MultiDimensional Signal Processing
Master Degree in “Ingegneria delle Telecomunicazioni”
A.A. 2015-2016

Pietro Guccione, PhD

DEI - DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL’INFORMAZIONE
POLITECNICO DI BARI

Pietro Guccione – Assistant Professor in Signal Processing
(pietro.guccione@poliba.it, http://dee.poliba.it/guccioneweb/index.html)
Basics of statistics I

- Required background of probability theory (random variable, vector of random variables, first and second order statistics, covariance matrix, joint pdf, multivariate Gaussian r.v., chi-square distribution, transformation of a r.v., random number generation)

- General concept of statistics.


- Family of exponential distributions

- Maximum Likelihood estimation, MLE applied to exponential distribution
Statistics is a mathematical science that tries to find regularity within data, supposing that:

- Data are generated by a random mechanism
- Some hypothesis can be done on the mechanism that generated the data

The final purpose of the statistics are, among others:

- Build a proper model, able to explain the data generation;
- Match the data with the hypotheses, so that prediction on future data can be done;
- Estimate the set of parameters of the model
- …
Introduction to Statistics

Statistics has a large fields of applications, since it is, firstly, an applied subject. However, a common set of statistical methods grounding on theory has been developed.

Statistics includes:
- Efficient summarization, tabulation and graphical display of data;
- Design of experiments;
- **Statistical inference**

The most application of research is in statistical inference. New recent fields of applications are:
- agricultural research, biostatistics, chemistry, climatology, image processing, quality control, …

In statistical inference, sample of data are used to draw inferences about some aspect of the *population*, from which the data derive (model construction and parameters estimation).
Point estimation
For each unknown parameter of the model, a single value is estimated from the data. No precision about such estimation is done, although precision can be estimated.

Interval estimation
It provides a range of values which has a pre-determined (high) probability of including the true, but unknown, value of parameters.

Hypothesis testing
Specific hypotheses are set regarding the parameters of interest and data observation are used to assess the validity of the hypotheses.

Underlying philosophy: *Frequentist approach*
The assumption is that we can take repeated samples of data under the same conditions, from the same population.
This approach may produce silly results in some circumstances. For this reason, the *Bayesian approach* is sometimes used, in alternative.
Alternative: *Bayesian inference*

\[
P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}
\]

**\(H\)** = any hypothesis whose probability may be affected by data. Often, there are competing hypotheses from which one chooses the most probable.

**\(P(H)\)** = prior probability, i.e. the probability of \(H\) before the discovering of \(E\).

**\(E\)** = the data (evidence).

**\(P(H \mid E)\)** the posterior probability, i.e. the probability of \(H\) after \(E\) is discovered.

**\(P(E \mid H)\)**, the likelihood, i.e. the probability of observing just \(E\), given the hypotheses \(H\).

**\(P(E)\)** = the probability of the evidence (model evidence).
Bayesian inference

No need to invoke the possibility to repeat the sampling.

The parameters of interest have a probability of distribution, so no fixed number is finally given as estimation of the parameters, instead a probability is given. The ‘dispersion’ of such probability is the *precision* with which the parameters are known.

*Difficulty*: how to choose the probability distribution of the parameters.
Background

Basics of Probability Theory. Main issues:

- random variable,
- vector of random variable, and joint pdf
- first and second order statistics,
- covariance matrix,
- multivariate Gaussian r.v.,
- chi-square distribution,
- transformation of a r.v.,
- random number generation
A random variable is a variable whose value is subject to variations due to chance (i.e. randomness).

It can take on a set of possible different values in a given interval, each of them with associated value of probability.

A continuous r.v. is associated to a continuous set of values taken on; a discrete r.v. is associated to a discrete set (i.e. numerable) set of values.

\[ f_X(x) \]

\[ F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x) \, dx \]

Some properties

\[ f_X(x) \geq 0, \quad \int_{-\infty}^{+\infty} f_X(x) \, dx = 1 \]

\[ f_X(x) = \frac{dF_X(x)}{dx} \]

\[ P(A) = \int_{A} f_X(x) \, dx \]

\[ x_1 > x_2 \implies F_X(x_1) \geq F_X(x_2) \]
A vector of random variables is a collection of r.v. jointly considered:

\[ X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \]

\[ F_X(x) = F_{X_1X_2...X_n}(x_1, x_2, ... x_n) \]

\[ f_X(x) = f_{X_1X_2...X_n}(x_1, x_2, ... x_n) = \frac{d^n F_{X_1X_2...X_n}(x_1, x_2, ... x_n)}{dx_1 dx_2 ... dx_n} \]

Joint probability cumulative function;
Joint probability density function
Vector of Random Variables

Case: \( n=2 \)

\[
\begin{bmatrix}
X \\
Y
\end{bmatrix}
\quad F_{XY}(x, y)
\]

\[
f_{XY}(x, y) = \frac{d^2 F_{XY}(x, y)}{dxdy}
\]

Marginal probabilities

\[
f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy
\]

\[
f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx
\]

Conditioned probabilities

\[
f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}
\]

\[
f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}
\]

Remember that

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}
\]
First and second order statistics

\[ \mu_X = E[X] = \int_{-\infty}^{+\infty} x \cdot f_x(x) \, dx \]

Case: \( n=1 \)

\[ \sigma_X^2 = E[(X - \mu_X)^2] \]

\[ \mu = E \begin{bmatrix} X \\ Y \end{bmatrix} \]

Case: \( n=2 \)

\[ \Sigma = \begin{bmatrix} \sigma_X^2 & c_{XY} \\ c_{XY} & \sigma_Y^2 \end{bmatrix}, \quad c_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \]

Case: \( n>2 \)

\[ \mu = E \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \]

\[ \Sigma = \begin{bmatrix} \sigma_{X_1}^2 & c_{X_1X_2} & \cdots & c_{X_1X_n} \\ c_{X_2X_1} & \sigma_{X_2}^2 & \cdots & c_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{X_nX_1} & c_{X_nX_2} & \cdots & \sigma_{X_n}^2 \end{bmatrix}, \quad c_{X_iX_j} = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})] \]
The **covariance matrix** is the matrix of the mutual co-variances. It is positive, semi-definite and symmetric

$$
\mu = E \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{X_1}^2 & c_{X_1X_2} & \cdots & c_{X_1X_n} \\ c_{X_2X_1} & \sigma_{X_2}^2 & \cdots & c_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{X_nX_1} & c_{X_nX_2} & \cdots & \sigma_{X_n}^2 \end{bmatrix}, \quad c_{X_iX_j} = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})]
$$

The correlation matrix

$$
R = \begin{bmatrix} m_{X_1}^2 & r_{X_1X_2} & \cdots & r_{X_1X_n} \\ r_{X_2X_1} & m_{X_2}^2 & \cdots & r_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_nX_1} & r_{X_nX_2} & \cdots & m_{X_n}^2 \end{bmatrix}, \quad m_{X_i}^2 = E[X_i^2] \quad r_{X_iX_j} = E[X_iX_j] \quad \Sigma = R - \mu\mu^T
$$

The Covariance matrix of uncorrelated r.v.:

$$
\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{X_2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{X_n}^2 \end{bmatrix} = \text{diag}\left\{ \sigma_{X_1}^2, \sigma_{X_2}^2, \ldots, \sigma_{X_n}^2 \right\}
$$
The multivariate Gaussian distribution is a generalization of the one-dimensional (univariate) Gaussian (or normal) distribution to higher dimensions. In this random vector, every linear combination of its $n$ components has a univariate normal distribution. The joint probability density function is known, once the vector of means and the covariance matrix are known:

$$x \sim N(\mu, \Sigma)$$

$$f_x(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$

The knowledge of the vector of means and the matrix of covariance is sufficient to infer the joint distribution. Uncorrelation implies independency; since independency always implies uncorrelation, the two concepts are equivalent for a multivariate Gaussian distribution.
The case for \( n=2 \) is particularly interesting. 

\( \rho \) is the correlation coefficient between \( X \) and \( Y \):

\[
\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
\]

\((x, y) \sim N([\mu_X, \mu_Y]^{T}, \Sigma),\)

\[
\Sigma = \begin{bmatrix}
\sigma_x^2 & \rho \sigma_x \sigma_y \\
\rho \sigma_x \sigma_y & \sigma_y^2
\end{bmatrix}
\]

\[
f_{XY}(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} \exp \left[-\frac{1}{2(1 - \rho^2)} \left[ \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} - 2\rho (x - \mu_x)(y - \mu_y) \right] \right]
\]
Transformation of a r.v. /1

A r.v. can be transformed into another r.v.

\[ Y = g(X) \]
\[ f_Y(y) = ? \]

Solution:

\[ F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int_{D_Y} f(x)dx, \quad D_Y \{ x \mid g(x) \leq y \} \]

If \( g(x) \) is invertible (strictly increasing):

\[ F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \]
\[ f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dx} \frac{dx}{dy} = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} \]

(strictly decreasing):

\[ F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \]
\[ f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{dF_X(g^{-1}(y))}{dx} \frac{dx}{dy} = -\frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} \]
A couple of r.v. can be transformed into another r.v.

\[ Z = g(X, Y) \]

\[ f_Z(z) = ? \]

**Solution:**

\[ F_Z(y) = P(Z \leq z) = P(g(X, Y) \leq z) = \int_{D_Z} f_{XY}(x, y) dxdy, \quad D_z \{ z \mid g(x, y) \leq z \} \]

**Example:**

\[ Z = X + Y \]

\[ F_Z(z) = P(Z \leq z) = P(X + Y \leq z) = P(X, Y \leq z - X) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x, y) dydx \]

And, if independent:

\[ F_Z(z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x, y) dydx \]

\[ f_Z(z) = \frac{d}{dz} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x, y) dydx = \int_{-\infty}^{+\infty} f_{XY}(x, z - x) dx = \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x) dx \Rightarrow \]

\[ f_Z(z) = f_X(z) \ast f_Y(z) \]
Examples

Transformation of a Gaussian r.v. with the function:

\[ Y = X^2 \]
\[ X \sim N \left(0, \sigma^2 \right) \]

Transformation of a Gaussian couple of r.v. with the function:

\[ Z = \sqrt{X^2 + Y^2} \]
\[ X, Y \sim N \left(0, \sigma^2 \right) \]

Transformation of a Gaussian couple of r.v. with the function:

\[ Z = X^2 + Y^2 \]
\[ X, Y \sim N \left(0, \sigma^2 \right) \]
The $\chi^2$ (said chi-square) distribution with $k$ degrees of freedom is the result of summing the square of $k$ standard normal variables, i.e.:

$$X = \sum_{n=1}^{k} Z_n^2$$

$Z \sim N(0,1)$

$X \sim \chi(k)$

This distribution is widely used in inferential statistics, e.g. in hypothesis testing or to build confidence intervals.

The chi-squared distribution is used also in the so-called chi-squared tests for the goodness of fit of a samples drawn from a population to agree with a theoretical distribution

$$f_X(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

$$E[X] = k$$

$$\sigma_X^2 = 2k$$

$$\Gamma(n) = (n-1)!$$

$$\Gamma(t) = \int_{-\infty}^{t} x^{t-1} e^{-x} dx$$
The chi-squared test is a statistical hypothesis test in which the sampling distribution of the test statistics is a chi-square distribution when the null hypothesis is true. The test is often generated by the sum of squared errors or through a sample variance. The test statistics follows also the hypothesis of independent data Gaussian distributed. This hypothesis, for the central limit theorem, is often valid for error data (remember that error is the unexplained residual of a model vs data and may be thought as due to our lacking in further detailing the model. It is generated by many secondary factors of, presumably, the same amount and independent among them).

The chi-squared test is used to determine whether a significant difference exists between the data and the model. An example is the Pearson chi-squared test.

\[ X^2 = \sum_{i=1}^{c} \frac{(n_i - E_i)^2}{E_i} \]

\[ X^2 = \sum_{i=1}^{c} \sum_{t=1}^{2} \frac{(n_{it} - E_{it})^2}{E_{it}} \]

**Suggestion for laboratory:**

Deepen the argument, implement (and justify the result) a Pearson chi-square to assess:

- Goodness of fit (a population with a distribution)
- Homogeneity (comparison of two distribution)
It may arise the need to generate random numbers from PC. The general ‘pseudo-random’ routines within the PC are related to register contents and clock and provide a random number uniformly generated between 0 and 1. What is the transformation that generates random number according to a given distribution $Y$, provided a uniform distribution $X$ in $[0,1]$?

$$X \sim U[0,1]$$

$$Y = \phi(X), \quad s.t. f_Y(y) \in [a,b] \quad \text{is given}$$

Solution of the problem is the function. $\phi = F_Y^{-1}$

Proof:

$$F_X(x) = x, \quad x \in [0,1]$$

$$F_Y(t) = P\left(F_Y^{-1}(X) \leq t\right) = P(X \leq F_Y(t)) = F_X \left(F_Y(t)\right) = F_Y(t)$$

$$Y = \phi(X) = F_Y^{-1}(X)$$
Example:
We need to generate a population of samples from a r.v. exponentially distributed

\[ f_Y(y) = \lambda \exp(-\lambda y) \]
\[ F_Y(y) = 1 - \exp(-\lambda y) \]
\[ y \geq 0 \]

The inverse of the probability distribution function is:

\[ \phi(x) = F_Y^{-1}(x) = -\frac{1}{\lambda} \log(1 - x) \]

If X is uniform in the interval [0,1], the r.v.

\[ Y = \phi(X) = -\frac{1}{\lambda} \log(1 - X) \]

is distributed accordingly to the wanted exponential
Collection of data samples.
Different methods of data collection may exist, based on the fact that data must be:
- Generated from an experiment (Design of Experiment),
- Are available, but there is a large choice (sampling),
- Are available, but the choice is limited (a model is needed),
- Are available, but we impose a model to select which data fit our interest.

Lecture 3 has been dedicated to the problem of data collection.
The likelihood

Let us suppose to have a set of n observations \( \{ x_1, x_2, \ldots, x_n \} \)

We want to draw conclusions about the value of one or more parameters \( \{ \theta \} \)

The quantities \( \{ x_1, x_2, \ldots, x_n \} \) are supposed particular outcomes of the random variables

\[
\{ X_1, X_2, \ldots, X_n \}
\]

These r.v. are supposed independent and identically distributed, with a distribution whose pdf is

\[
f_X(x; \theta)
\]

and depends on the parameters \( \{ \theta \} \).

The likelihood is the answer to the question: what is the joint probability that exactly those samples were extracted from the population?

This is

\[
f_X(x; \theta) = \prod_{k=1}^{n} f_X(x_k; \theta)
\]

If viewed as a function of \( \theta \) it is addressed as

\[
L(\theta; x) = f(x; \theta)
\]

Often, the log-likelihood is considered

\[
\ell(\theta; x) = \log L(\theta; x)
\]
Given the set of $n$ samples $\{x_1, x_2, ..., x_n\}$ taken from a probability distribution, $f_X(x; \theta)$, we use the observation to estimate the parameter, which is unknown.

Let $\hat{\theta}(x_1, x_2, ..., x_n)$ the *estimated* parameter.

The corresponding function that produced the estimation is $\hat{\theta}(X_1, X_2, ..., X_n)$ and it is a function of the r.v. and it is itself a r.v.. It is called *estimator* for $\theta$.

In a situation in which it is necessary to compare more estimators, we look for properties such as *unbiasedness*, *consistency*, *efficiency* and others.
Properties of estimators /1

Unbiasedness
The estimator $\hat{\vartheta}$ is said unbiased if $\mathcal{I} = E[\hat{\vartheta}]$. The bias is then $\mathcal{I} - E[\hat{\vartheta}]$.

The estimator is a distribution, if the estimator is unbiased, this distribution is centered around the true value of the parameter (this supposes that a true value exists, i.e. that the parameter is deterministic even if unknown).

Example:
Take samples from a Normal distribution with parameters $\{\mu, \sigma^2\}$
The most obvious estimator of the mean is:
$$\hat{\mu} = \frac{1}{n} \sum_{k} x_k$$
which is also unbiased (let’s try a proof).
Properties of estimators /2

Consistency
The estimator \( \hat{\theta} \) is said consistent if the variance of the estimator tends to zero as the size of the sample goes to infinity.

Said in equations:

\[
P\left( \left| \theta - \hat{\theta} \right| > \varepsilon \right) \to 0 \quad n \to \infty
\]

Example:
Take samples from a Normal distribution with parameters \( \{ \mu, \sigma^2 \} \)

The arithmetic mean is a consistent estimator of the mean, since the variance of the estimator is proportional to \( 1/n \)
Efficiency
A unbiased and consistent estimator are good properties. But they could be not enough. Two or more estimator may have these properties. How can we distinguish good estimators among them? Efficiency of an estimator is a possible answer.

An unbiased estimator is said efficient if it has, with that given size of sample, the minimum possible variance; the efficiency of all the other estimators is the ratio of the minimum possible variance with the actual variance.
Cramer-Rao bound

Cramer-Rao inequality
The Cramer-Rao inequality gives the lower bound for an estimator variance (i.e. the precision that we can get about an estimator)

Given a set of r.v. \( \{X_1, X_2, \ldots, X_n\} \) taken from the distribution \( f(x; \vartheta) \), under certain regularity conditions, the estimator \( \hat{\vartheta} \) is subject to:

\[
Var[\hat{\vartheta}] \geq I_{\vartheta}^{-1}
\]

\[
I_{\vartheta} = E \left[ \left( \frac{\partial \ln[L(\vartheta; x)]}{\partial \vartheta} \right)^2 \right] = E \left[ \left( \frac{\partial \ell}{\partial \vartheta} \right)^2 \right]
\]

The quantity \( I_{\vartheta} = E \left[ \left( \frac{\partial \ell}{\partial \vartheta} \right)^2 \right] \) is said Fisher information (matrix), since the lower is the variance, the more ‘information’ we have about the parameters and the larger is \( I_{\vartheta} \).
The Maximum Likelihood estimation is the best known, most used and most important method of estimation.

Given the likelihood function

\[ L(\theta; x) = f(x; \theta) \]

The Maximum Likelihood Estimator (MLE) is the value \( \hat{\theta} \) that maximizes \( L(\theta; x) \)

Since finding the maximum is an optimization problem (often very complicate), it may be difficult to find closed-form expressions for \( \hat{\theta} \).

Often, the log-likelihood is maximized:

\[ \hat{\theta} = \arg \max_{\theta} \{ \ell(\theta; x) \} \]
If an interval for $\theta$ is known (or intervals), it may be a good idea to plot the trend of the likelihood as a function of the parameter(s) $\theta$, to have some information regarding possible problems in finding the MLE using the optimization methods.
Application of properties

Family of exponential distributions: Gaussian, Beta distribution and Binomial distribution, Gamma distribution.

These are general kind of distributions that may be written in a general form

\[ f(x; \vartheta) = \exp \left\{ \sum_{j=1}^{k} A_j(\vartheta) B_j(x) + C(x) + D(\vartheta) \right\} \]

Often, the exponential distributions are preferred as models for the data, since this way the log-likelihood becomes a sum of relatively simple expression.

Example: the Gaussian distribution becomes the sum of a constant and a quadratic expression of \( x \)

\[ \ell(\vartheta; x) = \log L(\vartheta; x) = \log \left\{ \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right] \right\} = \]

\[ \log \left[ \frac{1}{\sqrt{2\pi\sigma}} \right] - \frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \]
Maximum Likelihood and exponential distribution

The Maximum Likelihood for exponential distribution can take particularly simple forms.

\[ \phi_j = A_j(\vartheta) \]

Natural (or canonical) parameters

\[ f(x; \phi) = \exp \left\{ \sum_{j=1}^{k} \phi_j B_j(x) + C(x) + D(\phi) \right\} \]

The MLE estimator is, in this case:

\[ t_j = \text{MLE}\{\phi_j\} = E[T_j] = E\left[ \sum_{i=1}^{n} B_j(x_i) \right] \]

\[ \ell(\phi; x) = \sum_{j=1}^{k} \phi_j t_j + \sum_{i=1}^{n} C(x_i) + nD(\phi) = \text{const} + \sum_{j=1}^{k} \phi_j t_j + nD(\phi) \]

\[ \frac{\partial \ell}{\partial \phi_j} = t_j + n \frac{\partial D(\phi)}{\partial \phi_j} = 0 \Rightarrow t_j = -n \frac{\partial D(\phi)}{\partial \phi_j} \]

\[ \left\{ E\left[ \frac{\partial \ell}{\partial \phi_j} \right] = 0 \iff E[T_j] = -n \frac{\partial D(\phi)}{\partial \phi_j} \right\} \]

Cramer-Rao inequality
Examples of Maximum Likelihood

Examples:
Estimation of the parameter of an exponential variable

\[ f(x; \theta) = \theta \cdot \exp(-\theta x), \quad x > 0 \]
\[ L(\theta; x) = \theta^n \cdot \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \]
\[ \ell(\theta; x) = n \log \theta - \theta \sum_{i=1}^{n} x_i \]

Estimation of the mean of a Gaussian variable

\[ f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right) \]
\[ L(\theta; x) = ... \]
\[ \ell(\theta; x) = ... \]

[to solve by yourself]
Computation of MLE

Sometimes, equations for MLE may not have explicit solutions and iterative methods are needed to solve the equations.

Newton-Raphson method

\[ \ell(\mathcal{G}; x) \Rightarrow g(\mathcal{G}) = \frac{\partial \ell}{\partial \mathcal{G}}, \quad H(\mathcal{G}) = \frac{\partial^2 \ell}{\partial \mathcal{G}_i \partial \mathcal{G}_j} \]

\[ g(\mathcal{G}) \approx g(\mathcal{G}_0) + H(\mathcal{G}_0) \cdot (\mathcal{G} - \mathcal{G}_0) + ... \Rightarrow \]

\[ g(\mathcal{G}) = 0 \Leftrightarrow 0 = g(\mathcal{G}_0) + H(\mathcal{G}_0) \cdot (\mathcal{G} - \mathcal{G}_0) \Rightarrow \]

\[ \mathcal{G}_{n+1} = \mathcal{G}_n - H^{-1}(\mathcal{G}_n) \cdot g(\mathcal{G}_n) \]

Fisher scoring:

it allows the computation of the first derivative only. The Fisher information matrix is always definite-positive, so there are no problem of convergence as in the N-R method.

\[ E \left[ \frac{\partial^2 \ell}{\partial \mathcal{G}_i \partial \mathcal{G}_j} \right] = -E \left[ \left( \frac{\partial \ell}{\partial \mathcal{G}_i} \right) \left( \frac{\partial \ell}{\partial \mathcal{G}_j} \right) \right] \]

In Newton-Raphson method, use

\[ -E \left[ \left( \frac{\partial \ell}{\partial \mathcal{G}_i} \right) \left( \frac{\partial \ell}{\partial \mathcal{G}_j} \right) \right] \text{ instead of } H(\mathcal{G}) \]
Simplex method

N-R and Fisher scoring are examples of gradient methods, since the derivative is used to decide the direction of change between successive estimates of $\theta$. The simplex method is a direct-search method that does not require the computation of derivatives.

At each stage, the value of $\ell(\theta; x)$ is calculated at the vertices of a simplex (which is a triangle for a 2D problem, a tetrahedron for a 3D problem and so on). These values determine position and size of the next simplex. The position is adjacent to the previous simplex, and the direction is the one that increases $\ell(\theta; x)$.

The size of the simplex shrinks as the maximum is approached, ensuring convergence.

This method is more efficient when multiple local maxima are present in the log-likelihood, since it avoids that the solution is trapped in local maxima.