Estimation Theory

Estimation theory deals with finding numerical values of interesting parameters from given set of data. We start with formulating a family of models that could describe how the data were generated. This model will usually belong to a family of models indexed by parameters of interest (i.e. each model from within this family will be uniquely identified by the numerical values of its parameters). We are interested in fitting a model from such a family to the data, i.e. finding values of parameters that define model describing data in the best way.

Denote the $T$-dimensional data vector as $x_T = (x(1), x(2), \ldots, x(T))^T$ and the vector of parameters as $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_m)^T$. An estimator $\hat{\Theta}$ of the parameter vector $\Theta$ is a function which allows to estimate the parameters from the data. A numerical value of the estimator $\hat{\Theta}$ is called estimate of $\Theta$. 
Well known examples of estimators are the sample mean and variance of the random variable $x$:

\[
\begin{align*}
\hat{\mu} &= \frac{1}{T} \sum_{i=1}^{T} x_i \\
\hat{\sigma}^2 &= \frac{1}{T-1} \sum_{i=1}^{T} (x_i - \hat{\mu})^2
\end{align*}
\]

- Classic estimation - the parameters are regarded as some unknown constants;

- Bayesian estimation - parameters treated as random variables with some \textit{a priori} pdf $p_{\Theta}(\Theta)$.
Classical estimation

Method of moments: Assume that there are $T$ independent measurements $x(1), \ldots, x(T)$ from a common probability distribution $p(x|\Theta)$ characterized by the parameters $\Theta = (\Theta_1, \ldots, \Theta_m)^T$. We want to estimate the values of the parameters $\Theta_i$ from the data.

Recall, that the $j$th moment of a probability distribution is defined as

$$\alpha_j = E[x^j|\Theta] = \int x^j p(x|\Theta) \, dx$$

- calculate the expressions of $m$ moments $\alpha_j$ as functions of parameters from their definitions;

- calculate sample moments from the data:
  $$d_j = \frac{1}{T} \sum_{i=1}^{T} [x(i)]^j$$
• form $m$ equations by equating the expressions for moments with their estimates $\alpha_j(\Theta) = d_j$;

• solve the resulting system of $m$ equations for the $m$ unknown parameters

Usually, first $m$ moments are sufficient. The solutions to the equation $\alpha_j(\Theta) = d_j$ are called moment estimators.

**Example:** Let the data sample $x(1), \ldots, x(T)$ be drawn independently from a pdf

$$p(x|\Theta) = \frac{1}{\Theta_2} \exp \left(-\frac{x - \Theta_1}{\Theta_2}\right)$$

with $\Theta_1 < x < \infty$ and $\Theta_2 > 0$. Calculate the first 2 moments

$$\alpha_1 = E[x|\Theta] = \int_{\Theta_1}^{\infty} \frac{x}{\Theta_2} \exp \left(-\frac{x - \Theta_1}{\Theta_2}\right) dx$$

$$= \ldots = \Theta_1 + \Theta_2$$
\[
\alpha_2 = E[x^2|\Theta] = \int_{\Theta_1}^{\infty} \frac{x^2}{\Theta_2} \exp \left( -\frac{x - \Theta_1}{\Theta_2} \right) dx
\]

\[
= \ldots = (\Theta_1 + \Theta_2)^2 + \Theta_2^2
\]

Estimating the sample moments, \(d_1, d_2\) from the data we can form the system of 2 equations

\[
\Theta_1 + \Theta_2 = d_1
\]
\[
(\Theta_1 + \Theta_2)^2 + \Theta_2^2 = d_2
\]

which gives the moment estimates

\[
\hat{\Theta}_1 = d_1 - (d_2 - d_1^2)^{1/2}
\]
\[
\hat{\Theta}_2 = (d_2 - d_1^2)^{1/2}
\]

This method can also be used with the central moments \(\mu_j = E[(x - \alpha_1)^j|\Theta]\) and their sample estimates \(s_j = \frac{1}{T-1} \sum_{i=1}^{T} (x(i) - d_i)^j\). The method of moments is simple; does not have statistically desired properties, hence is used if no better alternatives are available;
**Least-Squares estimation:** Consider a model

\[ y_T = H\Theta + v_T \]

where \( \Theta \) is the vector of parameters, \( v_T \) is the vector of measurement errors and \( H \) is an observation matrix. \( y_T \) is an observation vector. The number of observations is greater than the number of parameters.

The least squares criterion

\[ e_{LS} = \frac{1}{2} \| v_T \|^2 = \frac{1}{2}(y_T - H\Theta)^T(y_T - H\Theta) \]

allows to minimize the effect of the measurement errors. The solution minimising the \( e_{LS} \) is

\[ \hat{\Theta}_{LS} = H^+y_T \]

where \( H^+ = (H^TH)^{-1}H^T \) is a pseudoinverse of \( H \).
Example: Many nonlinear curve-fitting problems can be turned into a linear regression by applying an appropriate transform of the data, e.g. fitting the exponential relationship to the data

\[ y(t) = e^{at+b} \]

can be achieved by applying the log transform first and performing linear regression

\[ \log(y)(t) = at + b \]

Example: More generally least squares can be applied also if the model is linear in parameters but nonlinear in observations

\[ y(t) = \sum_{i=1}^{m} a_i \phi_i(t) + v(t) \]

where the basis functions \( \phi_i \) are nonlinear in \( t \).
Maximum likelihood (ML) estimation

- assumes that the unknown parameters are constant;
- no prior information available;
- good statistical properties;
- works well when there is a lot of data;
- ML solution \( \hat{\Theta} \) chooses the parameters defining the model under which the data are most likely.

The likelihood function

\[
p(x_T|\Theta) = p(x(1), \ldots, x(T)|\Theta)
\]

has the same form as the joint density of the measurements. Often, the log likelihood is used \( \ln p(x_T|\Theta) \).
The likelihood equation

$$\frac{\partial}{\partial \Theta} \ln p(x_T|\Theta) = 0$$

enables to find the maximum of the likelihood function.
Assuming that the measurements are independent, the likelihood factors out

$$p(x_T|\Theta) = \prod_{j=1}^{T} p(x_j|\Theta)$$

where $p(x_j|\Theta)$ is a conditional pdf of a measurement $x_j$. In this case, the log likelihood consists of a sum of logs of conditional pdfs.

The vector likelihood equation consists of $m$ scalar equations

$$\frac{\partial}{\partial \Theta_i} \ln p(x_T|\Theta) = 0, \quad i = 1, \ldots, m$$

Usually these are nonlinear and coupled - need numerical solution methods.
Example: Assume $T$ independent observations of a r.v. $x$ about which we assume its pdf to be $N(\mu, \sigma^2)$ with unknown $\mu$ and $\sigma^2$. We can use maximum likelihood method to determine the values of parameters that make the data most likely. The likelihood function

$$p(x_T|\mu, \sigma^2) = (2\pi \sigma^2)^{-T/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^{T} (x(j) - \mu)^2 \right)$$

The log likelihood function

$$\ln p(x_T|\mu, \sigma^2) = -\frac{T}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{T} (x(j) - \mu)^2$$

and the likelihood equations

$$\frac{\partial}{\partial \mu} p(x_T|\mu, \sigma^2) = 0$$
$$\frac{\partial}{\partial \sigma^2} p(x_T|\mu, \sigma^2) = 0$$
produce

\[ \frac{1}{\sigma^2} \sum_{j=1}^{T} (x(j) - \mu) = 0 \]

\[ -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{T} (x(j) - \mu)^2 = 0 \]

Thus the ML estimates of the unknown parameters are \( \hat{\mu}_{ML} = \frac{1}{T} \sum_{j=1}^{T} x(j) \) and \( \hat{\sigma}^2_{ML} = \frac{1}{T} \sum_{j=1}^{T} (x(j) - \hat{\mu}_{ML})^2 \).

Hence, the ML estimates of the Gaussian first 2 moments amount to sample mean and sample variance.

Note: If, in the least squares estimation, the measurement errors are i.i.d. r.v.’s with pdf \( N(0, \sigma^2) \), then the least squares estimation yields the maximum likelihood estimates.
Bayesian estimation

Classical estimation methods assumed that the parameters were unknown deterministic constants. In Bayesian estimation

- parameters are *random*;

- *a priori* parameter pdf \( p_{\Theta}(\Theta) \) assumed to be known;

- the *posterior density* \( p_{\Theta|X}(\Theta|x_T) \) contains all the information about the parameters \( \Theta \)

*Maximum a posteriori (MAP) estimator* is a popular method of estimation of parameters in Bayesian framework
• similar to the classical maximum likelihood;

• maximise the posterior pdf $p_{\Theta|x}(\Theta|x_T)$ of $\Theta$ given data $x_T$;

• the MAP estimator is the most probable value of the parameter as evidenced by data;

• From Bayes rule

$$p_{\Theta|x}(\Theta|x_T) = \frac{p_{x|\Theta}(x_T|\Theta)p_{\Theta}(\Theta)}{p_{x}(x_T)}$$

• the denominator - the prior density of data, independent of $\Theta$;

• hence, finding MAP estimator amounts to maximising the nominator - the joint pdf of parameters and data $p_{\Theta,x}(\Theta,x_T)$. 

12
• MAP estimator can be found from the log likelihood equation

\[
\frac{\partial}{\partial \Theta} \ln p(\Theta, x_T) = \frac{\partial}{\partial \Theta} \ln p(x_T|\Theta) + \frac{\partial}{\partial \Theta} \ln p(\Theta) = 0
\]

• in the above the first term in the sum is the same as in the maximum likelihood;

• the second term - prior information about the parameters

Note: If the prior pdf of \( \Theta \) is uniform for parameter values for which \( p(x_T|\Theta) > 0 \) (i.e. second term above equals zero), then MAP and ML estimators are equal. Thus, MAP and ML coincide when there is no available prior information about the parameters.

Example: Assume that \( x(1), \ldots, x(T) \) are independent observations drawn from
\( N(\mu_x, \sigma_x^2) \). The \( \mu_x \) is itself a rv with pdf \( N(0, \sigma_{\mu}^2) \).

Assuming that \( \sigma_x^2, \sigma_{\mu}^2 \) are known, find estimator \( \hat{\mu}_{MAP} \). Using log likelihood equation which yields

\[
\hat{\mu}_{MAP} = \frac{\sigma_{\mu}^2}{\sigma_x^2 + T\sigma_{\mu}^2} \sum_{j=1}^{T} x(j)
\]

- if we have no a priori knowledge about \( \mu \):

\[
\sigma_{\mu}^2 \to \infty \Rightarrow \hat{\mu}_{MAP} \to \frac{1}{T} \sum_{j=1}^{T} x(j)
\]

- Similarly, \( T \to \infty \Rightarrow \hat{\mu}_{MAP} \to \frac{1}{T} \sum_{j=1}^{T} x(j) \)

i.e. if the number of samples increases, the prior information about parameters becomes less important;

- MAP estimator combines the prior information and samples according to their reliability (i.e. \( \sigma_{\mu}^2, \sigma_x^2 \)).
Maximum Entropy Principle

- information on the pdf $p_x$ of a scalar r.v. $x$, e.g. having estimates of expectations of a number of functions of $x$,

$$E[F_i(x)] = c_i, \text{ for } i = 1, \ldots, m$$

may not be enough for unique (principled) choice of pdf.

- finite number of data points may not tell us exactly what the underlying probability distribution is.

- Maximum entropy principle - (regularisation) criterion stating that out of all probability distribution functions compatible with the constraints above one should choose one which has least structure.
• least structure = maximum entropy (where entropy is extended from rv’s to pdfs).

• entropy measures randomness - thus maximizing entropy means choosing pdf compatible with the constraints and making least assumptions about the data.

• under some conditions a maximum entropy pdf satisfying conditions above is

\[ p_0(\xi) = A \exp\left(\sum_i a_i F_i(\xi)\right) \]

where constants \( A, a_i \) are determined from constraints above;
• gaussian pdf has the largest entropy among all pdfs with zero mean and unit variance,

\[ p_0(\xi) = A \exp(a_1\xi^2 + a_2\xi) \]

• in many dimensions - gaussian pdf has maximum entropy among all pdfs with the same covariance matrix.

• thus - gaussian is the least structured of all distributions and entropy can be used as a measure of nongaussianity.