Aalto University

T-61.3040 Statistical Signal Modeling

Autumn 2011

Amaury Lendasse & Yoan Miche

Today's Lecture (15.9)

- Probability Theory
- Estimation Theory

But first, and quickly: the z-transform

- Converts discrete time-domain to frequency-domain
- Generalization of the Fourier transform
- Consider discrete set of numbers $x(n)$
- Fourier:

$$X(\omega) = \sum_{n=-\infty}^{+\infty} x(n) e^{-i\omega n}$$

- z:

$$X(z) = \sum_{n=-\infty}^{+\infty} x(n) z^{-n}, z \in \mathbb{C}$$

- So, the Fourier transform is the evaluation of the z-transform around the unit circle in $\mathbb{C}$

Probability Theory

- Random process: sequence of random variables $x(0), x(1), x(2) \dots$
- Denote by Ω the sample space (all possible outcomes)
- A random variable is a (measurable) function $x : \Omega \to \mathbb{R}$ typically
- It can be continuous or discrete

Probability Theory: CDF

- When $x : \Omega \rightarrow \mathbb{R}$, there exists the *cumulative distribution function (cdf)* F such that

$$F_x(a) = P(x \leq a)$$

- The cdf F_x has some properties:
  - F_x is monotonically increasing
  - $\lim_{a \rightarrow -\infty} F_x(a) = 0$
  - $\lim_{a \rightarrow \infty} F_x(a) = 1$
- Intuitively: "Area of the pdf up to a"

Probability Theory: PDF

- And obviously $P(a < x \leq b) = F_x(b) - F_x(a)$
- If F_x is absolutely continuous (der. exists and int. of the der. gives F_x), then x has a *probability density function (pdf)* f_x defined as

$$f_x(a) = \frac{dF_x(a)}{da}$$

Probability Theory: CDF of Normal distribution

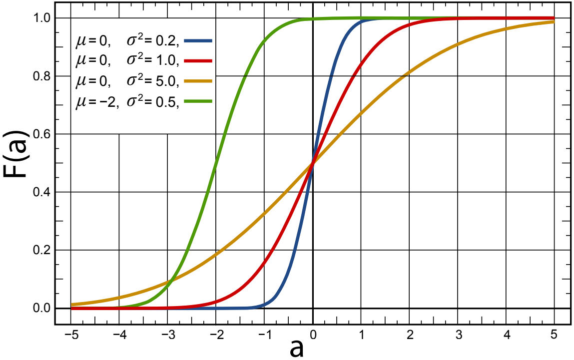

Figure: CDF of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ (from Wikipedia)

Probability Theory: PDF of Normal distribution

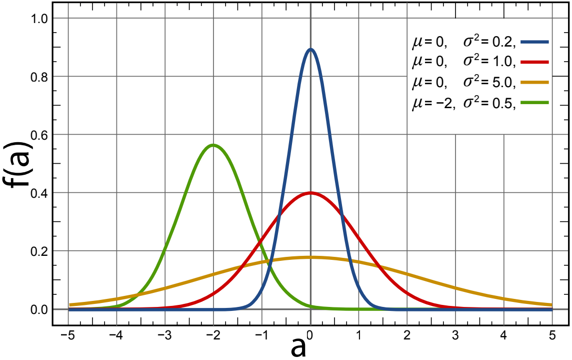

Figure: PDF of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ (from Wikipedia)

Probability Theory

- A distribution can be described by its parameters (for the normal distribution, μ and σ^2, e.g.)
- For some, the parameters can be calculated using the expectation $E(x)$
- Assuming the existence of f_x, then the expected value of x (or expectation), $E(x)$ is defined as

$$E(x) = \int_{-\infty}^{+\infty} a f_x(a) da$$

Probability Theory

- Examples of quantities using the expectation
  - $\mathrm{Var}(x) = E((x - E(x))^2)$, the variance
  - $r_{xy} = E(xy^*)$, the correlation
  - $J = E((x - \hat{x})^2)$, the mean squared error (MSE) (for estimation purposes)

Joint distributions

- Distribution of random process not only dependent on distributions of variables $x(0), x(1), \ldots$
- Usually $x(n)$ and $x(n-k)$ depend on each other (does not always appear in distributions $x(n)$ and $x(n-k)$)
- Random variables x_1 and x_2 have *joint distribution* and *density functions*:

$$F(a, b) = \mathrm{P}\left(x_1 \leq a, x_2 \leq b\right), \qquad f(a, b) = \frac{\partial^2}{\partial a \partial b} F(a, b)$$

- Joint distribution function for more variables defined similarly

Quantities based on Expectation

- For random variables x and y, *correlation* r_{xy} is

$$r_{xy} = E(xy^*)$$

and *covariance* c_{xy} is

$$\begin{aligned} c_{xy} &= \mathrm{Cov}(x, y) \\ &= E\left([x - E(x)]\,[y - E(y)]^*\right) \\ &= E(xy^*) - E(x)E(y^*) \end{aligned}$$

- If $E(x) = E(y) = 0$, then $r_{xy} = c_{xy}$

Independence

- x and y are (statistically) *independent* if

$$P_{xy}(a, b) = P_x(a)P_y(b)$$

- A similar, but weaker, property is *correlation*
- x and y are *uncorrelated* if

$$E(xy^*) = E(x)E(y^*)$$

Some properties

- Independent $\Rightarrow$ uncorrelated
- Uncorrelated $\nRightarrow$ independent
- x and y are said to be *orthogonal* if $E(xy^*) = 0$
- If $E(x) = E(y) = 0$ then orthogonal $\Leftrightarrow$ uncorrelated

The normal distribution

- A *normally distributed* (a.k.a., Gaussian) random variable has the probability density function

$$f_x(a) = \frac{1}{\sigma_x}\phi\left(\frac{a-m_x}{\sigma_x}\right) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left(-\frac{(a-m_x)^2}{2\sigma_x^2}\right)$$

- Properties of the normal distribution: with x and y jointly normally distributed (jointly Gaussian):
  - Any linear combination $ax + by$ is normally distributed
  - Independent $\Leftrightarrow$ uncorrelated

Estimation Theory

- Estimating: obtaining information about unknown quantity θ using data D
- Usually θ cannot be solved exactly from D
- Convenient (but inaccurate) to choose a single value for θ based on probability model and observations
- "*Estimating*" value of θ from observations

Estimating θ

- *Estimation* is done using *estimator*: function of the observations (considered random variables)
- Estimator is also a random variable
- Estimator should be "close" to parameter θ

Making the estimator

- Estimator distribution and parameters derived from observations (e.g., mean and variance)
- *Estimate*: an estimator where observations replace random variables
- Estimate is a *realization* of the estimator (numerical value)

How to select a good estimator?

- No miracle recipe, if all we have is probability model and observations
- Any particular value is always a wrong answer if parameter θ can have several different values
- Choosing "best" wrong answer requires more information than just statistical model
- In this course: find the best estimator according to a cost function (measure of error)

Let's Estimate

- We model a random process by

$$x(n) = \theta + v(n), n = 0, 1, \ldots, N-1, \text{ where } v(n) \sim \mathcal{N}(0, \sigma^2)$$

- How can θ be estimated from observations $x(n)$?
- E.g., by: $\hat{\theta} = x(5) + 3$ (it is a function of the observations, hence it is an estimator)
- Note that since $x(5)$ is a random variable, $\hat{\theta}$ also

Not a great estimator

- $\hat{\theta} = x(5) + 3$ is not likely to be a good estimator
- Constant 3 added to observation takes it further from "true value" (likely)
- How about $\hat{\theta} = x(5)$? Now

$$E\left(\hat{\theta}\right) = E\left(x(5)\right) = E\left(\theta + v(5)\right) = \theta$$

- Seems better: estimator gets correct value on average

Estimation bias

- *bias* = systematic error of an estimator (regarding the expected value)
- Estimator $\hat{\theta} = x(5) + 3$ provides estimates which differ from real value by an average of 3
- *An estimator* $\hat{\theta}$ *of a parameter* θ *is unbiased if* $E(\hat{\theta}) = \theta$
- Above, $\hat{\theta} = x(5)$ is an unbiased estimator

Asymptotical bias

- When an estimator $\hat{\theta}_N$ is formed by using N observations and

$$\lim_{N\to\infty} E\left(\hat{\theta}_N\right) = \theta$$

  the estimator $\hat{\theta}_N$ is *asymptotically unbiased*

- Unbiased $\not\Rightarrow$ better (you might want your estimator to be biased)

Back to our estimator

- Is the unbiased $\hat{\theta} = x(5)$ a good estimator?
- Variance $\mathrm{Var}\left(\hat{\theta}\right) = \mathrm{Var}\left(x(5)\right) = \sigma^2$ is large
- Form another unbiased estimator

$$\hat{\theta} = \frac{1}{N} \sum_{i=0}^{N-1} x(i)$$

- The variance is now σ^2/N

Mean Squared Error

- *Mean Squared Error (MSE)* of an estimator $\hat{\theta}$

$$\mathsf{MSE}(\hat{\theta}) = E\left(\left(\hat{\theta} - \theta\right)^2\right)$$

- Can be written as

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{var}\left(\hat{\theta}\right) + \left[E\left(\hat{\theta}\right) - \theta\right]^2 = \mathsf{variance} + (\mathsf{bias})^2$$

- Unbiased estimator: MSE = Variance

About the MSE

- MSE includes both bias and variance
- Should you always choose the estimator which minimizes the MSE?
- No, because estimator minimizing MSE may depend on estimated parameters: Then estimator is not feasible
- In addition, estimator minimizing MSE is often non-linear

Conditional expectation

- Estimator which minimizes MSE is *conditional expectation* $E(\theta|x)$, where x represents the observations
- Generally this is difficult to calculate, and may be impossible to implement
- Special case: if θ and x are jointly normally distributed then conditional expectation has certain properties

Properties of the conditional expectation

- $(\theta, x) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow E\,(\theta|x)$:
  1. Is unbiased
  2. Has the smallest variance of all estimators
  3. Is a linear function of x
  4. Is normally distributed
- Unfortunately, in practice, assumption of normal distribution usually not reasonable

Likelihood function

- With x a random variable with a probability distribution p depending on parameters θ
- $L(\theta|x_0) = p(x_0|\theta) = P_\theta(x = x_0)$ is called the *likelihood function of* θ *given the outcome* $x = x_0$
- In general form, $L(\theta|x) = p(x|\theta)$ is the *likelihood function* of θ

Likelihood function

- Observe $x = x_0$ and calculate $p(x_0|\theta)$ for a value of θ
- $p(x_0|\theta)$ small: observation x_0 is unlikely for this value of θ
- $p(x_0|\theta)$ large: likely to observe x_0
- Comparison should be carried out for different values of x_0 over same value of θ

Likelihood function, in practice

- In practice we make θ vary, not x_0
- For example, value of $p(x_0|\theta_1)$ compared with $p(x_0|\theta_2)$
- Talk about *likelihood* and not *probability*: it is not a probability distribution of θ
- Shape of the likelihood function $L(\theta|x) = p(x|\theta)$ indicates accuracy of estimate
- Sharp "peak" means most of θ values are unlikely

Using the log-likelihood

- When dealing with likelihood functions, easier to use the log-version of it
- Work with $\log p(x|\theta)$ instead of $p(x|\theta)$: often easier to maximize
- Since θ is for multiple parameters, $L(\theta|x)$ is usually a product of likelihood functions
- Often with exponentiated terms
- Hard to differentiate, work with
- log 'ing the likelihood makes it easier (at least a bit...)
- log being monotonically increasing, maximum values at the same points

Using the log-likelihood: an example

- Assume we have derived the likelihood $L(\theta_1, \theta_2|x)$ as the Gamma distribution:

$$L(\theta_1, \theta_2|x) = \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} x^{\theta_1 - 1} e^{-\theta_2 x}$$

- Now enjoy finding the maximum of $L(\theta_1, \theta_2|x)$ w.r.t. θ_2
- This "thing" looks obviously better with the log-likelihood:

$$\log L(\theta_1, \theta_2|x) = \theta_1 \log \theta_2 - \log \Gamma(\theta_1) + (\theta_1 - 1) \log x - \theta_2 x$$

- Now the derivative looks like

$$\frac{\partial}{\partial \theta_2} \log L(\theta_1, \theta_2|x) = \frac{\theta_1}{\theta_2} - x$$

Using the log-likelihood: an example

- And since we have that x is a sequence of observations $x(0), x(1), \ldots, x(N-1)$, the log-likelihood uses the sum of the $x(i)$ (the product, with the log)
- So, we have

$$\frac{\partial}{\partial \theta_2} \log L(\theta_1, \theta_2 | x) = (N-1)\frac{\theta_1}{\theta_2} - \sum_{i=0}^{N-1} x(i)$$

- And finally we have our estimator

$$\hat{\theta}_2 = \theta_1 \left(\frac{1}{N-1} \sum_{i=0}^{N-1} x(i) \right)^{-1}$$

Estimator variance: relation to the curvature

- Variance of (unbiased) estimator $\hat{\theta}$ is bounded by the *inverse of the Fisher Information (Cramer-Rao bound)*

$$var(\hat{\theta}) \geq I(\theta)^{-1}$$

- Fisher Information $I(\theta)$ is related to the curvature of the log-likelihood:

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta|x)\right)$$

- Small variance of the estimator $\Rightarrow$ large curvature and an accurate estimate

Using Likelihood for estimation

- Likelihood function can be used directly for estimating
- Choose θ so that likelihood function is maximized
- Value of θ that makes observations as likely as possible according to selected model
- This method is called *Maximum Likelihood (ML) method* and corresponding estimator is *ML estimator*

Maximum a posteriori estimator

- So, for Maximum Likelihood: $\hat{\theta}_{ML} = \arg\max_\theta L(\theta|x)$
- Now, if we have some information on θ, in the form of its distribution $p(\theta)$ (prior distribution)
- The *Maximum a posteriori (MAP)* estimator is θ which maximizes posterior distribution

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

Maximum a posteriori estimator

- Which means we have for MAP:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} L(\theta|x)p(\theta)$$

- Difference with ML estimator is that likelihood function is multiplied by prior $p(\theta)$
- More general case, since ML estimator is same as MAP when $p(\theta)$ is uniform

MAP: an example

- Let's take our usual sequence $x(0), \ldots, x(N-1)$, iid and following $\mathcal{N}(\mu_{\text{orig}}, \sigma^2_{\text{orig}})$
- And suppose we know (or assume) that $\mu_{\text{orig}} \sim \mathcal{N}(\mu_{\text{pri}}, \sigma^2_{\text{pri}})$ (prior)
- We want the MAP estimate of μ_{orig} given these assumptions, i.e.

$$\hat{\mu}_{\text{orig}} = \arg\max_{\mu_{\text{orig}}} L(\mu_{\text{orig}}|x)p(\mu_{\text{orig}})$$

MAP: an example

- We have to maximize (with $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$)

$$
\begin{aligned}
L(\mu_{\text{orig}} | x) p(\mu_{\text{orig}}) = \quad & \left[\prod_{i=0}^{N-1} \frac{1}{\sigma_{\text{orig}}} \phi\left(\frac{x(i) - \mu_{\text{orig}}}{\sigma_{\text{orig}}} \right) \right] \\
& \times \left[\frac{1}{\sigma_{\text{orig}}} \phi\left(\frac{\mu_{\text{orig}} - \mu_{\text{pri}}}{\sigma_{\text{pri}}} \right) \right]
\end{aligned}
$$

MAP: an example

- Which, using log-likelihood, is identical to maximizing (w.r.t. μ_{orig})

$$-\sum_{i=0}^{N-1}\left(\frac{x(i)-\mu_{\text{orig}}}{\sigma_{\text{orig}}}\right)^2-\left(\frac{\mu_{\text{orig}}-\mu_{\text{pri}}}{\sigma_{\text{pri}}}\right)^2$$

- Giving finally

$$\begin{aligned}\hat{\mu}_{\text{orig}}^{MAP} = \;& \frac{(N-1)\sigma_{\text{pri}}^2}{(N-1)\sigma_{\text{pri}}^2+\sigma_{\text{orig}}^2}\left[\frac{1}{N-1}\sum_{i=0}^{N-1}x(i)\right] \\ & +\frac{\sigma_{\text{orig}}^2}{(N-1)\sigma_{\text{pri}}^2+\sigma_{\text{orig}}^2}\mu_{\text{pri}}\end{aligned}$$

About Orthogonality

- Why speak of orthogonality and vector spaces here?
- How are vector spaces related to estimation?
- *Orthogonality principle* provides useful way to solve problems where MSE is minimized

Vector spaces and estimation

- Random variables can be considered as vectors in inner product space:
- Linear combinations of random variables are random variables
  - As an inner product one can use $x'y = E(xy^*)$
- MSE can be seen as inner product of $x - \hat{x}$ with itself, since

$$(x - \hat{x})'(x - \hat{x}) = E\left(|x - \hat{x}|^2\right)$$

Orthogonality principle

- Let vectors $x_1, \dots, x_k$ be in a vector space with inner product $x_i' x_j$
- We observe $y = \sum_{i=1}^{k} a_i x_i + e$
- *Orthogonality principle* states: *if we minimize squared norm of error $e'e$, then error is orthogonal to every vector x_i*
- So $\min e'e \Longrightarrow e'x_i = 0, \forall i = 1, 2, \dots, k$

Orthogonality used: linear case

- If we want to construct a linear estimator $\hat{y}$ of the random vector y as

$$\hat{y} = \sum_{i=0}^{N-1} a_i x(i) + \varepsilon$$

- We want to solve coefficients a_i so that MSE $E\left(|y - \hat{y}|^2\right)$ is minimized
- Then $\hat{y}$ is the linear estimator minimizing the MSE if and only if

$$\begin{cases} E\left((y - \hat{y})\, x^*(i)\right) = 0, \forall i = 0, \ldots, N-1 \quad \text{and} \\ E\left(y - \hat{y}\right) = 0 \end{cases}$$

Example of estimation for the linear case

- Example: estimate random variable y with estimator $\hat{y} = f(x)$
  - Want to find a "good" estimator
  - $y =$ quantity that you want to model
  - $x =$ variable that can be observed
  - $\hat{y} =$ quantity which can be calculated when x is observed

Restricting to linear estimators

- A good choice which minimizes the MSE

$$E\left((y-\hat{y})^2\right)$$

- We restrict to linear estimators

$$\hat{y} = ax + b$$

- then $E\left((y-\hat{y})^2\right) = E\left((y-ax-b)^2\right)$

Solving...

- Solve a and b from zeros of the derivative J_a and J_b of the MSE

$$\begin{array}{rcl} J_a & = & -2E\left(\left(y-ax-b\right)x\right)=0 \quad \Leftrightarrow \quad E\left(\left(y-\hat{y}\right)x\right)=0 \\ J_b & = & -2E\left(y-ax-b\right)=0 \quad \Leftrightarrow \quad E(y)=E(\hat{y}) \end{array}$$

- Equations can be interpreted as orthogonality conditions:
  - Error $y-\hat{y}$ is orthogonal to variables (x and the constant 1), which are used to model y
- In other words $E(ex)=0$ and $E(e1)=0$, where $e=y-\hat{y}$

Finally

- Orthogonality conditions can be solved to get an estimator which minimizes the MSE
- Later in the course we will encounter situations where we can apply the orthogonality principle
- We could always get the solution by differentiating, but the orthogonality principle is sometimes easier to use