as ○ and thicker edges (connections) represent stronger associations.

Representation

- An undirected graphical model represents a distribution $P(X_1, ..., X_d)$ defined by an undirected graph H, and a set of positive potential functions ψ_c associated with the cliques of H, s.t.

$$p(x_1, \cdots, x_d) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x_c})$$

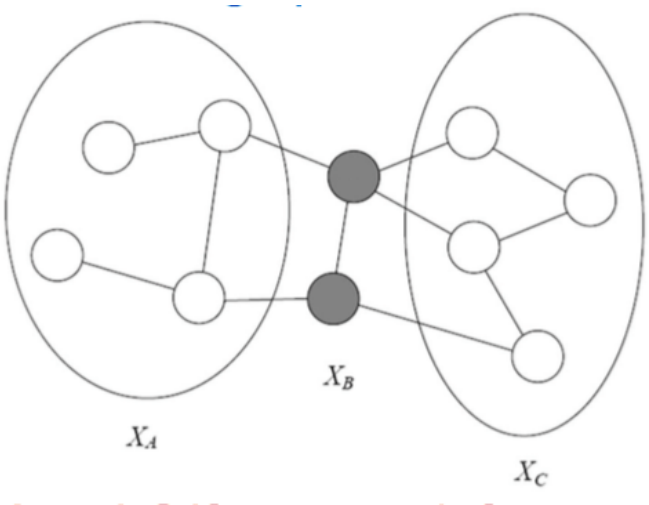
where Z is known as the partition function. It is a normalization constant.

$$Z = \sum_{x_1, \cdots, x_d} \prod_{c \in C} \psi_c(\mathbf{x_c})$$

- Also known as Markov Random Fields, Markov networks ...
- The potential function can be understood as a contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.

Global Markov properties

- Let H be an undirected graph



- B separates A and C if every path from a node in A into a node in C passes through a node in B :

$$sep_H(A; C|B).$$

- A probability distribution satisfies the global Markov property if for any disjoint A, B, C, such that B separates A and C, A is independent of C given B :

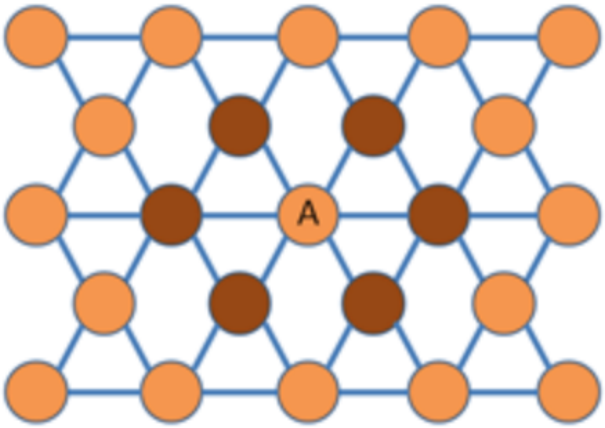
$$I(H) = \{A \perp C|B : sep_h(A; C|B)\}.$$

Local Markov independencies

- For each node $X_j \in V$, there is unique Markov blanket of X_j, denoted MB_{X_j}, which is the set of neighbors of X_j in the graph (those that share an edge with X_j)
- The local Markov independencies associated with H is :

$$I_j(H) : \left\{ X_j \perp V - \{X_j\} - MB_{X_j} | MB_{X_j} : \forall j \right\}$$

In other words, X_j is independent of the rest of the nodes in the graph given its immediate neighbors

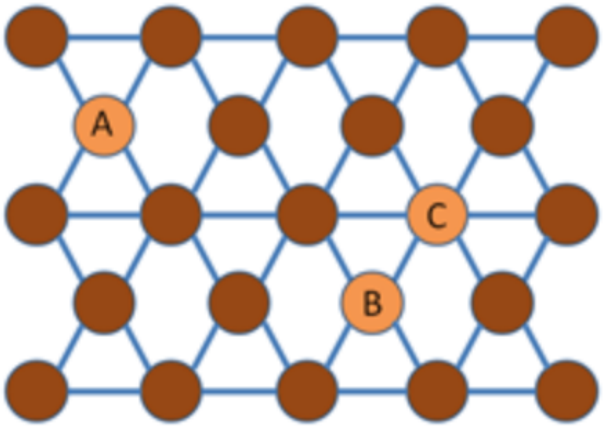


By the local property, A is conditionally independent from the rest of the network given its neighbors.

Pairwise properties

- Any two nodes in the Markov network are conditionally independent given the rest of the network if they are not neighbors.
- The pairwise property associated with H is :

$$X_j \perp X_i | X_i \notin MB_{X_j} : \forall j$$

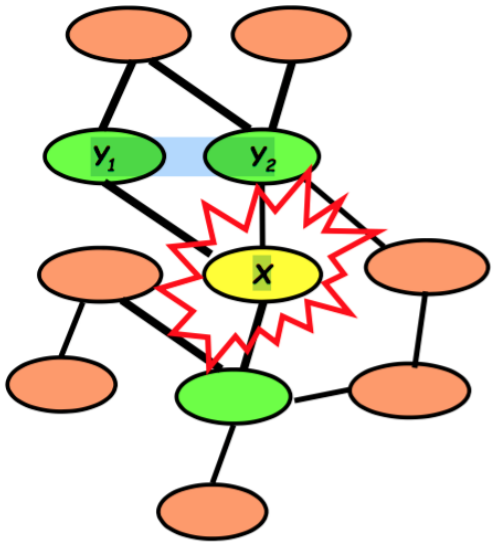


By the pairwise property, A is conditionally independent from B and C given the rest of the network, but B and C are still dependent on each other.

Summary : Conditional Independence Semantics in an MRF

Structure : an undirected graph

- Meaning : a node is conditionally independent of every other node in the network given its neighbors
- Local contingency functions (potentials) and the cliques in the graph completely determine the joint distribution.
- Give correlations between variables, but no explicit way to generate samples

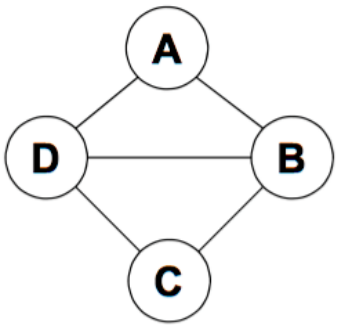


Outline

3 Undirected graph models
- I. Quantitative Specification
- II. Independence properties

Cliques

- A complete subgraph (clique) is a subgraph V' such that nodes in V' are fully interconnected.
- A (maximal) clique is a complete subgraph s.t. any superset is not complete.
- A sub-clique is a not-necessarily-maximal clique.
- Example
  - max-cliques : $\{A, B, D\}$, $\{B, C, D\}$
  - sub-cliques : $\{A, B\}$, $\{B, C\}$, $\cdots$ e.g. all edges and singletons



Gibbs Distribution and Clique Potential

- an undirected graphical model represents a distribution $P(X_1, ..., X_d)$ defined by an undirected graph H, and a set of positive potential functions ψ_c associated with the cliques of H, s.t.

$$p(x_1, \cdots, x_d) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x_c}) \text{ A Gibbs distribution}$$

where Z is known as the partition function. It is a normalization constant.

$$Z = \sum_{x_1, \cdots, x_d} \prod_{c \in C} \psi_c(\mathbf{x_c})$$

- Also known as Markov Random Fields, Markov networks ...
- The potential function can be understood as a contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.

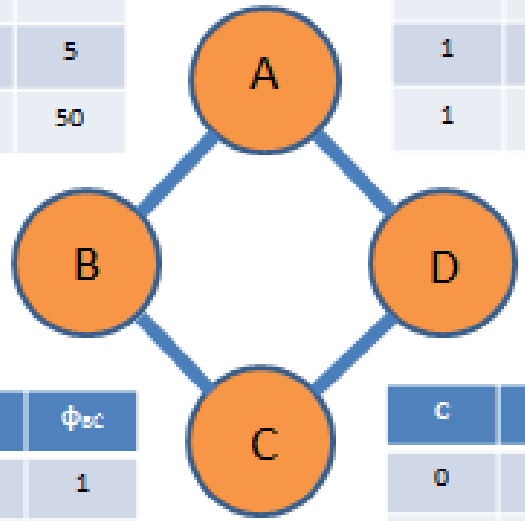
A first example

A	B	C	D	$\phi_{AB}\phi_{BC}\phi_{CD}\phi_{AD}$
0	0	0	0	250
0	0	0	1	37500
0	0	1	0	50000
0	0	1	1	625000
0	1	0	0	1125
0	1	0	1	168750
0	1	1	0	50000
0	1	1	1	625000
1	0	0	0	250
1	0	0	1	375
1	0	1	0	50000
1	0	1	1	6250
1	1	0	0	112500
1	1	0	1	168750
1	1	1	0	5000000
1	1	1	1	625000

$Z = 7520750$

A	B	ϕ_{AB}
0	0	50
0	1	5
1	0	5
1	1	50

A	D	ϕ_{AD}
0	0	5
0	1	50
1	0	50
1	1	5

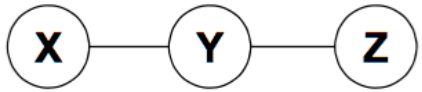


B	C	ϕ_{BC}
0	0	1
0	1	5
1	0	45
1	1	50

C	D	ϕ_{CD}
0	0	1
0	1	15
1	0	40
1	1	50

Source : `https://wiki.ubc.ca/Course:CPSC522/Markov_Networks`

Interpretation of Clique Potentials with simpler examples



- The model implies $X \perp Z|Y$. This independence statement implies (by definition) that the joint probability must factorize as :

$$p(x, y, z) = p(y)p(x|y)p(z|y)$$

- We can write it as

$$p(x, y, z) = p(x, y)p(z|y)$$

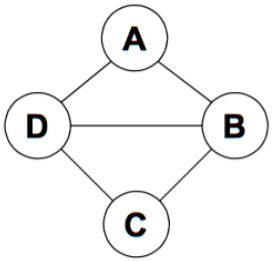
or

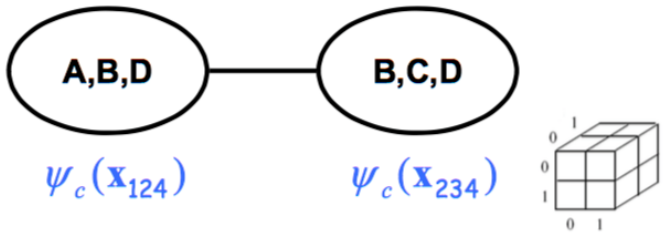
$$p(x, y, z) = p(x|y)p(z, y)$$

but
- cannot have all potentials be marginals
- cannot have all potentials be conditionals.

- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.
- Potential functions are more flexible than probabilities.

Example UGM – using max cliques



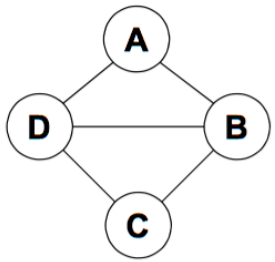


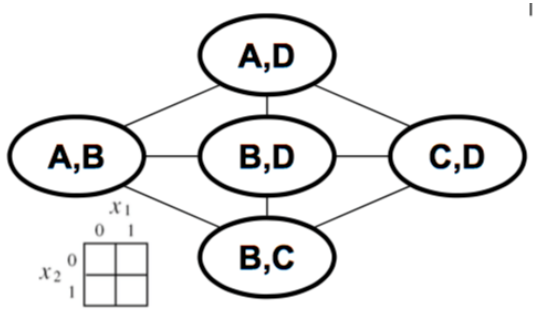
$$P'(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_c(\mathbf{x}_{124})\psi_c(\mathbf{x}_{234})$$

$$Z = \sum_{x_1,,x_2,x_3,x_4} \psi_c(\mathbf{x}_{124})\psi_c(\mathbf{x}_{234})$$

- For discrete nodes, we can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table.
- But computing Z is not easy.

Example UGM – using sub cliques





$$P''(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

$$Z = \sum_{x_1,,x_2,x_3,x_4} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

- We can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table.
- Pair MRF is a popular and simple special case.
- Are the two representations equivalent ? Are they of the same size ?

Outline

3 Undirected graph models
- I. Quantitative Specification
- II. Independence properties

Hammersley-Clifford Theorem

- If arbitrary potentials are utilized in the following product formula for probabilities,

$$P(x_1, \cdots, x_d) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{x_1, \cdots, x_d} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

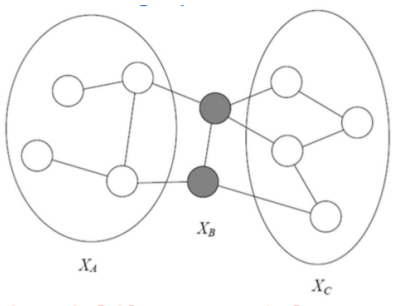
  then the family of probability distributions obtained is exactly that set which respects the qualitative specification (the conditional independence relations) described earlier.

- **Theorem** : Let P be a positive distribution over V, and H a Markov network graph over V. If H is an *Independence map* for P, then P is a Gibbs distribution over H.

Global and local Markov properties

- The global Markov properties of an undirected graph H are

$$I(H) = \{X_A \perp X_C | X_B : spe_H(X_A; X_C | X_B)\}$$



- The pairwise Markov independencies associated with an undirected graph $H = (V; E)$ are

$$I(H) = \{X \perp Y | V - \{X, Y\}; \{X, Y\}) \notin E\}$$

Example $X_1 \perp X_5 | \{X_2, X_3, X_4\}$

- A distribution has the local Markov property w.r.t. a graph $H = (V; E)$ if the conditional distribution of variable given its neighbours is independent of the remaining nodes

$$I(H) = \{X \perp V - (X \cup N_H(X)) | N_H(X); X \in V\}$$

Global and local Markov properties

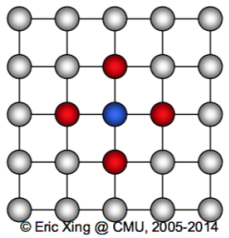
- **Theorem** (Hammersley-Clifford) : If the distribution is strictly positive and satisfies the local Markov property, then it factorizes with respect to the graph.

  A positive distribution P satisfies the conditional independence properties of an undirected graph H iff P can be represented as a product of factors, one per maximal clique, I ;e.

  $$P(y;\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(y_c;\theta_c)$$

  where C is the set of all maximal cliques and $Z(\theta)$ is the partition function.

- $N_H(X)$ is also called the Markov blanket of X.



Exponential form

- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential $\psi_c(\mathbf{x}_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(\mathbf{x}_c)$

$$\psi_c(\mathbf{x}_c) = \exp(-\phi_c(\mathbf{x}_c))$$

- This gives the joint probability a nice additive structure

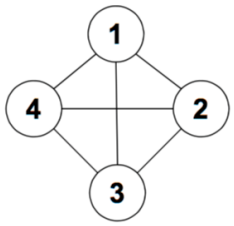
$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{ -\sum_{c \in C} \phi_c(\mathbf{x}_c) \right\}$$

  where the sum in the exponent is called the "free energy". The free energy can be positive or negative.
- In physics, this is called the "Boltzmann distribution".
- In computer science, this is called the "Gibbs distribution".
- In statistics, this is called a log-linear model.

Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

Boltzman machines



- A fully connected graph with pairwise (edge) potentials on binary-valued nodes ($x_i \in \{-1, +1\}$ or $x_i \in \{0, 1\}$) is called a Boltzmann machine

$$\begin{aligned} P(x_1, \cdots, x_4) &= \frac{1}{Z} \exp\left\{ \sum_{ij} \phi_{ij}(x_i, x_j) \right\} \\ &= \frac{1}{Z} \exp\left\{ \sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \right\}. \end{aligned}$$

- Hence the overall energy function has the form

$$H(x) = \sum_{ij} (x_j - \mu) \Theta_{ij} (x_j - \mu) = (\mathbf{x} - \mu)^T \Theta (\mathbf{x} - \mu)$$

Ising models

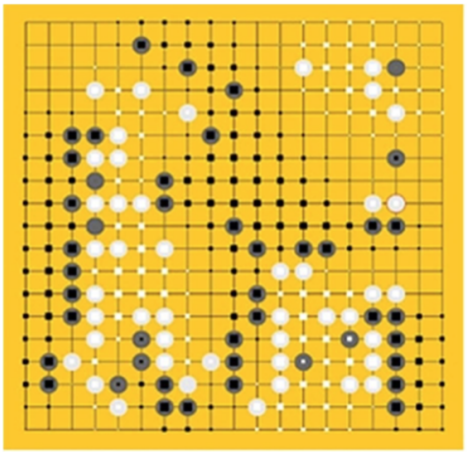
- Variables are binary.
- Nodes are arranged in a reguler topology (regular grid) and they are connected only to their geometric neighbors.

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{ \sum_{i,j \in N_i^j} \theta_{ij} x_i x_j + \sum_i \theta_i x_i \right\}$$

- Same as sparse Boltzmann machine where $\theta_{ij} \neq 0$ iff i and j are neighbors.
  - Ising's model described dependencies of order 2, like a multivariate Gaussian model.
  - e.g. nodes are pixels, potential function encourages nearby pixels to have the same intensities.
  - This model implies

$$P(X_j | X_{-j} = x_{-j}) = \frac{\exp\left(\theta_{j0} + \sum_{j,k \in E} \theta_{jk} x_k\right)}{1 + \exp\left(\theta_{j0} + \sum_{j,k \in E} \theta_{jk} x_k\right)}$$

    where θ_{jk} measures the dependence between x_j and x_k conditionally to the other variables.
  - Applications : Modeling Go, opinions in a social network



This is the middle position of a Go game. Overlaid is the estimate for the probability of becoming black or white for every intersection. Large squares mean the probability is higher.

- Potts model is a multi-state Ising model (categorical variables).

Ising models

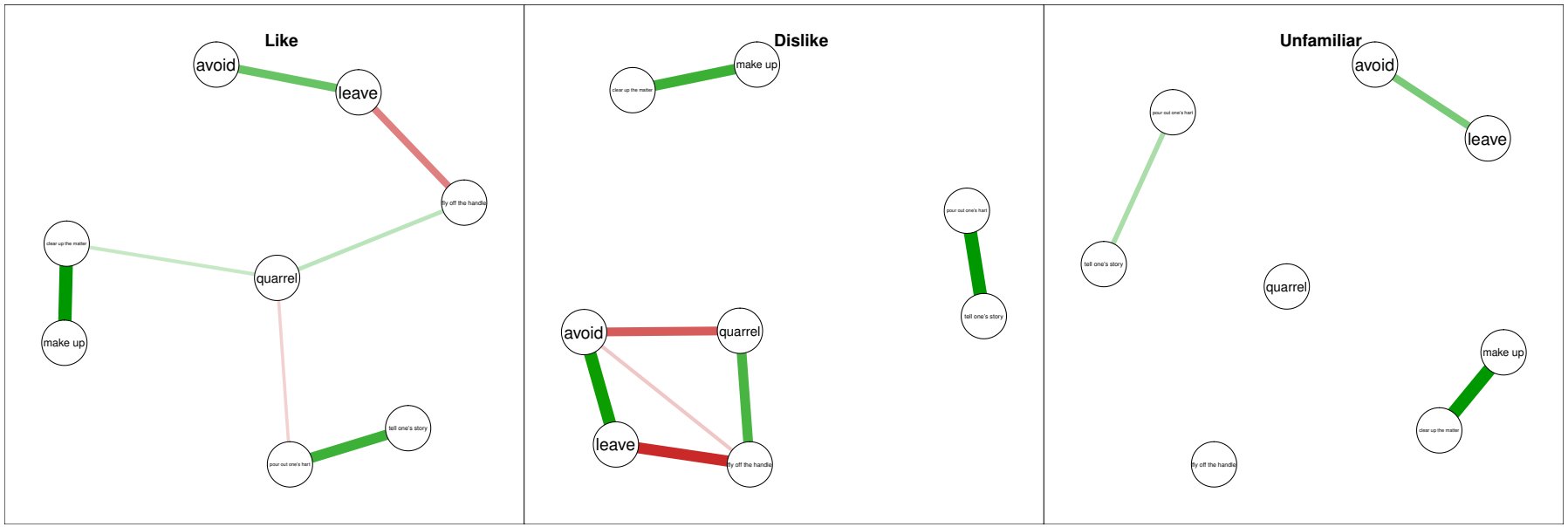
Exercice

- The raw data consist of the binary judgments of 101 first-year psychology students who indicated whether or not they would display each of 8 anger-related behaviors when being angry at someone in each of 6 situations.
- The 8 behaviors consist of 4 pairs of reactions that reflect a particular strategy to deal with situations in which one is angry at someone, namely,
  (1) fighting (fly off the handle, quarrel),
  (2) fleeing (leave, avoid),
  (3) emotional sharing (pour out one's heart, tell one's story),
  (4) making up (make up, clear up the matter).
- The six situations are constructed from two factors with three levels :
  (1) the extent to which one likes the instigator of anger (like, dislike, unfamiliar)
  (2) the status of the instigator of anger (higher, lower, equal).
  Each situation is presented as one level of a factor, without specifying a level for the other factor.

Ising models

Exercice

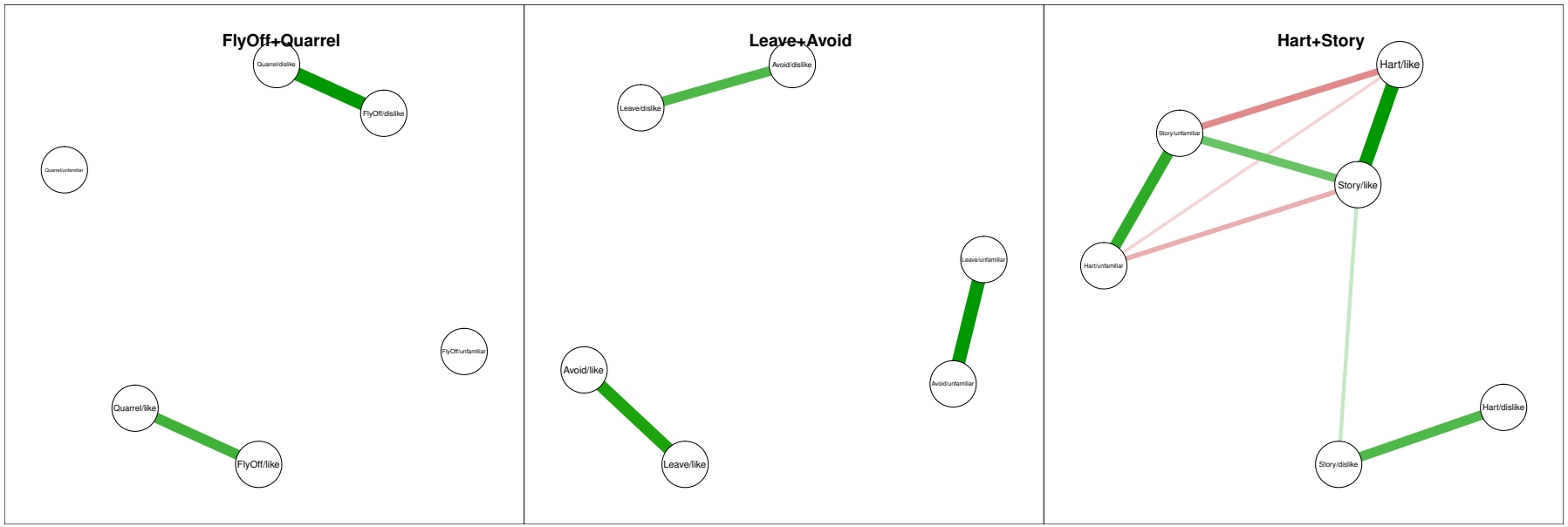
- Give an interpretation to these graphs, build for 3 situations.



Ising models

Exercice

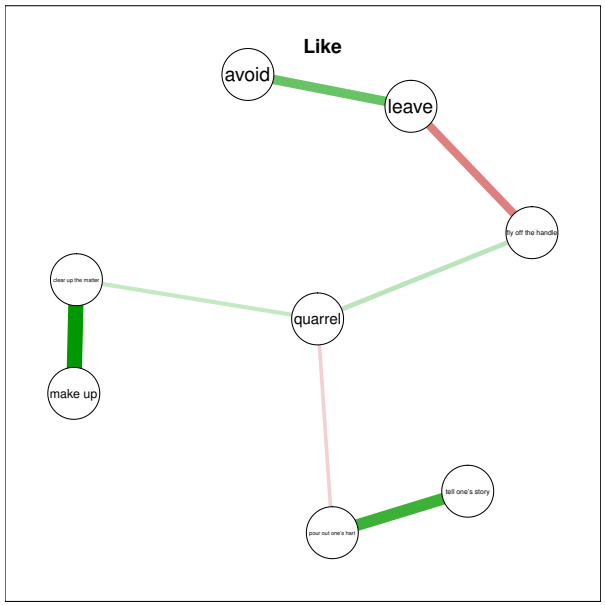
- Give an interpretation to these graphs, build for 3 behaviours.

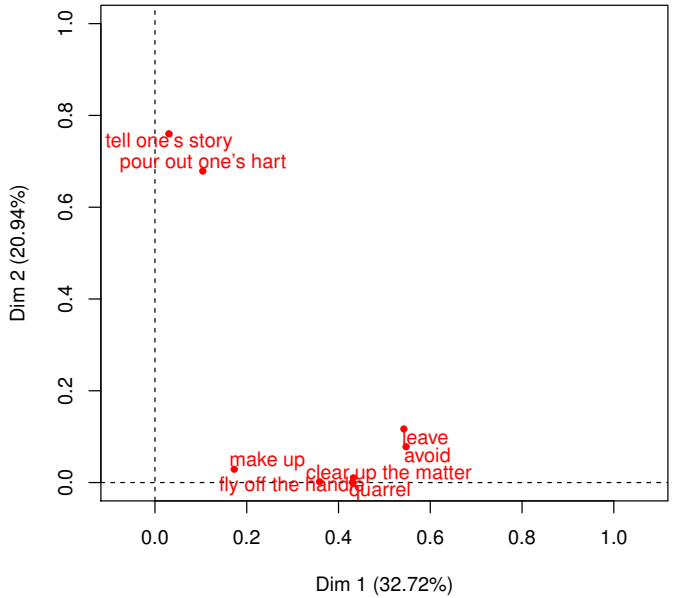


Ising models

Comparison with MCA

- MCA is another way to represent the variables.

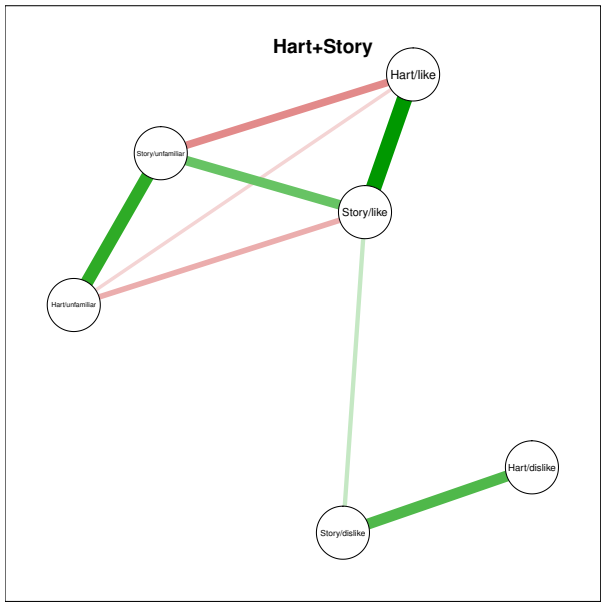


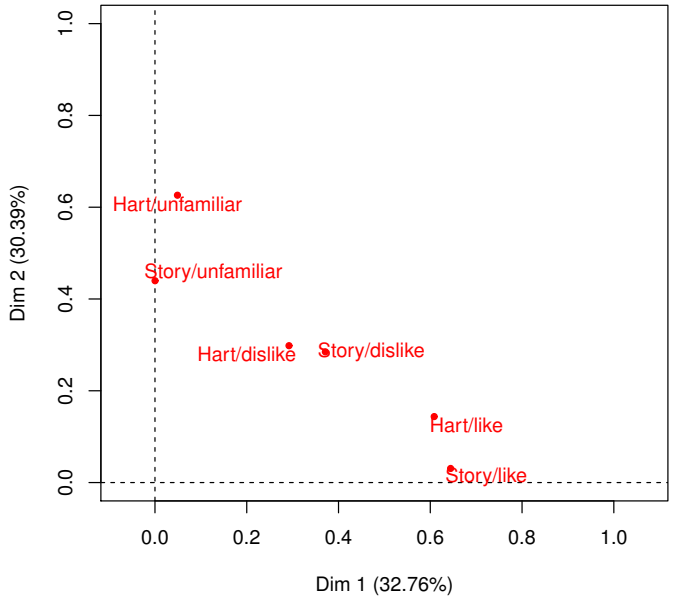


Ising models

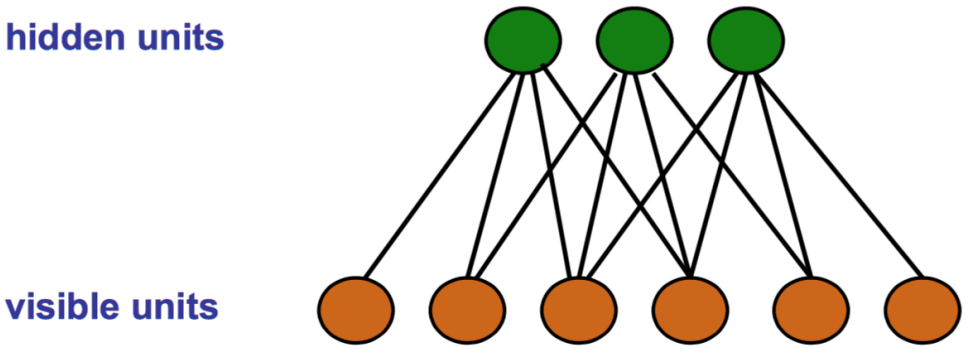
Comparison with MCA

- MCA is another way to represent the variables.





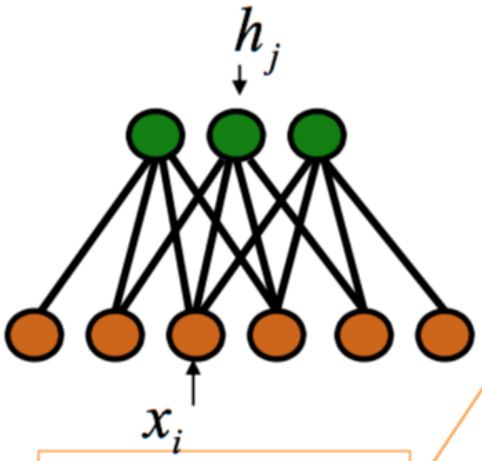
Restricted Boltzmann Machine



$$P(\mathbf{x}, \mathbf{h}; \theta) = \frac{1}{Z} \exp\left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{ij} \phi_{ij}(x_i, h_j) \right\}$$

- In Boltzmann machine the model is homogeneous.
- The introduction of latent (or hidden) variables allow to describe an heterogeneous space.
- For example, the visible units may be the pixels of an image and the hidden units the class of the pixel (ex : earth or sky).
  In deeplearning, the hidden units are interpreted as meta-features.
- RBM can be used to extract summarized information about the observations.

A constructive definition



$$p(\mathbf{x}\,|\,\mathbf{h}) = \prod_i p(x_i\,|\,\mathbf{h}),$$

$$p(x_i\,|\,\mathbf{h}) = \exp\Big\{ \sum_a \hat{\theta}_{ia} f_{ia}(x_i) + A_i(\{\hat{\theta}_{ia}\}) \Big\}$$

$$\hat{\theta}_{ia} = \theta_{ia} + \sum_{jb} W_{ia}^{jb} g_{jb}(h_j) = \theta_{ia} + \sum_j \vec{W}_{ia}^{j} \vec{g}_j(h_j)$$

$$p(\mathbf{h}\,|\,\mathbf{x}) = \prod_j p(h_j\,|\,\mathbf{x})$$

$$p(h_j\,|\,\mathbf{x}) = \exp\Big\{ \sum_b \hat{\lambda}_{jb} g_{jb}(h_j) + B_j(\{\hat{\lambda}_{jb}\}) \Big\}$$

$$\hat{\lambda}_{jb} = \lambda_{jb} + \sum_{ia} W_{ia}^{jb} f_{ia}(x_i) = \lambda_{jb} + \sum_i \vec{W}_i^{jb} \vec{f}_i(x_i)$$

coupling in the log-domain with shifted parameters

vector of local sufficient statistics (features)

- Firstly, one writes the conditional distributions.
- Then, they map to the RBM random field :

$$p(\mathbf{x}, \mathbf{h}; \theta) = \exp\left\{ \sum_i \tilde{\theta}_i f_i(x_i) + \sum_j \tilde{\theta}_j g_j(h_j) + \sum_{i,j} f_i(x_i)^T W_{ij} g_j(h_j) \right\}$$

RBM

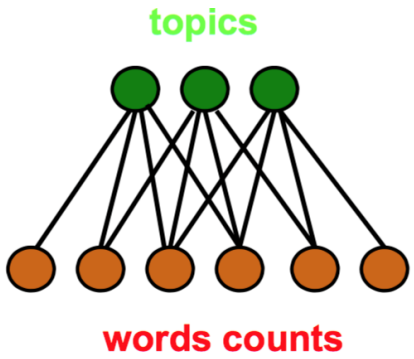
Exercice

- Fit a RBM with 2 units for the Anger "Like" data.
- Fit a RBM with 3 units for the Anger "Emotional sharing" data.

Go to https ://perso.univ-rennes1.fr/valerie.monbet/GM/Anger.html for an example.

RBM for text modeling

- Context



- $h_j \in \mathbb{R}$, $\bar{h}_j = \sum_i w_{ij} x_i$
  Ex : h_j = 3 : topic j has strength 3
- $x_i \in \mathbb{N}$
  Ex : x_i = n : world i occurred n times.

- Chosen conditional distributions

$$p(h_j|\mathbf{x}) \sim \mathcal{N}\left(\sum_i w_{ij} x_i, 1\right)$$

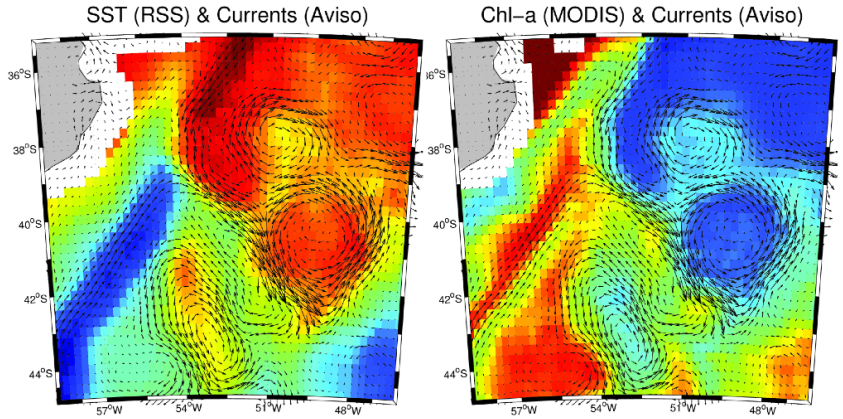
$$p(x_i|\mathbf{h}) \sim \textit{Binomial}\left(N, \frac{\exp(\alpha_j + \sum_j w'_{ij} h_i)}{1 + \exp(\alpha_j + \sum_j w'_{ij} h_i)}\right)$$

Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

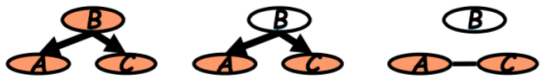
Examples

- Satellite images of sea surface temperatures and sea color close to Brazil.



- Gene networks.

- Gene expressions may be influenced by unobserved factors that are post-transcriptionally regulated.



- The unavailability of the state of B results in a constrain over A and C. This constraints can be for instance be interpreted as correlation and they will be highlighted by the graph.

Gaussian graphical models

- Let us consider Markov networks where all the variables are continuous.
- In this context, the Gaussian multivariate distribution is often used because it has nice analytical properties.

$$\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

$$p(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}(det\mathbf{\Sigma})^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

- Precision matrix : $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$

$$p(\mathbf{x}; \mu = 0, \mathbf{\Theta}) = \frac{(det\mathbf{\Theta})^{1/2}}{(2\pi)^{d/2}} \exp\left(\frac{1}{2}\sum_i \theta_{ii}x_i^2 - \sum_{i<j}\theta_{ij}x_i x_j\right)$$

  We can view this as a continuous Markov Random Field with potentials defined on every node and edge (ex : Ising's model).
- The Gaussian distribution only models dependencies of order 2. Then it describes Markovian graphs of order 1.

Covariance and precision matrices

- Covariance matrix $\mathbf{\Sigma}$
  $\mathbf{\Sigma}_{ij} = 0 \Rightarrow X_i \perp X_j$ or $p(X_i, X_j) = p(X_i)p(X_j)$
  Graphical model interpretation ?

- Precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$
  $\mathbf{\Theta}_{ij} = 0 \Rightarrow X_i \perp X_j | X_{-ij}$ or $p(X_i, X_j | X_{-ij}) = p(X_i | X_{-ij})p(X_j | X_{-ij})$
  Graphical model interpretation ?

Sparse precision *vs* sparse covariance in Gaussian graphical models

$$\Theta = \Sigma^{-1} = \begin{pmatrix} 1 & 6 & 0 & 0 & 0 \\ 6 & 2 & 7 & 0 & 0 \\ 0 & 7 & 3 & 8 & 0 \\ 0 & 0 & 8 & 4 & 9 \\ 0 & 0 & 0 & 9 & 5 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.10 & 0.15 & -0.13 & -0.08 & 0.15 \\ 0.15 & -0.03 & 0.02 & 0.01 & -0.03 \\ -0.13 & 0.02 & 0.10 & 0.07 & -0.12 \\ -0.08 & 0.01 & 0.07 & -0.04 & 0.07 \\ 0.15 & -0.03 & -0.12 & 0.07 & 0.08 \end{pmatrix}$$

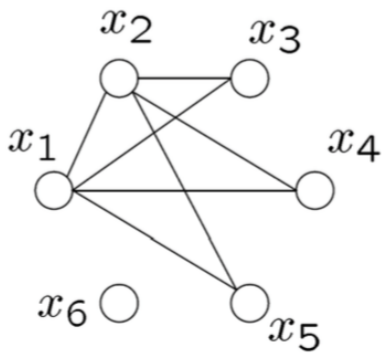
$$\Theta_{15} = 0 \Leftrightarrow X_1 \perp X_5 | X_{-\{15\}}$$

does not imply

$$X_1 \perp X_5 \Leftrightarrow \Sigma_{15} = 0$$

Sparse precision in Gaussian graphical models

$$Q = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$



Some nice properties of Gaussian distribution

- One of the nice properties of the multivariate Gaussian distribution is that all its conditional distributions are Gaussian too.
  Let Y and Z be such that $Z = (X_1, \cdots, X_{p-1})$ and $Y = X_p$,

$$Y|Z = z \sim \mathcal{N}\left(\mu_Y + (z - \mu_Z)^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}\right)$$

  where

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}.$$

- Remark : $\mu_Y + (z - \mu_Z)^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}$ is exactly of the same form as the linear regression of Y over Z with $\beta = \Sigma_{ZZ}^{-1} \sigma_{ZY}$.

More on the precision matrix

$$Y|Z = z \sim \mathcal{N}\left(\mu_Y + (z-\mu_Z)^T \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY}, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY}\right), \ \mathbf{\Sigma} = \begin{pmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

- Remark : $\mu_Y + (z-\mu_Z)^T \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY}$ is exactly of the same form as the linear regression of Y over Z with $\beta = \Sigma_{ZZ}^{-1}\sigma_{ZY}$.
- With the same partition for $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$, since

$$\mathbf{\Theta\Sigma} = I$$

$$\theta_{ZY} = -\theta_{YY}\mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY}$$

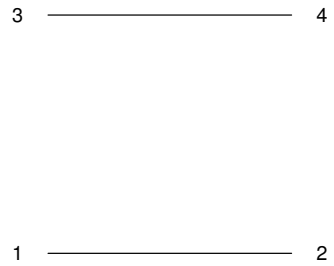
with $1/\theta_{YY} = \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} > 0$. So that,

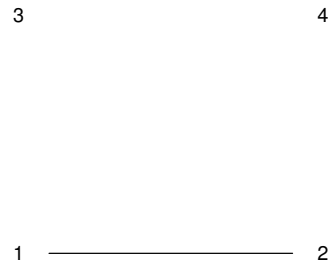
$$\beta = \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} = -\theta_{ZY}/\theta_{YY}$$

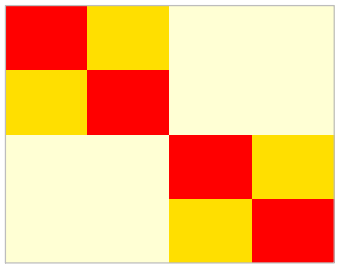
- The conditional dependencies between Y and Z are described by the mean. The zero coefficients of θ_{ZY} leads to zero coefficients in β (and reversely).
- Finally, $\mathbf{\Theta}$ captures all the second order structure that is needed to describe the conditional distribution.

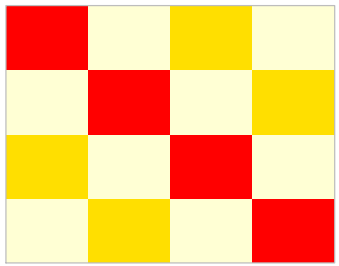
Toys examples

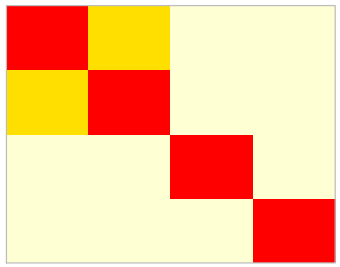
- 3 examples of graphs with their precision matrices.











Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
   - Learning when the structure is known
   - Estimation of the graph structure
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

Examples

- We have n iid observations (ex : expression of the genes, like/unlike from customers, students marks, etc.).

$$x_1, \cdots, x_n, \quad x_i \in \mathbb{R}^d$$

- We want to learn the structure and the parameters of the graphical models (e.g. the multivariate distribution).

$$\mu, \Theta$$

- Maximization of the likelihood.

Outline

6 Learning (continuous) undirected graph models
- Learning when the structure is known
- Estimation of the graph structure

Estimation for a full graph

- Observations $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ of $\mathbf{X}$
- Assumption : the graph is full : μ and $\mathbf{\Sigma}^{-1}$ have no zero coefficients.
- The empirical covariance matrix is given by

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

  where $\bar{\mathbf{x}}$ is the empirical mean.
- The log-likelihood is given by

$$\ell(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \text{trace}(\mathbf{S\Theta})$$

  up to an additive constant.
- And

$$\frac{\partial}{\partial\mathbf{\Theta}}\log\det(\mathbf{\Theta}) = \mathbf{\Theta}^{-1}, \quad \frac{\partial}{\partial\mathbf{\Theta}}\text{trace}(\mathbf{S\Theta}) = \mathbf{S}$$

- So that the estimator of the maximum likelihood is $\mathbf{\Sigma} = \mathbf{\Theta}^{-1} = \mathbf{S}$

Estimation for a sparse graph

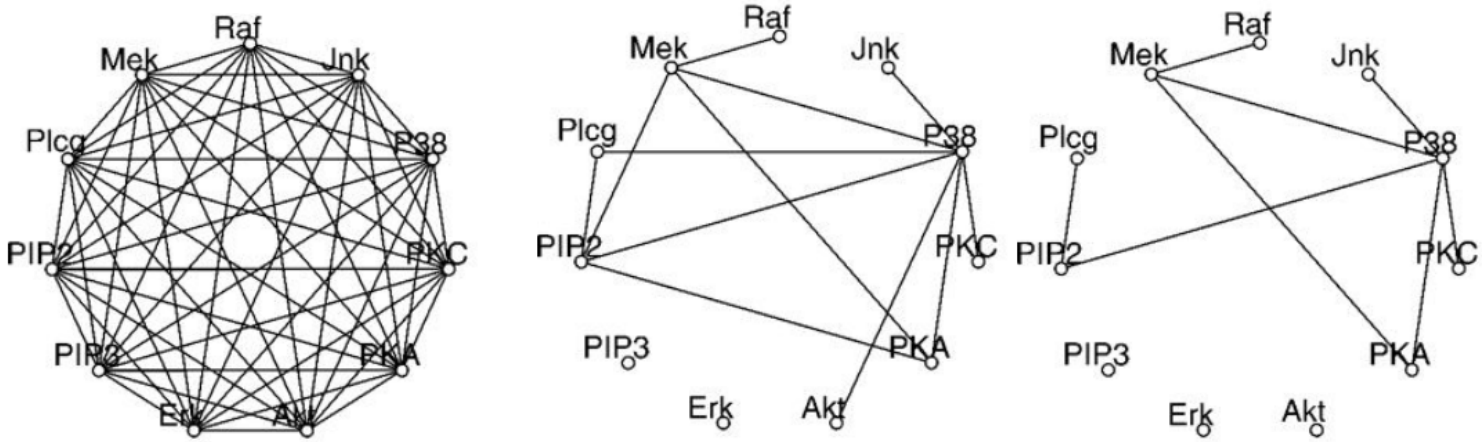
- Generally, the graphs are sparse.
- If the graph structure is known and denoted E, the maximization of the likelihood leads to a quadratic optimization problem under equality constraints.

$$\max_{\Theta,\Gamma} \left\{ \log\det(\Theta) - \mathrm{trace}(\mathbf{S}\Theta) - \sum_{(j,k)\notin E} \gamma_{jk}\theta_{jk} \right\}$$

- The gradient equation is

$$\Theta^{-1} - \mathbf{S} - \Gamma = 0$$

Γ is a matrix of Lagrange parameters with non-zero values for all pairs with edges absent.



Ideas of the modified regression algorithm

- We partition the matrices $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ into two parts
  - Part 1 : the first $d-1$ rows and columns
  - Part 2 : the last row and column

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \begin{pmatrix} \mathbf{\Theta}_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{d-1} & 0 \\ 0 & 1 \end{pmatrix}$$

- This implies

$$w_{12} = -W_{11}\theta_{12}/\theta_{22} = W_{11}\beta$$

- Then, with empirical estimates, $W_{11}\beta - \hat{\sigma}_{12} - \gamma_{12} = 0$.
- There are $d-q$ non-zero elements in γ_{12} (edges constrained to be zero), so that the previous equation is equivalent to

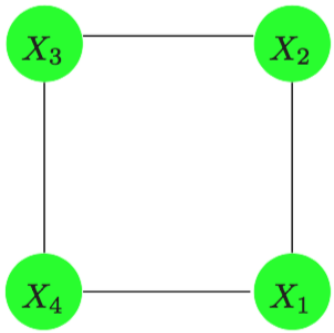
$$\beta^* = (W_{11}^*)^{-1}\,\hat{\sigma}_{12}^*.$$

- Furthermore, $\frac{1}{\theta_{22}} = w_{22} - w_{12}^T\beta$ and $w_{22} = \hat{\sigma}_{22}$ since the diagonal of $\mathbf{\Gamma}$ is zero.
- This leads to a simple iterative algorithm.

A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.

1. Intialize $W = \widehat{\mathbf{\Sigma}}$
2. Repeat for $j = 1, 2, \cdots, p$, until convergence
   1. Partition the matrix W into part 1 : all but the jth row and column, and part 2 : the jth row and column.
   2. Solve $W_{11}^{*}\beta^{*} = \hat{\sigma}_{12}^{*}$ for the unconstrained edge parameters β^{*}, using the reduced system of equations. Obtain $\hat{\beta}$ by padding $\hat{\beta}^{*}$ with zeros in the appropriate positions.
   3. Update $w_{12} = W_{11}\hat{\beta}$
3. In the final cycle (for each j) solve $\hat{\theta}_{12} = -\hat{\beta}\hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = \hat{\sigma}_{22} - w_{12}^{T}\hat{\beta}$

A simple graph for illustration



$$\boldsymbol{S} = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 10.00 & 1.00 & 1.31 & 4.00 \\ 1.00 & 10.00 & 2.00 & 0.87 \\ 1.31 & 2.00 & 10.00 & 3.00 \\ 4.00 & 0.87 & 3.00 & 10.00 \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix} 0.12 & -0.01 & 0.00 & -0.05 \\ -0.01 & 0.11 & -0.02 & 0.00 \\ 0.00 & -0.02 & 0.11 & -0.03 \\ -0.05 & 0.00 & -0.03 & 0.13 \end{pmatrix}$$

Outline

6 Learning (continuous) undirected graph models
- Learning when the structure is known
- Estimation of the graph structure

Estimation of the graph structure

- In most problems, the graph structure is unknown.
- A natural solution is to introduce a Lasso penalty. The new optimization problem is written

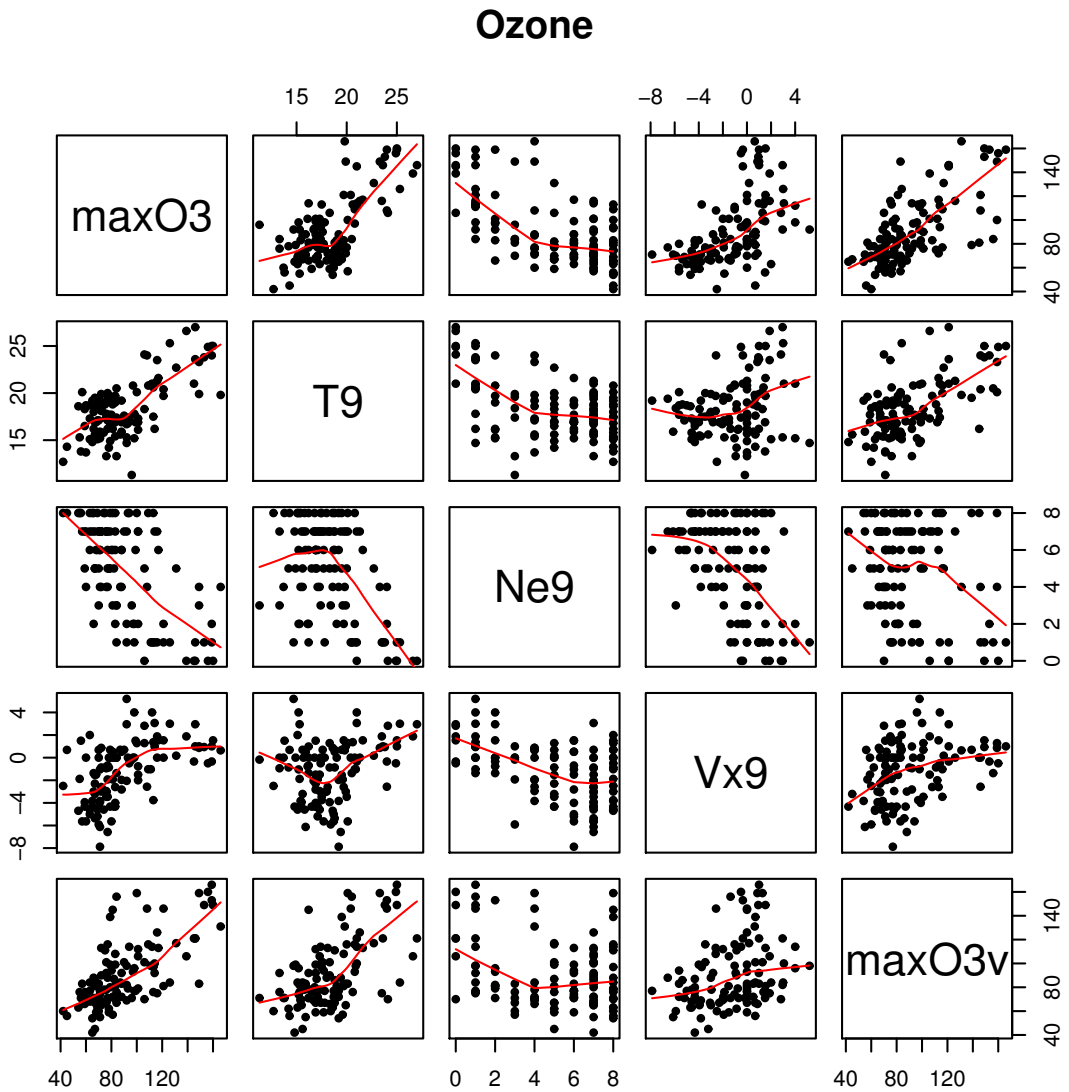
$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \{\log\det(\mathbf{\Theta}) - \text{trace}(\mathbf{S\Theta}) - \lambda\|\mathbf{\Theta}\|_1\}$$

  with $\|\mathbf{\Theta}\|_1 = \sum_{ij} |\theta_{ij}|$.
- As in Lasso regression, the non significant parameters are shrunk to zero.
- The function to be minimized is convex. The algorithm called "graphical lasso" is similar to the one proposed for sparse graph estimation when the structure is known. The gradient equation is replaced by

$$\mathbf{\Theta}^{-1} - \mathbf{S} - \lambda Sign\mathbf{\Theta} = 0$$

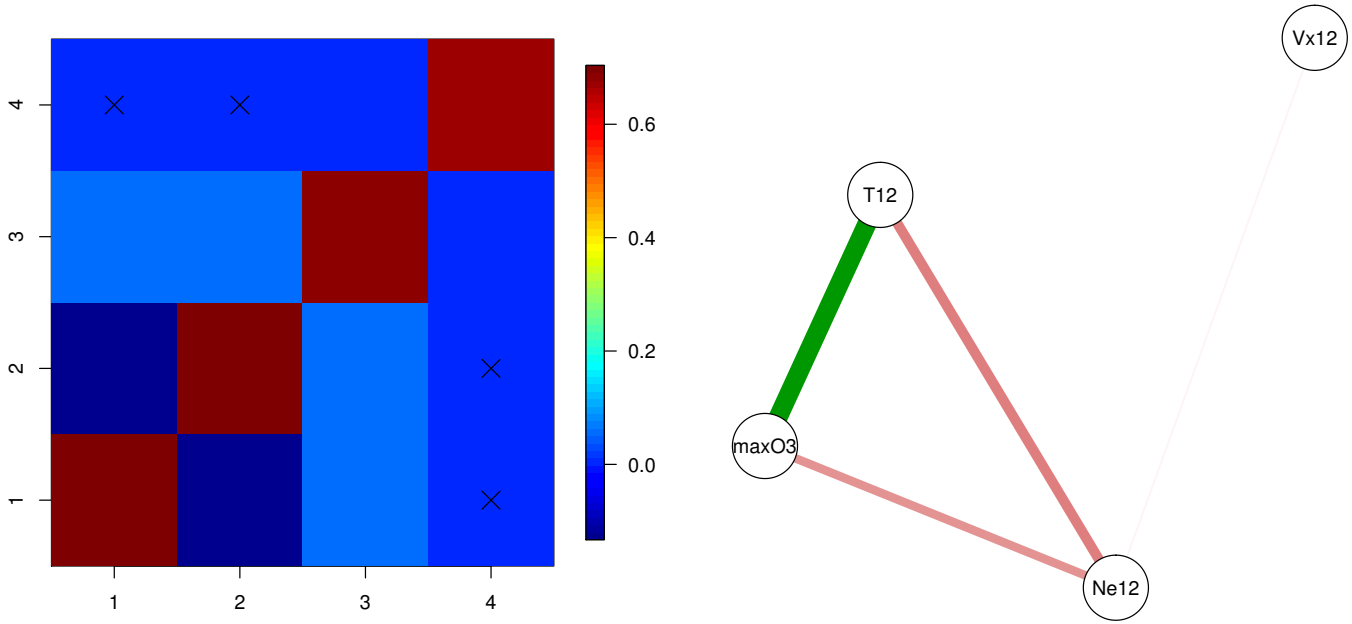
  where $Sign(\theta_{ij}) = sign(\theta_{ij})$ if $\theta_{ij} \neq 0$ and $Sign(\theta_{ij}) \in [-1, 1]$ if $\theta_{ij} = 0$
- There are some consistency results for this algorithm.

Example, Ozone concentrations



Ozone concentrations

- Correlation matrix and graph on a subset of variables



Ozone concentrations

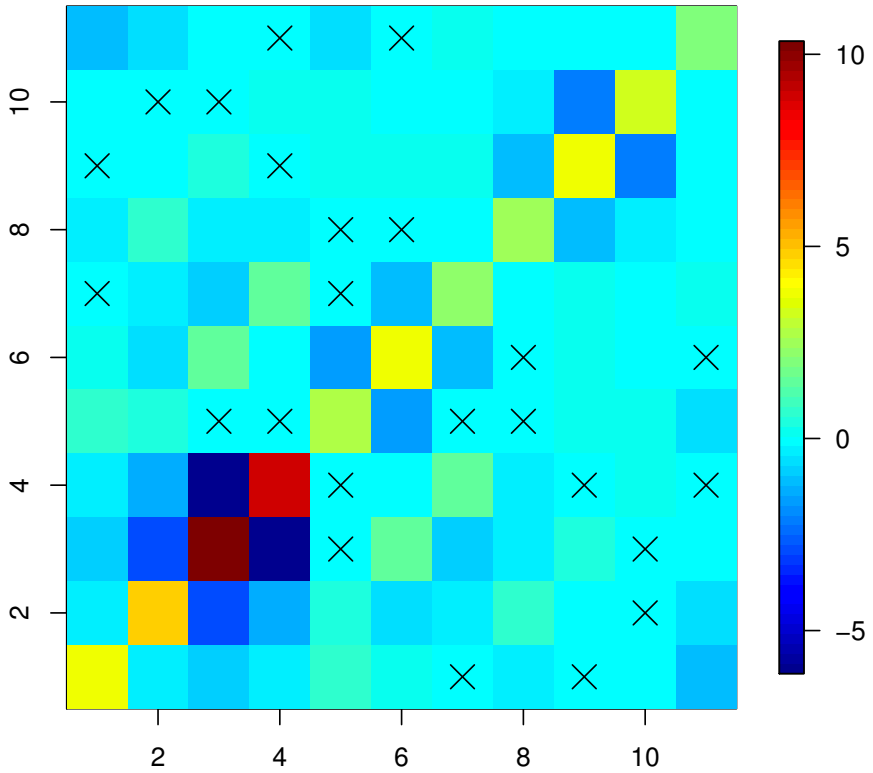
Exercice

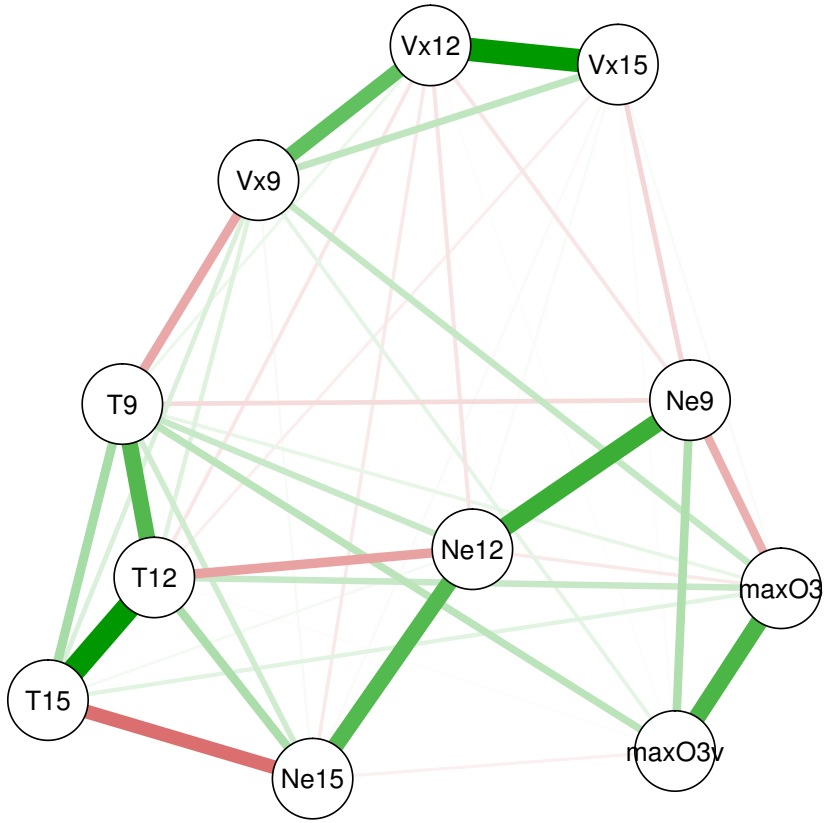
- Fit a Gaussian Graphical Model on the qualitative variables of the Ozone data set.
- Use BIC and/or AIC to find the "best" regularization constant.
- Plot the precision matrix and the associated graph.
- Interpret the graph.

See `https://perso.univ-rennes1.fr/valerie.monbet/GM/Ozone_GGM.Rmd` for an example.

Ozone concentrations

- Correlation matrix and graph on continuous variables





Customer's satisfaction survey

Exercice

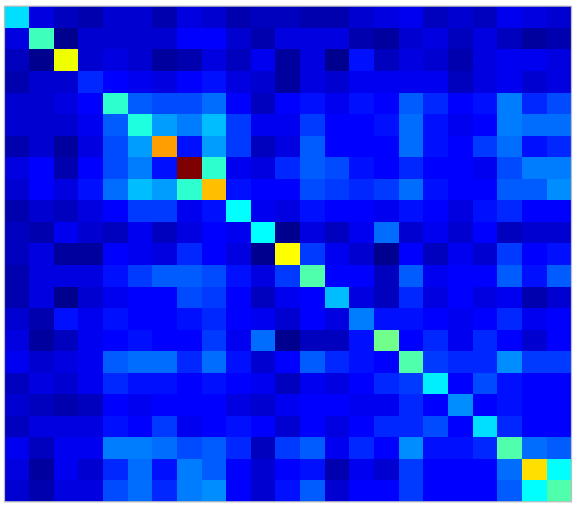
- Fit a Gaussian Graphical Model on the customer's satisfaction survey. You can use `glasso` package available on CRAN.
- The data describe customer's satisfaction for mobile phones.
  $n = 250$, $p = 23$.
  Customers give marks between 1 and 10.
- Use BIC and/or AIC to find the "best" regularization constant.
- Plot the precision matrix and the associated graph.
- Interpret the graph.

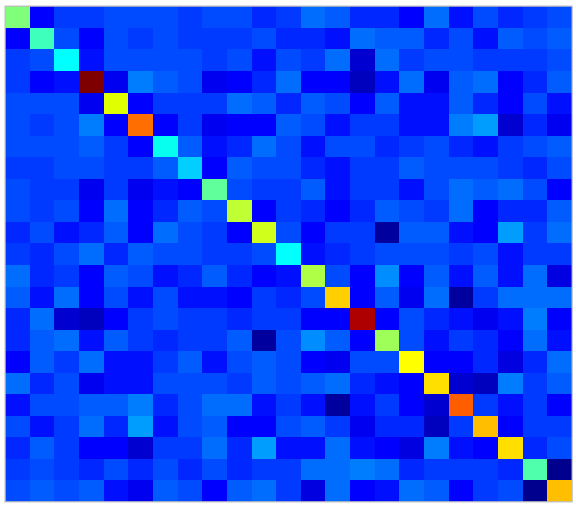
Data available at
`https://perso.univ-rennes1.fr/valerie.monbet/GM/mobi.Rdata`

Example, customer's satisfaction survey

- Customer's satisfaction for mobile phones. $n = 250$, $p = 23$. Customers give marks between 1 and 10.
- Estimation of $\mathbf{S}$ and $\widehat{\mathbf{\Sigma}}$ for the complete graph.
- Following $\mathbf{S}$, variables 8 and 9 are highly correlated : *BuyAgain*, *Recommend*. Same for 22 and 23 : *FairPrice*,*GoodValue*.
- The precision matrix is difficult to interpret.

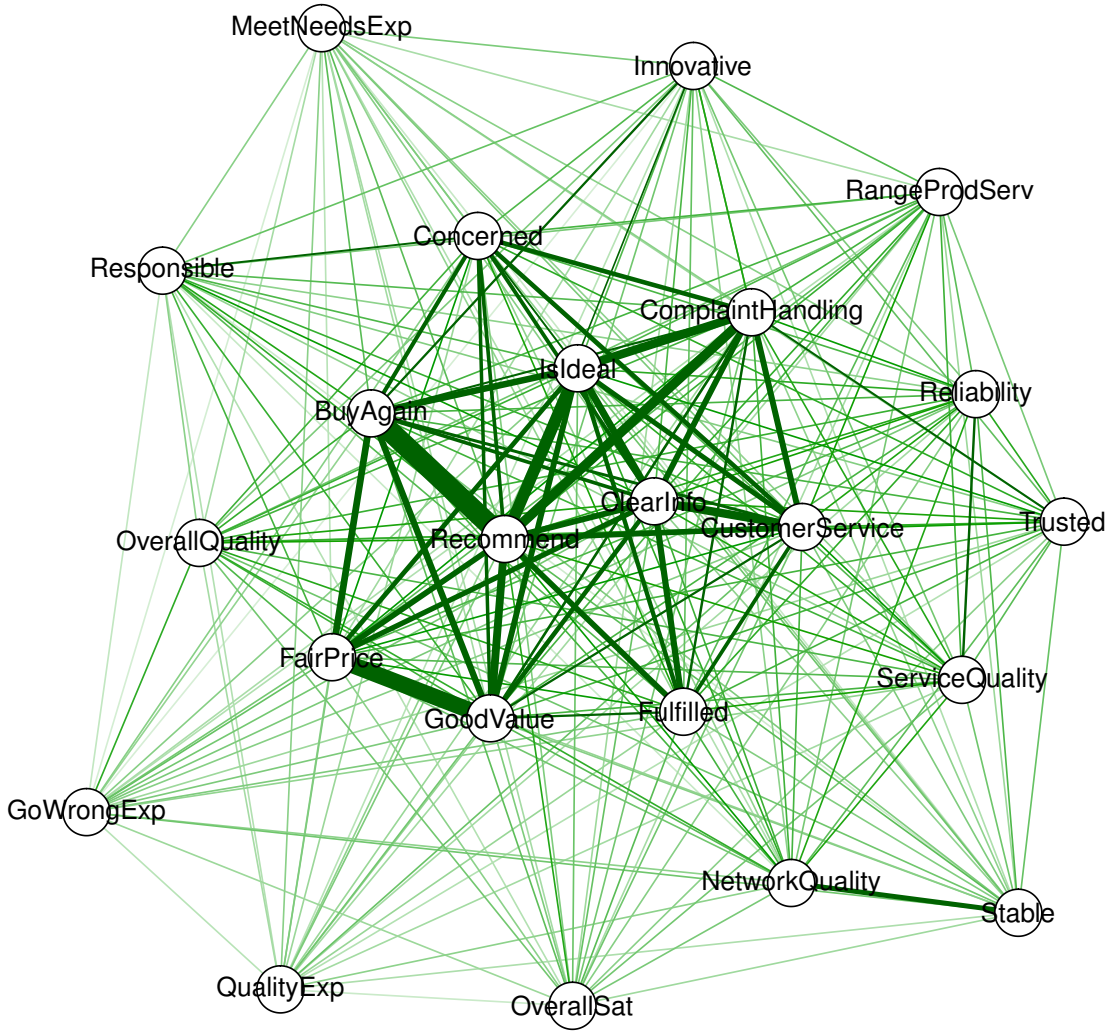
$\mathbf{S}$ $\qquad\qquad$ $\widehat{\mathbf{\Theta}}$





Example, customer's satisfaction survey

- Full graph

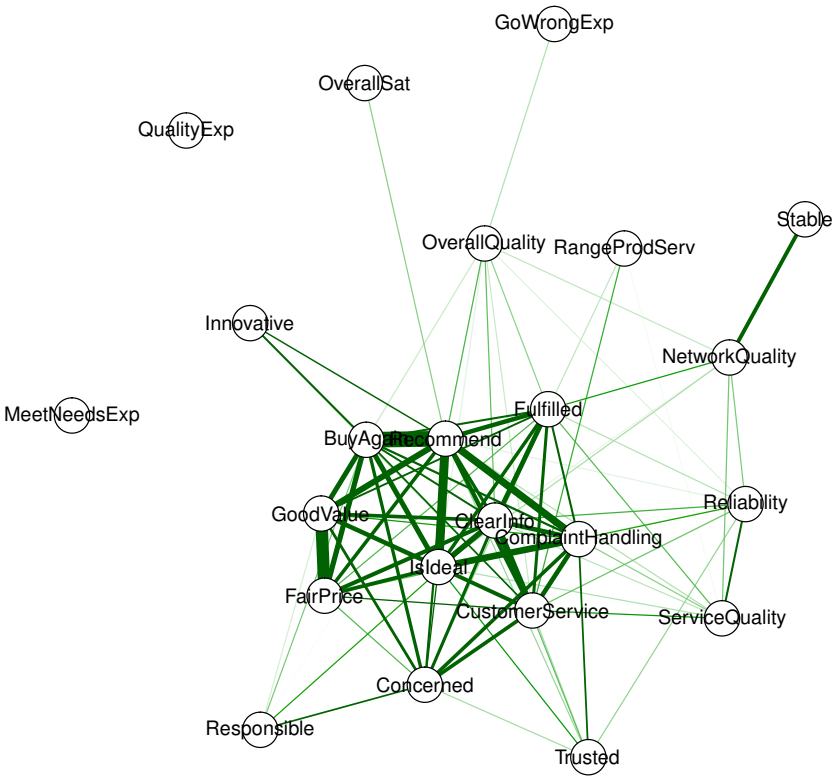


Example, customer's satisfaction

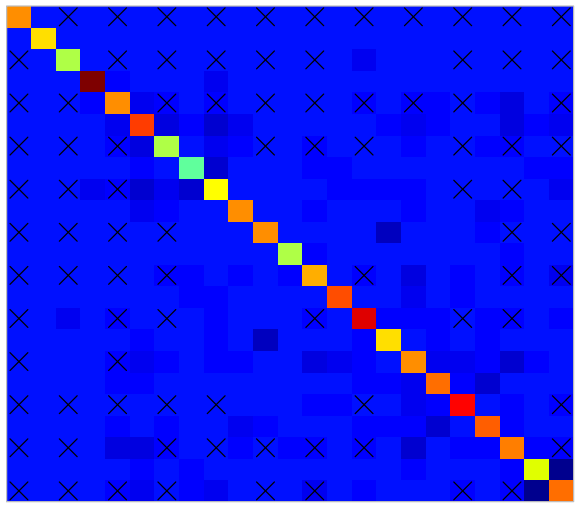
- Satisfaction survey. $n = 250$, $p = 23$.
- Sparse model ?
- λ is selected according to AIC and BIC criteria.

λ	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
AIC	1483	1463	1436	1442	**1430**	1441	1458	1475	1493	1508
BIC	1793	1681	1560	1534	1462	**1455**	1469	1482	1500	1511

Graphe



$\widehat{\Theta}$



Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

Learning Ising models (pairwise MRF)

- Assuming the nodes are discrete, and edges are weighted, then for a sample $\mathbf{x}$, we have

$$P(\mathbf{x}|\Theta) = \exp\left(\sum_{i\in V}\theta_{ii}^T\mathbf{x}_i + \sum_{(i,j)\in E}\theta_{ij}\mathbf{x}_i\mathbf{x}_j - \Phi(\Theta)\right)$$

  Φ plays the role of the normalisation constant.

- Recall that Ising model implies that

$$P(X_j|X_{-j} = x_{-j}) = \frac{\exp\left(\theta_{j0} + \sum_{j,k\in E}\theta_{jk}x_k\right)}{1 + \exp\left(\theta_{j0} + \sum_{j,k\in E}\theta_{jk}x_k\right)}$$

- It can be shown following the same logic that we can use L_1 regularized logistic regression to obtain a sparse estimate of the neighbourhood of each variable in the discrete case (graph structure).
- Once the graph structure is known, non zero parameters can be estimated.

Estimation when the graph structure is known

- For undirected graphical models, the log likelihood does not decompose, because the normalization constant Z is a function of all the parameters

$$P(x_1, \cdots, x_d) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c), \quad Z = \sum_{x_1, \cdots, x_d} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

- In general, we will need to do inference (i.e. marginalization or conditional prediction) to learn parameters for undirected models, even in the fully observed case.

Estimation when the graph structure is known, Ising model

- Observations $\mathbf{x}_i = (x_{i1}, \cdots, x_{id}) \in \{0,1\}^d$, $i = 1, \cdots, N$.
- Log likelihood of Ising model

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \left\{ \sum_{(j,k)\in E} \theta_{jk} X_j X_k - \Phi(\boldsymbol{\Theta}) \right\}$$

- Gradient of log-likelihood

$$\frac{\partial \ell(\boldsymbol{\Theta})}{\partial \theta_{jk}} = \sum_{i=1}^{N} x_{ij} x_{ik} - N \frac{\partial \Phi(\boldsymbol{\Theta})}{\partial \theta_{jk}}$$

$$\frac{\partial \Phi(\boldsymbol{\Theta})}{\partial \theta_{jk}} = \sum_{\mathbf{x}\in\mathcal{X}} \mathbf{x}_j \mathbf{x}_k p(\mathbf{x}, \boldsymbol{\Theta}) = E_{\boldsymbol{\Theta}}(X_j X_k)$$

- The gradient is zero when

$$\widehat{E}(X_j X_k) = E_{\boldsymbol{\Theta}}(X_j X_k)$$

where $\widehat{E}$ is the empirical expectation.

Estimation when the graph structure is known, Ising model

- To find a solution of
$$\widehat{E}(X_j X_k) = E_{\Theta}(X_j X_k)$$
gradient descent methods can be used if d is not to large ($d \leq 30$).
- Problems becomes intractable when d is large. To compute $E_{\Theta}(X_j X_k)$, the computation of $p(X; \Theta)$ is required for all 2^{d-2} possible values of $|\mathcal{X}| = 2^d$.
- The more common approach is then to use Gibbs sampling to approximate $E_{\Theta}(X_j X_k)$ from several successive sampling of the conditional distributions $P_{\Theta}(X_j | X_{-j})$.
- Indeed, if samples of p_{Θ} are available, $E_{\Theta}(X_j X_k)$ can be approximated by empirical estimates.

Gibbs sampler

- Let $\mathbf{X} = (X_1, \cdots, X_d)$ be a vector of random variables.
- We want to generate a sample of the joint distribution of $\mathbf{X}$.
- Sometimes, it is difficult to sample according to the joint distribution of $\mathbf{X}$ but easy to sample according to the conditional distributions $P(X_j|X_{-j})$. It is the case for graphical models.
- The Gibbs sampler simulates alternatively according to $P(X_j|X_{-j})$, $j \in \{1, \cdots, d\}$.
- Under some regularity assumptions, one can show that when (and if) the process is stabilized the obtained sample is distributed as $\mathbf{X}$.
- Algorithm of Gibbs sampler

Gibbs Sampler

1. Choose initial values $X_j^{(0)}$, $j \in \{1, \cdots, d\}$.
2. Repeat for $t = 1, 2, \cdots$
   For $j = 1, \cdots, p$
   Sample $X_j^{(t)}$ according to $P(X_j|X_{-j}^{(t-1)})$
3. Continue step 2. until the distribution of $(X_1^{(t)}, \cdots, X_p^{(t)})$ does not change anymore.

Gibbs sampler, Markov chain

- A Markov chain is defined by a (finite) set of states and a transition matrix Q.
- If the Markov chain is ergodic, it admits a stationary distribution π such that $\pi = Q\pi$.
- The Gibbs sampler is an ergodic Markov chain with transition matrix

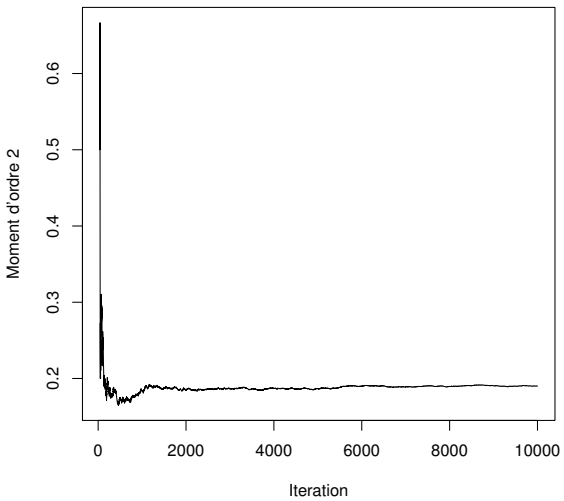
$$Q_j = p(x_j | X_{-j})$$

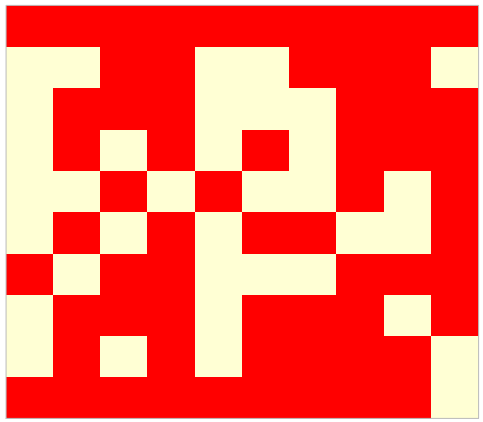
  and its stationary distribution is the joint distribution of $(X_1, \cdots, X_d)$.
- In general, a large number of iterations if required to achieve the stationary distribution. This number corresponds to the mixing time of the Markov chain. [1]

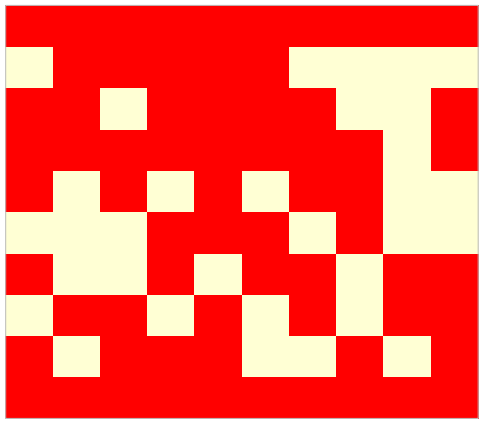
1. Roberts, G. O., Smith, A. F. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic processes and their applications, 49(2), 207-216.

Examples

- One simulates a Markov field over a 10 by 10 grid, with parameters $\boldsymbol{\Theta}$ such that $\theta_{ij} = 1/4$ if X_i and X_j are neighbours and 0 else.
- Mixing time + 2 simulated fields





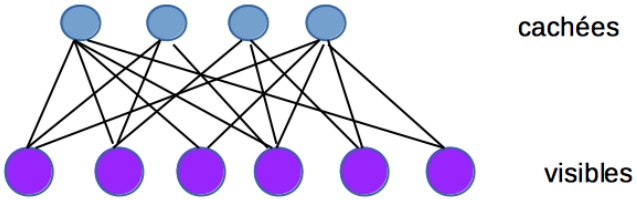


Outline

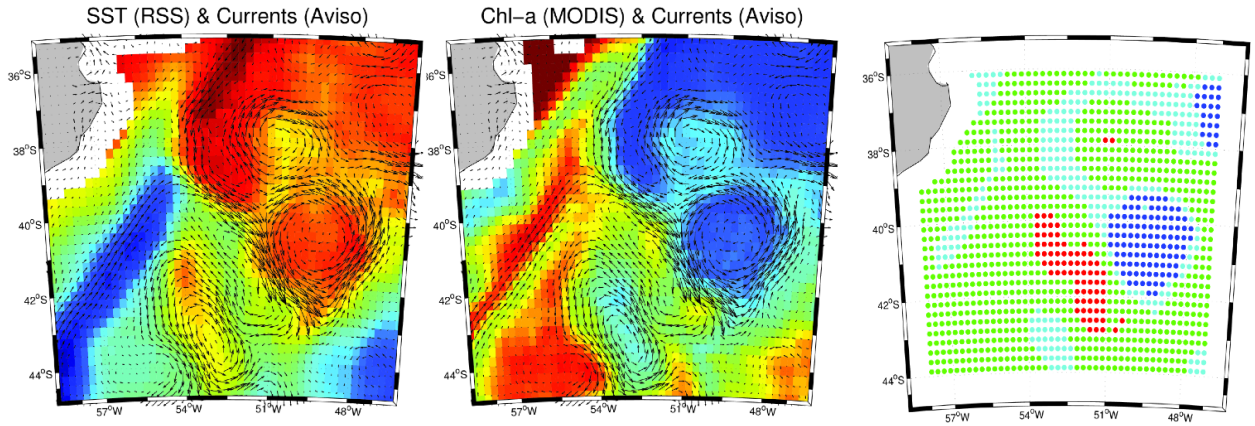
1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
   - When visible variables are continuous
   - Mixture models
   - When visible variables are discrete
9. Directed graph models
10. Some concluding remarks

Latent variables

- In some applications, latent (or hidden) variables are introduced.
- Discrete/discrete variables : Boltzmann machines.



- Continuous/discrete variables : segmentation.



Pierre Tandéo, Telecom Bretagne, Brest.

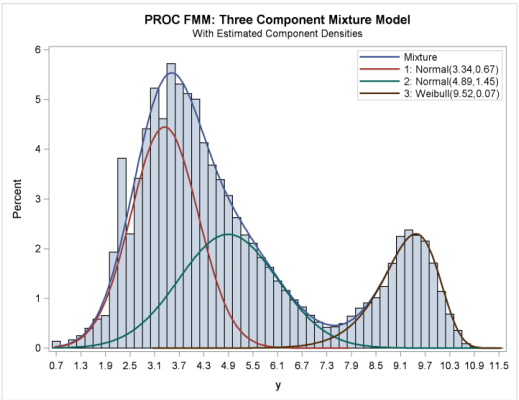
- Continuous or discrete/continuous variables : latent factors.
  Well known example : Principal component analysis.

Learning with latent variables

- When a part of the variables is latent, there are 2 problems
  1. Prediction of the hidden variables given the observations and the model. Usually, the posterior probabilities are estimated

$$p(X_{\mathcal{H}}|X_{\mathcal{V}};\boldsymbol{\Theta})$$

  2. Estimation of model parameters $\boldsymbol{\Theta}$.
- It is clear that the two problems are linked. If $\boldsymbol{\Theta}$ is unknown, parameters and posterior probabilities have to be estimated simultaneously.
- Example : clustering, mixture models.

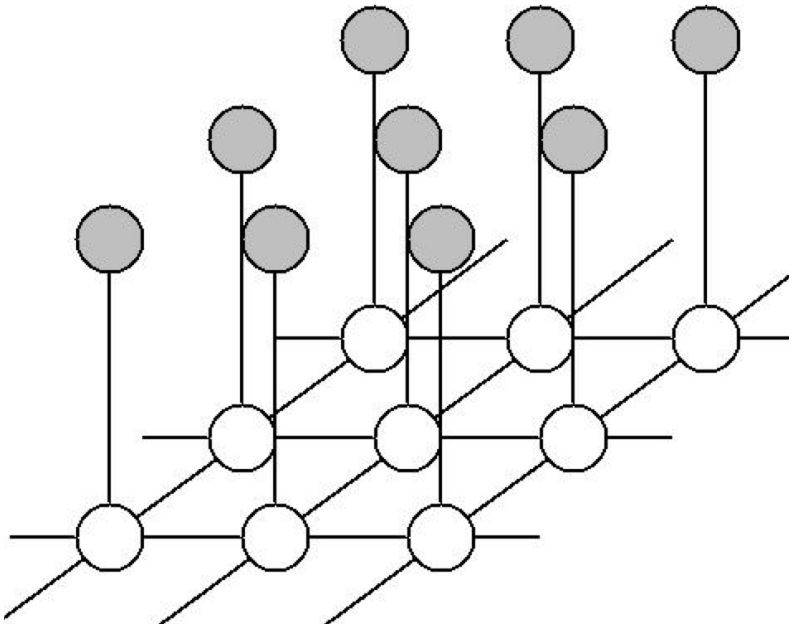


Outline

8 Graphical models with latent variables
- When visible variables are continuous
- Mixture models
- When visible variables are discrete

Graphe with visible continuous variables

- Image restoration and image segmentation, the considered graphs have specific architecture.
- Usually, the edges corresponding to the observed variables are connected.
- A hidden variable is associated to each observed variable. For image restoration the hidden variables are in the same space as the observed variables. For image segmentation, the hidden variables can take only a finite number of values $\{1, \cdots, K\}$.
- Each hidden variable is connected only to its neighbours.



Example in vision, image segmentation

- The problem is to discriminate the foreground from the background.
- Observation $x \in \mathcal{X}$ = image.
- For each pixel i, a variable $Y_i \in \{0, 1\}$ is defined such that $y_i = 1$ if the pixel belongs to the foreground and 0 elsewhere.
- Then a score function g_i is defined such that $g_i(1, x) > g_i(0, x)$ if the pixels around i belong to the foreground. Ex : $g(y_i, x) = p(y_i|x)$ computed by a color model.
- In order to restore a good spatial coherence, g_i (or y_i) is connected to its neighbors

$$y^* = arg \max_{y \in \{0,1\}^n} \sum_{i=1}^{n} g_i(y_i, x) + \sum_{i,j \in \mathcal{I}}^{n} g_{i,j}(y_i, y_j)$$

with $\mathcal{I}$ the neighborhood, and g_{ij} a function which returns a large value if all the points of the neighborhood are in the same ground. Ex : $g_{ij}(y_i, y_j) = \exp(-\gamma \|y_i - y_j\|^2)$, γ a positive constant.

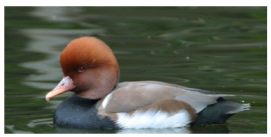


Fig. 1.1 Input image to be segmented into foreground and background. (Image source: http://pdphoto.org)

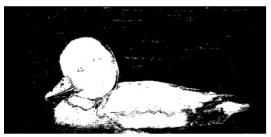


Fig. 1.2 Pixelwise separate classification by g_i only: noisy, locally inconsistent decisions.

Fig. 1.3 Joint optimum y^* with spatially consistent decisions.

http ://www.nowozin.net/sebastian/papers/nowozin2011structured-tutorial.pdf

Example in vision, image segmentation

$$y^* = arg \max_{y \in \{0,1\}^n} \sum_{i=1}^{n} p(y_i|x) + \lambda \sum_{i,j \in \mathcal{I}}^{n} \exp(-\gamma \|y_i - y_j\|^2)$$

Fig. 4.2 A natural image to be segmented. (Image source: `http://pdphoto.org`)

Fig. 4.3 Resulting foreground region.

Fig. 4.4 Left: heatmap of unary potential values. Right: segmentation masks for large w.

Fig. 4.5 Segmentation masks for medium and small w.

http ://www.nowozin.net/sebastian/papers/nowozin2011structured-tutorial.pdf

Example in vision, image segmentation

- The model can depend on unknown parameters θ (and γ).
- Problem

$$y^* = arg \max_{y \in \{0,1\}^n} \sum_{i=1}^{n} p_\theta(y_i|x) + \lambda \sum_{i,j \in \mathcal{I}}^{n} \exp(-\gamma \|y_i - y_j\|^2)$$

  only allows to infer Y given the observations x.
- In order to learn the parameters, an EM algortihm can be used. It is an iterative algorithm which consists in 2 steps
  - E step : inference of Y given x and $\hat{\theta}$
  - M step M : estimation of θ given the complete data (x, Y)
- The simplest use of EM algorithm is probably the mixture of distributions.

Outline

8 Graphical models with latent variables
- When visible variables are continuous
- Mixture models
- When visible variables are discrete

Mixture models

- Mixture models are used for clustering.

$$\begin{array}{rc} \textit{discrete} & S \\ & \downarrow \\ \textit{continuous} & X \end{array}$$

- It may be interpreted as a soft version of the k-means.
- A mixture model characterizes the distribution of X of a pair (S, X) such that
  - S is a discrete random variable defined on $\{1, \cdots, K\}$; S is hidden.
  - X is a random variable defined on $\mathbb{R}^p$ such that $P(X|S = k)$ admit a density $f_k(.; \theta_k)$ for all $k \in \{1, \cdots, K\}$.
  - From the theorem of total probabilities $P(X \in A) = \sum_{k=1}^{K} P(X \in A|S = k)P(S = k)$ then

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)$$

with $\pi_k = P(S = k)$.

Learning of mixture models

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)$$

- Parameters π_k and θ_k, $k \in \{1, \cdots, K\}$ are estimated by maximum likelihood using the Expectation-Maximization algorithm (EM algorithm).
- The algorithm proceeds in two step which are iterated.
  - In E step, the parameter is fixed to its current values and the probability that a sample i belongs to class k is estimated.
  - In M step, parameters are estimated given probabilities computed at the previous step.
- EM algorithms for Gaussian mixture looks like a k-means algorithm.

EM algorithm ..... kmeans

Initialization Choose the number of classes K and initialize the parameter vector $\rightarrow \theta^{(0)}$

Iterate until the estimated parameter does not change significantly any more

- **E-step** For $i = 1, \cdots, n$, estimated the posteriori probabilities

$$\begin{aligned} T_{k,i} &= P(S = k | X = x_i, \theta^{(t)}) \\ &= \frac{\pi_k^{(t)} f_k(x_i; \theta_k^{(t)})}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} f_\ell(x_i; \theta_\ell^{(t)})} \end{aligned}$$

- **M-step** Estimate parameters for $k = 1, \cdots, K$

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} T_{k,i}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} T_{k,i} x_i}{\sum_{i=1}^{n} T_{k,i}}$$

$$\sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} T_{k,i} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} T_{k,i}}$$

Initialization Choose the number of classes K then sample randomly K class center in the observations.

Iterate until the within class variance criteria does not grow significantly any more

- **Allocation** For $i = 1, \cdots, n$, Alocate i to class k such that $d(x_i, g_k) \leq d(x_i, g_\ell)$ for all $\ell = 1, \cdots, K$

- **Estimation** Compute class centers g_k of K classes.

$$g_k = \frac{1}{N_k} \sum_{i \in C_k} x_i$$

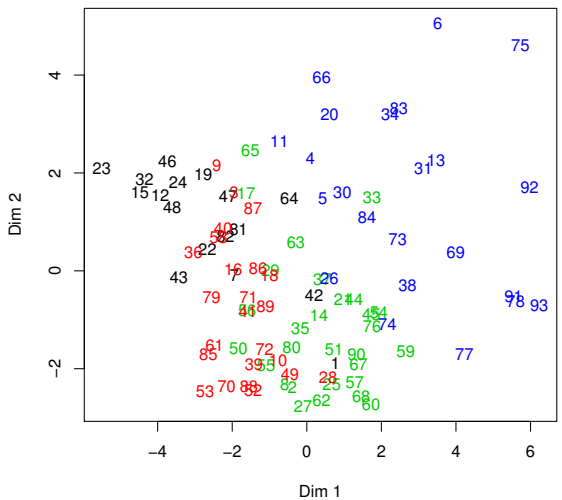
Example

- Data from french departments
- Choice of the number of classes according to BIC criteria

$$BIC = -\ln(\mathcal{L}) - k\ln(n)$$

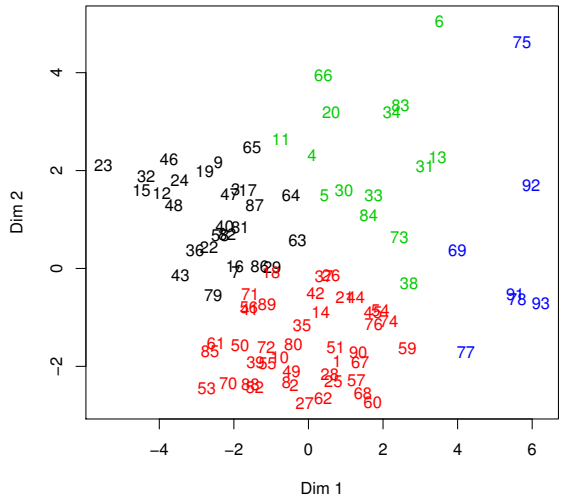
Number of classes	1	2	3	4	5
BIC	-2412	-2390	-2479	**-2505**	-2458

Gaussian mixture model



Soft frontiers.

kmeans



Frontiers are hyperplans.

Outline

8 Graphical models with latent variables
- When visible variables are continuous
- Mixture models
- When visible variables are discrete

Graphes with discrete variables

- As for continuous variables, a discrete Markovian network can include hidden edges.
- Let us denote again $X_{\mathcal{H}}$ the hidden variables and $X_{\mathcal{V}}$ the observed ones.
- The log-likelihood is given by

$$\begin{aligned}\ell(\boldsymbol{\Theta}) &= \sum_{i=1}^{n} \log P_{\boldsymbol{\Theta}}(X_{\mathcal{V}} = x_{i\mathcal{V}}) \\ &= \sum_{i=1}^{n} \left(\log \sum_{x_{\mathcal{H}} \in \mathcal{X}_{\mathcal{H}}} \exp\left(\theta_{jk} x_{ij} x_{ik} - \Phi(\boldsymbol{\Theta})\right) \right)\end{aligned}$$

- The sum over $x_{\mathcal{H}}$ means that the sum is over all the possible realizations of the hidden variables.
- The gradient is given by

$$\frac{\partial \ell(\boldsymbol{\Theta})}{\partial \theta_{jk}} = \widehat{E}_{\mathcal{V}} E_{\boldsymbol{\Theta}}(X_j X_k | X_{\mathcal{V}}) - E_{\boldsymbol{\Theta}}(X_j X_k)$$

Graph with discrete visible variables

- The gradien of the log-likelihood is

$$\frac{\partial \ell(\Theta)}{\partial \theta_{jk}} = \widehat{E}_{\mathcal{V}} E_{\Theta}(X_j X_k | X_{\mathcal{V}}) - E_{\Theta}(X_j X_k)$$

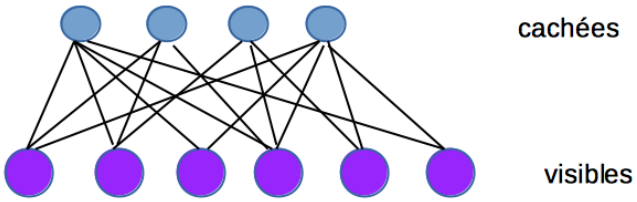
- The second term is an expectation given the model. It can by approximated by simulation with a Gibbs sampler.
- The first term is an empirical expectation conditionally to the visible edges.

$$E_{\Theta}(X_j X_k | X_{\mathcal{V}}) = \begin{cases} x_{ij} x_{ik} & \text{if } j, l \in \mathcal{V} \\ x_{ij} P_{\Theta}(X_k = 1 | X_{\mathcal{V}} = x_{i\mathcal{V}}) & \text{if } j \in \mathcal{V}, k \in \mathcal{H} \\ P_{\Theta}(X_j = 1, X_k = 1 | X_{\mathcal{V}} = x_{i\mathcal{V}}) & \text{if } j, k \in \mathcal{H} \end{cases}$$

- And two separate runs of Gibbs sampling are required ; the first to estimate $E_{\Theta}(X_j X_k)$ by sampling from the model, and the second to estimate $E_{\Theta}(X_j X_k | X_{\mathcal{V}} = x_{i\mathcal{V}})$.
- The learning task can have a very large computational costeven if the network is quite small.
  Indeed, expectations have to be computed at each iteration of the optimization algortihm. And each expectation computation needs the convergence of a Gibbs sampler...
- For some graphes with a particular architecture, the algorithms can be improved.

Boltzmann machines

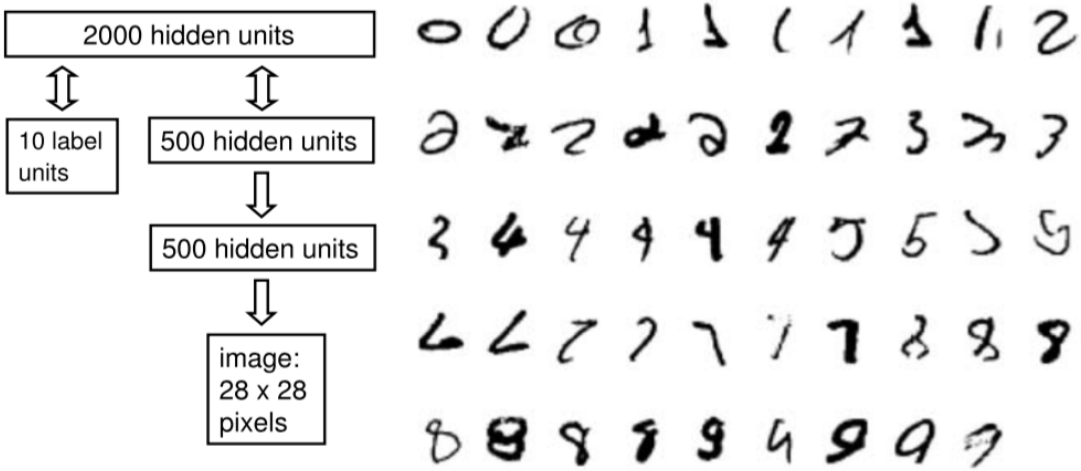
- The Restricted Boltzmann Machines are graphical models with a particular architecture inspired from neural networks. The units are organized in 2 layers, one is visible and the other one is hidden. Units belonging to the same layer are not connected to each other.



- This model is used (for instance, in deeplearning) to extract interesting features from images.
- Since the unit of a given layer are independent given the other variables, the Gibbs sampler can be simplified.
- Furthermore, it has been observed by Hinton (2006), that the Gibbs sampler does not need to converge to the stationary distribution to give the right direction to the optimization algorithm. His algorithm is called contrastive divergence.

Boltzmann machines

- Using contrastive divergence, it is possible to train an RBM to recognize hand-written digits from the MNIST dataset (LeCun et al., 1998)
- the RBM achieves an error rate of 1.9% on the test set without the two 500 units layers and 1.25% with the two 500 units layers
- The figure shows the nerwork architecture and some difficult examples which are correctly classify.



Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
   - Bayesian networks
   - Example in biology
   - Hidden Markov models
   - Factor analysis
   - State space model
10. Some concluding remarks

Outline

9 Directed graph models
- Bayesian networks
- Example in biology
- Hidden Markov models
- Factor analysis
- State space model

Bayesian networks

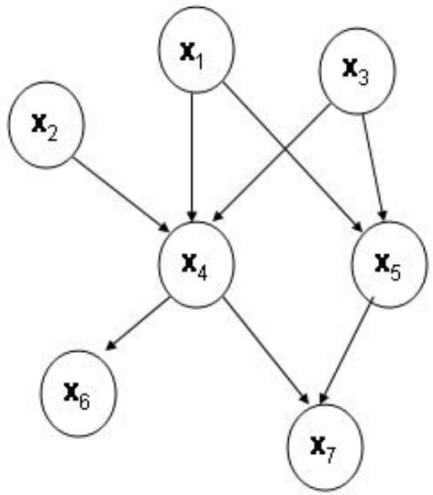
- Directed graph models are known as Bayesian networks or belief network.

$$p(x_1, \cdots, x_p) = \prod_j p(x_j | pa_j)$$

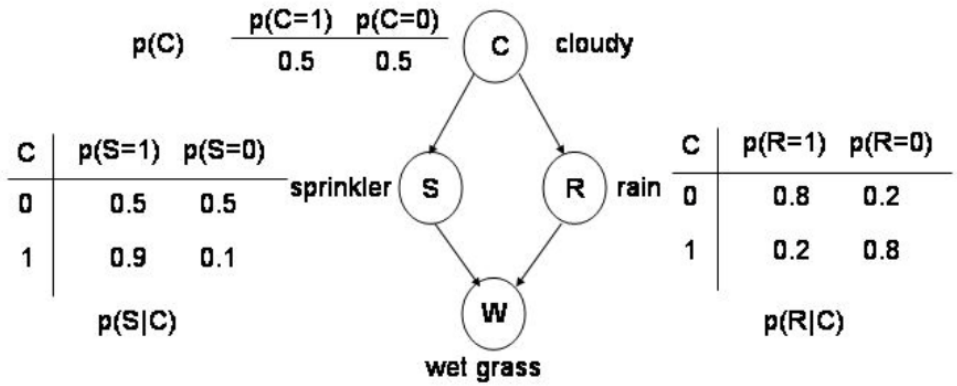
  where pa_j denotes the parents of x_j.
- For the graph below,

$$p(x_1, \cdots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

- The learning task reduces to the estimation of the marginal probabilities and conditional probabilities.



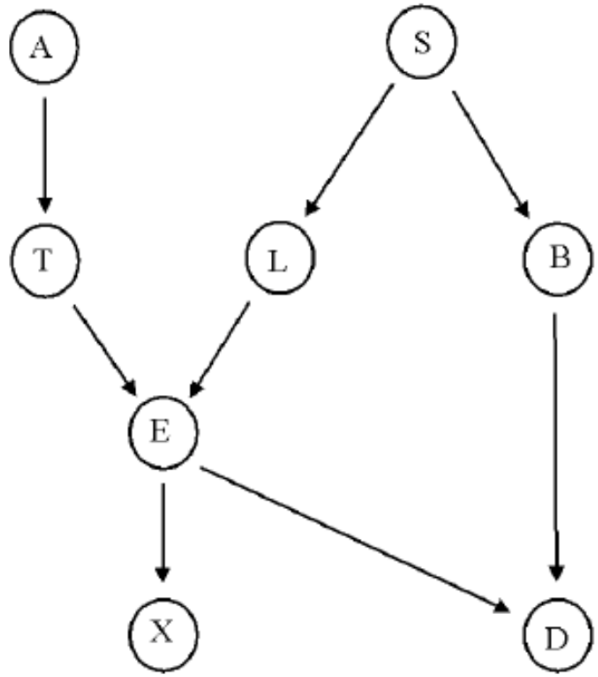
Bayesian networks, example



- The grass is wet can have 2 causes : rain or sprinkler.
- For example, $P(W = T|S = T, R = F) = 0.9$ et donc $P(W = F|S = T, R = F) = 0.1$.

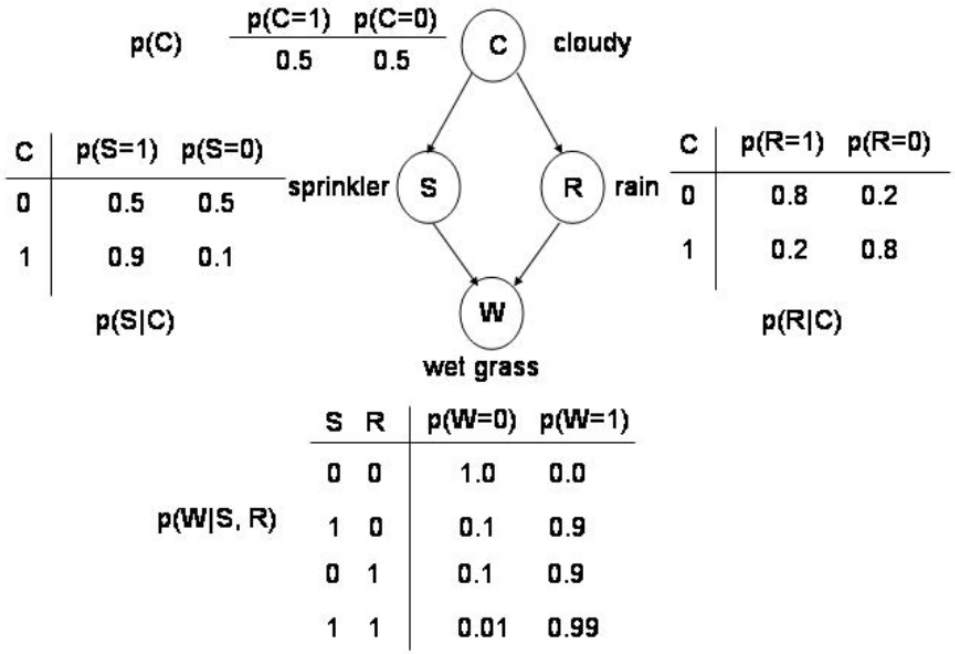
An other example

- `Asia Network'
  - A: trip to Asia
  - T: turbeculosis
  - S: smoking
  - L: Lung Cancer
  - B: Bronchitis
  - E: Turbeculosis/Lung Cancer
  - X-ray results
  - D: Dyspnea



- Given: X-rays, Dyspnea, patient went to Asia, patient smokes
- Wanted: posterior probability of Bronchitis

Bayesian networks, example



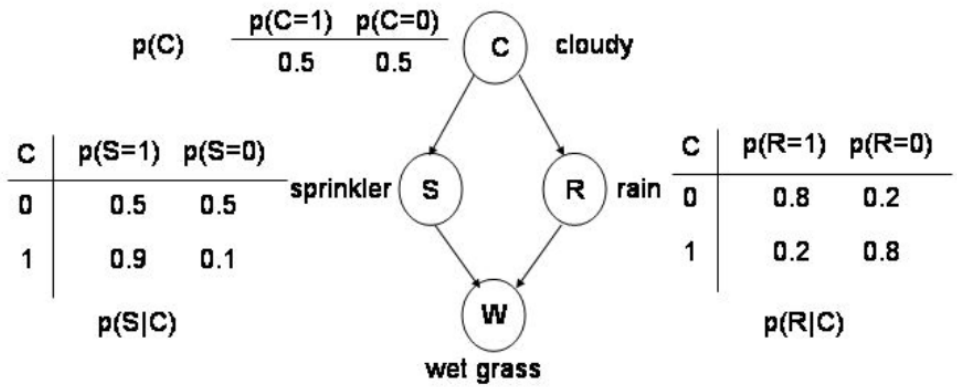
- From the Bayes formula,

$$P(C, S, R, W) = P(W|C, S, T)P(R|C, S)P(S|C)P(C).$$

- With the conditional dependencies properties, it simplifies to

$$P(C, S, R, W) = P(W|S, T)P(R|C)P(S|C)P(C).$$

Bayesian networks, example



p(W|S, R)

S	R	p(W=0)	p(W=1)
0	0	1.0	0.0
1	0	0.1	0.9
0	1	0.1	0.9
1	1	0.01	0.99

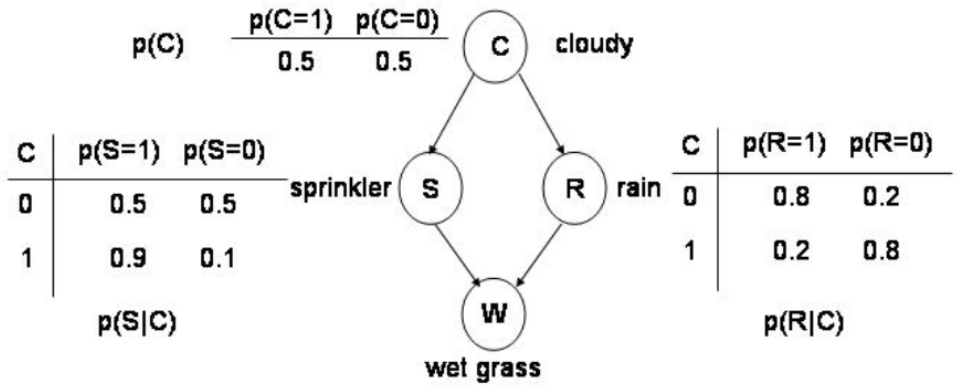
- What is the most probable cause of wet grass ?

$$P(S=1|W=1) = \frac{P(S=1,W=1}{P(W=1)} = \frac{\sum_{c,r} P(C=c,S=1,R=r,W=1)}{P(W=1)} = 0.28/0.65$$

$$P(R=1|W=1) = \frac{P(R=1,W=1}{P(W=1)} = \frac{\sum_{c,s} P(C=c,S=s,R=1,W=1)}{P(W=1)} = 0.46/0.65$$

- Rain is more probable than sprinkler.

Bayesian networks, example



p(W|S, R)

S	R	p(W=0)	p(W=1)
0	0	1.0	0.0
1	0	0.1	0.9
0	1	0.1	0.9
1	1	0.01	0.99

- Conditional probabilities can be computed, even if only a part of the variables is observed.
- For instance, if "wet grass" and "rain" are observed
- The posteriori probability of sprinkler on given wet grass and rain

$$P(S = 1|W = 1, R = 1) = 0.19$$

Bayesian networks, learning

- In Bayesian network the structure of the network is usually known.
- But some variables are hidden.
- Parameters are estimated by maximum likelihood.
- If some variables are hidden, EM algorithm is used.
- Algorithme EM

  Repeat until convergence
  1. Compute the posterior probabilities of the hidden variables given the observed ones and the paramter estimation obtain at the previous iteration.
  2. Estimation of the parameters by maximum likelihood with each observation weighted by its posterior probability in each group.

- The properties of conditionnal dependancies usually allow to factorized the optimization problem in smaller sub problems.

Directed graph models in the framework of linear models

- For multivariate Gaussian distribution, the Structural Equation Model (SEM) give a common framework for directed and undirected (or bidirected) graph models.
- Assume that the joint distribution of $X = (X_1, \cdots, X_p)^T$ is multivariate Gaussian.
- Each X_j is a linear function of $(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_p)^T$ and a stochastic noise term ϵ_j

$$X = a_0 + A^T X + \epsilon$$

  where $\epsilon \sim \mathcal{N}(0, Q)$.
- $X \sim \mathcal{N}((\mathbf{I} - A)^{-T} a_0, (\mathbf{I} - A)^{-T} Q (\mathbf{I} - A)^{-1})$
- The graph associated to this model will contain
  - the directed edge $k \to j$ when A_{kj} is required to be not zero,
  - the bidirected edge $k \leftrightarrow j$ when q_{kj} is required to be not zero.

Outline

9 Directed graph models
- Bayesian networks
- Example in biology
- Hidden Markov models
- Factor analysis
- State space model

Proteine data

- Following the work of Friedman et al. (2000), the expression level or the allele frequency of each gene is associated with one node. In addition, we can include additional nodes denoting other attributes that affect the system, such as experimental conditions, temporal indicators, and exogenous cellular conditions.
- As a result, we can model in a single, comprehensive BN both the biological mechanisms we are interested in and the external conditions influencing them at the same time.
- BNs were, for instance, used to represent complex direct and indirect relationships among multiple interacting molecules while accommodating biological noise.
- Les us consider Sachs et al. (2005) protein-signalling data.
- The data consist in the simultaneous measurements of 11 phosphorylated proteins and phospholipids derived from thousands of individual primary immune system cells, subjected to both general and specific molecular interventions.
- We will consider only the 853 data manipulated with general interventions

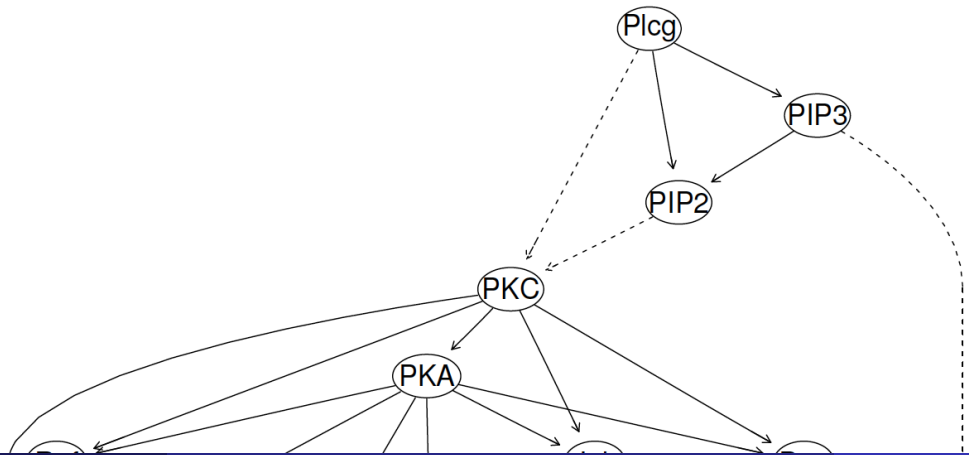
```
> library(bnlearn)
> sachs <- read.table("sachs.data.txt", header = TRUE)
> head(sachs)
   Raf   Mek  Plcg  PIP2  PIP3   Erk  Akt PKA   PKC  P38  Jnk
1 26.4 13.20  8.82 18.30 58.80  6.61 17.0 414 17.00 44.9 40.0
2 35.9 16.50 12.30 16.80  8.13 18.60 32.5 352  3.37 16.5 61.5
3 59.4 44.10 14.60 10.20 13.00 14.90 32.5 403 11.40 31.9 19.5
4 73.0 82.80 23.10 13.50  1.29  5.83 11.8 528 13.70 28.6 23.1
5 33.7 19.80  5.19  9.73 24.80 21.10 46.1 305  4.66 25.7 81.3
6 18.8  3.75 17.60 22.10 10.90 11.90 25.7 610 13.70 49.1 57.8
```

Sachs' data

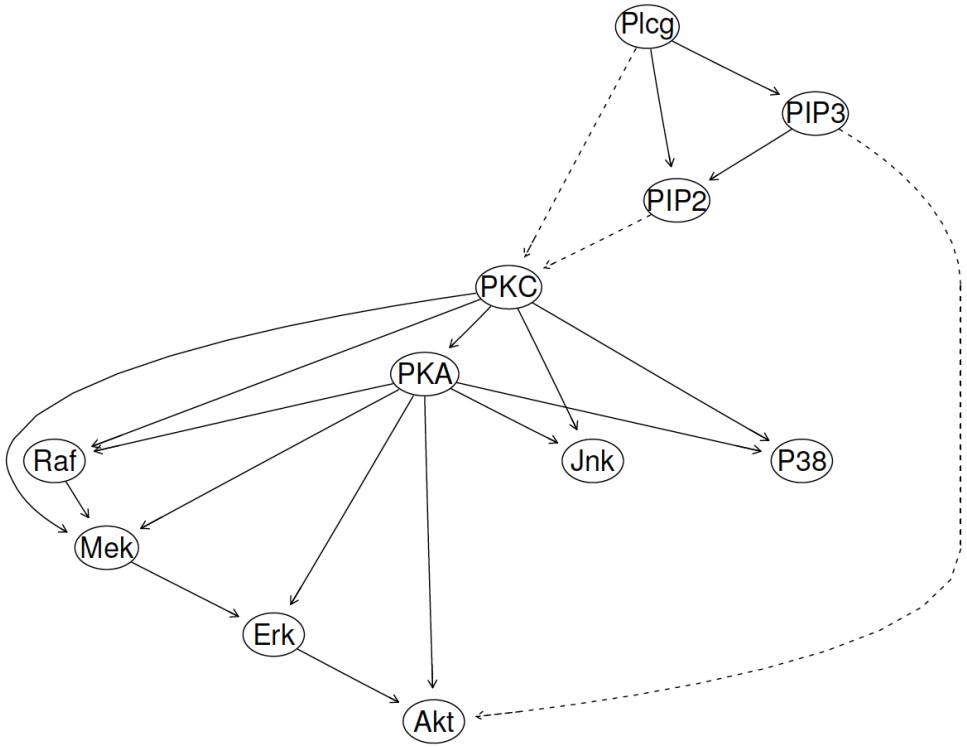
- The data consist in the simultaneous measurements of 11 phosphorylated proteins and phospholipids derived from thousands of individual primary immune system cells, subjected to both general and specific molecular interventions.
- We will consider only the 853 data manipulated with general interventions

```
> library(bnlearn)
> sachs <- read.table("sachs.data.txt", header = TRUE)
> head(sachs)
   Raf   Mek  Plcg  PIP2  PIP3   Erk  Akt PKA   PKC  P38  Jnk
1 26.4 13.20  8.82 18.30 58.80  6.61 17.0 414 17.00 44.9 40.0
2 35.9 16.50 12.30 16.80  8.13 18.60 32.5 352  3.37 16.5 61.5
3 59.4 44.10 14.60 10.20 13.00 14.90 32.5 403 11.40 31.9 19.5
4 73.0 82.80 23.10 13.50  1.29  5.83 11.8 528 13.70 28.6 23.1
5 33.7 19.80  5.19  9.73 24.80 21.10 46.1 305  4.66 25.7 81.3
6 18.8  3.75 17.60 22.10 10.90 11.90 25.7 610 13.70 49.1 57.8
```

- The data are continuous, as they represent the concentration of the molecules under investigation.



Sachs' data



Protein-signalling network from Sachs et al. (2005). Signalling pathways that are known from literature but were not captured by the BN are shown with a dashed line.

- Exercice $\rightarrow$ *Rmarkdown*

Outline

9 Directed graph models
- Bayesian networks
- Example in biology
- Hidden Markov models
- Factor analysis
- State space model

Hidden Markov models

- HMM are models for time series with regimes : speech and writing recognition, dynamical systems, meteorological time series, etc.

- Definition
  $\{X_t\} = \{S_t, Y_t\} \in \{\mathbf{S} \times \mathbf{Y}\}$ with $\{S_t\}$ hidden and
  - $P(S_t|S_0 = s_0, \cdots, S_{t-1} = s_{t-1}, Y_1 = y_1, \cdots, Y_{t-1} = y_{t-1}) = P(S_t|S_{t-1} = s_{t-1})$
  - $P(Y_t|S_0 = s_0, \cdots, S_t = s_t, Y_1 = y_1, \cdots, Y_{t-1} = y_{t-1}) = P(Y_t|S_t = s_t)$

Recall
Mixture model (e.g. mixture of Gaussian)

$$\begin{array}{rc} \textit{discrete} & S \\ & \downarrow \\ \textit{continuous} & Y \end{array}$$

$$\begin{array}{lccccccccc} \textit{State} & \cdots & \rightarrow & S_{t-1} & \rightarrow & S_t & \rightarrow & S_{t+1} & \rightarrow & \cdots \\ & & & \downarrow & & \downarrow & & \downarrow & & \\ \textit{Observation} & \cdots & & Y_{t-1} & & Y_t & & Y_{t+1} & & \cdots \end{array}$$

- Parametrization of an HMM
  - $p_\theta(s_t|s_{t-1})$ transition probabilities
  - $p_\theta(y_t|s_t)$ emission probabilities

Learning and inference

- Likelihood

$$\begin{aligned} p_\theta(Y_1 = y_1, \cdots, Y_t = y_T) &= \int_{\mathbf{S}^{\mathbf{T}+\mathbf{1}}} p_\theta(s_0)\Pi_{t=1}^T p_\theta(s_t|s_{t-1})p_\theta(y_t|s_t)ds_0 ds_1 ... ds_T \\ &= \Pi_{t=1}^T \int_{\mathbf{S}} p_\theta(y_t|s_t)p_\theta(s_t|y_1, \cdots, y_{t-1})ds_t \end{aligned}$$

- *Prediction* : compute $p_\theta(s_t|y_1, \cdots, y_{t-1})$

$$p_\theta(s_t|y_1, \cdots, y_{t-1}) = \int_{\mathbf{S}} p_\theta(s_t|s_{t-1})p_\theta(s_{t-1}|y_1, \cdots, y_{t-1})ds_{t-1}$$

- *Filtering* : compute $p_\theta(s_t|y_1, \cdots, y_t)$

$$p_\theta(s_t|y_1, \cdots, y_t) \propto p_\theta(y_t|s_t)p_\theta(s_t|y_1, \cdots, y_{t-1})$$

- *Smoothing* : compute $p_\theta(s_t|y_1, \cdots, y_t, \cdots, y_T)$

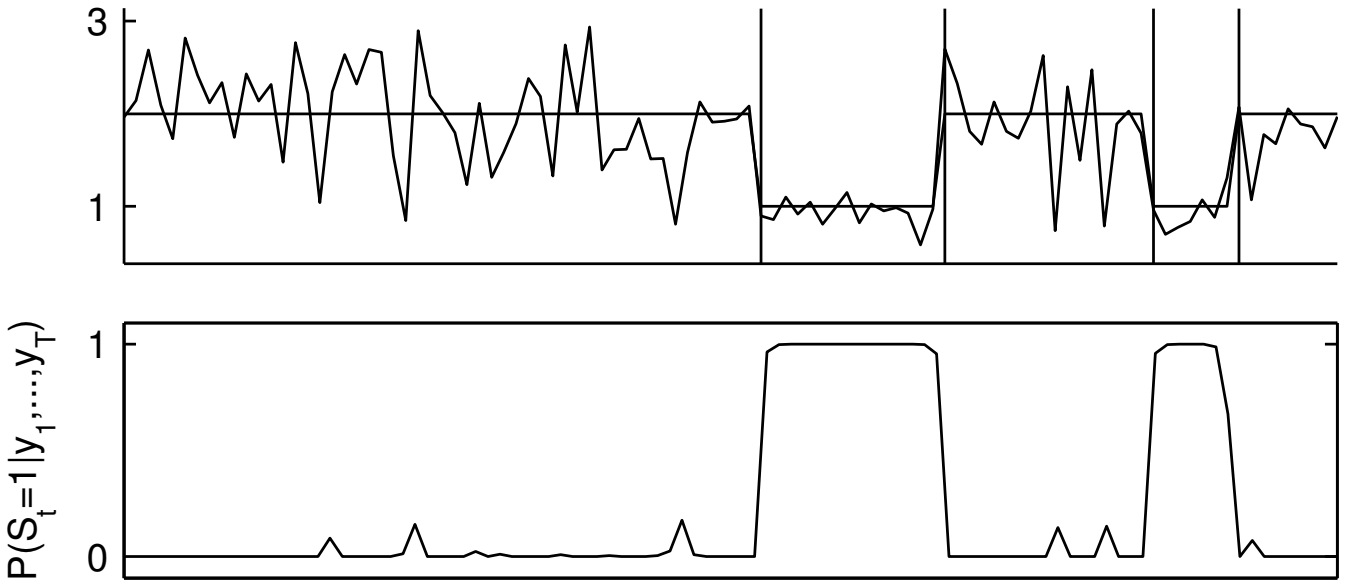
$$\begin{aligned} &p_\theta(s_t|y_1, \cdots, y_T) \\ &= p_\theta(s_t|y_1, \cdots, y_t) \int_{\mathbf{S}} \frac{p_\theta(s_{t+1}|s_t)}{p_\theta(s_{t+1}|y_1, \cdots, y_t)} p_\theta(s_{t+1}|y_1, \cdots, y_T)ds_{t+1} \end{aligned}$$

HMM with finite state space

- Example of simulation

$$Y_t = m^{(S_t)} + \sigma^{(S_t)} W_t$$

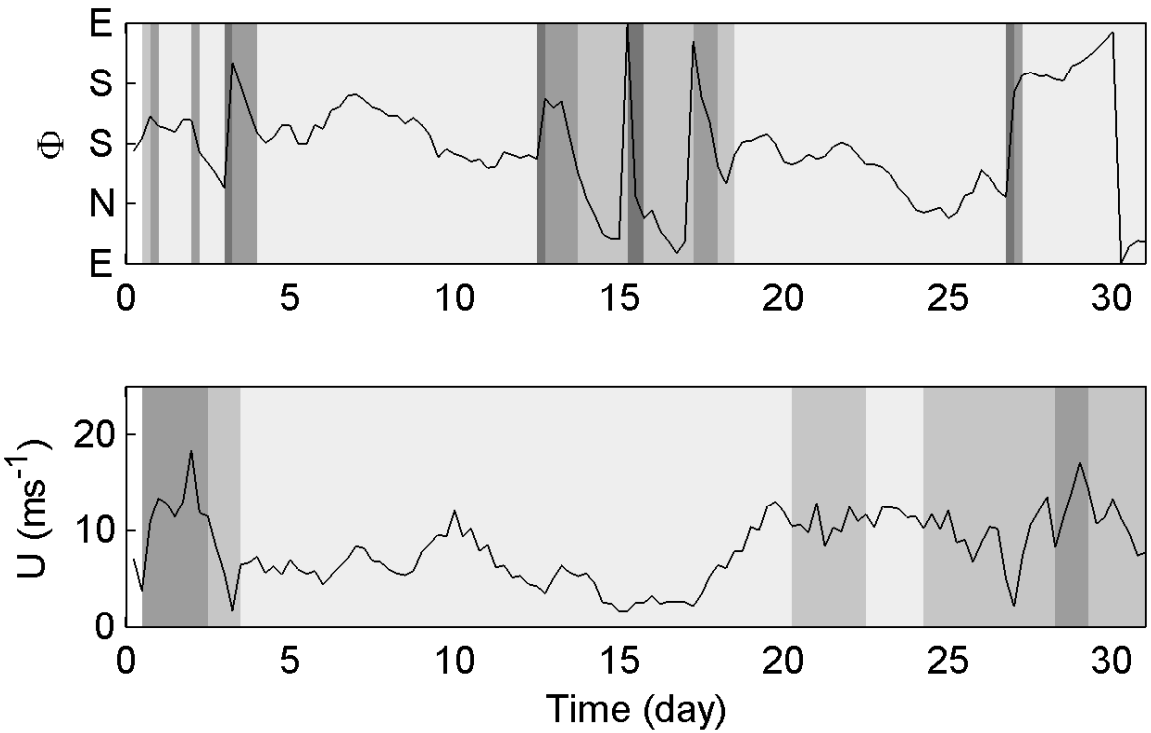
  - $W_t \sim iid\mathcal{N}(0,1)$
  - $m^{(1)} = 1$, $m^{(2)} = 2$, $\sigma^{(1)} = 0.2$, $\sigma^{(2)} = 0.5$, $Q = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$



- For an HMM, forward-backward algorithm allow to compute the log-likelihood and solve the filtering and smoothing problems.
- For likelihood maximization, EM algortihms or gradient methods can to be implemented.

Example : modeling of wind time series

- Motivations : Meteo, Pollution, Fatigue of structures, Covariate (waves)
- Data (Ouessant, 44 ans, Δt = 6 h, january month)
- Regimes are observed, but we remark that the time seriess are smoother tahn the ones of HMM.



Example : modeling of wind time series

- It suggests a graph with connected observations.

$$\begin{array}{lccccccccc} \textit{State} & \cdots & \to & S_{t-1} & \to & S_t & \to & S_{t+1} & \to & \cdots \\ & & & \downarrow & & \downarrow & & \downarrow & & \\ \textit{Observation} & \cdots & \to & Y_{t-1} & \to & Y_t & \to & Y_{t+1} & \to & \cdots \end{array}$$

- It leads to a Markov switching autoregressive model e.g. in each regime the time series are described by an auto-regressive process

$$Y_t = \alpha_0^{(s)} + \alpha_1^{(s)} Y_{t-1} + \Sigma^{(s)} \epsilon_t$$

- Parameters are $\alpha_0^{(s)}$, $\alpha_1^{(s)}$, $\alpha_1^{(s)}$ for each regime s and the transition matrix of the Markov chain.
- Estimation used an EM algorithm.

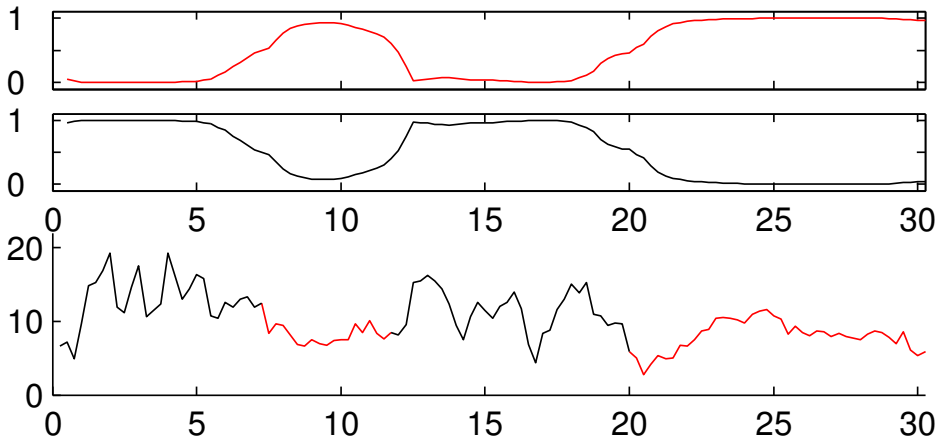
Example : modeling of wind time series

- The obtain model is interpretable (rresults for january in Ouessant)

$$Y_t = \begin{cases} 1.46 + 0.79\,Y_{t-1} + 1.37\epsilon_t & (S_{t-1} = 1) \\ 2.24 + 0.77\,Y_{t-1} + 2.4\epsilon_t & (S_{t-1} = 2) \end{cases}$$

  - **Regime 1** : variability is lower , more stable regime, anticyclonic conditions
  - **Regime 2** :variability is higher, cyclonic conditions
  - Transition matrix : $\begin{bmatrix} 0.98 & 0.02 \\ 0.03 & 0.97 \end{bmatrix}$, stationary distribution $\begin{bmatrix} 0.40 \\ 0.60 \end{bmatrix}$

- Smoothing probabilities $P[S_t | y_1, ..., y_T]$ for Jan. 2000



Outline

9 Directed graph models
- Bayesian networks
- Example in biology
- Hidden Markov models
- Factor analysis
- State space model

Generalities

- The factor analysis is a simple latent variable model. It can be view as an unsupervized linear regression model.
- The latent variable is assumed to lie in a lower-dimensional linear subspace of the space of the observed variables.
- The model is similar to the one of mixture except that the hidden variables are now continuous (usually assumed to have Gaussian distribution).
- Multivariate Gaussian distribution (recall)

$$p\left(\mathbf{x}_1, \mathbf{x}_2; \mu, \mathbf{\Sigma}\right) = \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{21} \\ \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{22} \end{pmatrix}\right)$$

then

$$p(\mathbf{x}_1) = \mathcal{N}\left(\mu_1, \mathbf{\Sigma}_{11}\right)$$

and

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}\left(\mathbf{m}_{1|2}, \mathbf{V}_{1|2}\right)$$

with

$$\mathbf{m}_{1|2} = \mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \mu_{22}), \quad \mathbf{V}_{1|2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$

Factor analysis model

- The observed variables **Y** follow a Gaussian distribution conditionally to the hidden variables **X**.

$$p(\mathbf{x}) = \mathcal{N}(0, I)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mu + \Lambda\mathbf{x}, \Psi)$$

  Λ is the *loading matrix*.

- The second probability correspond to the model

$$\mathbf{y} = \mu + \Lambda\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Psi)$$

- Ψ is a diagonal covariance matrix. It means that the components of **y** are independent conditionally to **x**.
- It is easy to deduce the expression of the parameters of - the distribution of the complete data
  - the marginal distributions
  - the conditional distributions .
- Learning is performed by an EM algorithm.

EM algorithm, Factor analysis model

- For the **E-step**, one need to compute $Q_i(\mathbf{x}^{(i)}) = p(\mathbf{x}^{(i)}|(\mathbf{y}^{(i)}; \mu, \Lambda, \Psi)$
  This probability is Gaussian with

$$\mathbf{m}_{\mathbf{x}|\mathbf{y}} = \mu_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}(\mathbf{y} - \mu), \quad \mathbf{V}_{\mathbf{x}|\mathbf{y}} = \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}$$

- For the **M-step**, we need to maximize the expectation of the log-likelihood of the complete data

$$\ell(\mu, \Lambda, \Psi) = \sum_{i=1}^{n} E_{\mathbf{x}^{(i)} \sim Q_i} \left[\log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \mu, \Lambda, \Psi) + \log \left(p(\mathbf{x}^{(i)}) - Q_i(\mathbf{x}^{(i)}) \right) \right]$$

  with respect to μ, Λ and Ψ.

- Only the first term depends on Λ. Gradient of this term equals to zero leads to

$$\Lambda = \left(\sum_{i=1}^{n} (\mathbf{y}^{(i)} - \mu) E_{\mathbf{x}^{(i)} \sim Q_i} \left[\mathbf{x}^{(i)T} \right] \right) \left(\sum_{i=1}^{n} E_{\mathbf{x}^{(i)} \sim Q_i} \left[\mathbf{x}^{(i)} \mathbf{x}^{(i)T} \right] \right)^{-1}$$

  One recognize a form close to the estimation of the parameters of a linear model.

EM algorithm, Factor analysis model (continuing)

- The expectations

$$E_{\mathbf{x}^{(i)} \sim Q_i}\left[\mathbf{x}^{(i)}\right] = \mu_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}} = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(\mathbf{y}^{(i)} - \mu)$$

$$E_{\mathbf{x}^{(i)} \sim Q_i}\left[\mathbf{x}^{(i)}\mathbf{x}^{(i)^T}\right] = \mu_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}}\mu^T_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}} + I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda$$

  In practice the estimation of the parameters obtained at the previous iterations of the EM algorithms are used to update Λ.

- The estimation of Ψ is the diagonal of Φ where

$$\Phi = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}^{(i)}\mathbf{y}^{(i)^T} - \mathbf{y}^{(i)}\mu_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}}\Lambda^T - \Lambda\mu_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}}\mathbf{y}^{(i)^T} + \Lambda(\mu_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}}\mu^T_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}} + \Sigma_{\mathbf{x}^{(i)}|\mathbf{y}^{(i)}})\Lambda^T$$

- Estimating μ is trivial

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}^{(i)}$$

Outline

9 Directed graph models
- Bayesian networks
- Example in biology
- Hidden Markov models
- Factor analysis
- State space model

State space model

- The state space model (SSM) is a chain of factor models.
- It has the same structure as the HMM, but the hidden variables are now continuous.

$$\begin{array}{lccccccccc} \textit{State} & \cdots & \rightarrow & X_{t-1} & \rightarrow & X_t & \rightarrow & X_{t+1} & \rightarrow & \cdots \\ & & & \downarrow & & \downarrow & & \downarrow & & \\ \textit{Observation} & \cdots & & Y_{t-1} & & Y_t & & Y_{t+1} & & \cdots \end{array}$$

- A standard form for SSM is

$$\begin{cases} X_0 \sim p(x_0) & \\ X_t = f(X_{t-1}) + \eta_t & (\textit{hidden}) \\ Y_t = h(X_t) + \epsilon_t & (\textit{observed}) \end{cases}$$

- In the linear (Gaussian) case, it reduces to

$$\begin{cases} X_0 \sim \mathcal{N}(\mu_0, Q_0) \\ X_t = MX_{t-1} + \eta_t \\ Y_t = HX_t + \epsilon_t \end{cases}$$

with η_t and ϵ_t centered Gaussian noises independant from the intial condition, with variances Q and R. R is inversible.

- Parametrization of an HMM
  - $p_\theta(x_t|x_{t-1})$ transition probabilities : $X_t|X_{t-1} = x_{t-1} \sim \mathcal{N}(Mx_{t-1}, Q)$
  - $p_\theta(x_t|x_t)$ emission probabilities : $Y_t|X_t = x_t \sim \mathcal{N}(Hx_t, R)$

Inference in SSM

- *Prediction* : compute $p_\theta(x_t|y_1,\cdots,y_{t-1})$

$$p_\theta(x_t|y_1,\cdots,y_{t-1}) = \int_{\mathcal{X}} p_\theta(x_t|x_{t-1})p_\theta(x_{t-1}|y_1,\cdots,y_{t-1})ds_{t-1}$$

- *Filtering* : compute $p_\theta(x_t|y_1,\cdots,y_t)$

$$p_\theta(x_t|y_1,\cdots,y_t) \propto p_\theta(y_t|x_t)p_\theta(x_t|y_1,\cdots,y_{t-1})$$

- *Smoothing* : compute $p_\theta(x_t|y_1,\cdots,y_t,\cdots,y_T)$

$$\begin{aligned} &p_\theta(x_t|y_1,\cdots,y_T) \\ &= p_\theta(x_t|y_1,\cdots,y_t)\int_{\mathbf{S}} \frac{p_\theta(x_{t+1}|s_t)}{p_\theta(x_{t+1}|y_1,\cdots,y_t)} p_\theta(x_{t+1}|y_1,\cdots,y_T)dx_{t+1} \end{aligned}$$

Kalman filter

- Filtering problem : estimation of $p_\theta(x_t|y_1,\cdots,y_t)$
- In the linear (Gaussian) case, it reduces to the computation of $\widehat{X}_t = x_{t|t} = E[X_t|Y_{1:t} = y_{1:t}]$ and $P_{t|t} = Var(X_t|Y_{1:t} = y_{1:t})$
- The **Kalman filter** is the Best Linear Unbiaises Estimator (BLUE) of X_t given a sequence of observations $\{y_1,\cdots,y_t\}$

$$x_{t|t} = (I - KH)x_{t|t-1} + Ky_t \text{ with the predicted state } x_{t|t-1} = Fx_{t-1|t-1}$$

- We remark that

$$Var(x_{t|t}) = P_{t|t} = x_{t|t}x_{t|t}^T = P_{t|t-1} - 2P_{t|t-1}H^TK^T + K(R + HP_{t|t-1}H^T)K^T$$

and the Kalman gain is defined by

$$K^* = \arg\min_{K\in\mathbb{R}^{d\times m}} \left(P_{t|t-1} - 2P_{t|t-1}H^TK^T + K(R + HP_{t|t-1}H^T)K'^T\right)$$

which leads to

$$K^* = P_{t|t-1}H^T(R + HP_{t|t-1}H^T)^{-1}$$

- The filtering covariance $P_{t|t}$ is also computed by the Kalman filter.

Filtre de Kalman

Filtre de Kalman

- Initialisation
  $X_0 \sim \mathcal{N}(0, Q_0)$
- Prediction
  $x_{t|t-1} = F x_{t-1|t-1}$
  $P_{t|t-1} = F P_{t-1|t-1} F' + Q$
- Correction
  $K_t = P_{t|t-1} H^T (H P_{t|t-1} H^T + R)^{-1}$
  $x_{t|t} = x_{t|t-1} + K_t (y_t - H x_{t|t-1})$
  $P_{t|t} = (I - K_t H) P_{t|t-1}$

Ref : Kalman (1966)

Remark : the Kalman gain is a variance "rapport", which measures the confidence according to the state approximations (prediction of the model and observation).

Kalman filter, example

- A example is described in ...
- The robot has a state $x = (p, u)$
- The robot also has a GPS sensor, which is accurate to about 10 meters, which is good, but it needs to know its location more precisely than 10 meters.
- We might also know something about how the robot moves : It knows the commands sent to the wheel motors, and its knows that if it's headed in one direction and nothing interferes, at the next instant it will likely be further along that same direction.
- But of course it doesn't know everything about its motion : It might be buffeted by the wind, the wheels might slip a little bit, or roll over bumpy terrain ; so the amount the wheels have turned might not exactly represent how far the robot has actually traveled, and the prediction won't be perfect.

- Position sensor is combined to the stateupdate to give and estimation of the position and the velocity of the robot.

Outline

1. Introduction
2. Fondamentals of graphical models
3. Undirected graph models
4. Examples of discrete undirected graph models
5. Examples of undirected graph models for continuous observations
6. Learning (continuous) undirected graph models
7. Learning (discrete) undirected graph models
8. Graphical models with latent variables
9. Directed graph models
10. Some concluding remarks

Conluding remarks

- Probability theory povides the glue whereby the parts are combined, ensuring that the system as a whole consistent, and providing ways to interface models to data.
- The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- Many of the classical multivariate probabilistic systems studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism.
- The graphical model framework provides a way to view all of these systems as instances of a common underlying formalism.

- If you want to go further : lectures of Eric Xing