ELEC2146

Electrical Engineering Modelling and Simulation

Parameter Estimation

Dr Ray Eaton

S2, 2016

Lecture notes written by Dr Julien Epps

Overview

- Estimators
  - What are they ?
  - Properties
  - Examples
- Method of moments
- Maximum likelihood
- Other estimators
  - MMSE
  - Jackknife
  - Bayesian

Motivation

- Have already seen problems where
  - Model structure is known, or assumed
  - Have (experimental) data of some kind
    - Usually inputs/outputs
  - Parameter values are unknown
- Stochastic models
  - Correct framework for parameter estimation
    - Data almost always noisy (contains random or stochastic component)

Parameter Estimation

- Objective
  - Determine a statistic

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

  - $X_1, X_2, \ldots, X_n$ are random variables representing samples from an overall population *X*
  - *h* is estimation *function* (no numerical value)
  - $\hat{\Theta}$ is an *estimator*
  - The observed *estimate* of parameter θ is

$$\hat{\theta} = h(x_1, x_2, \ldots, x_n)$$

    - This is a numerical value
    - $x_1, x_2, \ldots, x_n$ are the observed samples

Estimator Properties

- Bias
  - An estimator $\hat{\Theta}$ is unbiased for θ if

$$E(\hat{\Theta}) = \theta$$

  - otherwise biased, with bias $b(\theta) = E(\hat{\Theta}) - \theta$
  - i.e. if on average, $\hat{\Theta}$ is close to the true parameter value θ
  - i.e. sampling distribution of the estimator is centred over the parameter being estimated

Estimator Properties

- Minimum variance
  - Would like the sampling distribution of an estimator to have minimum variance
    - Estimates fall close to θ
  - An unbiased minimum-variance estimator $\hat{\Theta}$ has the property
$$\mathrm{var}(\hat{\Theta}) < \mathrm{var}(\tilde{\Theta})$$
  - for all other estimators $\tilde{\Theta}$ of θ for the same sample
  - In practise:
    - Want variance as small as possible
    - Comparing biased and unbiased estimators: use MSE between $\hat{\theta}$ and θ

Estimator Properties

- Consistency
  - An estimator $\hat{\Theta}$ is a consistent estimator for θ if

$$\lim_{n\to\infty} P\left(\left|\hat{\Theta}-\theta\right| \geq \varepsilon\right) = 0 \quad \forall \varepsilon > 0$$

  - i.e. estimator converges to the true parameter value θ when more observed data are used during estimation
  - i.e. sampling distribution of the estimator is centred over the parameter being estimated
  - An unbiased estimator is consistent if

$$\lim_{n\to\infty} \mathrm{var}(\hat{\Theta}) = 0$$

Example Estimators

- Estimator: $\bar{X}$ Estimate: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample mean
  - Uses entire observed sample
- Estimator: $\tilde{X}$ Estimate: $\tilde{x} = \frac{1}{|M|}\sum_{M} x_i$, $M \subset \{1,2,\ldots,n\}$
  - Uses some part of the sample
- Estimator: $\frac{1}{2}\left(\min(X_i) + \max(X_i)\right)$ Estimate: $\frac{1}{2}\left(\min(x_i) + \max(x_i)\right)$
- Estimator: $\bar{X}_{tr(\alpha)}$ Estimate:
  - mean of observed sample excluding smallest and largest α%, i.e. excluding extreme values

Maximum Likelihood

- What is likelihood ?
  - If $x_1, x_2, \ldots, x_n$ are *n* independent sample values, the likelihood is defined as:

$$L(\theta \mid x_1, x_2, \ldots, x_n) = f(x_1 \mid \theta) f(x_2 \mid \theta) \ldots f(x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

  - Same for discrete random variable with pmf $P_X(x)$
  - Since by definition

$$f(x_i \mid \theta) \leq 1$$

    L is small number; *very* small if *n* is large
  - Often we use the log likelihood instead

$$\log L(\theta \mid x_1, x_2, \ldots, x_n) = \log f(x_1 \mid \theta) + \ldots + \log f(x_n \mid \theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

  - *L* has no meaning, only use it for comparison

Maximum Likelihood

- Chooses estimate $\hat{\theta}$ of θ that maximises *L*
  - i.e.

$$\theta = \arg\max_{\theta} \left\{ L(\theta \mid x_1, x_2, \ldots, x_n) \right\}$$

  - How ?
  - Differentiate *L* wrt θ, set to zero

$$\frac{dL(\hat{\theta} \mid x_1, x_2, \ldots, x_n)}{d\hat{\theta}} = 0$$

  - Alternatively, differentiate $\ln\left(L(\hat{\theta} \mid x_1, x_2, \ldots, x_n)\right)$ if this is easier
    - Maximum occurs at same value $\hat{\theta}$

Maximum Likelihood

- Interpretation
  - What is really going on here ?
  - We can use any estimate we like
    - e.g. $\hat{\theta}_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\theta}_2 = \tilde{x} = \frac{1}{|M|}\sum_{M} x_i$

      $\hat{\theta}_3 = \frac{1}{2}\left(\min(X_i) + \max(X_i)\right) \qquad \hat{\theta}_4 = \bar{X}_{tr(\alpha)}$
    - etc . . .
  - Here we find the estimate $\hat{\theta}$ that maximises the likelihood of the observed data $x_1, x_2, \ldots, x_n$, given the model (whose parameter is θ)

Maximum Likelihood

- More than one parameter:

$$\frac{\partial L(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m | x_1, x_2, \ldots, x_n)}{\partial \hat{\theta}_j} = 0, \quad j = 1, 2, \ldots, m$$

- Some properties:
  - Large sample behaviour ($n \rightarrow \infty$):
    - ML estimator is approximately unbiased
    - ML estimator is approximately the minimum variance estimator

Minimum Mean Square Error

- Covered in sufficient detail in LS topic

Jackknife Estimators

- For observed sample values $x_1, x_2, \ldots, x_n$
  - Compute the *i*th estimator $\hat{\theta}_i$ as a function of all samples *except* x_i
  - Repeat for i = 1, 2, . . . , n, to produce n estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$
  - The jackknife estimate $\hat{\theta}$ is a linear combination of the n estimates

Bayesian Estimators

- Maximum likelihood does not take into account *prior* information
  - i.e. the distribution of the parameter
- Demonstrate Bayesian estimation by example:
  - Suppose we have a coin, and want to predict the probability of heads, based on our observation of coin tosses
  - The true probability of heads is a random variable *P*, which could be anywhere from 0 to 1

$$f_P(p) = \begin{cases} 1 & p \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Bayesian Estimators

- $f_P(p)$ is the marginal density of *P*
  - Possible outcomes due to variation in *P* alone
  - "prior" distribution (=what we assume before we observe *X*)
- Now toss the coin, create a sample *X*
- Conditional density of *X* given *P* = *p* is

$$f_{X|P}(x \mid p) = p^x (1-p)^{1-x}$$

  - *X* = 1 denotes a head
- Probability theory → expression for joint density of *X*, *P*:

$$f_{X,P}(x, p) = f_{X|P}(x \mid p) f_P(p) = p^x (1-p)^{1-x}$$

  - Possible outcomes due to variation in *X*, *P*

Bayesian Estimators

– Also need the marginal density of *X*:

$$f_X(x) = \int_p f_{X,P}(x,p)dp = \int_p p^x (1-p)^{1-x} dp$$

$$= x\int_0^1 pdp + (1-x)\int_0^1 (1-p)dp = x\frac{1}{2} + (1-x)\frac{1}{2}$$

$$= \frac{1}{2}$$

– Where does all this get us ?

– Want to use prior information to come up with an estimator

– i.e. want conditional density of *P* given *X*

Bayesian Estimators

– Want conditional density of *P* given *X*

$$f_{P|X}(p \mid x) = \frac{f_{X|P}(x \mid p) f_P(p)}{f_X(x)} = \frac{f_{X,P}(x,p)}{f_X(x)}$$

– In our example:

$$f_{P|X}(p \mid x = 1) = \frac{p.1}{\frac{1}{2}} = 2p$$

$$f_{P|X}(p \mid x = 0) = \frac{(1-p).1}{\frac{1}{2}} = 2(1-p)$$

– The conditional density allows us, **having observed *X* = *x***, to determine the probability of a head *p*

- If *X* = 1, $f_{P|X}(p \mid x)$ is 'tilted' towards ↑ *p*
- If *X* = 0, $f_{P|X}(p \mid x)$ is 'tilted' towards ↓ *p*

Bayesian Estimators

- If we toss a coin and observe a head, think a head is more likely
- If we toss a coin and observe a tail, think a tail is more likely
- $f_{P|X}(p \mid x)$ reflects what we have seen, what we know, about the coin
- What is the estimate for *p* ?
- E.g. conditional mean:

$$\hat{p} = \int_0^1 p f_{P|X}(p \mid x = 1)\,dp = \tfrac{2}{3}$$

$$\hat{p} = \int_0^1 p f_{P|X}(p \mid x = 0)\,dp = \tfrac{1}{3}$$