Computer Science Year 2

Algorithms & Data

Estimation, Regression, Classification

Prof Alin Achim

Last time …

- If a MVU estimator does not exist, or can not be found, the parameters can be obtained from the likelihood function.

- A Maximum Likelihood (ML) parameter estimate is found by maximising the likelihood function $p(x;\theta)$ which is essentially the probability of the data given the parameters;

- ML estimators are asymptotically efficient, as the number of observations increase and the covariance of the estimates tends to CRLB

- A major advantage of the MLE is that we can find an estimate from the given data numerically since it requires only the maximum of a known function. The Newton-Raphson iterative techniques or the Expectation-Maximisation (EM) algorithm can be used for iterative estimation of the parameters.

Objectives

- Least Square Estimation (LSE)
- Method of Moments (MoM)

Least Squares - The intuition

- Consider (again) the multiple observations:

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

- Or more generally:

$x[n] = f(\theta_1, \theta_2, \ldots, \theta_M) + w[n]$, i.e. f is a function of M parameters

- The least square estimator (LSE) minimizes:

$$J = \sum_{n=0}^{N-1} \left(x[n] - f(\theta_1, \theta_2, \ldots, \theta_M)\right)^2$$

Least Squares - Properties

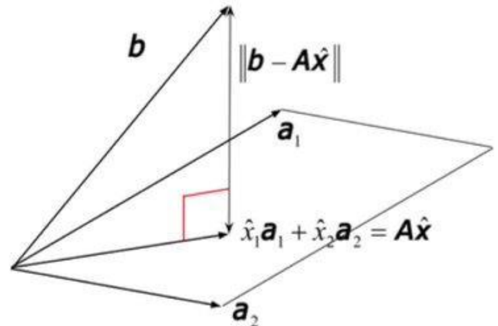
- LSE is widely used when estimating parameters for linear models
- No assumptions about the data are made
- If $w[n] \sim N(0, \sigma^2)$, LSE coincides with the MLE!
- Geometric interpretation: the LS estimate is an orthogonal projection of the data vector onto the space defined by the independent variable.

- Inverse problems formulation:
  - $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}, \boldsymbol{n} \sim N(0, \sigma^2)$
  - $\widehat{\boldsymbol{x}} = min_{\boldsymbol{x}} ||\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}||^2$

Geometric interpretation

• $\boldsymbol{A}\hat{\boldsymbol{x}}$ is the orthogonal projection of $\boldsymbol{b}$ onto *range(**A**)*

$$\Leftrightarrow \boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{x}}) = \boldsymbol{0} \Leftrightarrow \boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b}$$



Least Squares – A much simpler example

- Suppose we have a random sample $\{x_n\}, n = 0, \ldots, N-1$, drawn from a population with mean μ_x and standard deviation σ_x.
  - We can express x_n using a linear model:

$$x_n = \varepsilon_n + \mu_x, E[\varepsilon_n] = 0 \; and \; E[\varepsilon_n^2] = \sigma_x^2$$

  - Estimate μ_x via LSE, that is minimise:

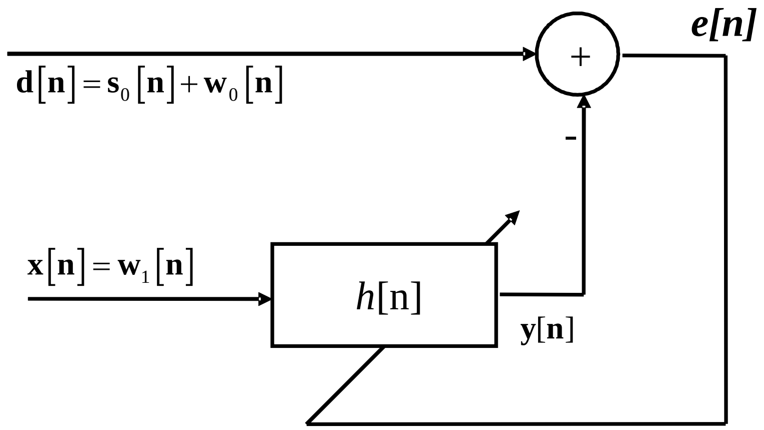
$$J(\mu_x) = \sum_{n=0}^{N-1} (x_n - \hat{\mu}_x)^2$$

$$\frac{\partial V(\mu_x)}{\partial \mu_x} = -2 \sum_{n=0}^{N-1} (x_n - \hat{\mu}_x)^2 = 0$$

$$\Rightarrow \sum_{n=0}^{N-1} x_n = N\hat{\mu}_x \Rightarrow \hat{\mu}_x = \frac{1}{N} \sum_{n=0}^{N-1} x_n$$

Least Squares – A signal processing example

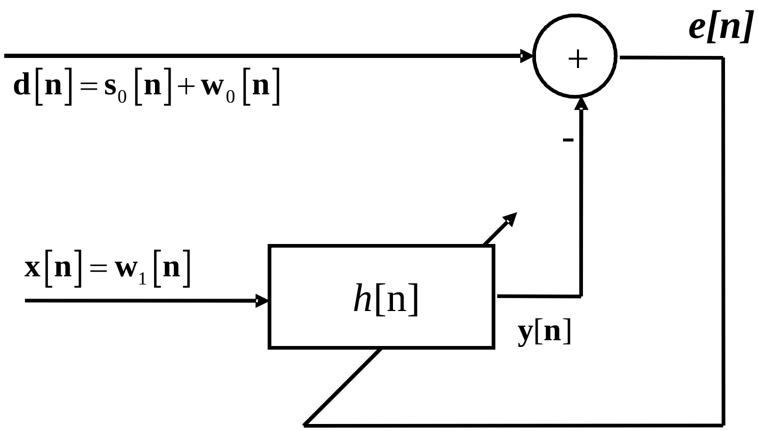
- One of the main applications of LSE is in Adaptive Noise Cancellation (ANC).



- A signal s_0 is received with some uncorrelated noise w_0. A second signal contains noise w_1 uncorrelated with s_0 but correlated with w_0 in some unknown way. The problem is to enhance the signal s_0 in the primary input.

- The principle of ANC is to estimate the noise in the primary input by filtering the reference noise signal with a linear filter. *How is the filter designed??*

Least Squares – A signal processing example

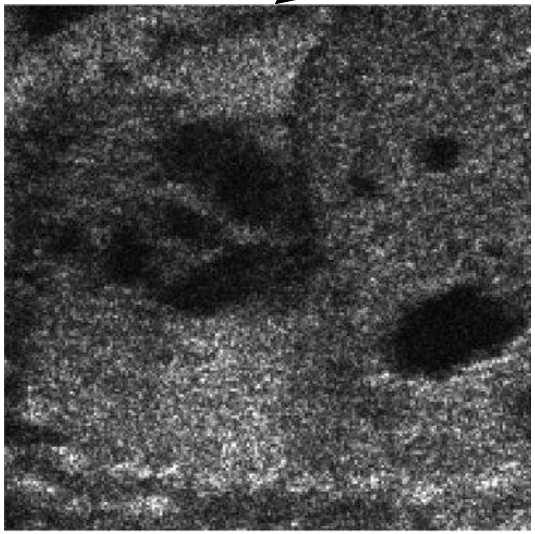


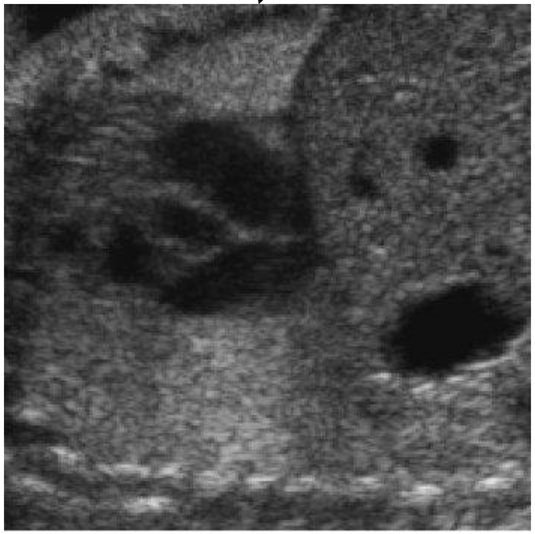
- Choose the weights at time *n* to minimize:

$$J[n] = \sum_{k=0}^{n} e^2[k] = \sum_{k=0}^{n} (d[k] - y[k])^2 = \sum_{k=0}^{n} \left(d[k] - \sum_{l=0}^{p-1} h[l] x[k-l] \right)^2$$

Real world application – image denoising

$$I(x, y) = S(x, y) \cdot \eta(x, y)$$





Speckle Noise (multiplicative): unit-mean, log-normal distributed.

Need to balance between speckle suppression and signal detail preservation!!!

The Symmetric Alpha-Stable (SαS) Model

SαS Characteristic Function:

$$\varphi(\omega) = e^{-\gamma|\omega|^{\alpha}}$$

α: characteristic exponent, 0<α<=2 *(determines thickness of the distribution tails, α=2: Gaussian, α=1: Cauchy)*

γ: dispersion parameter

for Gaussian → variance = 2 x γ

for Cauchy → γ behaves like variance

Parameter Estimation

- After applying the DWT: $\boxed{d = s + \xi}$

$$\boxed{\phi_d(\omega) = \exp(-\gamma_s|\omega|^{\alpha_s}) \cdot \exp(-\frac{\sigma^2}{2}|\omega|^2)}$$

- Noise Estimation: $\boxed{\hat{\sigma} = \frac{1}{0.6745} MAD(\{d\})}$

- Signal parameter estimation method: After estimating the level of noise σ we find the parameters α_s and γ_s by regressing

$$\boxed{y = \log[-(\log|\Phi_d(\omega)|^2 + \sigma^2\omega^2)]}$$

on $w = \log|\omega|$ in the model: $y_k = \mu + \alpha \cdot w_k + \varepsilon_k$

where: μ=log(2γ), ε_k – error term, and $(\boldsymbol{\omega_\kappa}, \boldsymbol{\kappa} = 1,\ldots,\mathrm{K}) \in \boldsymbol{R}$

Method of Moments (MoM) – The intuition

- The method of moments seeks to equate the moments as implied by the underlying model of the population distribution $p(x)$ (mean, variance, skewness, kurtosis, etc) with the actual moments in the sample.

$$population\ moment = sample\ moment$$

- The population probability density function (pdf) is a function of the parameter we want to estimate, θ.

$$p(x) = f(x|\theta), \mathrm{e.g.}\, f(x|\theta) \propto e^{-\frac{x^2}{2\theta^2}}$$

Method of Moments (MoM) – The intuition

$$population\ moment = sample\ moment$$

and

$$p(x) = f(x|\theta)$$

1. $E[x^n] = \int_{-\infty}^{\infty} x^n\, \mathrm{p}(x) = \int_{-\infty}^{\infty} x^n\, f(x|\theta) = g(\theta)$

2. $\theta = g^{-1}(E[x^n])$

3. $\theta = g^{-1}\left(n^{th}\ sample\ moment\right)$

MoM – Properties

- Main advantage: extremely easy to determine and implement
- The order of the moment(s) used depends on the parameter to be estimated, i.e. population mean/maximum/correlation is estimated using sample mean/maximum/correlation, etc
- To obtain estimates of several population parameters, several moments need to be employed
- As an example, consider again the case of the 2 parameter Gaussian distribution, in which case, we use the first and second order moments, i.e. the sample mean and sample variance:

$$E[x] = g_1(x|\theta_1) = \frac{1}{N}\sum_{n=1}^{N} x_n \; and \; E[x^2] = g_2(x|\theta_2) = \frac{1}{N}\sum_{n=1}^{N} x_n^2$$

MoM – Another simple example

- Suppose the population distribution follows an uniform distribution with unknown parameter θ:

$$p(X) = \begin{cases} \frac{1}{\theta}, for\ 0 \le X \le \theta \\ 0, \qquad otherwise \end{cases}$$

- Given a random sample $\{x_n\}, n = 1, \dots, N$, estimate θ.

1. $E[X] = \int_0^\theta x \frac{1}{\theta} dx = \frac{\theta}{2}$
2. $\theta = 2E[X]$
3. $\hat{\theta} = 2\bar{X}$

$$E[X^n] = g(\theta)$$

$$\theta = g^{-1}(E[X^n])$$

$$\hat{\theta} = g^{-1}\left(n^{th}\ sample\ moment\right)$$

Real world application – image fusion

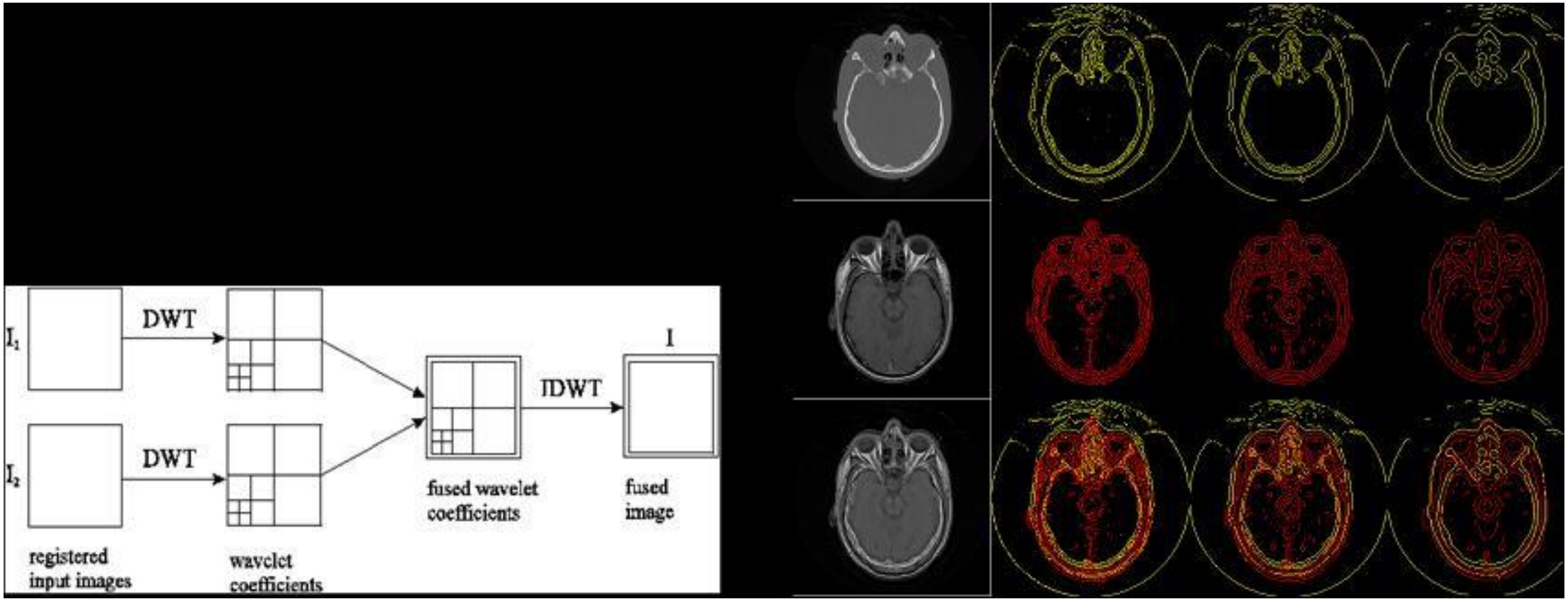
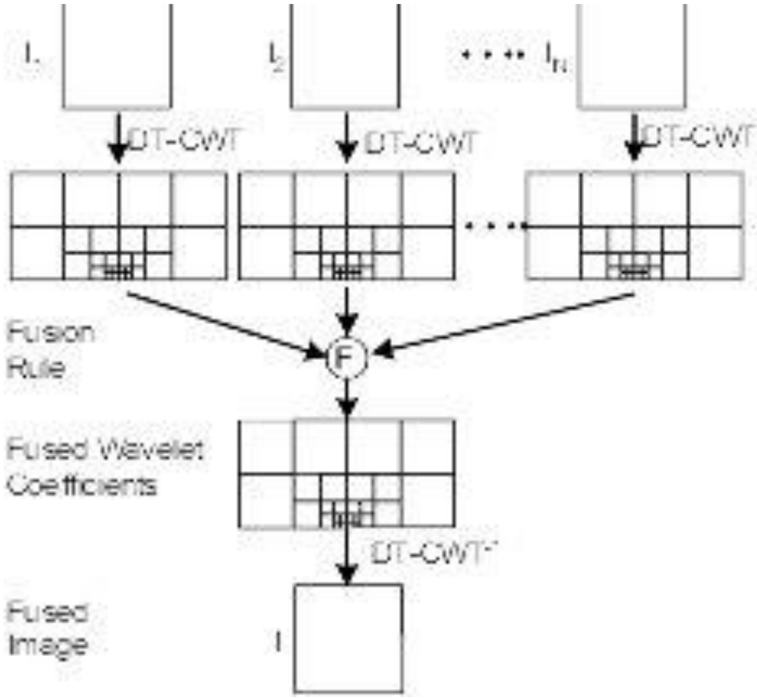


Image fusion algorithms

- Pixel – level fusion
- Region – level fusion
- Compressive fusion



Pixel level approaches:

- Maximum selection scheme
- Average scheme
- Weighted average scheme

The FLOM weighted average (FLOM-WA) scheme

I. DTCWT of both images

II. For each subband pair

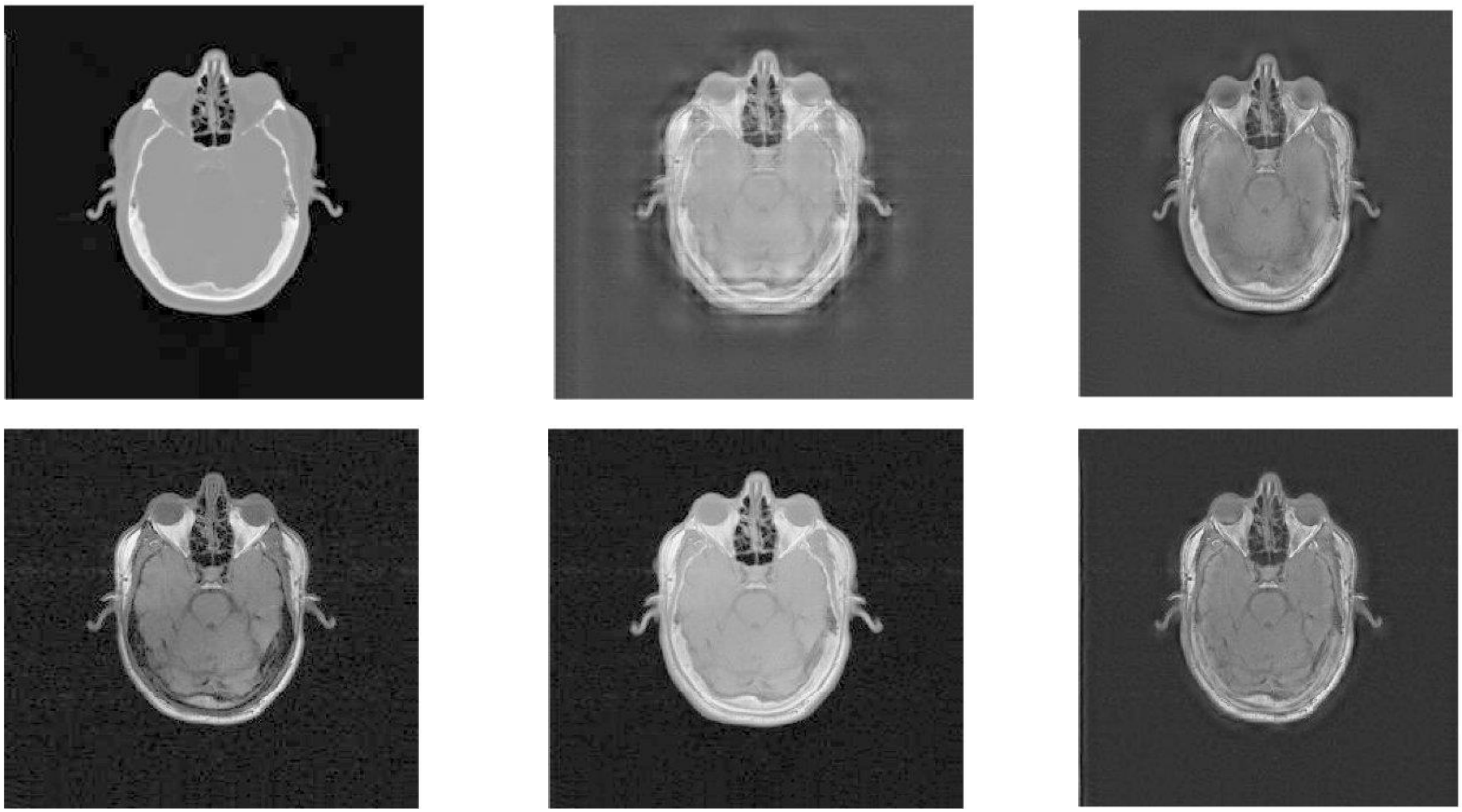
1. Estimate γ_{x_1} and γ_{x_2}
2. Compute $Corr_\alpha(X_1, X_2)$
3. Calculate the fused coefficient using

$$Y = w_1 X_1 + w_2 X_2$$

III. Average coefficients in lowpass residual

IV. IDTCWT

Example WA-FLOM Fusion of MR & CT images



Summary of LSE & MoM

- LSE:
  - minimises the sum of squares between the measurements and the model
  - no assumption about the data — generally applicable estimators
  - best in the context of linear models
- MoM:
  - equate the sample with population moments
  - the simplest, intuitive, works well in straightforward cases
  - estimators not always with good properties (e.g. for small sample size)