Brief Review on Estimation Theory

K. Abed-Meraim

ENST PARIS, Signal and Image Processing Dept.

abed@tsi.enst.fr

This presentation is essentially based on the course ‘BASTA’ by E. Moulines

Presentation Outline

- Basic concepts and preliminaries
- Parameter estimation
- Asymptotic theory
- Estimation methods (ML, moment, ...)

Basic Concepts and Preliminaries

Definition and applications

The statistics represent the set of methods that allow the analysis (and information extration) of a given set of observations (data). Application examples include:

- The determination of the production quality by a probing study.
- The measure of the visibility impact of a web site (i.e. number of readed pages, visiting strategies, ...).
- The modelisation of the packets flow at a high-speed network gate.
- The descrimination of important e-mails from spam.
- The prediction of missing data for the restoration of old recordings.
- The estimation and tracking of a mobile position in a cellular system.
- etc, etc, ...

Some history...

One can distinguish 3 phases of development:

- Begining of XIX-th century, apprition of the first data analysis experiments (Prony, Laplace) and the first canonical method in statistics (Gauss, Bayes).
- In the first part of the XX-th century (until the 1960s) the basis of the statistical inference theory have been established by (Pearson, Fisher,, Neyman, Cramer,...). However, due to the lack of powerful calculation machines, the applications and the impact of the statistics were quite limited.
- With the fast development of computers and data bases, the statistic has seen a huge expansion and the number of its applications covers a very large number of domains either in the industry or in research labs.

Statistical model

- In statistics, the observation $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ are seen as a realization of a random vector (process) $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ which law P is *partially* known.

- The observation model translates the *a priori knowledge* we have on the data.

- The nature and complexity of the model varies considerably from one application to another...

Parametric model

- *Parametric model*: is a set of probability laws $(P_\theta, \theta \in \Theta)$ indexed by scalar or vectorial parameter $\theta \in \mathbb{R}^d$.

- *Observation*: the observation X is a random variable of distribution P_θ, where the parameter θ is unknown.

- The probability of a given event is a function of θ and hence we'll write: $P_\theta(A), E_\theta(X)$, ...

Objectives

When considering parametric models, the objectives are often:

- *The estimation*: which consists to find an approximate value of parameter θ.

- *The testing*: which is to answer the following type of questions... Can we state, given the observation set, that the proportion of defective objects θ is smaller that $).)1$ with a probability higher than 99%?

Example: Gaussian model

- A random variable X is said standard gaussian if it admits a p.d.f.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

which is referred to as $X = \mathcal{N}(0, 1)$.

- X is a gaussian random variable of mean μ and variance σ^2 if

$$X = \mu + \sigma X_0$$

where X_0 is a standard gaussian.

- *Gaussian model*: the observation $(X_1, \cdots, X_n)$ are n gaussian iid random variables of mean μ and variance σ^2 (i.e. $\theta = (\mu, \sigma)$).

Statistic's concept

- To build statistical estimators or tests, one has to evaluate certain function of the observation: $T_n = T(X_1, \cdots, X_n)$. Such a function is called *statistic*.
- It is crucial that the defined statistic is not a function of the parameter θ or the exact p.d.f. of the observations.
- A statistic is a random variable which distribution can be computed from that of the observations.
- Note that a statistic is a random variable but not any random variable is a statistic.

Examples

- Empirical mean: $T_n = \sum_{i=1}^{n} X_i/n$.
- Median value: $T_n = (X)_n$.
- Min + Max: $T_n = 0.5\,(\max(X_1, \cdots, X_n) + \min(X_1, \cdots, X_n))$.
- Variance: $T_n = \sum_{i=1}^{n} X_i^2/n$.

Parametric Estimation

Parametric versus non-parametric

- *Non-parametric*: The p.d.f. f of X is unknown but belongs to a known function space $\mathcal{F}$, e.g.

$$\mathcal{F} = \{f : \mathbb{R} \rightarrow \mathbb{R}^+, \text{twice differentiable and} f" \leq M\}.$$

  leads to difficult estimation problems !!

- *Semi-parametric*: Consider for example a set of observations $\{(X_i, z_i)\}$ following the regression model $X_i = g(\theta, z_i) + \epsilon_i$ where g is a known function and ϵ_i are iid random variables. This model is said semi-parametric if the p.d.f. of ϵ_i is completely unknown.

- *Parametric*: The previous model is parametric if the p.d.f. of ϵ_i is known (up to certain unknown point parameters).

Parametric estimation

- Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be an observation of a statistical model $(P_\theta, \theta \in \Theta)$.

- An estimator is a function of the observation

$$\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$$

  used to infer (approximate) the value of the unknown parameter.

Example: Estimation of the mean value

- Let $(X_1, X_2, \cdots, X_n)$ be a n-sample iid observation given by $X_i = \theta + X_{i0}$, $\theta \in \mathbb{R}$ and X_{i0} are iid zero-mean random variables.
- Mean estimators:

1- Empirical mean: $\hat{\theta}_n = \sum_{i=1}^{n} X_i/n$.

2- Median value: $\hat{\theta}_n = (X)_n$.

3- (Min + Max)/2: $\hat{\theta}_n = \frac{\max(X_1,\cdots,X_n)+\min(X_1,\cdots,X_n)}{2}$.

Estimator

- A statistic is referred to as 'estimator' to indicate that it is used to 'estimate' a given parameter.
- The estimation theory allows us to characterize 'good estimators'.
- For that one needs 'performance measures' of a given estimator.
- Different performance measures exist that sometimes might lead to different conclusions: i.e. an estimator might be 'good' for a first criterion and 'bad' for another.

Bias

- An estimator T of parameter θ is said *unbiased* if θ is the mean-value of the distribution of T (θ being the exact value of the parameter): i.e. $E_\theta(T) = \theta$.

- Otherwise, the estimator T is said 'biased' and the difference $b(T, \theta) = E_\theta(T) - \theta$ represents the estimation bias.

Example: variance estimation

- Let $(X_1, \cdots, X_n)$ be an iid observation of pdf $p_\theta(x) = \frac{1}{\sigma} p(x - \mu)$, $\theta = (\mu, \sigma^2)$, and p satisfies $\int x^2 p(x) dx = 1$ and $\int x p(x) dx = 0$.
- $S_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$ is an unbiased estimator of σ^2.
- $V_n = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2$ is a biased estimator of σ^2 which bias is given by $b = -\sigma^2/n$. It is however said *asymptotically* unbiased as the bias goes to zero when n tends to infinity.

Unbiased estimator

- Instead of θ, one might be interested by a function of this parameter... For example in the previous example, the objective can be to estimate $\sigma = \sqrt{\theta_2}$ instead of $\sigma^2 = \theta_2$. When θ is a parameter vector, one might, in particular, be interested in estimating only a sub-vector of θ.

- T is an unbiased estimator of $g(\theta)$ if $E_\theta(T) = g(\theta)$ for all $\theta \in \Theta$.

- Otherwise, $b(T, \theta, g) = E_\theta(T) - g(\theta)$ would represent the bias of this estimator.

Bias and transforms

- Non-lineat transforms of unbiased estimators are not necessarily unbiased: i.e. if T is an unbiased estimator of θ, $g(T)$ is not in general an unbiased estimate of $g(\theta)$.

- For example, if T is an unbiased estimate og θ then T^2 is not an unbiased estimate of θ^2. Indeed, we have

$$E_\theta(T^2) = var_\theta(T) + (E_\theta(T))^2 = var_\theta(T) + \theta^2.$$

Mean squares error

Another pertinent performance measure is the mean squares error (MSE). The MSE measures the dispersion of the estimator arround the ‘true’ value of the parameter:

$$MSE(T, \theta) = R(T, \theta) = E(T(X) - \theta)^2.$$

The MSE can be decomposed into:

$$MSE(T, \theta) = (b(T, \theta))^2 + var_\theta(T).$$

Example: MSE of the empirical mean

- $(X_1, \cdots, X_n)$ n-sample iid observation of law $\mathcal{N}(\mu, \sigma^2)$.

- Empirical mean: $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$.

- Unbiased estimator and

$$var(\overline{X}) = \frac{\sigma^2}{n}.$$

Estimator's comparison: Risk measure

- We have considered previously the *quadratic risk (loss) function*:

$$l(\theta, \alpha) = (\theta - \alpha)^2.$$

- Other risk (loss) functions are possible and sometimes more suitable:

1- Absoluve-value error: $l(\theta, \alpha) = \theta - \alpha$,

2- Truncated quadratic risk function: $l(\theta, \alpha) = \min((\theta - \alpha)^2, d^2)$.

3- The 0-1 risk function: $l(\theta, \alpha) = 0$ if $\theta - \alpha \leq \epsilon$ and $l(\theta, \alpha) = 1$ otherwise.

- The mean risk value for an estimator is defined as $E_\theta(l(T(X), \theta))$.

Estimator's comparison

One can compare 2 estimators w.r.t. their risk values.

- An estimator T is said 'better' than another estimator T' if

$$R(T, \theta) \leq R(T', \theta), \quad \forall \; \theta \in \Theta$$

with strict inequality for at least one value of the parameter θ.

- Except for 'very particular cases', it does not exist an estimator *uniformly* better than all other estimators.

Reducing the class of estimators

- *Unbiased estimators*: we seek for the unbiased estimator with the minimum quadratic risk value.

- *Invariance*: One might be interested in estimators satisfying certain invariance property. For example, in a translation model, one is interested in the estimators that satisfy:
  $T(X_1 + c, \cdots, X_n + c) = c + T(X_1, \cdots, X_n)$.

- *Linearity*: One seeks here for the best linear estimator. This is the case, for example, in the linear regression problem (e.g. Theorem of Gauss-Markov).

Cramer Rao Bound: regular model

- For ‘regular’ statistical models it is possible to determine a lower bound for the quadratic risk (MSE). It is the Cramer-Rao Bound (CRB).

- A statistical model is regular if:

  1- The model is dominated: i.e. $P_\theta(A) = \int_A p_\theta(x)\mu(dx) \; \forall \; A \in \mathcal{B}(X)$.

  2- Θ is an open set of $\mathbb{R}^d$ and $\partial p(x;\theta)/\partial\theta$ exists for all x and all θ.

  3- The pdfs have the same support for all values of θ, i.e. for $A \in \mathcal{B}(X)$, we have either $P_\theta(A) = 0 \; \forall \; \theta$ or $P_\theta(A) > 0 \; \forall \; \theta$.

  4- $\int_X \frac{\partial}{\partial\theta} p(x;\theta)\mu(dx) = \frac{\partial}{\partial\theta}\int_X p_\theta(x)\mu(dx) = 0$.

Cramer Rao Bound: likelihood & score function

- The function $\theta \rightarrow p(x;\theta)$ is called *likelihood* of the observation.

- For a regular model, the function $\theta \rightarrow S(x;\theta) = \nabla_\theta \log p(x;\theta)$ is called *score* function of the observation.

- When for all θ, $E(S(X;\theta)^2) < \infty$, one define the *Fisher Information Matrix (FIM)* as:

$$I(\theta) = E_\theta[S(X;\theta)S(X;\theta)^T].$$

Fisher information: Properties

- Additivity for iid observations:

$$I_n(\theta) = Cov_\theta(\nabla_\theta \log p(X_1, \cdots, X_n; \theta)) = ni(\theta)$$

  where

$$i(\theta) = Cov_\theta(\nabla_\theta \log p(X_1; \theta))$$

  in other words, each new information contributes in an identical way to the global information.

- When the score function is twice differentiable, we have:

$$I_n(\theta) = -E_\theta(\nabla_\theta^2 \log p(X_1, \cdots, X_n; \theta)).$$

Cramer Rao Bound

- Let $T(X)$ be a statistic such that $E_\theta(T(X)^2) < \infty, \forall\, \theta$ and assume that the considered statistical model is regular.

- Let $\psi(\theta) = E_\theta(T(X))$. Then

$$var_\theta(T(X)) \geq \nabla_\theta \psi(\theta)^T I^{-1}(\theta) \nabla_\theta \psi(\theta).$$

- If T is an unbiased estimator of θ, then the CRB becomes:

$$var_\theta(T(X)) \geq I^{-1}(\theta).$$

Example: Empirical mean for gaussian process

- $(X_1, \cdots, X_n)$ n-sample iid observation of law $\mathcal{N}(\mu, \sigma^2)$ (σ^2 known).
- The Fisher information for the mean parameter is given by:

$$I_n(\theta) = n/\sigma^2.$$

- The empirical mean MSE reaches the CRB and hence it is the best estimator (for the quadratic risk) in the class of unbiased estimates.

Example: Linear model

- Observation model: $X = Z\theta + \epsilon$ where $X = [X_1, \cdots, X_n]^T$ is the observation vector, Z is a full rank known matrix and ϵ is the error vector of zero-mean and covariance $E(\epsilon\epsilon^T) = \sigma^2 I$.
- The least squares estimate of θ given by

$$\hat{\theta} = Z^{\#} X$$

  is unbiased and of MSE

$$Var_\theta(\hat{\theta}) = \sigma^2 (Z^T Z)^{-1}.$$

- If ϵ is a gaussian noise, then the FIM is given by $I(\theta) = (Z^T Z)/\sigma^2$ and hence the LS estimate is the best unbiased estimate w.r.t. the quadratic risk.

Efficiency

- An unbiased estimate of θ which reaches the CRB is said *efficient*. It is an unbiased estimate with minimum error variance.

- Efficient estimators exist for the class of exponential distributions where

$$p(x;\theta) \propto \exp(A(\theta)T(x) - B(\theta)).$$

Asymptotic Theory

Asymptotic approach

- Study of the estimators in the limit of ‘large sample sizes’, i.e. $n \to \infty$.
- For usual models, the estimates converge to the exact value of the parameter: *consistency*.
- We then study the dispersion of the estimators around the limit value θ.
- Our tools are: the law of large numbers and the central limit theorem.

Consistency

- Let $(X_1, \cdots, X_n)$ be an observation of a statistical model $(P_\theta, \theta \in \Theta)$.

- $T_n = T_n(X_1, \cdots, X_n)$ is a sequence of consistent estimators of θ if for all θ the sequence of random variables T_n converges in probability to θ:

$$\lim_{n \to \infty} P_\theta(T_n - \theta \geq \delta) = 0 \quad \forall \theta \in \Theta, \delta > 0.$$

.

Large numbers law

- The consistency is often a consequence of the *large numbers law*.

- *Large numbers law*: Let $(X_1, \cdots, X_n)$ be a sequence of iid random variables such that $E(X_1) < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow_P E(X).$$

Consistency & continuous transform

- Let T_n be a consistent sequence of estimators of θ, $T_n \to_p \theta$.
- Let ϕ be a continuous function in Θ.
- $\phi(T_n)$ is then a sequence of consistent estimators of $\phi(\theta)$.

Convergence rate

- The consistency is an interesting property but does not give us information on how fast the estimator converges to the limit value.

- In the case of the empirical mean one can easily verify that $\sqrt{n}(\overline{X}_n - \mu)$ is bounded in probability which gives us a rough idea on the convergence speed!!

Asymptotically normal estimator

- An estimator sequence T_n of θ is said asymptotically normal if

$$\sqrt{n}(T_n - \theta) \rightarrow_d \mathcal{N}(0, \sigma^2(\theta)).$$

  where $\sigma^2(\theta)$ is the *asymptotic variance* of the considered estimator.

- This asymptotic result allows us to evaluate (often in a simpler way) the dispersion of the estimators aroud the true value of the parameter.

Convergence in distribution

Let $(X_n, n \geq 0)$ be a sequence of random variables. X_n is said to converge in distribution to X (i.e. $X_n \rightarrow_d X$) if one of the following equivalent properties is verified:

- For any bounded continuous function f:
  $\lim_{n\rightarrow\infty} E(f(X_n)) = E(f(X))$.
- For all u, $\lim_{n\rightarrow\infty} E(e^{iuX_n}) = E(e^{iuX})$
- For all subsets $A \in \mathcal{B}(\mathbb{R})$ such that $P(X \in \partial A) = 0$ we have
  $\lim_{n\rightarrow\infty} P(X_n \in A) = P(X \in A)$.

Confidence interval

- Let $(T_n, n \geq 0)$ be a sequence of random variables such that $\sqrt{n}(T_n - \theta) \rightarrow_d T\tilde{\mathcal{N}}(0, \sigma^2)$.

- Let $A = [-a,\ a]$ such that $P(T \in \{a, -a\}) = 0$, then we have

$$\lim_n P_\theta(\sqrt{n}(T_n - \theta) \in [-a,\ a]) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-a}^{a} \exp(-x^2/2\sigma^2)dx = \alpha,\ \forall\, \theta.$$

- Consequently,

$$\lim_n P_\theta(\theta \in [T_n - a/\sqrt{n},\ T_n + a/\sqrt{n}]) = \alpha,\ \ \forall\ \theta$$

which represents a confidence interval of level α for θ.

Central limit theorem

The asymptotic normality of the estimators comes from the *central limit theorem* that can be stated as follows:

Let $(X_1, \cdots, X_n)$ a sequence of iid random variables of mean μ and variance $\sigma^2 = E(X^2) < \infty$. Then,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \rightarrow_d \mathcal{N}(0, \sigma^2).$$

The δ-method

- Let T_n a consistent sequence of estimators of θ.
- The continuity theorem states that $g(T_n)$ is a consistent estimate of $g(\theta)$.
- However, this result does not give any information about the convergence rate nor about the asymptotic normality of the estimator $g(T_n)$??

The δ-method

- Suppose that $\sqrt{n}(T_n - \theta) \rightarrow_d T$ and let g be a locally differentiable function at θ. Then:

$$\sqrt{n}(g(T_n) - g(\theta)) \rightarrow_d g'(\theta)T.$$

- If $T = \mathcal{N}(0, \sigma^2), then \sqrt{n}(g(T_n) - g(\theta))$ is asymptotically normal $\mathcal{N}(0, g'(\theta)^2\sigma^2)$.

Relative asymptotic efficiency

- Let T_n and S_n be two asymptotically normal estimators of θ:

$$\sqrt{n}(T_n - \theta) \to_d \mathcal{N}(0, \sigma_T^2(\theta))$$
$$\sqrt{n}(S_n - \theta) \to_d \mathcal{N}(0, \sigma_S^2(\theta))$$

- T_n is said 'asymptotically better' that S_n if

$$\sigma_T^2(\theta) \le \sigma_S^2(\theta) \;\; \forall\, \theta.$$

Estimation methods

Moments method

- $(X_1, \cdots, X_n)$ n iid random variables $(P_\theta, \theta \in \Theta)$.

- Let $\mu_i(\theta) = E_\theta(g_i(X))$ ($g_i, i = 1, \cdots d$ are given functions).

- Moments method consists in solving in θ the equations

$$\mu_i(\theta) = \hat{\mu}_i, \quad i = 1, \cdots d.$$

where $\hat{\mu}_i$ are empirical (sample averaged) moments.

Moments method

- Several moment choices exist. They should be chosen such that:

  1- One can express explicitely the considered moment function in terms of θ.

  2- Insure a bi-univoque relation between the moments and the desired parameter θ.

- The method is applicable in simple cases only where we have a small number of parameters and there is no ambiguity w.r.t. the choice of the statistics.

Consistency of the moment's estimator

- Using the large numbers law, we have:

$$\frac{1}{n}\sum_{i=1}^{n} g_l(X_i) \rightarrow_d E_\theta(g_l(X)).$$

- If the function μ: $\Theta \rightarrow \mathbb{R}^d$ is invertible with a continuous inverse function, then the continuity theorem states that

$$\hat{\theta} = \mu^{-1}(\hat{\mu})$$

  is a consistent estimate of θ. Similarly, one can establish the asymptotic normality of the moment's estimator using the central limit theorem and the δ-method.

Maximum likelihood method

- Let $X = (X_1, \cdots, X_n)$ a sequence of random variables corresponding to the model $(P_\theta, \theta \in \Theta)$. Let p_θ represents the pdf of X.
- *Likelihood*: $\theta \rightarrow p(x;\theta)$ seen as a function of θ.
- *Maximum likelihood estimation*: estimation of $\hat{\theta}$ such that

$$p(x;\hat{\theta}) \geq \max_\theta p(x;\theta).$$

- If $p(x;\theta)$ is differentiable, then $\hat{\theta}$ is a solution of

$$\Delta_\theta logp(x;\hat{\theta}) = 0.$$

Log-likelihood function

- *Log-likelihood*: $L(x;\theta) = \log p(x;\theta)$.

- In the case of iid observations:

$$\frac{1}{n}\log p(x;\theta)_p - K(\theta_0,\theta)$$

where $K(\theta_0,\theta)$ is the Kullback-Leibler information defined by

$$K(\theta_0,\theta) = -E_{\theta_0}\left[\log\frac{p(X;\theta)}{p(X;\theta_0)}\right].$$

Kullback information

The Kullback-Leibler information is a ‘distance’ measure between two pdf satisfying:

- $K(p_{\theta_0}, p_\theta) \geq 0$
- $K(p_{\theta_0}, p_\theta) = 0$ iff

$$P_{\theta_0}(x : p(x;\theta_0) = p(x;\theta)) = 1.$$

Mean and variance of a gaussian

- Log-likelihood:

$$\log p(x;\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

- Likelihood equations:

$$\frac{\partial p}{\partial \mu}(x;\hat{\mu};\hat{\sigma}^2) = 0, \quad \frac{\partial p}{\partial \sigma^2}(x;\hat{\mu};\hat{\sigma}^2) = 0.$$

- Solutions:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2.$$

Non-unicity of ML estimate: Uniform distribution

- $(X_1, \cdots, X_n)$ iid random variables of uniform distribution in $[\theta - 0.5 \;\; \theta + 0.5]$.

- Likelihood

$$p(x;\theta) = \begin{cases} 1 \quad if \;\; \theta \in [\max(X_i) - 0.5, \min(X_i) + 0.5] \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

- The likelihood is constant in the interval

$$[\max(X_i) - 0.5, \min(X_i) + 0.5].$$

Other methods

- Minimum contrast method,
- M-estimation,
- Z-estimation
- Robust estimation
- $\vdots$