Classic Estimation

Summary

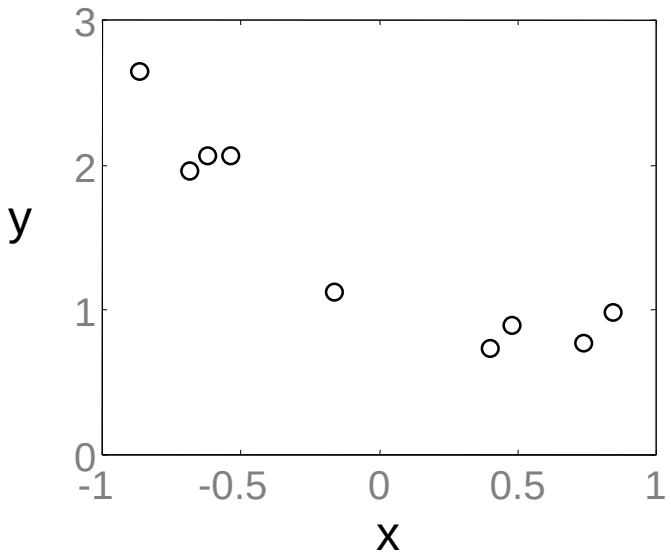
- Motivation
- Deterministic Methods
- Classic Probabilistic Methods
- Examples

Challenge

Given the data points shown in the figure we wish to approximate them by a 2nd order polynomial

$$y=c_2 x^2+ c_1 x+ c_0$$

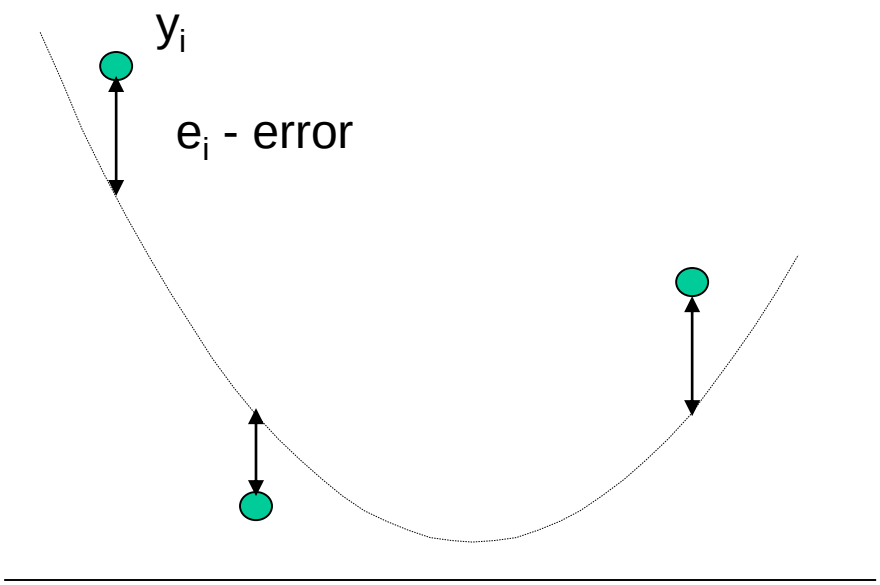
Problem: how to compute c_0, c_1, c_2.



Note: we assume that the x values are accurately known but the y values are corrupted by measurement noise.

Gauss (1777-1855)

Parabolic Fit



The absolute value of the errors must be small

Quadratic cost:

$$E = \sum_i e_i^2$$

<e,e>

Estimation of the coefficients

What coefficients minimize E ?

Stationarity condition:

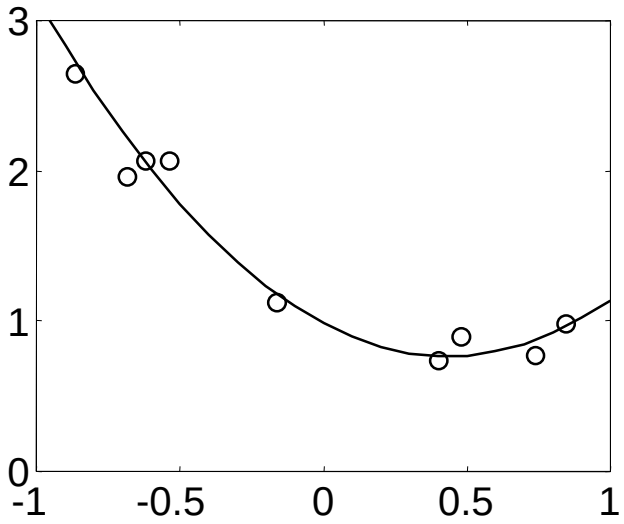
$$\frac{\partial E}{\partial c_p} = 0 \Rightarrow \frac{\partial}{\partial c_p} \sum_i e_i^2 = \sum_i 2 e_i x_i^p = \sum_i 2(y_i - c_2 x_i^2 - c_1 x_i - c_0) x_i^p = 0$$

$$\begin{bmatrix} \sum_i 1 & \sum_i 1 x_i & \sum_i 1 x_i^2 \\ \sum_i x_i 1 & \sum_i x_i x_i & \sum_i x_i x_i^2 \\ \sum_i x_i^2 1 & \sum_i x_i^2 x_i & \sum_i x_i^2 x_i^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_i \\ \sum_i y_i x_i^2 \end{bmatrix}$$

internal products between basis functions

internal products between the observations and the basis functions

Results



Least Squares Method

Given a set of observations (x_i, y_i). We wish to approximate y_i by $f(x_i,\theta)$, θ being an unknown vector of parameters. The error is

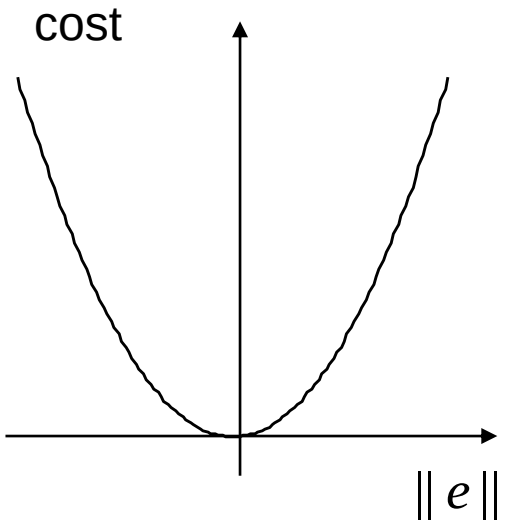
$$e_i = y_i - f(x_i,\theta)$$

How to obtain θ ?

Least squares method:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i} \| e_i \|^2$$

$\| . \|$ is the Euclidean norm



Hypothesis: it is assumed that x_i is accurately known.

Linear Model

Model: $y_i = \phi(x_i)'\theta + e_i$

LS estimate: $\theta = A^{-1}b$

$$A = \sum_{i=1}^{n} \phi(x_i)\phi(x_i)'$$

$$b = \sum_{i=1}^{n} \phi(x_i)y_i$$

Proof

$$E = \sum_i (y_i - \phi(x_i)'\theta)'(y_i - \phi(x_i)'\theta)$$

Computing the derivative with respect to q and using the appropriate properties

$$\frac{dE}{d\theta} = 0 \Rightarrow -2\sum_i \phi(x_i)(y_i - \phi(x_i)'\theta) = 0 \Rightarrow \sum_i \phi(x_i) y_i = \sum_i \phi(x_i)\phi(x_i)'\theta$$

$\hat{\theta} = A^{-1}b$ is a stationary point.

If the matrix

$$\frac{d^2 E}{d\theta^2} = 2\sum_i \phi(x_i)\phi(x_i)'$$ is positive definite,

E is minimized by $\hat{\theta}$

Geometric Interpretation

Observations $y=(y_1, ..., y_n)$, belong to a vector space E of dimension n.

Model sequence $\hat{y}=(\phi(x_1)'\theta, ..., \phi(x_n)'\theta)$, belongs to a subspace $S \subset E$ with a lower dimension.

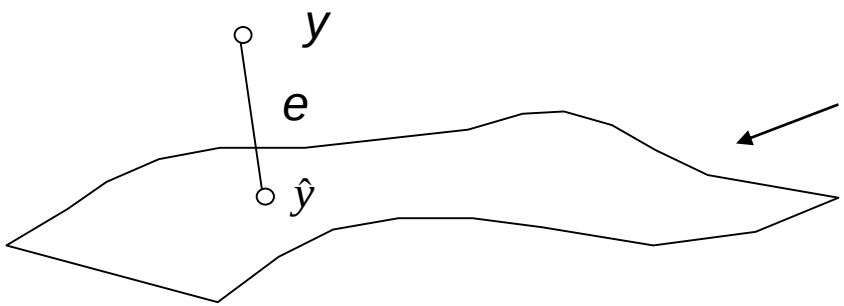
The least squares method computes the coefficients of the orthogonal projection of y onto the subspace S, using the internal product:

$$< x,y >= \sum_i x_i y_i$$

$\hat{y}$ projection is such that the projection $e=y-\hat{y}$ error is orthogonal to all the vectors in S:

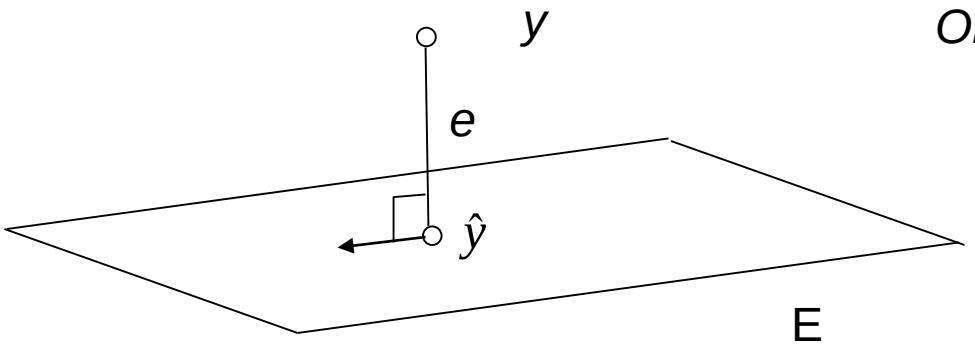
$<e,b>=0, \ \forall b \in S$ (orthogonality condition)

Geometric Interpretation



Set of signals generated by the model

If it is a linear subspace S (linear model), $\hat{y}$ is the orthogonal projection of y onto S



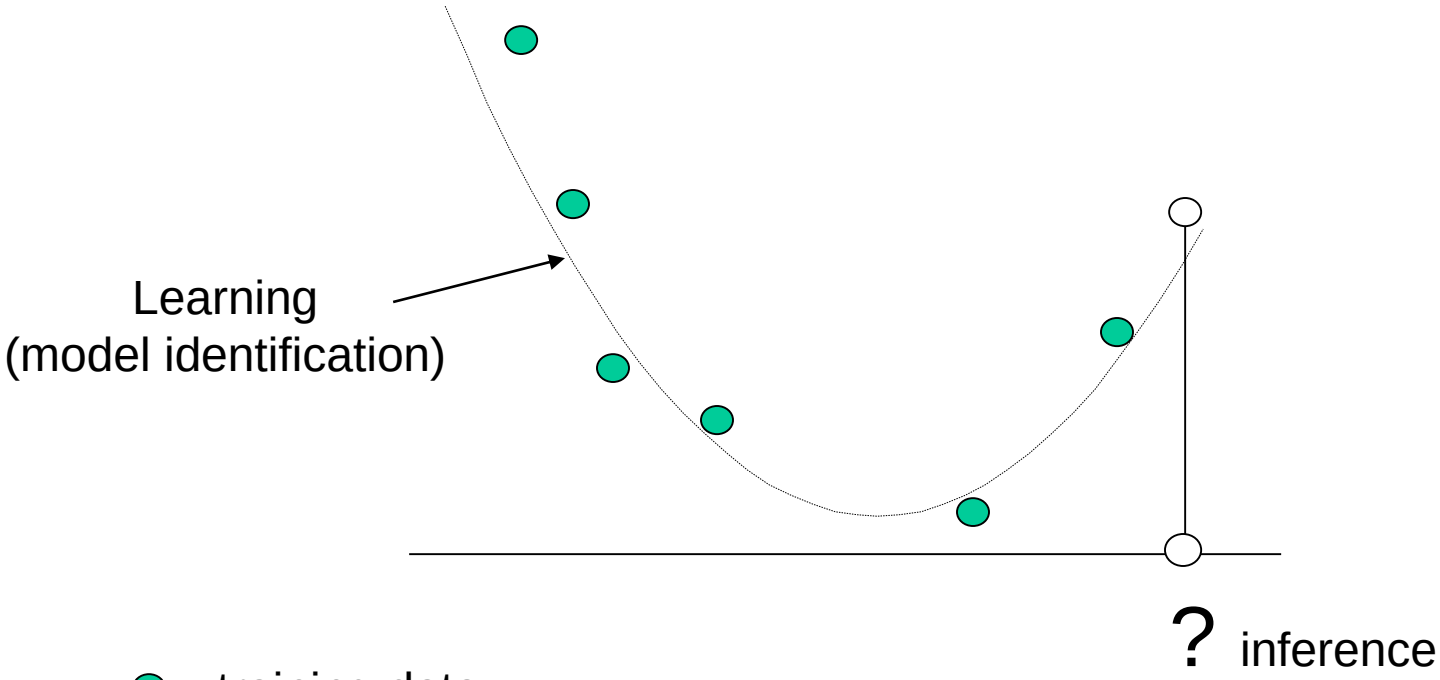
Orthogonality conditions

<e,b>=0

b – any vector of S

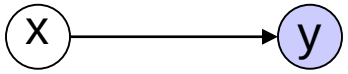
Learning vs Inference

The least squares method allows the solution of several problems.



it does not provide an uncertainty measure!

Estimation



y depends on x and on a vector θ
x is known or unknown

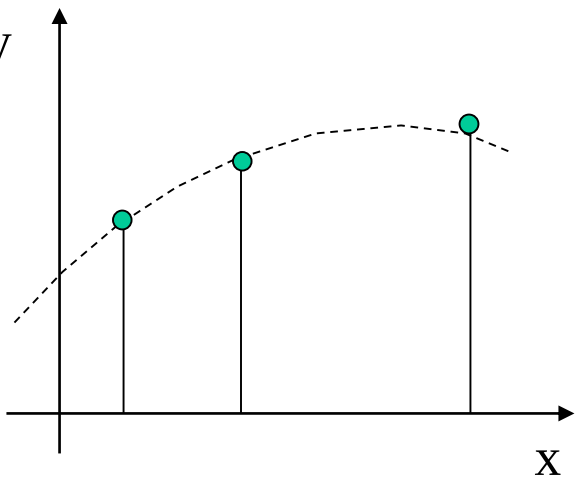
General problem: how to estimate x, θ, given y ?

Subproblems:

obtain x from y, θ	(inference)
obtain θ from y, x	(learning, identification)

The learning problem is often based on observations obtained at different instants of time or even different experiments.
The estimation of the unknown variable x is usually performed in each experiment.

Regression



Given (x_i, y_i), we wish to approximate y_i by $f(x_i, \theta)$ with θ unknown.

Hypothesis (asymmetric): x is accurately known.

Least squares estimate:

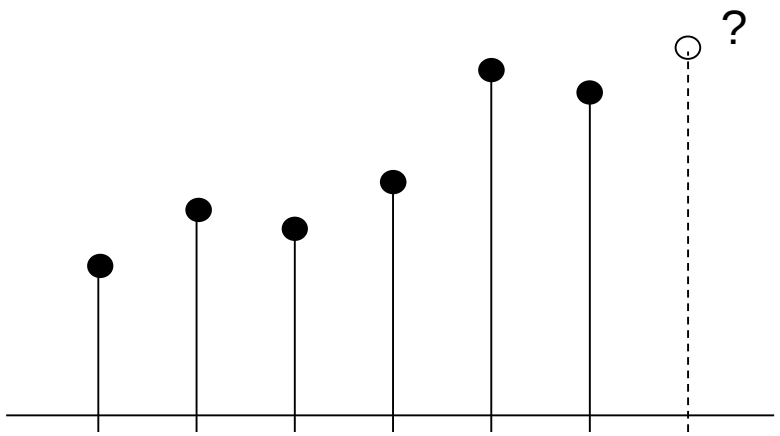
$$\hat{\theta} = \arg\min_{\theta} \sum_{i} \| e_i \|^2 \qquad e_i = y_i - f(x_i, \theta)$$

Examples: linear combination of basis functions, neural networks.

Prediction

Given y=(y_1 , ..., y_{t-1}), we wish to predict y_t using a linear combination of past observations:

$$y_1 = a_t\, y_{t-1} + \ldots + a_p\, y_{t-p} + e_t$$



The coefficients a_i are estimated by least squares.

Frequency Estimation

Given a model

$$y_i = A\cos(\omega t_i + \varphi) + e_i$$

We wish to estimate A, ω, φ ?

Available data: (t_i, y_i) i=1, ...,N

Idea: choose A, ω, φ minimizing $E = \sum_i e_i^2$

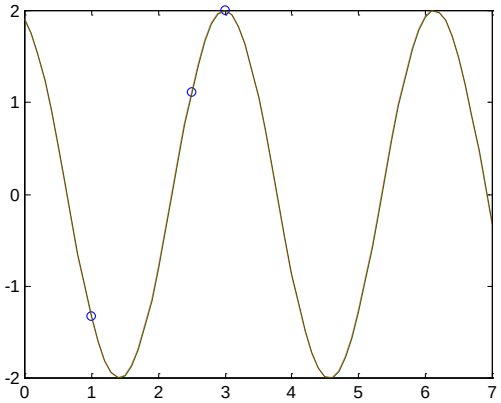
non linear problem

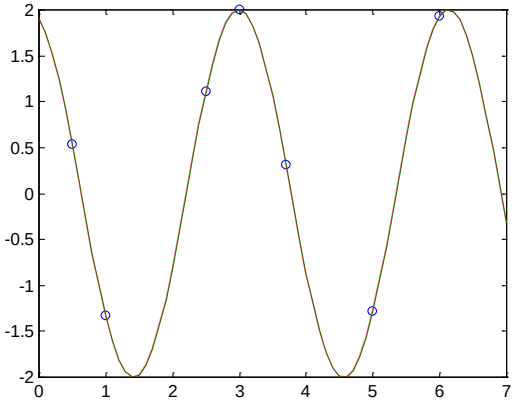
Estimation of a Sinusoid

3 points

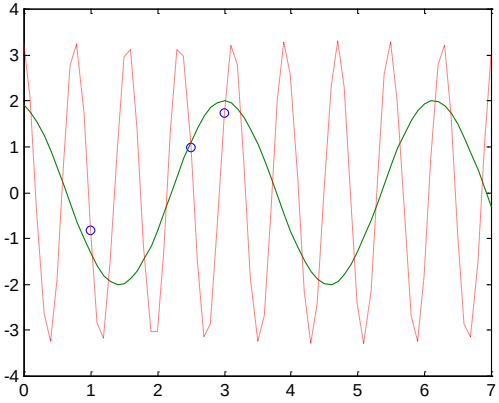
7 points

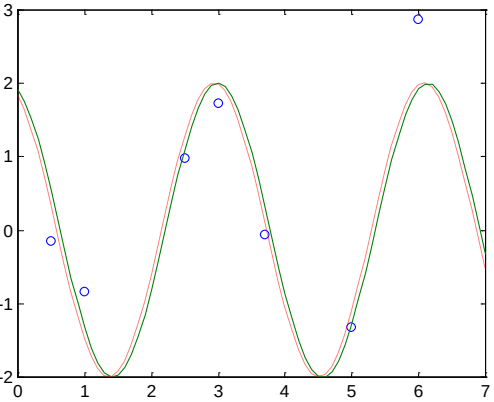
without noise





with noise





Optimization in the interval: $(A,\omega,\phi) \in]0,5[\times]0,10[\times]0,2\pi[$

(white noise with σ=.5)

Restrictions

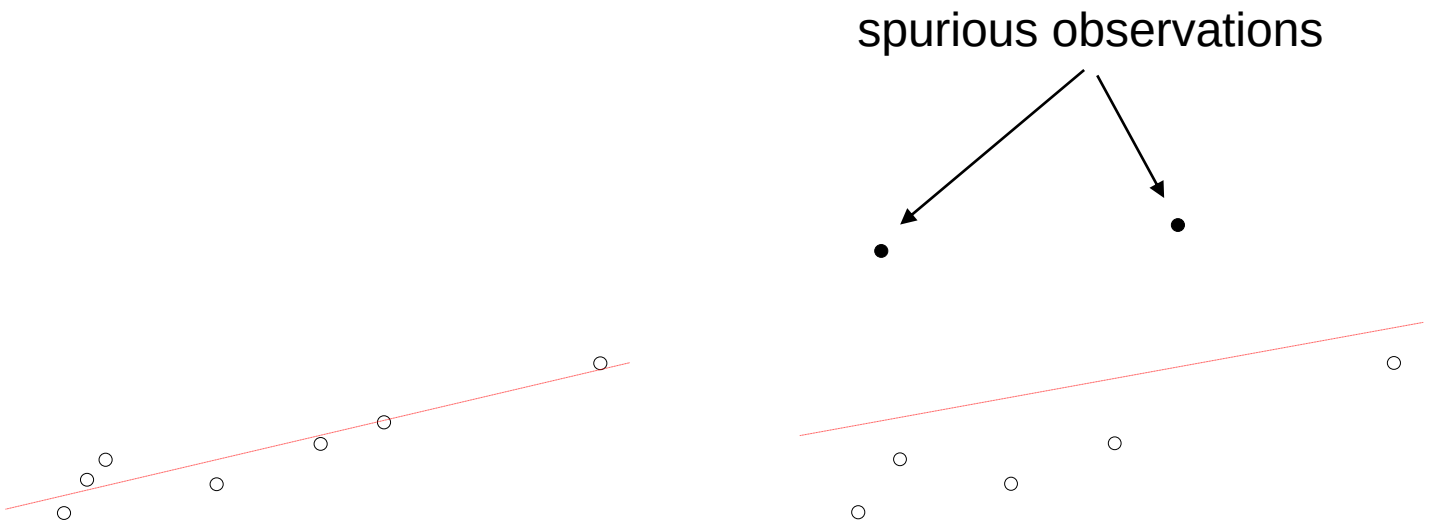
The least squares method has the following restrictions:

• *does not allow a statistical description of the error*: nothing is known about the error we obtain in the next experiment.

•*Not robust*

•Leads to *difficult optimization problems* when non linear models are used. In general, only local minima of E can be obtained.

Robustness



Not robust

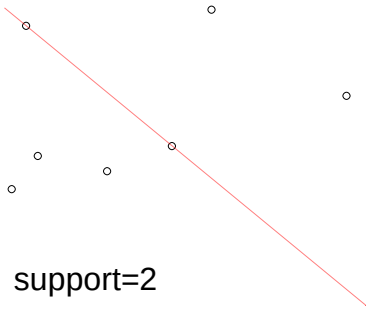
(why ?)

Alternatives:

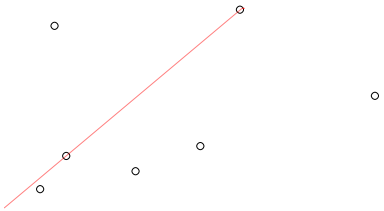
- weighted LS (errors are weighted by confidence degrees)
- robust estimation methods

RANSAC

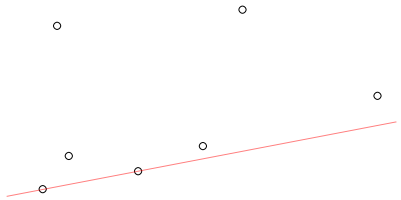
Generation of random hypothesis



support=2

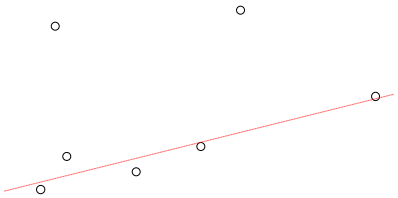


support=3



support=5

Refinement



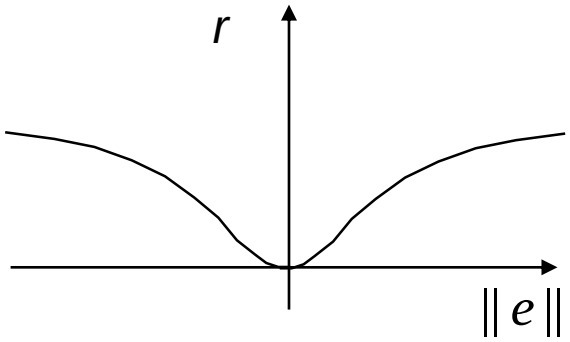
Procedure

- *choose a minimal set of data allowing to estimate the parameters*
- *compute the number of observations well approximated by the model*
- *repeat the previous steps N times and at the end choose the estimate with bigger support (the estimate is refined using the support observations)*

Robust Methods

Robust estimator:

$$\hat{x} = \arg\min_{x} \sum_{i} \rho(\| e_i \|)$$



Example: LMed

•requires recursive numeric optimization

Exercises

1. Given a signal y=(y_1 , ..., y_N), determine the coefficients of the linear predictor

$$\hat{y}_t = a_1 y_{t-1} + ... + a_p y_{t-p}$$

N>>p

using the least squares method.

2. Generalize the previous problem to the case in which we want to predict the signal k steps ahead.

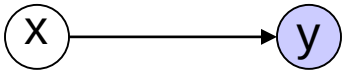
Work

Consider two parabolas and observations close to each of them. However, we do not know which parabola fits each observation. Estimate the coefficients of the parabolas using the least squares method and robust methods.

Characterize the performance of both methods

Classic Probabilistic Methods

Classic vs Bayesian Methods



y – realization of a random variable

x - unknown

Estimation methods:

- classic: LS, EM
- Bayesian: MAP, MSE

Classic methods consider x as a **deterministic variable**.

Bayesian methods consider x as a **random variable** characterized by an *a posteriori* distribution *p(x|y)*

Classic Estimation

•classic estimation methods are not based on general principles.

• An estimator is a map from the observation space into the parameter space.

• Each estimator is a random variable which can be statistically characterized. In general we wish to define estimators with a set of desired properties (e.g., unbiased and with low variance).

Maximum Likelihood Method

(Fisher, 1921)

Maximum likelihood estimate (ML)

$$\hat{x} = \arg\max_{x} L(x) \qquad L(x) = p(y \mid x) \qquad \text{likelihood function}$$

or

$$\hat{x} = \arg\max_{x} l(x) \qquad l(x) = \log p(y \mid x) \qquad \text{log likelihood function}$$

If y= $y_1,\ldots, y_N$ is a set of independent and identically distributed (iid) observations, then

$$L(x) = \prod_i p(y_i \mid x) \qquad l(x) = \sum_i \log p(y_i \mid x)$$

Normal Distribution

Problem: estimate the mean and covariance matrix of a normal distribution N(μ,R) from a set of n independent random variables y_1, ..., y_n.

Log likelihood function:

$$l(\mu,R) = K - \frac{n}{2}\log|R| - \frac{1}{2}\sum_i (y_i - \mu)' R^{-1} (y_i - \mu)$$

Mahalanobis distance

Optimizing l, we get

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \hat{R} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\mu})(y_i - \hat{\mu})'$$

Proof

Mean estimation:

$$\frac{\partial l}{\partial \mu} = 0 \Rightarrow \frac{\partial}{\partial \mu} \sum_i (x-\mu)' R^{-1} (x-\mu) = -2 \sum_i R^{-1} (x-\mu) = 0$$

multiplying by R we get

$$\hat{\mu} = \frac{1}{n} \sum_i x$$

Covariance estimation :

$$\frac{\partial l}{\partial R} = 0 \Rightarrow -\frac{n}{2} \frac{\partial}{\partial R} \log |R| - \frac{1}{2} \sum_i \frac{\partial}{\partial R} (y_i - \mu)' R^{-1} (y_i - \mu) = 0$$

$$-nR^{-1} + \sum_i R^{-1} (y_i - \mu)(y_i - \mu)' R^{-1} = 0$$

Multipying by R on the left and on the right

$$\hat{R} = \frac{1}{n} \sum_i (y_i - \mu)(y_i - \mu)'$$

Note: see the derivative rules presented in Lecture 1

Properties of ML estimates

•Invariance to nonlinear 1-1 transformations of the parameter:

$$\text{if } z = f(x) \text{ then } \hat{z}_{ML} = f(\hat{x}_{ML})$$

In the sequel we consider that we observed a sequence of *n* iid variables, with density *p*, such that *p* belongs to the family of functions adopted as model, *i.e.* $p=p_{x_0}$, where x_o is the correct value of the parameter *x*.

• the ML estimator is consistent ($\hat{x}$ converges to x_o when *n* tends to infinity)

• has a normal asymptotic distribution:

$$\sqrt{n}(\hat{x} - x_o) \xrightarrow[d]{} N(0, J)$$

where J is the Fisher information matrix associated to a single observation:

$$J = E\{\frac{dl}{dx}\frac{dl}{dx}^T\}$$

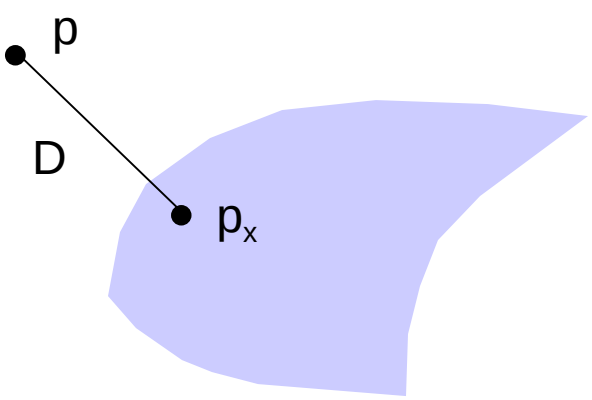
Approximation

How does the ML method work if the model is wrong ?

Proposition

Let *y* be a sequence of *n* iid variables with density *p* and $\{p_x\}$ a family of functions depending on *x* (model). If the ML estimates converge to $\hat{x}$ when n tends to infinity, then $\hat{x}$ minimizes the Kullback-Leibler divergence between *p* and $\{p_x\}$ (model):

$$D[p, p_X] = \int p(y) \log \frac{p(y)}{p_X(y)} dy$$



The divergence is a non symmetric distance.

ML vs LS

When the observations are normally distributed

$$p(y \mid \theta) = Ce^{-E(\theta)}$$

The maximum likelihood method is equivalent to the least squares method.

Polynomial Approximation

Data: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

Model: $y = \phi(x_i)'\theta + w_i \qquad \phi(x_i)' = [1 \quad x_i \quad x_i^2] \qquad w_i \sim N(0, \sigma^2)$

Log likelihood function:

$$l(\theta) = C - \frac{1}{2\sigma^2} \sum_i [y_i - \phi(x_i)'\theta]^2 = C - E(\theta)$$

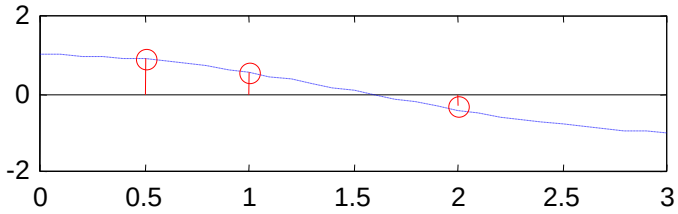
$$\theta = A^{-1}b \qquad A = \sum_{i=1}^{n} \phi(x_i)\phi(x_i)' \qquad b = \sum_{i=1}^{n} \phi(x_i)y_i$$

where *E* is the least squares energy. The maximum likelihood estimate is equal to the least squares estimate.

Did we gain anything ? Yes. $\hat{\theta}$ can be statistically characterized.

Example – Frequency estimation

observations

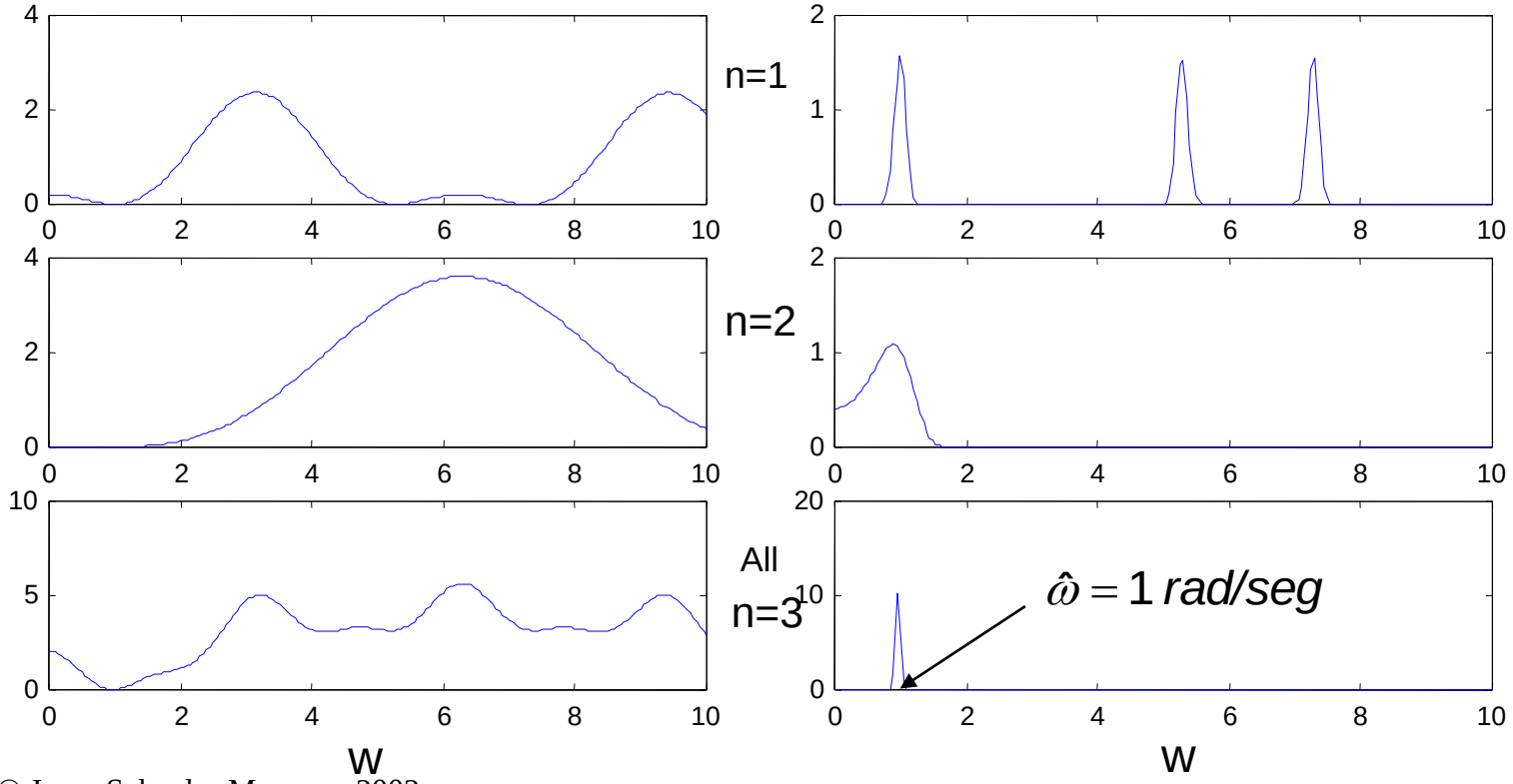


$$y_i = \cos(\omega t_i) + w_i$$

$$w_i \sim N(0,.01) \quad \omega = 1\ rad/seg$$

Energy

Likelihood function



Binary Detection

A binary symbol $\theta \in \{-A, A\}$ was transmitted in an analog channel in base band.

At the receiver, the observed signal is

$$y_i = \theta + w_i \quad i = 1,...,n$$

We wish to estimate θ assuming that the noise is uncorrelated and

$$w_i \sim N(0, \sigma^2)$$

The log likelihood function is

$$l(\theta) = -\frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2 = C + \frac{1}{\sigma^2} \sum_i y_i \theta$$

This is the expression of the *matched filter*.

Linear Prediction

Prediction is a regression problm. In linear prediction the regression vector consists of the past values of the signal.

$$y_i = a_1 y_{i-1} + ... + a_p y_{i-p} + v_i \qquad i=p+1, ...,n$$

$$y_i = \phi_i' \theta + v_i \qquad v_i \sim N(0,\sigma^2) \qquad \theta = [a_1 ... a_p]' \qquad \phi_i = [y_{i-1} ... y_{i-p}]'$$

The solution of this problem is similar to the estimation of parameters with a linear model (see polynomial approximation)

Characterization of Estimators

An estimator is a random variable which depends on x, y.

2nd order properties

Mean

$$\mu = E\{\hat{x}\}$$

Covariance matrix

$$R = E\{(\hat{x} - \mu)(\hat{x} - \mu)^T\}$$

Bias

$$B = x - E\{\hat{x}\}$$

Notes:
•ideal goal: B=0 e zero covariance. This goal is impossible in most problems.
•Expected values are computed using p(y/x).

Crámer-Rao bound

Can the covariance matrix of an unbiased estimator be the null matrix ?

Crámer-Rao theorem (unbiased estimator)

$$Cov\{\hat{x}\} \geq J^{-1}$$

J^{-1} Crámer-Rao bound, J is the Fisher information matrix

$$J = E\{\frac{dl}{dx}\frac{dl}{dx}^{T}\}$$

l log likelihood function

If $Cov\{\hat{x}\} = J^{-1}$ the estimator is called *eficient*.

definition: A>B if and only if A-B positive definite.

Example

In practice it is not always easy to compute the Crámer-Rao bound. One case which can be easily addressed concerns the estimation of the mean vector of a normal distribution, given N observations y_1,..., y_N.

$$Cov\{\hat{\mu}\} \geq \frac{1}{N} R$$

Proof: since *p(y|x)=N(μ,R),*

$$l(\mu) = C - \frac{1}{2}\sum_i (y_i - \mu)^T R^{-1} (y_i - \mu) \qquad \frac{dl}{d\mu} = R^{-1}\sum_i (y_i - \mu)$$

$$J = E\{\frac{dl}{d\mu}\frac{dl}{d\mu}^T\} = R^{-1} E\{\sum_i (y_i - \mu)\sum_j (y_j - \mu)\} R^{-1} = R^{-1} NR\, R^{-1} = NR^{-1}$$

$$CRB = J^{-1} = \frac{1}{N} R$$

Note: show that the ML estimator of m is efficient.

Monte Carlo Method

Monte Carlo Method:
Numeric evaluation of an estimator based on a large number of estimation experiments and statistical analysis of the results.

Exercises

1. Let x be a binary variable such that $P(0)=P_o$, where P_o is an unknown parameter. Determine the ML estimate of P_o from n independent realizations of x.
2. Determine the ML estimator of the mean and covariance matrix of a normal distribution $N(\mu,R)$, given n independent observations. (see Apendix).
3. Given n samples of a random signal y described by an autoregressive model $y_t= a\, y_{t-1}+ b\, w_t$, where w_t is a white process with distribution $w_t \sim N(0,1)$, determine coefficients a, b by the ML method.
4. Compute the Crámer-Rao bound for the estimation of the parameter of a Rayleigh distribution, knowing N realizations $y_1, ..., y_N$.

Appendix – Matrices (I)

Operations

Transpose:	$C = A'$	$c_{ij} = a_{ij}$
Sum:	$C = A + B$	$c_{ij} = a_{ij} + b_{ij}$
Multiplication:	$C = AB$	$c_{ij} = \sum_k a_{ik} b_{kj}$
Trace:	$c = tr(A)$	$c = \sum_k a_{ii}$
Inversion:	$C = A^{-1}$	$A^{-1}A = AA^{-1} = I$
eigen values and eigen vectors:	λ_i, v_i	$Av_i = \lambda_i v_i$

Properties

- a square matrix is non singular if all its eigen values are non zero;
- the rank of a matrix is equal to the number of eigen values different from zero; a non singular matrix has full rank.
- The eigen values of a symmetric matrix are real.

Appendix Matrices (II)

- the trace of a matrix is equal to the summ of all its eigen values;
- the determinant of a matrix is equal to the product of all its eigen values
- the inverse of a matrix has the following properties:

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$$

em que $[A] = n_1 \times n_1$ $[B] = n_1 \times n_2$
$[C] = n_2 \times n_1$ $[D] = n_2 \times n_2$

$$E = (A - BD^{-1}C)^{-1}$$

$$F = -EBD^{-1}$$

$$G = -D^{-1}CE$$

$$H = D^{-1} + D^{-1}CEBD^{-1}$$

Derivatives

The derivative of a scalar function f(X) with respect to matrix X is a matrix

$$\left[\frac{df(X)}{dX}\right]_{ij} = \frac{df(X)}{d[X]_{ij}}$$

Properties

trace

$$\frac{d}{dX}tr\{AX\} = A'$$

$$\frac{d}{dX}tr\{AX'\} = A$$

$$\frac{d}{dX}tr\{AXB\} = A'B'$$

$$\frac{d}{dX}tr\{AX'B\} = BA$$

$$\frac{d}{dX}tr\{XX'\} = 2X$$

$$\frac{d}{dX}tr\{AXBX\} = A'X'B' + B'X'A'$$

$$\frac{d}{dX}tr\{AXBX'\} = A'XB' + AXB$$

$$\frac{d}{dX}tr\{AX^{-1}B\} = -(X^{-1}BAX^{-1})'$$

determinant

$$\frac{d}{dX}|X| = |X|(X^{-1})'$$

$$\frac{d}{dX}\log|X| = (X^{-1})'$$

$$\frac{d}{dX}|AXB| = |AXB|(X^{-1})'$$

$$\frac{d}{dX}|X^n| = n|X|^n(X^{-1})'$$

Note: adapted from Sage & Melsa

Bibliography

Duda, Hart, Stork, Pattern Classification, Wiley, 2001.

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