MAP 531 : Statistics refresher

15 septembre 2017

Organisation

Lectures

Stéphanie Allassonnière, Université Paris Descartes, Ecole Polytechnique
stephanie.allassonniere@polytechnique.edu

Training sessions

Stéphanie Allassonnière and Elodie Vernet
(elodie.vernet@polytechnique.edu)

Tutorat

Antoine Havet, Thomas Kerdreux and Thomas Lartigue

Organisation

Lectures

September 25, 27, 29 : 9h - 12h
September 26 : 13h30 - 15h30
Ocober 6, 13, 20 : 9h - 12h

Training sessions

September 25, 27, 29 : 13h30 - 15h30
September 26 : 15h30 - 17h30
October 6, 13, 20 : 14h30 - 16h30

Tutorat

October 6, 13, 20 : 13h30 - 14h30 and 16h30 - 17h30
October 11, 13, 16, 20, 23 : 17h15 - 18h15

Presentation of the lectures

- Introduction to statistical modeling.
- Estimation.
- Confidence intervals.
- Introduction to hypothesis testing.
- Hypothesis testing for r.v. distributions

Statistic

Statistics is :

- collecte data,
- process them,
- analyse them,
- interpret the results
- and present them in order to make it accessible to everybody.

Nowadays : statistics is part of Data Science (together with computer science as well)

Example : Medical applications

- "Omic" data : gene regulation networks, genetic factors related to diseases, etc..
- Medical study : efficiency of treatments, prediction of treatment outcome, pharmacokinetic, etc,..

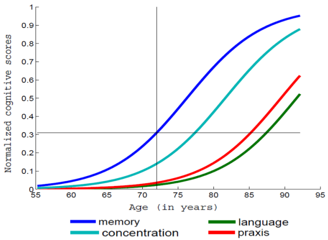
•Neuropsychological tests ADAS-Gog from ADNI

•248 subjects who converted from MCI to AD

•6 time-points per subjects on average (min 3, max 11)

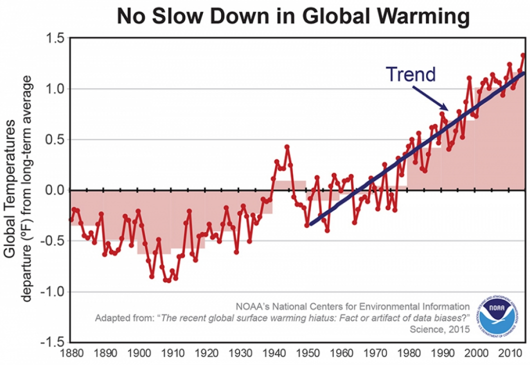
•Data points $y_{ij} \in]0,1[^4$ with propagation logistic model

The average trajectory of data changes



Example : Environnement

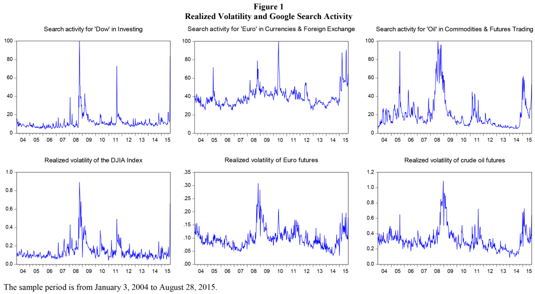
- Geophysics :
  weather forcasts, climatology, pollution forcast, etc..
- Ecologie :
  ecosystem evolution, population survival, ressources management, etc..



Contrary to much recent discussion, the latest corrected analysis shows that the rate of global warming has continued, and there has been no slow down.

Example : Insurance and finance

- Finance : modeling and prediction of financila asset, etc..
- Insurance : risk evaluation, etc..

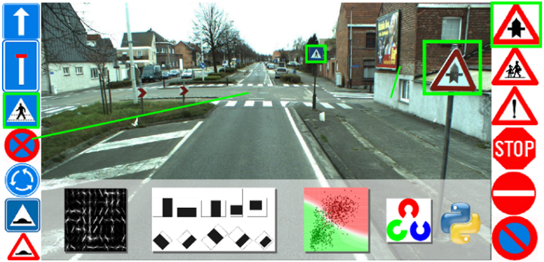


Example : Marketing

- Marketing : advertisement efficiency, "targeting", etc...
- Social networks

Example : Statistical learning

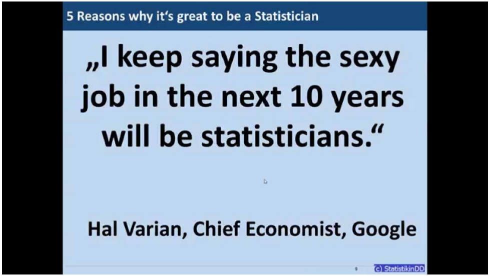
- Shape (or voice) recognition
- Computer vision
- Automatic translation



Forbes ranking 10 Best Jobs for 2016

`http://www.forbes.com/sites/karstenstrauss/2016/04/14/the-best-jobs-in-2016/`

- 1 Data Scientist
- 2 Statistician
- 4 Mathematician
- 6 Actuary



Statistical issue

- Starting point : observations

$$(x_1, \ldots, x_n).$$

Statistical issue

- Starting point : observations

$$(x_1, \ldots, x_n).$$

- Statistical modeling :
  - the observations is a the realisation of a random vector

$$Z = (X_1, \ldots, X_n) \ , \quad Z(\omega) = (X_1(\omega), \ldots, X_n(\omega)) = (x_1, \ldots, x_n) \ .$$

  - Model has to take into account the variability of the observations
    - noise (from measures or simply because it is a model)
    - sampling
    - individual effects
    - etc..

Statistical issue

- The probability distribution of

$$Z = (X_1, \ldots, X_n)$$

is partilly known : it summarises our prior knowledge of the phenomenon.

Statistical issue

- The probability distribution of

$$Z = (X_1, \ldots, X_n)$$

 is partilly known : it summarises our prior knowledge of the phenomenon.
- Mathematically speaking : we assume that the distribution of Z belongs to a family $\mathcal{P}$ of probability distributions.

Statistical issue

- The probability distribution of

$$Z = (X_1, \ldots, X_n)$$

is partilly known : it summarises our prior knowledge of the phenomenon.

- Mathematically speaking : we assume that the distribution of Z belongs to a family $\mathcal{P}$ of probability distributions.
- Question : Only from the observations

$$\mathtt{x}_1, \ldots, \mathtt{x}_\mathtt{n}$$

and the model $\mathcal{P}$, try to increase our knowledge of the distribution of Z.

Difference (one !) between Statistics and Probability

- Probability : the probability distribution (assumed to be) known...
  - Given this $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$ and a random vector $Z = (X_1, \ldots, X_n)$, one can compute several quantities such as :

$$\mathbb{P}(f(Z) \geq c), \quad \mathbb{E}[f(Z)].$$

Difference (one !) between Statistics and Probability

- Probability : the probability distribution (assumed to be) known...
  - Given this $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$ and a random vector $Z = (X_1, \ldots, X_n)$, one can compute several quantities such as :

$$\mathbb{P}(f(Z) \geq c), \quad \mathbb{E}[f(Z)].$$

- Statistics : Provide methods to solve the inverse problem :
  - Given one (or more) realisation of the radom vector $Z = (X_1, \ldots, X_n)$, determine characteristics of its law. This is called statistical inference.

Challenges :

From the observations and the model :

- Estimate : give an approximation of quantities related to the distribution (for example : its mean and variance or more complex ones) and evaluate the estimation erreur (called confidence intervals).

Challenges :

From the observations and the model :

- Estimate : give an approximation of quantities related to the distribution (for example : its mean and variance or more complex ones) and evaluate the estimation erreur (called confidence intervals).

- Test an hypothesis related to the distribution (example : $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$ can we say that $\mathbb{P} \in \mathcal{P}_0$?)

Challenges :

From the observations and the model :

- Estimate : give an approximation of quantities related to the distribution (for example : its mean and variance or more complex ones) and evaluate the estimation erreur (called confidence intervals).

- Test an hypothesis related to the distribution (example : $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$ can we say that $\mathbb{P} \in \mathcal{P}_0$?)

- Prefict a new outcome and evaluate the prediction error.

Challenges

Model complexity depends on :

- the studied phenomenon,

Challenges

Model complexity depends on :

- the studied phenomenon,
- Our knowledge (a priori)
  - may require the use of PDEs (ex : physics)
  - complex dependencies (ex : “omic”),

Challenges

Model complexity depends on :

- the studied phenomenon,
- Our knowledge (a priori)
  - may require the use of PDEs (ex : physics)
  - complex dependencies (ex : "omic"),
- inference complexity we can deal with
  - Quality of information sources and sampling (ex : MRIs, astronomic signals)
  - Inhomogeneous data to consider all together

Survey

One of the simplest statistical model

- population with N subjects
- Each subject decides either A or B
- $N\theta$ vote for A.
- the proportion $\theta \in \Theta = [0, 1]$ is unknown.
- N is extremly large.
- $\rightarrow$ Survey : sample $n \ll N$ subjects among this population.

Statistical model

- Population of size N.
- Sample of size n.
- Observations : 1 corresponds to A and 0 to B.
- Probability space : $\Omega = \{0,1\}^n$ with the corresponding $\sigma-$algebra.
- Observations : $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ one realisation of $Z = (X_1, \ldots, X_n)$, where X_i is the result of the i^{th} subject :

Proportion estimation

- Looks natural to "count" the number $\hat{\theta}_n$ of subject whose results are A in the sample

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(X_i) \, .$$

Proportion estimation

- Looks natural to“count” the number $\hat{\theta}_n$ of subject whose results are A in the sample

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(X_i) \; .$$

- Such a function of the observation is called an estimator [Any measurable function of the observations is an estimator !].

Proportion estimation

- Looks natural to “count” the number $\hat{\theta}_n$ of subject whose results are A in the sample

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(X_i) .$$

- Such a function of the observation is called an estimator [Any measurable function of the observations is an estimator !].

- The image distribution $\mathbb{P}_\theta$ by the statistic $n\hat{\theta}_n$ is, for all $\theta \in \Theta = [0, 1]$

$$\mathbb{P}_\theta(n\hat{\theta}_n = k) = \frac{\binom{\lfloor N\theta \rfloor}{k} \binom{N - \lfloor N\theta \rfloor}{n-k}}{\binom{N}{n}} .$$

This is the hypergéometric law.

Survey

The distribution of $(x_1, \ldots, x_n) \in \{0,1\}^n$ is given by

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}_\theta(X_1 = x_1)\mathbb{P}_\theta\left(X_2 = x_2 \,|\, X_1 = x_1\right) \times \mathbb{P}_\theta\left(X_n = x_n \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) \, .$$

$\rightarrow$ depends on $\theta \in \Theta$.

Survey

- Whole population : N, $N\theta$ for A, $(N - N\theta)$ for B

$$\mathbb{P}_\theta(X_1 = x_1) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \quad x_1 \in \{0, 1\},$$

Survey

- Whole population : N, $N\theta$ for A, $(N - N\theta)$ for B

$$\mathbb{P}_\theta(X_1 = x_1) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \quad x_1 \in \{0, 1\},$$

- Reduced population $N - 1$, $N\theta - x_1$ for A, $(N - 1 - (N\theta - x_1))$ for B

$$\mathbb{P}_\theta(X_2 = x_2 | X_1 = x_1) = \frac{(N\theta - x_1)^{x_2}(N - 1 - (N\theta - x_1))^{1-x_2}}{N - 1},$$

Survey

With n subjects : population $N-n$, $N\theta - \sum_{i=1}^{n-1} x_i$ for A, $(N\theta - \sum_{i=1}^{n-1} x_i)$ for B

$$\mathbb{P}_\theta(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n} \left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

Parametric model

- $\Omega = \{0,1\}^n$, set of observations.
- $\mathcal{F} = \mathcal{P}(\{0,1\}^n)$, corresponding $\sigma-$algebra.
- Observations $(X_1, \ldots, X_n)$: called here canonical variables since we can set

$$\omega = (x_1, \ldots, x_n) \in \{0,1\}^n , \quad X_i(\omega) = x_i .$$

- Observation distribution : for $(x_1, \ldots, x_n) \in \{0,1\}^n$:

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = p(\theta, x_1, \ldots, x_n),$$

where

$$p(\theta, x_1, \ldots, x_n) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \times \ldots$$
$$\times \; \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n} \left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

- Observation distribution : for $(x_1, \ldots, x_n) \in \{0,1\}^n$:

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = p(\theta, x_1, \ldots, x_n),$$

where

$$p(\theta, x_1, \ldots, x_n) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \times \ldots$$
$$\times \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n}\left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

- Counting measure μ on $\{0,1\}^n$ is defined : $\forall\ A \in \mathcal{P}(\{0,1\}^n)$, $\mu(A) = \mathrm{card}(A)$ then renormalized to produce a probability measure.

- Observation distribution : for $(x_1, \dots, x_n) \in \{0,1\}^n$:

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = p(\theta, x_1, \dots, x_n),$$

where

$$p(\theta, x_1, \dots, x_n) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \times \dots$$
$$\times \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n} \left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

- Counting measure μ on $\{0,1\}^n$ is defined : $\forall$ $A \in \mathcal{P}(\{0,1\}^n)$, $\mu(A) = \mathrm{card}(A)$ then renormalized to produce a probability measure.
- $\forall \theta \in \Theta$, $p(\theta, x_1, \dots, x_n)$ is the image density $\mathbb{P}_\theta$ by $X_1, \dots, X_n$

$$\mathbb{P}_\theta((X_1, \dots, X_n) \in A) = \sum_{(x_1, \dots, x_n) \in A} p(\theta, x_1, \dots, x_n).$$

- Observation distribution : for $(x_1, \dots, x_n) \in \{0,1\}^n$:

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = p(\theta, x_1, \dots, x_n),$$

where

$$p(\theta, x_1, \dots, x_n) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \times \dots$$
$$\times \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n} \left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

- Counting measure μ on $\{0,1\}^n$ is defined : $\forall\ A \in \mathcal{P}(\{0,1\}^n)$, $\mu(A) = \mathrm{card}(A)$ then renormalized to produce a probability measure.
- $\forall \theta \in \Theta$, $p(\theta, x_1, \dots, x_n)$ is the image density $\mathbb{P}_\theta$ by $X_1, \dots, X_n$

$$\mathbb{P}_\theta((X_1, \dots, X_n) \in A) = \sum_{(x_1, \dots, x_n) \in A} p(\theta, x_1, \dots, x_n).$$

Other estimators can be computes (max, mean, etc..)

- $\hat{\theta}_n$ is a point estimator of θ.
- Questions :
  - Can we quantify the error between θ and its estimation $\hat{\theta}_n$ (confidence intervals) ?
  - Which estimator is the best ?
  - Can we test “$\theta > 1/2$” ?

Michaelis-Menten model

- Example in biochemistry and pharmacology
- Two quantities of interest (v, s) satisfying

$$v = \frac{\alpha s}{s + \beta}$$

  with
  - v : answer (really if interest),
  - s : explanatory (controled by the user),
  - α, β unknown parameters.
- Ex : v inital velocity of a chemical reaction and s substrate concentration.
- In this example α is the maximal velocity and β is called Michaelis-Menten constant.

Statistical model

- To estimate α and β one drives n experiences with different substrate concentrations $s_1, \dots, s_n$.
- Measures are noisy.
- Most popular model :

$$V_i = \frac{\alpha s_i}{s_i + \beta} + \sigma \epsilon_i \ , \quad i = 1, \dots, n$$

with $\{\epsilon_i\}_{i=1}^n$ i.i.i $\mathcal{N}(0,1)$ and σ a scale parameter.

Statistical model

- Observation set : $\Omega = \mathbb{R}^n$.
- $\sigma-$algebra $\mathcal{B}(\mathbb{R}^n)$ borel sets of $\mathbb{R}^n$.
- Canonical variables
- Parameter $\theta = (\alpha, \beta, \sigma) \in \Theta = \mathbb{R}_+ \times \mathbb{R}^+ \times \mathbb{R}_+^*$.

Statistical model

- The observation distribution has a density w.r.t the Lebesgue measure on $\mathbb{R}^n$ given by

$$p(\theta, v_1, \ldots, v_n) = \prod_{i=1}^{n} p_{s_i}(\theta, v_i) \ ,$$

$$p_s(\theta, v) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left(v - \frac{\alpha s}{s+\beta}\right)^2\right) \ .$$

Parameter estimation

Estimator of α, β : solution of a non-linear mean square problem.

$$(\hat{\alpha}_n, \hat{\beta}_n) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left(V_i - \frac{\alpha s_i}{\beta + s_i} \right)^2 .$$

Michaelis-Menten model

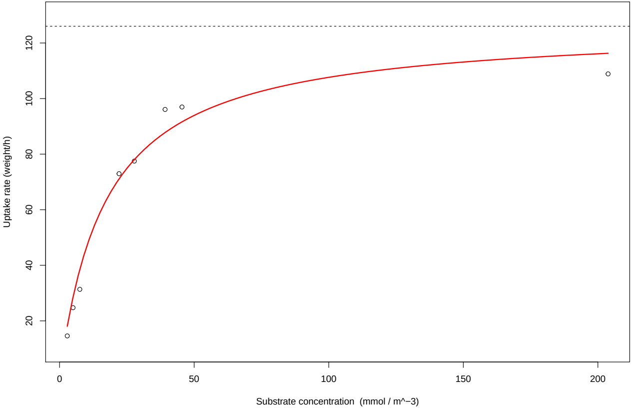


FIGURE – (black) : observations of Michaelis-Menten model, (red) : estimated kinetic

Statistical modeling

Building a statistical experiment requires to identify 3 elements :

Observations

$$\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$$

$=$ realisation of an experiment, starting point of the statistician.

Stochastic model

A question

Statistical modeling

Building a statistical experiment requires to identify 3 elements :

Observations

$$x_1, x_2, \ldots, x_n$$

Stochastic model associates to the experiement which mimic the generation of the observations.
Observations = realisation of random.
Their distribution translate mathematically our knowkedge of the generation process.
In general, this distribution is partielly known
Observations enable to increase our comprehension of this process.

A question

Statistical modeling

Building a statistical experiment requires to identify 3 elements :

Observations

$$\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$$

Stochastic model

A question associated to both the [observations, and the model]. In particular, estimation of unknown parameters, prediction, error estimation and hypothesis testing.

Statistical model produced by an observation

Définition

A statistical model is given by :

- a measurable space $(\Omega, \mathcal{F})$,
- a family $\mathcal{P}$ of probability distribution on $(\Omega, \mathcal{F})$,
- a measurable space $(\mathsf{Z}, \mathcal{Z})$, called observation space,
- a random variable Z defined on $(\Omega, \mathcal{F})$ with values in $(\mathsf{Z}, \mathcal{Z})$.

This writes

$$\{(\Omega, \mathcal{F}), (\mathsf{Z}, \mathcal{Z}), \mathcal{P}, Z\} .$$

If $\mathcal{P} = \{\mathbb{P}_\theta \, , \; \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^d$, the model is parametric, non-parametric otherwise.

Canonical statistical model

Defining $(\Omega, \mathcal{F})$ and Z rarely needed ; take :

$$(\Omega, \mathcal{F}) = \big(\mathsf{Z}, \mathcal{Z}\big) \text{ et } Z(\omega) = \omega, \quad \omega \in \Omega,$$

such that for all $\mathbb{P} \in \mathcal{P}$, $\mathbb{P}^Z = \mathbb{P}$.

Définition

A canonical model is given by :

- A measureable space $(\mathsf{Z}, \mathcal{Z})$, l'*Observation set*,
- A family of probability distributions $\mathcal{C}$ on $(\mathsf{Z}, \mathcal{Z})$.

Parametrical and non-parametrical models

- If $\mathcal{P}$ only depends on finite dimensional parameter :

$$\mathcal{P} = \left\{\mathbb{P}_\theta, \ \ \theta \in \Theta\right\},$$

where $\Theta \subset \mathbb{R}^d$, $\rightarrow$ *parametric model*, *non-parametric* otherwise.

Parametrical and non-parametrical models

- If $\mathcal{P}$ only depends on finite dimensional parameter :

$$\mathcal{P} = \left\{\mathbb{P}_\theta, \;\; \theta \in \Theta\right\},$$

where $\Theta \subset \mathbb{R}^d$, $\rightarrow$ *parametric model*, *non-parametric* otherwise.

- For a parametric model, , for all $\theta \in \Theta$,

$$\int_{\mathsf{Z}} h(z) \mathbb{P}_\theta^Z(\mathrm{d}z) = \int_\Omega h \circ Z(\omega) \mathbb{P}_\theta(\mathrm{d}\omega) = \mathbb{E}_\theta[h(Z)] \,.$$

Translation - scaling model

- Observations $(\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_\mathtt{n})$: n independent real observations.
- Model
  - Observations are independent and identically distributed.
  - Each is modeled by

$$X_i = \mu + \sigma\zeta_i \ ,$$

where $\{\zeta_i\}_{i=1}^n$ are r.v. with known density q w.r.t the Lebesgue measure on $\mathbb{R}$. μ : translation parameter and $\sigma > 0$ scaling parameter.

Translation - scaling model

- Observation space : $Z = \mathbb{R}^n$ (possible values of the observations).
- $\sigma-$algebra : $\mathcal{Z} = \mathcal{B}(\mathbb{R}^n)$ Borel sets on $\mathbb{R}^n$.
- Parameters : $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.
- Law : Law of $Z = (X_1, \ldots, X_n)$ has a density $p_n(\theta, x_1, \ldots, x_n)$ w.r.t. the Lebesgue measure on $\mathbb{R}^n$ given by

$$p_n(\theta, x_1, \ldots, x_n) = \prod_{i=1}^{n} p(\theta, x_i) \ ,$$

with

$$p(\theta, x) = \frac{1}{\sigma} q\left(\frac{x-\mu}{\sigma}\right) \ .$$

Particular cases :

- Gaussian density :

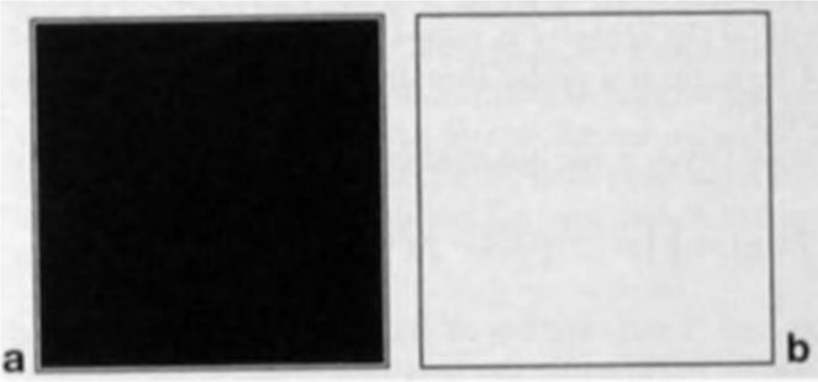
$$p(\theta, x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) .$$

- Laplacian density :

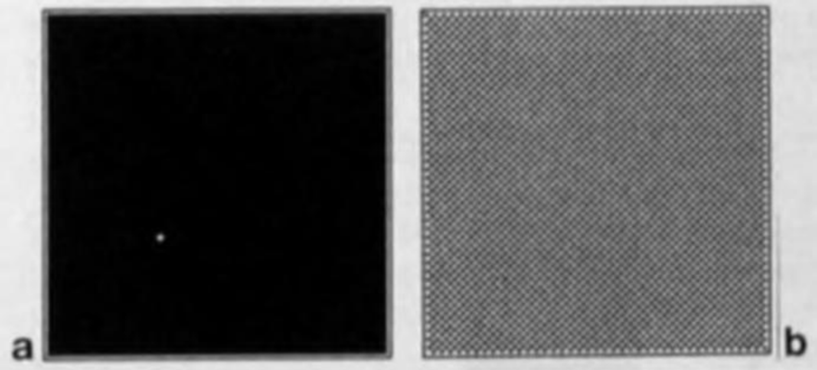
$$p(\theta, x) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right) .$$

Observation

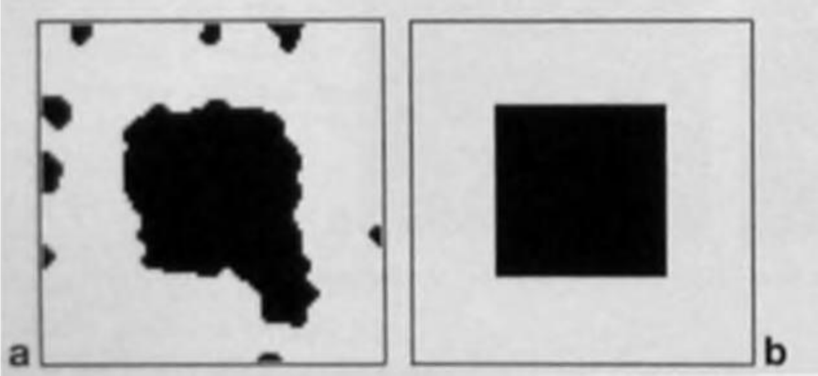
Trivial solution using only the prior



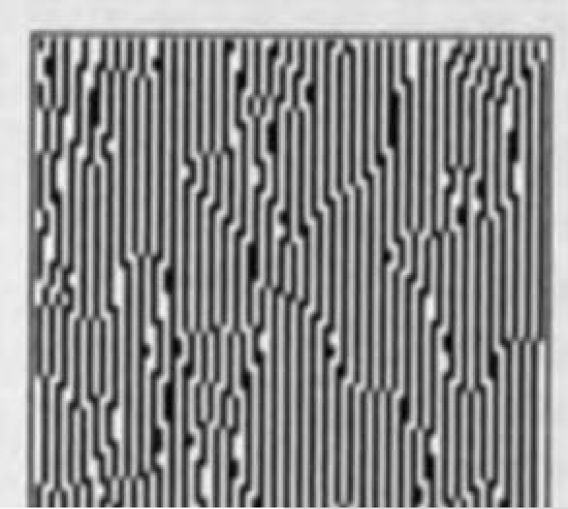
Images with high (left) and low (right) prior



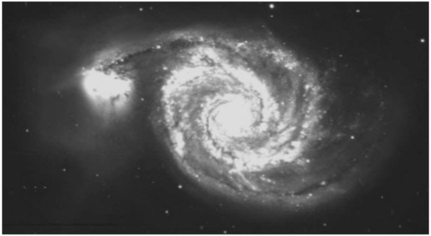
One solution and true image

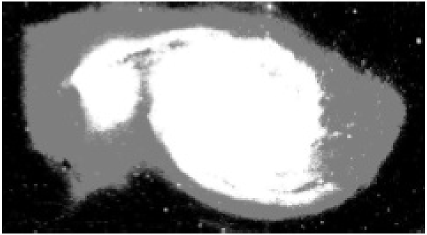


In case of wrong prior

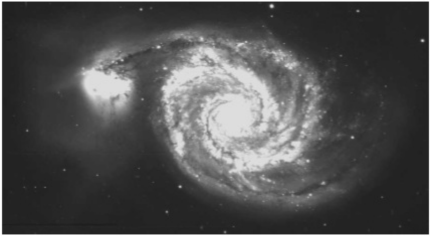


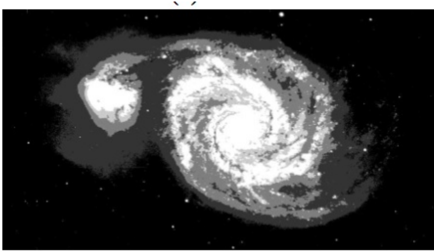
Other example with more complex model (HMM)



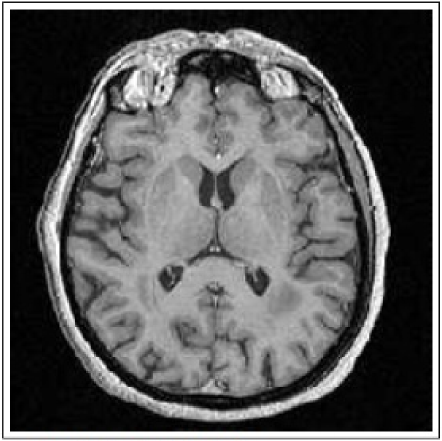


Other example with more complex model (HMM)

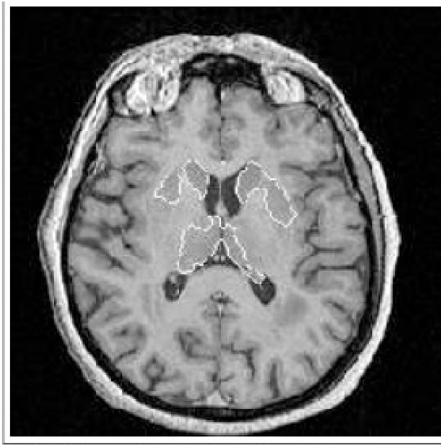




Segmentation



a. Image cérébrale IRM.



Résultat de la segmentation