University of BRISTOL

Computer Science Year 2

Algorithms & Data

Estimation, Regression, Classification

Prof Alin Achim

University of BRISTOL

Last time …

- Least squares (LS) estimation
  - Minimizes sum of squares between measurements and a model
  - Generally applicable estimator as no assumption is made about the data
  - Best for linear models
- Method of Moments (MoM)
  - Based on equating sample and population moments
  - Simplest estimation approach, intuitive, works well in straightforward cases
  - Not always leading to good results, especially in small sample sizes

Objectives

- Bayesian Estimation
  - Motivation
  - The Bayesian paradigm
  - The MMSE estimator
  - The MAE estimator
  - The MAP estimator
  - Examples

Classical vs Bayesian estimation

- Classical methods
  - The assumptions leading to asymptotic results may not apply sometimes;
  - Asymptotic approximations are not always reliable, even for medium sample sizes. For small sample sizes, estimators like the MLE (asymptotically justified) can even lead to absurd results;
  - Frequentist estimators work well *on average*, but not necessarily for the data at hand;
  - They are not able to account for any kind of extra-information that may be available;
  - Classical approach to estimation assumes that the parameter to be determined is a deterministic but unknown constant.

Classical vs Bayesian estimation

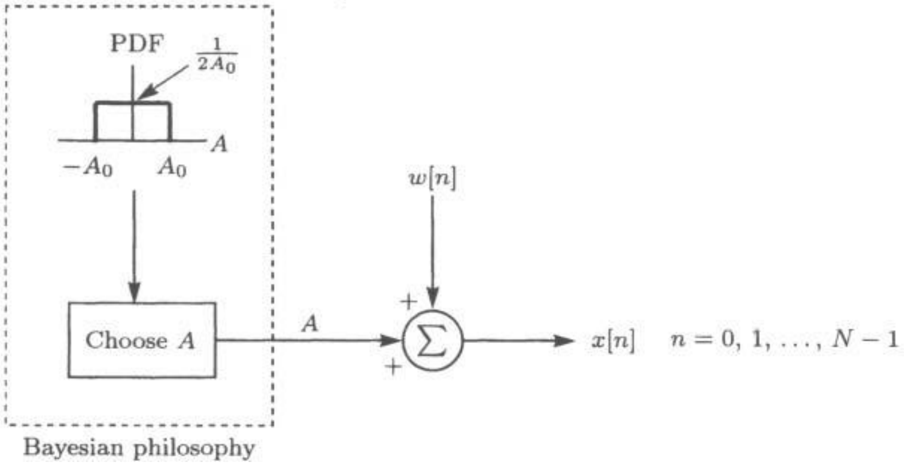
- Bayesian methods
  - In Bayesian approach the unknown parameter is assumed to be a random variable;
  - They enable prior information about the parameters to be incorporated in the estimation procedure;
  - They do not need to be justified by any asymptotic approximation;
  - Bayesian techniques are based on modelling the uncertainty with respect to the parameter θ through a probability distribution.

The Bayesian MSE

- Remember the DC level in WGN example:

$$x[n] = A + w[n],\ \text{where}\ n = 0,1,\ldots,N-1\ and\ w[n] \sim N(0,\sigma^2)$$

- The MVUE of A was found to be the sample mean, assuming -∞<A<∞ (deterministic unknown) …
- However, by assigning a particular PDF to the *random variable* (!) A:



- We can attempt to find an estimator of A that would minimize the MSE:

$$B_{MSE}(\hat{A}) = E\left[\left(A - \hat{A}\right)^2\right]$$

The Bayesian MSE

- Classical MSE:

$$mse(\hat{A}) = \int (\hat{A} - A)^2 p(x; A)\, dx$$

- Bayesian MSE:

$$Bmse(\hat{A}) = \iint (A - \hat{A})^2 p(x, A) dx dA$$

- Whereas the classical MSE depends on A (and hence estimators that attempt to minimize it will usually depend on A), the Bayesian MSE does not! That's because the parameter dependence is integrated away!

Elements of Bayesian analysis

- ***Bayes rule***: $$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

  where: $p(\theta|x)\ -\ posterior$ $p(x|\theta)\ -\ likelihood$ $p(\theta)\ -\ prior$

  $p(x)\ -\ evidence$

- **Definition: *Bayesian statistical model***

  A statistical model composed of a data generation model, p(x|θ), and a prior distribution on the parameters, p(θ).

- Joint distribution: $p(x,\theta) = p(x|\theta)p(\theta)$

- Marginal distributions: $$p(x) = \int p(x|\theta)\,p(\theta)d\theta$$ $$p(\theta) = \int p(x|\theta)\,p(\theta)dx$$

Example: DC level in WGN (continued)

$$Bmse(\hat{A}) = \iint (A - \hat{A})^2 \, p(x, A) dx dA$$

$$p(x, A) = p(A|x)p(x)$$

$$Bmse(\hat{A}) = \int \left[\int (A - \hat{A})^2 p(A|x) dA \right] p(x) dx$$

- The Bayesian MSE will be minimized if the integral in brackets can be minimized for each x.
- Taking the derivative:

$$\frac{\partial}{\partial \hat{A}} \int (A - \hat{A})^2 p(A|x) dA = \int \frac{\partial}{\partial \hat{A}} (A - \hat{A})^2 p(A|x) dA$$

$$= \int -2(A - \hat{A}) p(A|x) dA$$

$$= -2 \int A p(A|x) \, dA + 2\hat{A} \int p(A|x) dA$$

Example: DC level in WGN (continued)

- Setting to zero

$$-2\int Ap(A|x)\,dA + 2\hat{A}\int p(A|x)dA = 0$$

- And since the conditional PDF must integrate to 1

$$\hat{A} = \int Ap(A|x)dA$$

- Finally

$$\hat{A} = E(A|x)$$

University of BRISTOL

Bayesian estimators

- In general, a Bayesian estimator minimizes the conditional risk, which is the loss (cost function) averaged over the conditional (posterior) distribution of θ, given the observation (measurement) x:

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \int C\left[\theta, \hat{\theta}(x)\right] p(\theta|x) d\theta$$

- Definition: *The Bayes risk R* is the average cost E[C(ε)] and measures the performance of a given estimator.

$$R = E[C(\varepsilon)]$$

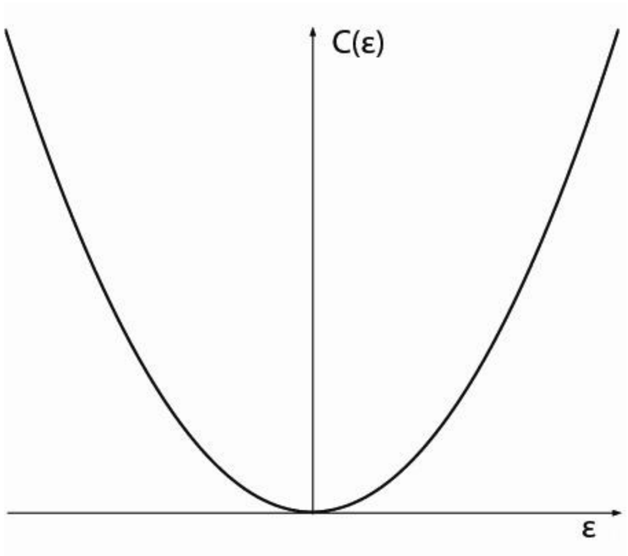
The Minimum Mean Square Error (MMSE) estimator

- Quadratic error cost function

$$C[\theta,\hat{\theta}(x)] = C(\varepsilon) = \varepsilon^2$$

- The corresponding optimal estimator is the *mean of the posterior PDF*

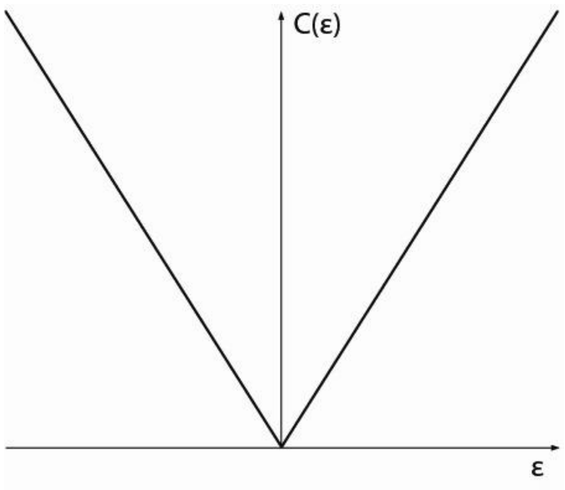
$$\hat{\theta} = \int \theta p(\theta|x)d\theta = E(\theta|x)$$



The Minimum Absolute Error (MAE) estimator

- Absolute error cost function

$$C[\theta, \hat{\theta}(x)] = C(\varepsilon) = |\varepsilon|$$



- General Bayesian estimator

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \int C[\theta, \hat{\theta}(x)] p(\theta|x) d\theta$$

➢ Using the two equations above, the MAE is obtained as

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \int |\theta - \hat{\theta}| p(\theta|x) d\theta$$

The MAE estimator

- The integral can be split into

$$g(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)\, p(\theta|x) d\theta + \int_{\hat{\theta}}^{-\infty} (\theta - \hat{\theta})\, p(\theta|x) d\theta$$

- In order to differentiate one can use Leibnitz's rule yielding

$$\frac{dg(\hat{\theta})}{d\hat{\theta}} = \int_{-\infty}^{\hat{\theta}} p(\theta|x) d\theta - \int_{\hat{\theta}}^{-\infty} p(\theta|x) d\theta$$

- And setting to 0 :-

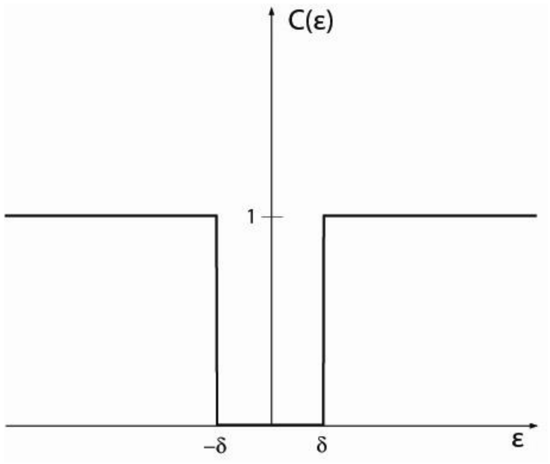
$$\int_{-\infty}^{\hat{\theta}} p(\theta|x) d\theta = \int_{\hat{\theta}}^{-\infty} p(\theta|x) d\theta$$

that is by definition the *median of the posterior PDF.*

The Maximum a Posteriori (MAP) Estimator

- Hit-or-miss cost function

$$C(\varepsilon) = \begin{cases} 0, & |\theta - \hat{\theta}| < \delta \\ 1, & \text{otherwise} \end{cases}$$



- General Bayesian estimator

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \int C[\theta, \hat{\theta}(x)] p(\theta|x) d\theta$$

➤ Using the two equations above, the MAP is obtained as

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \int_{|\theta - \hat{\theta}| \geq \delta} p(\theta|x) d\theta$$

The MAP Estimator

- Or

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmin}} \left[1 - \int_{|\theta - \hat{\theta}| < \delta} p(\theta|x) d\theta \right]$$

- In order to minimize the expected cost, when δ →0 one should select (the MAP equation)

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}} p(\theta|x)$$

that is, the mode of the posterior pdf.

- Using Bayes theorem together with the last equation, we can also write the MAP equation as (more useful in practice)

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}}\, p(x|\theta) p(\theta)$$

Example

- Assume that $$p(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

  and that the prior pdf is $$p(\theta) = \frac{\gamma}{\pi(\theta^2+\gamma^2)}$$

- The MAP estimator can be found as follows:-

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}}[\ln p\,(x|\theta) + \ln p\,(\theta)]$$

$$\Downarrow$$

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}}\left[-\frac{(x-\theta)^2}{2\sigma^2} + \ln\frac{\gamma}{\pi(\theta^2+\gamma^2)}\right]$$

Example (continued)

$$\hat{\theta}(x) = \underset{\theta}{\operatorname{argmax}} \left[-\frac{(x-\theta)^2}{2\sigma^2} + \ln \frac{\gamma}{\pi(\theta^2+\gamma^2)} \right]$$

- Differentiating with respect to θ

$$\frac{d}{d\theta} \left[-\frac{(x-\theta)^2}{2\sigma^2} + \ln \frac{\gamma}{\pi(\theta^2+\gamma^2)} \right] = \frac{x-\theta}{\sigma^2} - \frac{2\theta}{\theta^2+\gamma^2}$$

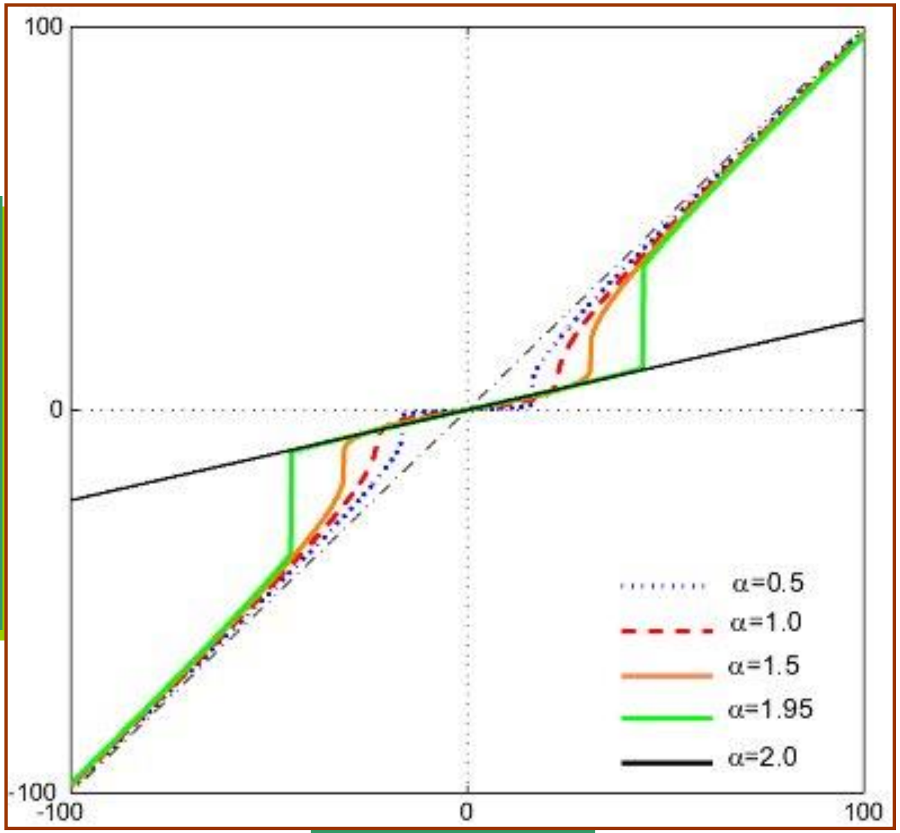
- Setting equal to 0 yields

$$\frac{x-\theta}{\sigma^2} = \frac{2\theta}{\theta^2+\gamma^2}$$

- Finally, rearranging

$$\theta^3 - x\theta^2 + (\gamma^2+\sigma^2)\theta - \gamma^2 x = 0$$

Example: MAP "Processor" I/O Curves



Processor Output

Input

Summary of Bayesian Estimation

- The Bayesian approach to estimation is fundamentally different from the classical (frequentist) approach;
- It consists of modelling the uncertainty with respect to the parameter θ through a probability distribution;
- It is able to provide answers to any statistical question in terms of probabilities.
  - *Disadvantages*:-
    - A prior distribution must be specified. This presupposes more work and can be subjective
    - Except for some special cases of prior distributions (e.g. Gaussian, Cauchy, exponential, Laplacian), the derivation of the posterior distribution is cumbersome and requires numerical methods.