Bayesian Inference

Summary

- Motivation
- A Posteriori Distribution
- Bayesian Estimation Methods
- Model Selection

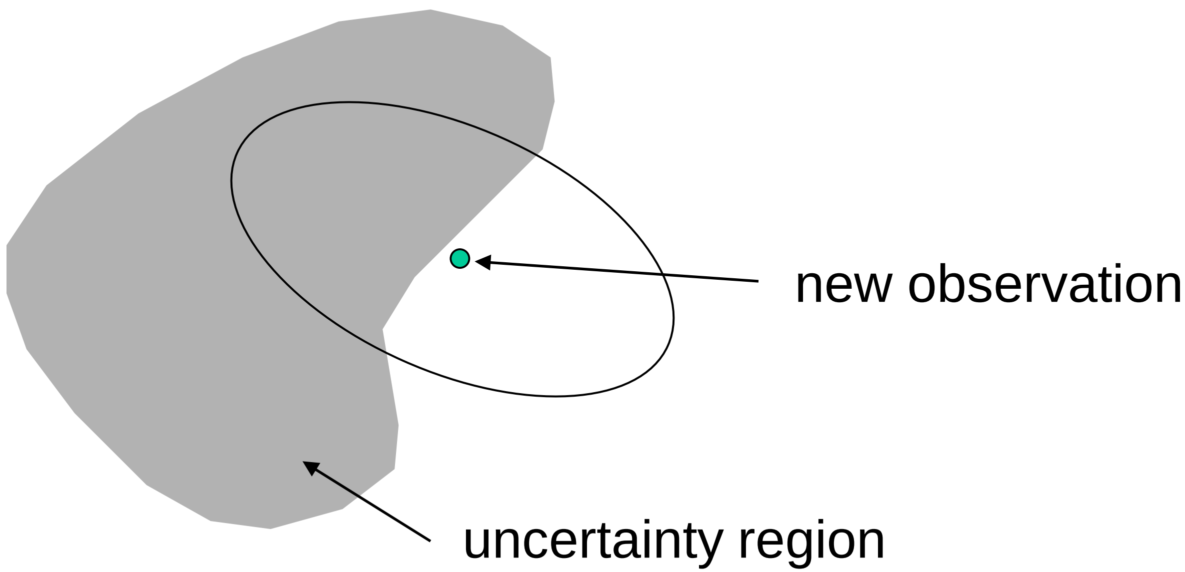
Question

Let x be a random variable with values in IR^2 and let y be a linear combination of the x components, corrupted by additive noise:

$$y = x_1 + x_2 + w$$

Is it possible to estimate x from y ?

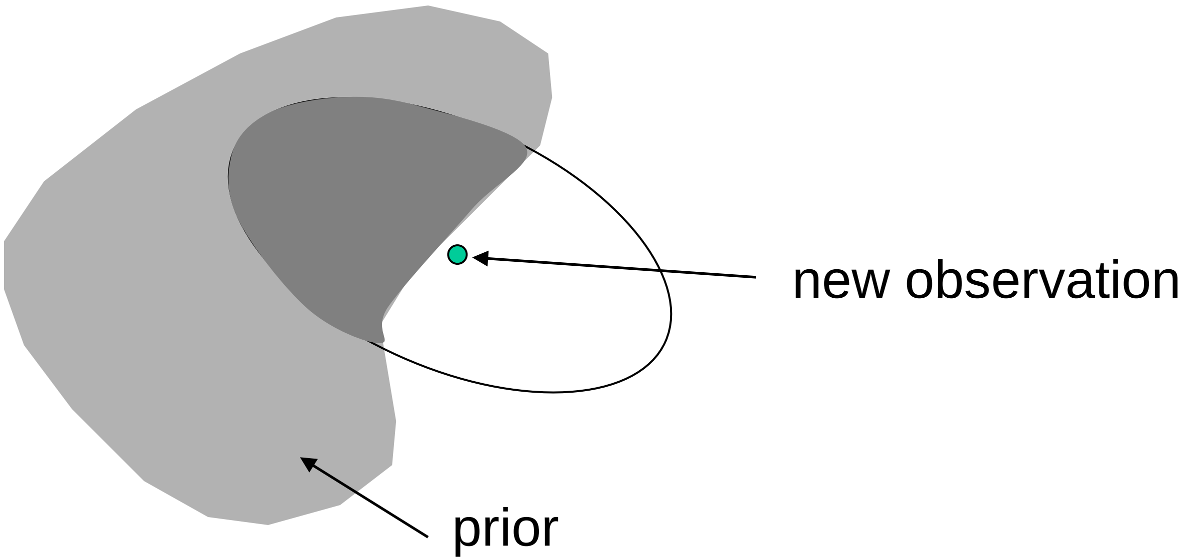
Data Fusion



Where is the boat ?

Bayes (1702-1761)

Bayesian Inference



- inicial location: p(x) prior
- sensor model: p(y|x)
- final location: p(x|y) a posteriori density function

The final result is a distribution!

A Posteriori Distribution
(known model)

How to compute the a posteriori distribution ?

Bayes law

likelihood function

prior

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)}$$

constant

Conjugate Prior

The prior represents the knowledge available about the unknown variables before any observation is made.

It should allow an easy computation of the a posteriori distribution.

A conjugate prior is a prior such that the *a posteriori* distribution has the same analytic expression as the prior, with different values of the parameters.

Exponencial Family

It is easy to obtain conjugate priors if the sensor model p(y|x) belongs to the exponencial family.

Definition: p(y|x) belongs to the **exponential family** if and only if

$$p(y \mid x) = h(y) g(x) \exp\{t(y) c(x)\} \qquad \text{e} \qquad \int p(y \mid x) dy = 1$$

conjugate prior:

$$p(x) = g(x)^d \exp\{b c(x)\}$$

a posteriori density:

$$p(x \mid y) = g(x)^{\tilde{d}} \exp\{\tilde{b} c(x)\}, \quad \tilde{d} = d + n, \ \tilde{b} = b + \sum_{i=1}^{n} t(y_i)$$

Several well known distributions e.g., normal (with known covariance), gamma, binomial, Poisson, belong to the exponential family.

Proof

Let y = y_1, ..., y_n be a sequence of independent observations.

likelihood function

$$p(y \mid x) = g(x)^n \prod_i h(y_i) \exp\{t(y_i)c(x)\}$$

a posteriori density

$$p(x \mid y) \;\alpha\; p(y \mid x)p(x)$$

$$\alpha\; g(x)^n \prod_i h(y_i) \exp\{t(y_i)c(x)\} \times g(x)^d \exp\{bc(x)\}$$

$$\alpha\; g(x)^{n+d} \exp\left\{\left(b + \sum_i t(y_i)\right)c(x)\right\}$$

$$\alpha\; g(x)^{\tilde{d}} \exp\{\tilde{b}c(x)\}$$

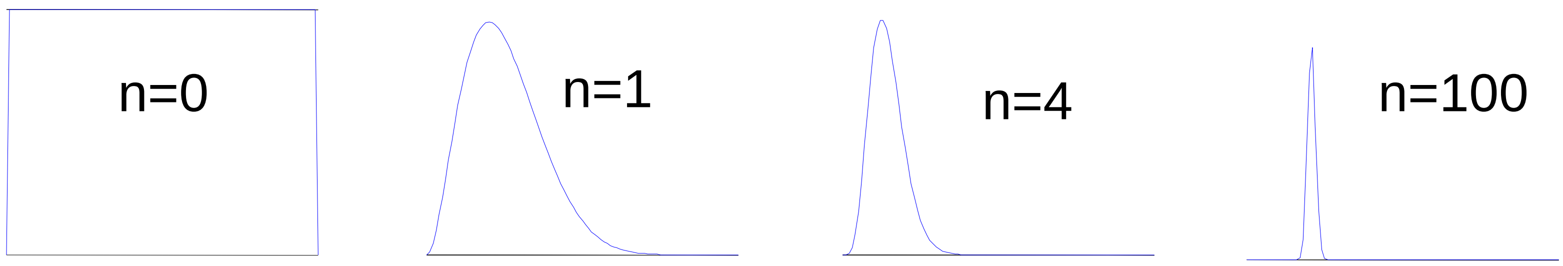
Binomial Distribution

The binomial distribution B(α) belongs to the exponential family.

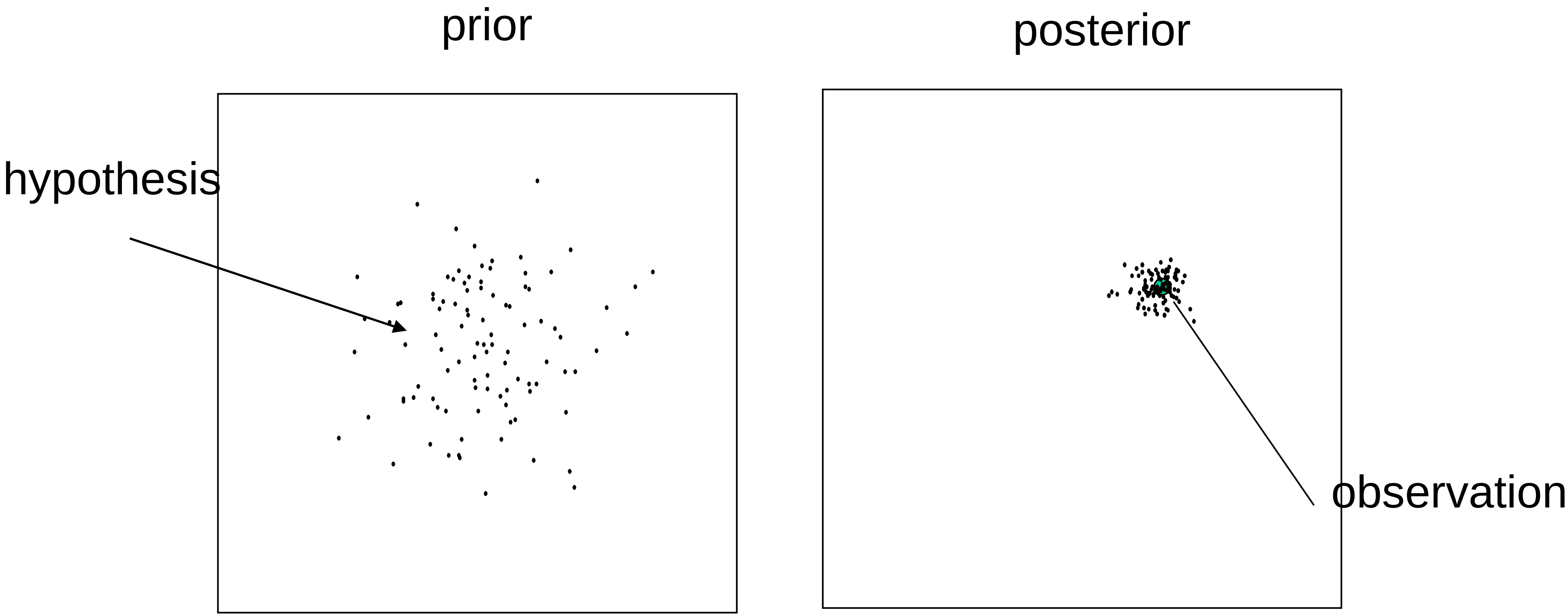
Conjugate prior: $P(\alpha) = c\alpha^{b}(1-\alpha)^{md-b}$ Beta distribution

A posteriori distribution: the same with $\tilde{b} = b + k,\ \tilde{d} = d + 1$

Example: α=.2



Example



This example considers $p(x)=N(0,I)$, $p(y/x)=N(x,.04I)$

Recursive Computation

Suppose we obtain n independent observations $y=(y_1, ..., y_n)$.

Then

$$p(x \mid y) = cp(y_1, y_2 \mid x)p(x) = cp(y_2 \mid x)p(y_1 \mid x)p(x)$$

This suggests the following recursion:

$$p(x \mid y_{1:k}) \;\; \alpha \;\; p(y_k \mid x)p(x \mid y_{1:k-1})$$

where $y_{1:k}=(y_1,..., y_k)$

This procedure is very useful when conjugate priors are used.

A Posteriori Distribution
(unknown model)

Let us assume that x depends on an unknown variable θ.

In this case

$$p(x\,|\,y) = \int p(x\,|\,\theta)p(\theta\,|\,y)d\theta$$

where

$$p(\theta\,|\,y) = \frac{p(y\,|\,\theta)p(\theta)}{p(y)}$$

When the model is unknown the Bayesian approach considers all possible models weighted by their confidence degrees $p(\theta/y)$, instead of using a single (best) model.

MAP and MMSE Estimates

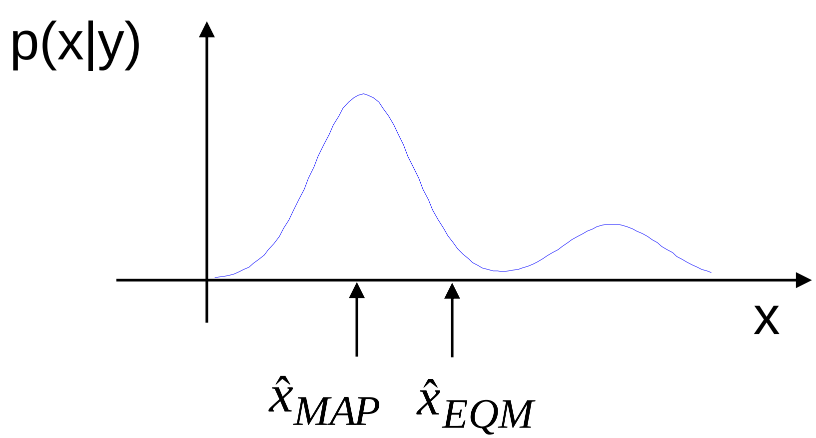
How to obtain an estimate of x from the *a posteriori* distribution ?

MAP estimate (maximum a posteriori)

$$\hat{x} = \arg\max_{x} p(x \mid y) = \arg\max_{x} p(y \mid x) p(x)$$

MMSE estimate (minimum mean squared error)

$$\hat{x} = E\{x \mid y\} = \int x \, p(x \mid y) \, dx$$



MAP vs ML

ML estimator: $$\hat{x} = \arg\max_{x} p(y \mid x)$$

MAP estimator: $$\hat{x} = \arg\max_{x} p(y \mid x)\underline{p(x)}$$

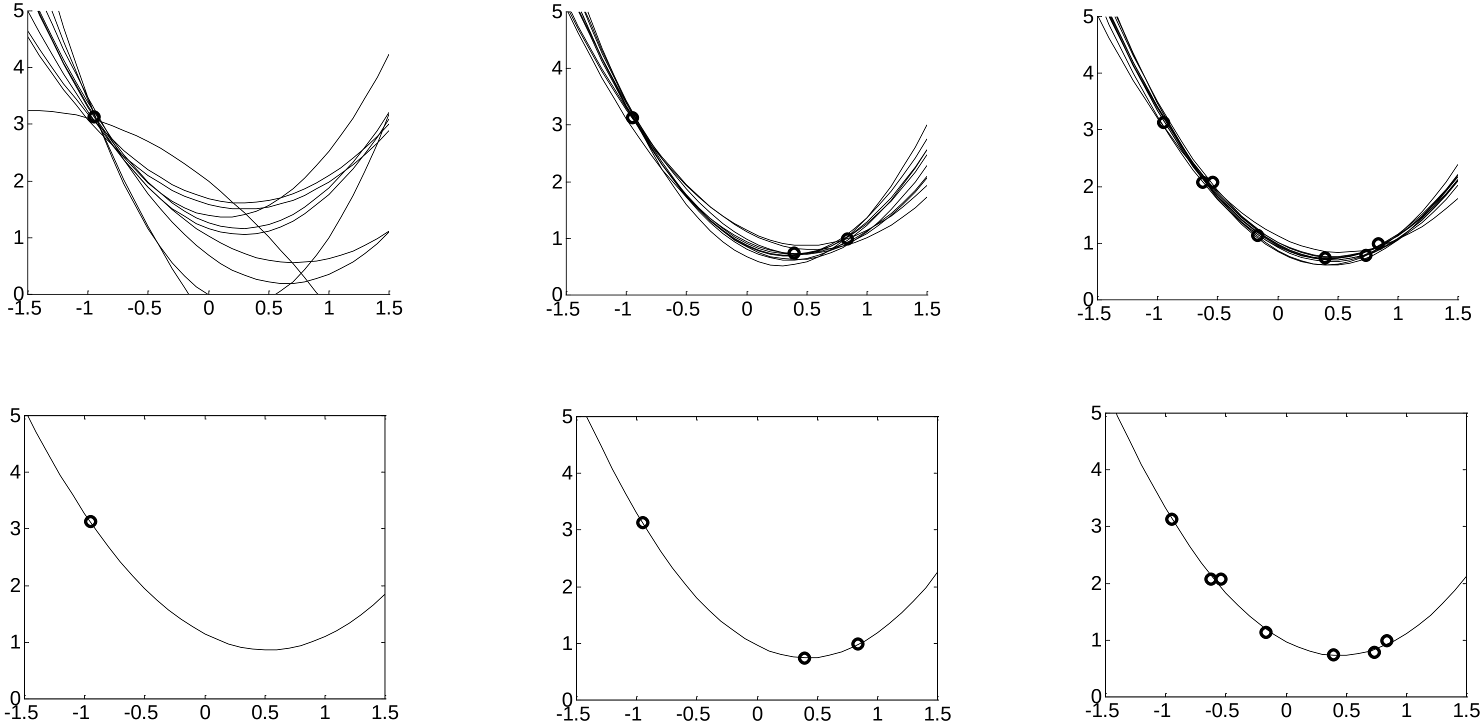
prior

The prior has an important role when there is few data.

(simple rule:the should be 10 observations for each parameter to be estimated.)

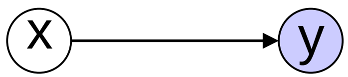
Parábolic Fit

n=1 n=3 n=7



MAP estimate

Gaussian Variables



Hypothesis: x,y have normal distribution.

Question: what is the distribution of *x* given *y* ?

Answer: $p(x \mid y) = N(\hat{x}, P)$

$$\hat{x} = \bar{x} + P_{xy} P_{yy}^{-1} (y - \bar{y})$$

$$P = P_{xx} - P_{xy} P_{yy}^{-1} P_{yx}$$

Notation: $\bar{a} = E\{a\}, P_{ab} = E\{(a - \bar{a})(b - \bar{b})'\}$

Proof

Lemma:
$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix} \qquad \begin{aligned} E &= (A\text{-}BD^{-1}C)^{-1} \\ F &= -EBD^{-1} \end{aligned}$$

$p(x/y) = N(\hat{x}, P)$. The argument of the exponential is

$$q = \begin{bmatrix} \mathrm{x} - \bar{\mathrm{x}} \\ \mathrm{y} - \bar{\mathrm{y}} \end{bmatrix}' \begin{bmatrix} P_{xx} & P_{yx} \\ P_{xy} & P_{yy} \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{x} - \bar{\mathrm{x}} \\ \mathrm{y} - \bar{\mathrm{y}} \end{bmatrix} = (\mathrm{x} - \bar{\mathrm{x}})'\mathrm{E}(\mathrm{x} - \bar{\mathrm{x}}) + 2(\mathrm{x} - \bar{\mathrm{x}})\mathrm{F}(\mathrm{y} - \bar{\mathrm{y}}) + \mathrm{c}$$

$$= \mathrm{x}'\mathrm{E}\mathrm{x} - \bar{\mathrm{x}}) - 2(\mathrm{x} - \bar{\mathrm{x}})(\mathrm{E}\bar{\mathrm{x}} - \mathrm{F}(\mathrm{y} - \bar{\mathrm{y}})) + \mathrm{c}'$$

Comparing with the exponent of $N(\hat{x}, P):\ \ x'P^{-1}x - 2x'P^{-1}\hat{x} + \hat{x}P^{-1}\hat{x}$

we conclude $P^{-1} = E,\ P^{-1}\hat{x} = E\bar{x}\text{-}F(y\text{-}\bar{y})$

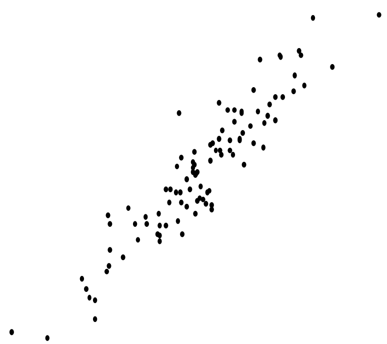
Therefore, $P = (P_{xx} - P_{xy}P_{yy}^{-1}P_{yx})^{-1},\ \hat{x} = \bar{x}\text{-}P_{xy}P_{yy}^{-1}(y\text{-}\bar{y})$

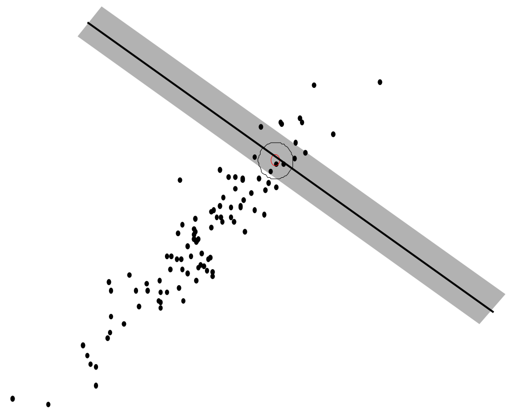
Example

Let x~N(0,R) be a random variable with values in R^2 and y a linear combination of x components, corrupted by white noise:

$$y = x_1 + x_2 + w$$

Is it possible to estimate x fom y ?





This example was obtained with x~N(0,P), w~N(0,.1), y=2 $P = \begin{bmatrix} .8 & .75 \\ .75 & .8 \end{bmatrix}$

Linear Model

Let us consider a linear model with additive Gaussian noise:

$$y = Cx + v \qquad x \sim N(\bar{x}, \bar{P}), \quad v \sim N(0, Q)$$

What is the distribution of *x*, after observing y ?

Answer: $p(x \mid y) = N(\hat{x}, P)$

$$\hat{x} = \bar{x} + K(y - C\bar{x}) \qquad K = \bar{P}C'S^{-1}$$

$$P = (I - KC)\bar{P} \qquad S = C\bar{P}C' + R$$

This result suggests an incremental update of the parameters when y is a sequence of independent observations.

Bayesian Estimation

Principles:

• The unknown parameters are random variables with known distribution.

• The observations allow to reduce uncertainty of the parameter estimates and to update their distribution. The updated distribution is denoted as **a posteriori distribution**.

• The update is done by the Bayes law.

Notes:

• Bayesian methods provide objective criteria for the design of estimators.

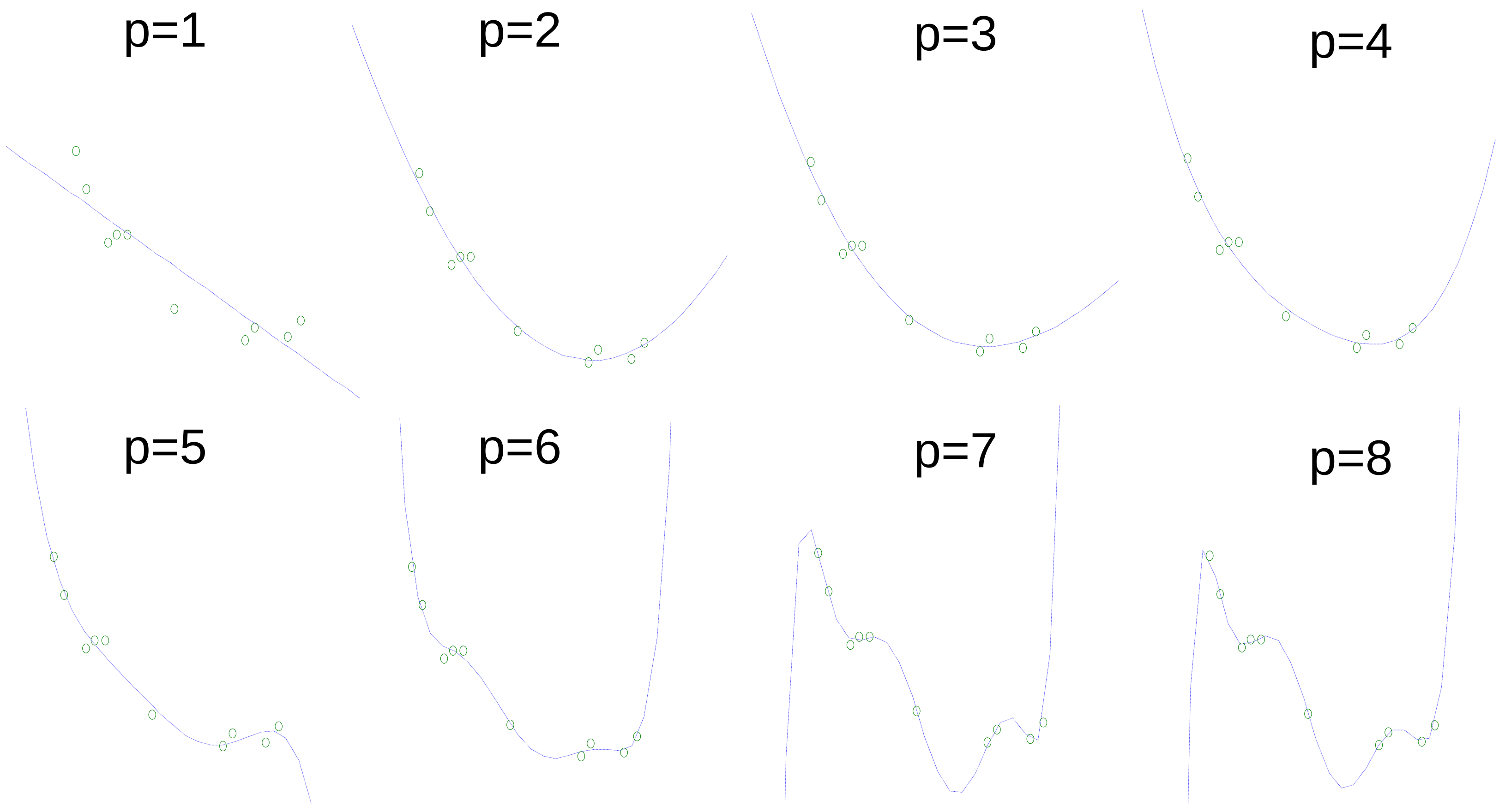
• They have better performance that classic methods when there is few data points.

Difficulties

Inference is more difficult in the following cases:

- *invalid data (outliers);*
- *incomplete data (hidden variables);*
- *need of model validation/selection;*
- *multiple models*

Model Selection



What is the best model ?

Model Selection

Let us consider all the available models $M_1, ..., M_c$ to represent a sequence of observations y.

What is the best ?

There are several criteria: MV, MAP, MDL, AIC, etc

Occam Razor

In XIV century Occam the following principle:

Choose the simplest model which describes the data with the desired accuracy.

Exercises

1. Let x_1 , ..., x_n be a sequence of independent and identically distributed observations. Knowing that

$$p(x_i \mid \alpha) = \alpha e^{-\alpha x_i} \quad p(\alpha) = c e^{-c\alpha} \quad \alpha, x_i > 0$$

compute the MAP estimate of a.

2. Consider a signal y_t generated by the model $y_t = a\, y_{t-1} + b\, u_t + w_t$ Determine a Bayesian estimate of a,b coefficients assuming that the inputs and outputs $y_1 \ldots y_n$, $u_1 \ldots u_n$ are available and the noise sequence $w_1 \ldots w_n$ consists of uncorrelated variables $w_i \sim N(0,\sigma^2)$.

3. Show that a density $p(x)=C\exp[-0.5\,(x'Ax+b'x)]$ is normal $N(\mu,P)$ with $P=A^{-1}$ e $\mu=-0.5\,A^{-1}\,b$.

4. Show that the product of two normal densities $N(\mu_i,P_i)$, i=1,2, is a normal density (apart from a scale factor).

Work

Consider data generated by two probabilistic models

a) x ~ $N(\mu,\sigma^2)$ with known σ^2 and $\mu \sim N(\mu_o, \sigma_o^2)$

b) $p(x/\alpha) = \alpha e^{-\alpha x} \quad x > 0, \quad p(\alpha) = \alpha_0 e^{-\alpha_0 \alpha} \quad \alpha > 0,$

Given an observation x, determine a criteria for the selection of the model.

Characterize the performance of the previous method computing the error probability experimentally.

Bibliography

J. Marques, Reconhecimento de Padrões. Métodos Estatísticos e Neuronais, IST Press, 1999