SGN 21006 Advanced Signal Processing: Lecture 8 Parameter estimation for AR and MA models. Model order selection

Ioan Tabus

Department of Signal Processing
Tampere University of Technology
Finland

Overview

- Basic notions of estimation theory
- Statistically efficient estimates
- ARMA, AR and MA models
- Estimation of AR and ARMA parameters
- Selection of the model order for AR models

Basic notions of estimation theory

Assume that a set of data $x_1, x_2, \ldots, x_N$ is given, and we postulate a model in the form of the pdf $p(x|a)$ as a function of a parameter a. An estimate of a is denoted $\hat{a}$ and it is a function of the data $x_1, x_2, \ldots, x_N$.

Example 1: Consider that the data $x_1, x_2, \ldots, x_N$ are "*independent and identically distributed*", i.e., were generated as realizations of the random variable $X = a + \varepsilon$. We assume that ε_i and ε_j are independent for all pairs of i, j, with $i \neq j$. Additionally we assume that ε has a Gaussian distribution $\mathcal{N}(0, \sigma)$.

- The intuitive significance of a is that it is the mean of the variables X_i.
- Having available a realization, $x_1, x_2, \ldots, x_N$, an estimate of the mean is

$$\hat{a} \quad = \quad \frac{1}{N} \sum_{i=1}^{N} x_i$$

  The estimate is a function of the data. The particular function used to compute the estimate is called estimator. Many estimators can be proposed.
- If the N realizations are changing, the value of $\hat{a}$ also changes. Hence $\hat{a}$ changes with the realizations of X_i and it is itself a random variable.
- By independence assumption the joint pdf is $p(x_1, x_2, \ldots, x_N|a) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{\frac{-\sum_{i=1}^{N}(x_i-a)^2}{2\sigma^2}}$.

The likelihood function

- The likelihood function is denoted $\mathcal{L}(a|x_1, x_2, \ldots, x_N)$ and is seen as a function in the variable a, given the data $x_1, x_2, \ldots, x_N$. It is defined formally as $\mathcal{L}(a|x_1, x_2, \ldots, x_N) = p(x_1, x_2, \ldots, x_N|a)$, although the pdf is defined for a fixed a and has the variables $x_1, x_2, \ldots, x_N$.
- In many cases, when the *pdf*s contain an exponential function, it is more convenient to operate with the logarithm of the likelihood, named log-likelihood $\log \mathcal{L}(a|x_1, x_2, \ldots, x_N)$.
- For Example 1, the likelihood is $\mathcal{L}(a|x_1, x_2, \ldots, x_N) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{\frac{-\sum_{i=1}^{N}(x_i-a)^2}{2\sigma^2}}$ and the log-likelihood is

$$\log \mathcal{L}(a|x_1, x_2, \ldots, x_N) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - a)^2 - \frac{N}{2}\log(2\pi) - N\log\sigma$$

The maximum likelihood (ML) estimate

- The maximum likelihood (ML) estimate is the estimate that maximizes the likelihood function. One may want to estimate the unknown parameter by the value having the highest "likelihood".
- This is just one possible estimate, of the many existing ones (moment fitting estimators are quite popular also). However, in general, when ML is easy to compute, it is likely to have very useful properties.
- The formal definition is

$$\begin{aligned} \hat{a}_{ML} &= \arg\max_a \mathcal{L}(a|x_1, x_2, \ldots, x_N) \\ \hat{a}_{ML} &= \arg\max_a \log \mathcal{L}(a|x_1, x_2, \ldots, x_N) \end{aligned}$$

The two definitions are equivalent, since log function is monotone increasing, and maximizing the function, or its logarithm, are achieved at the same value of a.

- For example 1, to find the maximizing a take the derivative of the log-likelihood with respect to a

$$\frac{\partial}{\partial a} \log \mathcal{L}(a|x_1, x_2, \ldots, x_N) = \frac{\partial}{\partial a} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - a)^2 - \frac{N}{2} \log(2\pi) - N \log \sigma \right\} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - a)$$

In order to find the extremum point equate now the derivative to 0, to get

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \hat{a}_{ML}) &= 0 \\ N\hat{a}_{ML} &= \sum_{i=1}^{N} x_i \\ \hat{a}_{ML} &= \frac{1}{N} \sum_{i=1}^{N} x_i \end{aligned}$$

Properties of a good estimator

The most interesting properties of an estimate are the following

- Small bias:

$$Bias(\hat{a}) = E[\hat{a}] - a$$

When $Bias(\hat{a}) = 0$ the estimate is said "unbiased".

- Small variance:

$$Variance = E[(\hat{a} - a)^2]$$

To quantify how small a variance can be, several bounds are available, the most important being Cramer-Rao bound.

- Consistency:

$$\lim_{N \to \infty} \hat{a} = a$$

When the number of available data points is very large one wishes to have a very precise estimation, very close to the true a, and the more data is provided, the closer the estimate becomes to the true value.

ML estimate of a in Example 1

- In the case of Example 1, the bias in the ML estimate of a can be evaluated easily:

$$E[\hat{a}_{ML}] = E[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \frac{1}{N}\sum_{i=1}^{N} E[a+\varepsilon_i] = \frac{1}{N}\sum_{i=1}^{N} a + \frac{1}{N}\sum_{i=1}^{N} E[\varepsilon_i] = a$$

  so that the bias $B(\hat{a}_{ML}) = E[\hat{a}_{ML}] - a = 0$ and the estimator is unbiased.

- The variance of the estimate $\hat{a}_{ML}$ results from

$$Var[\hat{a}_{ML}] = E[(\hat{a}_{ML}-a)^2] = E[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - a\right)^2] = \frac{1}{N^2}\sum_{i=1}^{N} Var(\varepsilon_i) = \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{N}$$

  where we have used the property that the variance of the sum of independent random variables $\varepsilon_1, \ldots, \varepsilon_N$ is the sum of variances of the variables, each variance being σ^2.

- When the number of measurements N grows to infinite, the variance of the ML estimate tends to 0, $\lim_{N\to\infty} Var[\hat{a}_{ML}] = \lim_{N\to\infty} \frac{\sigma^2}{N} = 0$. This ensures that the fluctuations of $\hat{a}_{ML}$ about mean are becoming extremely small, and is one form of convergence for random variables ($\hat{a}_{ML}$ is said to converge in mean square to a).

ML estimate of σ in Example 1

- When computing the ML estimate of a we assumed σ is known. In fact the estimate of a does not depend on σ.
- If we want to find an estimation of σ in addition to the estimate of a, we have to solve the system:

$$\frac{\partial}{\partial a} \log \mathcal{L}(a, \sigma | x_1, x_2, \ldots, x_N) = 0$$
$$\frac{\partial}{\partial \sigma} \log \mathcal{L}(a, \sigma | x_1, x_2, \ldots, x_N) = 0$$

and denote the solution $(\hat{a}_{ML}, \hat{\sigma}_{ML})$.

- The maximizing $\hat{a}_{ML}$ will not depend on σ and is as before $\hat{a}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$.
- To find the maximizing σ take the derivative of the log-likelihood with respect to σ

$$\begin{aligned} \frac{\partial}{\partial \sigma} \log \mathcal{L}(a, \sigma | x_1, x_2, \ldots, x_N) &= \frac{\partial}{\partial \sigma} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - a)^2 - \frac{N}{2} \log(2\pi) - N \log \sigma \right\} \\ &= \frac{2}{2\sigma^3} \sum_{i=1}^{N} (x_i - a)^2 - \frac{N}{\sigma} \end{aligned}$$

Equating this to 0, and taking now also a to be the ML estimate,

$$\begin{aligned} \frac{2}{2\hat{\sigma}_{ML}^3} \sum_{i=1}^{N} (x_i - \hat{a}_{ML})^2 - \frac{N}{\hat{\sigma}_{ML}} &= 0 \\ \hat{\sigma}_{ML}^2 &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{a}_{ML})^2 \end{aligned}$$

ML estimate of σ in Example 1

- The bias in the ML estimate of σ^2 can be evaluated as:

$$E[\hat{\sigma}^2_{ML}] = \frac{1}{N}\sum_{i=1}^{N} E[(x_i - \hat{a}_{ML})^2] = \frac{1}{N}\sum_{i=1}^{N} E[(a + \varepsilon_i - \hat{a}_{ML})^2]$$

and using that $\hat{a}_{ML} = a + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i$:

$$\begin{aligned} E[\hat{\sigma}^2_{ML}] &= \frac{1}{N}\sum_{i=1}^{N} E[(\varepsilon_i - \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i)^2] = \frac{1}{N}\sum_{i=1}^{N} E\left[\left(\frac{(N-1)\varepsilon_i - \sum_{j\neq i}\varepsilon_i}{N}\right)^2\right] \\ &= \left(\frac{(N-1)^2 E\varepsilon_i^2 + \sum_{j\neq i} E\varepsilon_i^2}{N^2}\right) = \frac{(N-1)N}{N^2}\sigma^2 = \frac{N-1}{N}\sigma^2 \end{aligned}$$

so that the bias $B(\hat{\sigma}^2_{ML}) = E[\hat{\sigma}^2_{ML}] - \sigma^2 = \frac{-1}{N}\sigma^2$ and the estimator is biased.

- An unbiased estimator of σ^2 can be found simply by scaling the ML estimator:

$$\hat{\sigma}^2_{unbiased} = \frac{N}{N-1}\hat{\sigma}^2_{ML} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{a}_{ML})^2$$

- The ML estimate, although is biased, is preferred to be use in some applications (in spectrum estimation)

Cramér-Rao bound

- Consider a set of data $\underline{x} = [x_1 \ \ldots \ x_N]$ generated by a distribution $p(\underline{x}|a)$ and take any estimate $\hat{a}(\underline{x})$. Denote the bias $Bias = B(a) = E[\hat{a}(\underline{x}) - a|a]$.
- The MSE can be defined as

$$\begin{aligned} MSE &= E[(\hat{a}(\underline{x}) - a)^2|a] = E[\hat{a}(\underline{x})^2|a] - 2aE[\hat{a}(\underline{x})|a] + a^2 \\ Var &= E[(\hat{a}(\underline{x}) - E[\hat{a}(\underline{x})|a])^2|a] = E[\hat{a}(\underline{x})^2|a] - (E[\hat{a}(\underline{x})|a])^2 \\ E[\hat{a}(\underline{x})^2|a] &= Var + (E[\hat{a}(\underline{x})|a])^2 \\ MSE &= Var + (E[\hat{a}(\underline{x})|a])^2 - 2aE[\hat{a}(\underline{x})|a] + a^2 = Var + (Bias)^2 \end{aligned}$$

Then the MSE of the estimate can be bounded by using the mean square error (MSE) inequality:

$$MSE = E[(\hat{a}(\underline{x}) - a)^2|a] \geq \frac{[1 + \frac{\partial}{\partial a} B(a)]^2}{E\left\{\left[\frac{\partial}{\partial a} \log p(\underline{x}|a)\right]^2 |a\right\}} = CRB$$

- The denominator in CRB is known as the Fisher information $I(a) = E\left\{\left[\frac{\partial}{\partial a} \log p(\underline{x}|a)\right]^2 |a\right\}$
- If the estimator is unbiased, $Bias = B(a) = 0$, then

$$Var(\hat{a}(\underline{x})) = E[(\hat{a}(\underline{x}) - a)^2|a] \geq \frac{1}{E\left\{\left[\frac{\partial}{\partial a} \log p(\underline{x}|a)\right]^2 |a\right\}} = \frac{1}{I(a)} = CRB$$

Efficient estimate

- An estimator is efficient if
  - it is unbiased $Bias = B(a) = E[\hat{a}(\underline{x}) - a|a] = 0$.
  - Its variance achieves the CRB

$$Var(\hat{a}(\underline{x})) \quad = \quad E[(\hat{a}(\underline{x}) - a)^2|a] \geq \frac{1}{E\left\{\left[\frac{\partial}{\partial a}\log p(\underline{x}|a)\right]^2|a\right\}} = \frac{1}{I(a)} = CRB$$

- When an estimator is efficient, there is no other unbiased estimator that can have smaller variance.

Asymptotic properties of the ML estimator

- the estimator is consistent: $\hat{a}(\underline{x}) \to a$ when $N \to \infty$
- the estimator is asymptotically Gaussian, with mean a and variance inverse of the Fisher information $I(a)$.
- the estimator is asymptotically efficient
- When an estimator is efficient, there is no other unbiased estimator that can have smaller variance.

ML estimates of uniform distribution parameters

- Consider a uniform distribution $\mathcal{U}(0, a)$. Here the parameter is a, which is the upper bound of the interval for which the distribution is defined. The pdf is $p(x|a) = 1/a$ if $x \leq a$ and $p(x|a) = 0$ if $x > a$.
- Assume that a single realization x_1 is available, so $N = 1$. What is the ML estimate of a?
- We need to maximize with respect to a the likelihood function $L(a|x_1)$. If $a < x_1$, $L(a|x_1) = 0$. If $a \geq x_1$, then $L(a|x_1) = 1/a$. The maximum $L(a|x_1) = 1/a$ is obtained when $\hat{a}_{ML} = x_1$.
- When $N = 1$ the ML is very biased. We have $E[x_1] = a/2$ and since $\hat{a}_{ML} = x_1$, we have $E[\hat{a}_{ML}] = a/2$, and the bias is $Bias(\hat{a}_{ML}) = a/2$.
- Assume two realizations, x_1, x_2 are available. Take x_1 to be the largest of them. Repeating the reasoning above, it results that $\hat{a}_{ML} = max\{x_1, x_2\}$.
- Assume that the number of realizations goes to infinity. Then the value $\hat{a}_{ML} = max\{x_1, \ldots, x_N\}$ tends to a and the ML estimate is consistent.
- ML for uniform distribution shows that not always partial derivatives are needed to find the ML estimates.

ML estimates for Gaussian distributed model errors

- Most often we assume Gaussian identically and independent distribution for the errors e_t in some models involving a vector of parameters $\underline{\theta}$ (see the AR models later in the lecture)

$$y_t \quad = \quad \underline{u}_t^T \underline{\theta} + e_t$$

Then maximizing the likelihood

$$\max_{\underline{\theta}} \log \mathcal{L}(\theta | x_1, x_2, \ldots, x_N) = \max_{\underline{\theta}} \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \underline{u}_i^T \underline{\theta})^2 - \frac{N}{2} \log(2\pi) - N \log \sigma \right)$$

becomes equivalent to minimizing the sum of squares

$$\min_{\underline{\theta}} \sum_{i=1}^{N} (y_i - \underline{u}_i^T \underline{\theta})^2$$

which explains why the LS estimator has very good statistical properties.

ARMA models: Simplest example AR(1) model

- Consider a first order autoregressive model

$$y_n = ay_{n-1} + e_n$$

where e_n is a white noise sequence of zero mean and variance σ_e^2, initialized at $n = 0$ with $y_0 = 0$.

- Using the delay in time operator we can obtain the dependence of y_n on the inputs $e_1, \ldots, e_n$

$$\begin{aligned} y_n &= aq^{-1}y_n + e_n \\ y_n &= \frac{1}{1 - aq^{-1}} e_n \\ y_n &= (1 + aq^{-1} + a^2q^{-2} + a^3q^{-3} + \ldots)e_n \\ y_n &= e_n + ae_{n-1} + a^2e_{n-2} + a^3e_{n-3} + \ldots + a^{n-1}e_1 \end{aligned}$$

- The cross-correlation Ey_ne_{n+k} for all $k \geq 0$ is zero:

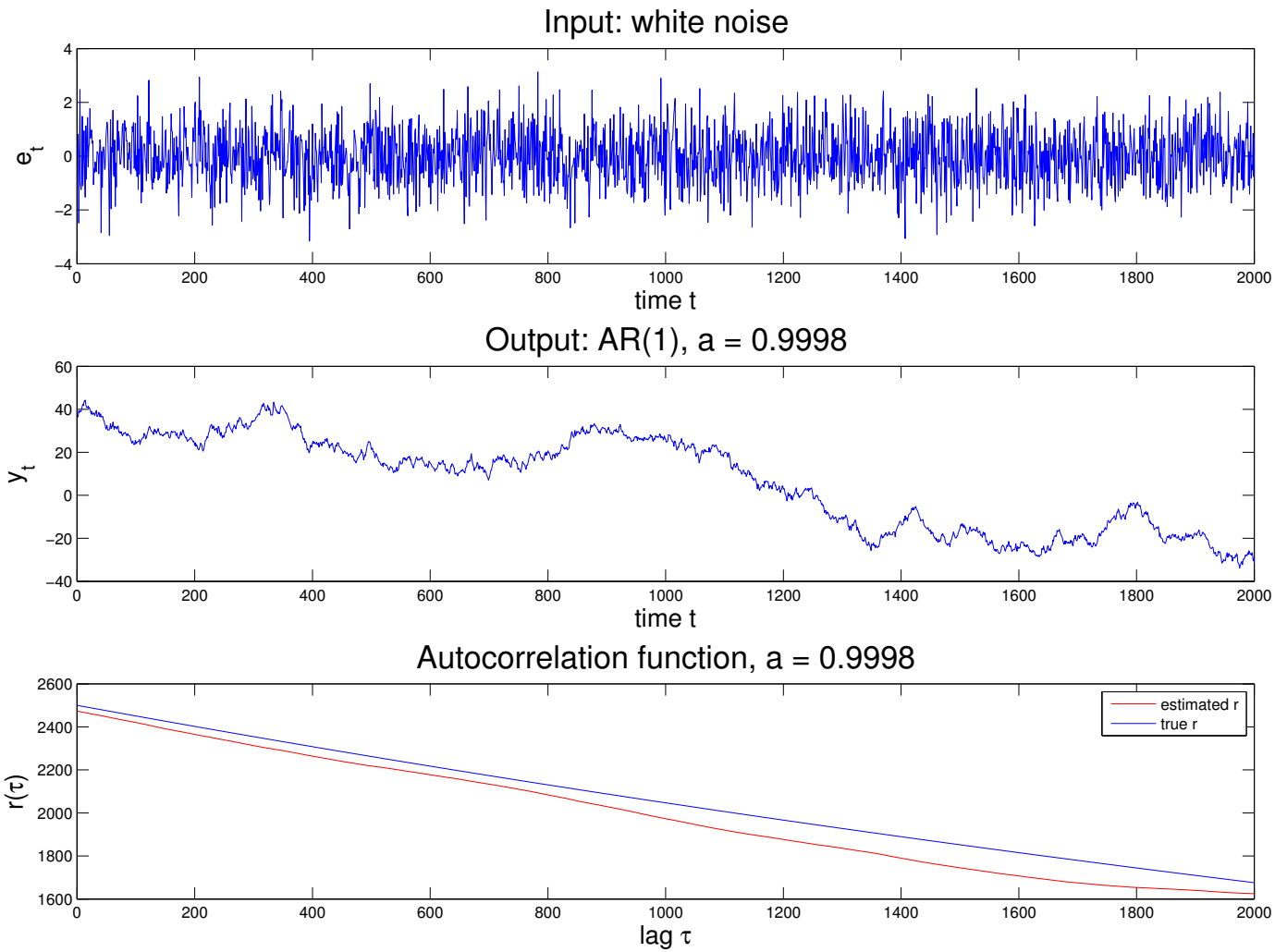
$$\begin{aligned} y_n &= e_n + ae_{n-1} + a^2e_{n-2} + a^3e_{n-3} + \ldots + a^{n-1}e_1 \\ Ey_ne_{n+k} &= Ee_{n+k}e_n + aEe_{n+k}e_{n-1} + a^2Ee_{n+k}e_{n-2} + a^3Ee_{n+k}e_{n-3} + \ldots + a^{n-1}Ee_{n+k}e_1 = 0 \end{aligned}$$

since e_n is not correlated with any e_{n+k} with $k \neq 0$.

- Given $y(0) = 0$ and $e_1, \ldots, e_N$ the values of $y_1, \ldots, y_N$ can be computed
- Given same $y(0) = 0$ and another sequence $e_1, \ldots, e_N$ results in other sequence $y_1, \ldots, y_N$. Hence the sequence $y_1, \ldots, y_N$ is random and depends on the realization of $e_1, \ldots, e_N$.

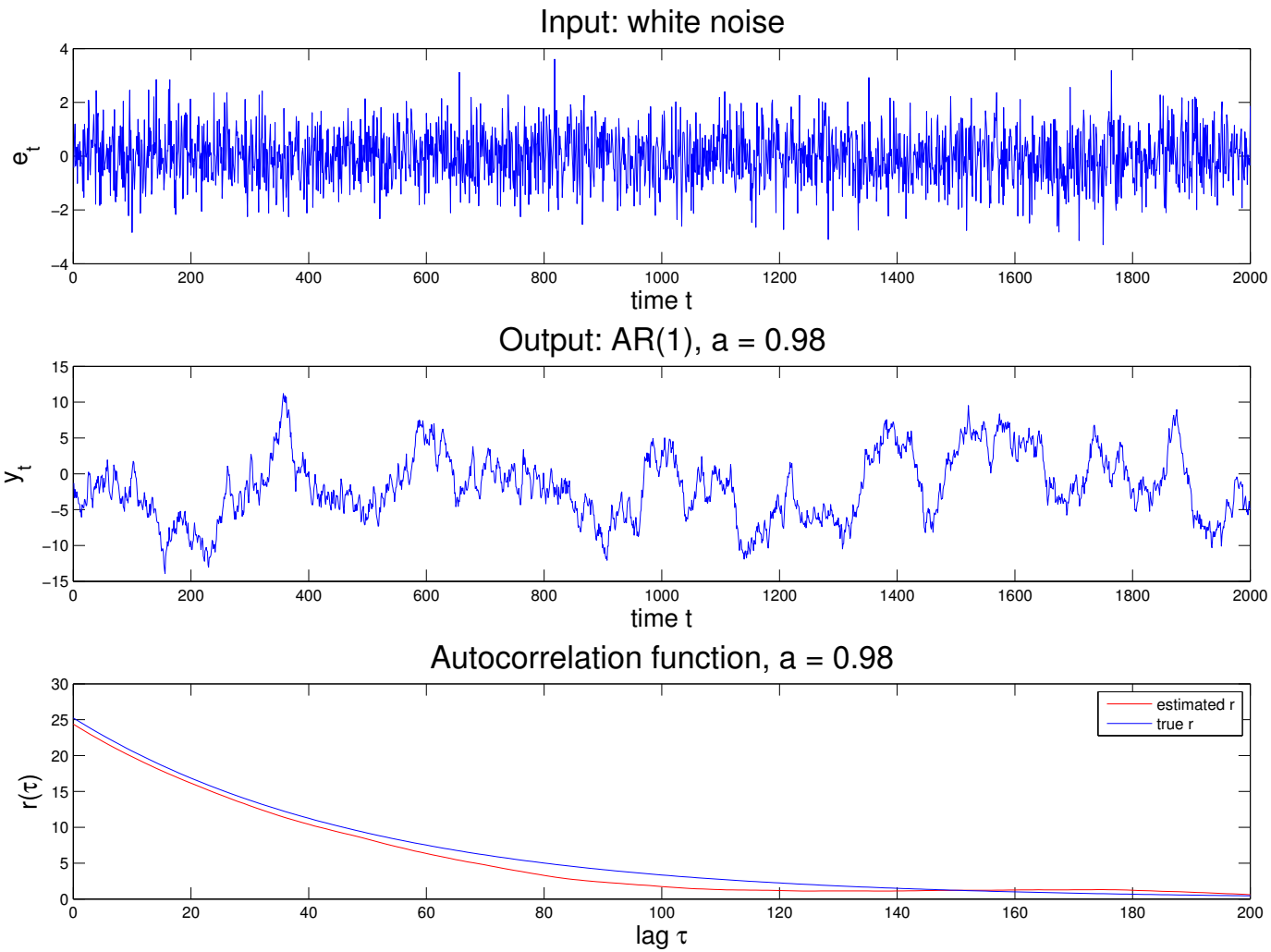
AR(1) model

Input and output signals of the model and the autocorrelation function, for $a = 0.9998$



AR(1) model

Input and output signals of the model and the autocorrelation function, for $a = 0.98$



Stability and stationarity of AR(1) model

- Property: The mean of y_n is zero, for all n. We have $y_1 = e_1$ and the mean $Ey_1 = 0$. Then $Ey_2 = aEy_1 + Ee_2 = 0$ and by induction we see that $Ey_n = 0$.
- We want to evaluate the variance of y_i, i.e., the autocorrellation function $r_{yy}(i, i) = Ey_i^2$. We have

$$r_{yy}(n, n) = Ey_n^2 = E(ay_{n-1} + e_n)^2 = a^2 Ey_{n-1}^2 + 2aEy_{n-1}e_n + Ee_n^2 = a^2 r_{yy}(n-1, n-1) + \sigma_e^2$$

- Iterating the equation $r_{yy}(n, n) = a^2 r_{yy}(n-1, n-1) + \sigma_e^2$

$$\begin{aligned} r_{yy}(n, n) &= a^2 r_{yy}(n-1, n-1) + \sigma_e^2 = a^4 r_{yy}(n-2, n-2) + (1 + a^2)\sigma_e^2 \\ &= a^{2n} r_{yy}(0, 0) + (1 + a^2 + a^4 + \ldots + a^{2n-2})\sigma_e^2 = a^{2n} r_{yy}(0, 0) + \frac{1 - a^{2n}}{1 - a^2}\sigma_e^2 \\ &= \frac{1 - a^{2n}}{1 - a^2}\sigma_e^2 \end{aligned}$$

Stability and stationarity of AR(1) model

- In general the autocorrelations depend on n,

$$r_{yy}(n,n) = \frac{1-a^{2n}}{1-a^2}\sigma_e^2$$

and hence the process AR(1) is not stationary.

- If the parameter a is smaller than 1 in modulus, after an initial transient regime the autocorrelation values are converging to the value

$$r_{yy}(n,n) = \frac{1}{1-a^2}\sigma_e^2$$

and the process is asymptotically stationary (but in practice the transient regime can be short).

- If the parameter a is larger than 1 in modulus, the autocorrelations are diverging, and the AR(1) process is non-stationary.
- In general, for a process AR(k) $y_n = a_1 y_{n-1} + \ldots a_k y_{n-k} + e_n$, the condition of stability is that all the roots of the polynomial $A(z) = 1 - a_1 z - a_2 z^2 - \ldots - a_k z^k$ are inside the unit circle. For the AR(1) this requires that the root of $A(z) = 1 - a_1 z$, which is a, be smaller than 1 in modulus.
- Exercise If one initializes the AR(1) randomly, with y_0 which has zero mean and variance $\frac{1}{1-a^2}\sigma_e^2$, then the AR(1) model is wide sense stationary, i.e all the autocorrelation functions $r_{yy}(i,k) = Ey(i)y(k)$ depend only on the difference between i and j, for $i, j > 0$.

AR, MA and ARMA models

- The general ARMA(n,m) model is described by the equation with differences

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + e_t + b_1 e_{t-1} + \ldots + b_m e_{t-m}$$

- Applying the Z transform to both members and assuming zero initial conditions

$$\mathcal{Z} y_t = a_1 \mathcal{Z} y_{t-1} + a_2 \mathcal{Z} y_{t-2} + \ldots + a_n \mathcal{Z} y_{t-n} + \mathcal{Z} e_t + b_1 \mathcal{Z} e_{t-1} + \ldots + b_m \mathcal{Z} e_{t-m}$$

$$Y(z) = a_1 z^{-1} Y(z) + a_2 z^{-2} Y(z) + \ldots + a_n z^{-n} Y(z) + E(z) + b_1 z^{-2} E(z) + \ldots + b_m z^{-m} E(z)$$

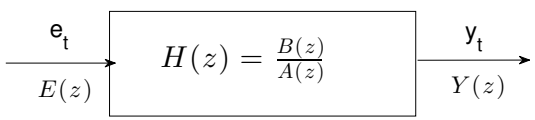
$$(1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_n z^{-n}) Y(z) = (1 + b_1 z^{-1} + \ldots + b_m z^{-m}) E(z)$$

$$Y(z) = \frac{1 + b_1 z^{-1} + \ldots + b_m z^{-m}}{1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_n z^{-n}} E(z) = \frac{B(z)}{A(z)} E(z)$$

- The model written in the time domain using time shift operator is

$$(1 - a_1 q^{-1} - a_2 q^{-2} - \ldots - a_n q^{-n}) y_t = (1 + b_1 q^{-1} + \ldots + b_m q^{-m}) e_t$$

$$y_t = \frac{1 + b_1 q^{-1} + \ldots + b_m q^{-m}}{1 - a_1 q^{-1} - a_2 q^{-2} - \ldots - a_n q^{-n}} e_t$$



AR, MA models

- The autoregressive AR(n) model is described by the equation with differences or the transfer function $H(z) = \frac{1}{A(z)}$

$$\begin{aligned} y_t &= a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + e_t \\ H(z) &= \frac{1}{A(z)} \end{aligned}$$

- The moving average MA(m) model is described by the equation with differences

$$\begin{aligned} y_t &= e_t + b_1 e_{t-1} + \ldots + b_m e_{t-m} \\ H(z) &= B(z) \end{aligned}$$

- Equivalence between models:
  - An MA(m) model is equivalent to an AR(∞) model
  - An AR(n) model is equivalent to an MA(∞) model
  - An ARMA(n,m) model has an equivalent AR(∞) model and an equivalent MA(∞) model

Equivalence of an ARMA model to AR(∞) model

- Example: consider an ARMA(1,1) model

$$
\begin{aligned}
H(z) &= \frac{1-0.8z^{-1}}{1+0.9z^{-1}} = ARMA(1,1) \\
H(z) &= \frac{1}{(1+0.9z^{-1})\frac{1}{(1-0.8z^{-1})}} \\
H(z) &= \frac{1}{(1+0.9z^{-1})(1+0.8z^{-1}+0.8^2z^{-2}+0.8^3z^{-3}+\ldots)} \\
&= AR(\infty)
\end{aligned}
$$

- The infinitely long AR(∞) can be truncated to a finite number of coefficients and then we have an approximate equivalence, between an ARMA(n,m) model and a very long AR(L) model. The long model has L coefficients, much more than the $n+m$ coefficients of the ARMA model. If two models are equivalent, in principle we prefer models having less parameters. However, the parameters in the AR model are easier to estimate (we already discussed the Levinson-Durbin and LS methods) while for ARMA models the parameter estimation is a nonlinear optimization problem, difficult to solve.

Estimation of AR models

- The autoregressive AR(n) model is described by the equation with differences

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + e_t$$

which is a prediction model of order n. Levinson-Durbin algorithm can be used to estimate its parameters, for all possible orders $n = 1, \ldots, n_{max}$.

- The LS method can also be applied to estimate the AR parameters. Denote the unknowns $\underline{\theta} = [a_1 \ a_2 \ \ldots a_n]^T$. Denoting the data vector $\underline{u}_t = [y_{t-1} \ y_{t-2} \ \ldots y_{t-n}]^T$ and since the "desired" values are $d_t = y_t$ the following system of equations results

$$\begin{cases} y_1 = \underline{u}_1^T \underline{\theta} + e_1 \\ \ldots \\ y_N = \underline{u}_N^T \underline{\theta} + e_N \end{cases}$$

The system can be solved in LS sense, and has the solution

$$\underline{\hat{\theta}} = \left(\sum_{i=1}^{N} \underline{u}_i \underline{u}_i^T \right)^{-1} \sum_{i=1}^{N} \underline{u}_i d_i \qquad (1)$$

Estimation of MA models

- The moving average MA(m) model is described by the equation with differences

$$\begin{aligned} y_t &= e_t + b_1 e_{t-1} + \ldots + b_m e_{t-m} \\ H(z) &= B(z) \\ r_y(k) &= E[y_t y_{t-k}] = E(e_t + b_1 e_{t-1} + \ldots + b_m e_{t-m})(e_{t-k} + b_1 e_{t-k-1} + \ldots + b_m e_{t-k-m}) \end{aligned}$$

- If $k > m$ (or $k < -m$) then $r_y(k) = 0$.
- If $0 < k \leq m$ (or $0 \leq k \leq -m$) then

$$r_y(k) = \sigma_e^2 \sum_{i=0}^{m-k} b_i b_{i+k}$$

- collecting all equations together

$$\begin{aligned} r_y(0) &= \sigma_e^2(1 + b_1^2 + b_2^2 + \ldots + b_m^2) \\ r_y(1) &= \sigma_e^2(b_1 + b_1 b_2 + b_2 b_3 + \ldots + b_{m-1} b_m) \\ &\ldots \\ r_y(m-1) &= \sigma_e^2(b_{m-1} + b_1 b_m) \\ r_y(m) &= \sigma_e^2 b_m \end{aligned}$$

which is a nonlinear system of $m+1$ equations and $m+1$ unknowns $\sigma_e^2, b_1, \ldots, b_m$. Nonlinear solvers may have troubles due to local minima in the criterion. More often used solver is the two stage least squares method, used also for the more general ARMA models, described next.

Estimation of ARMA models: Two-stage LS method

- The general ARMA(n,m) model is described by the equation with differences

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + e_t + b_1 e_{t-1} + \ldots + b_m e_{t-m}$$

- An ARMA(n,m) model has an equivalent AR(∞) model, $\frac{B(z)}{A(z)} = \frac{1}{C(z)}$ where $C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + \ldots$ has an infinite number of coefficients.
- We estimate first a truncated form of $C(z)$, where only the coefficients $c_1, c_2, c_3, \ldots, c_L$ are retained, with a very large L, but $L < N$, e.g. $L = N/4$. We use for estimation the LS method (this is the first LS stage) in the model

$$y_t = c_1 y_{t-1} + c_2 y_{t-2} + \ldots + c_L y_{t-L} + \varepsilon_t$$

using the values $y_1, \ldots, y_N$ and we obtain the estimated coefficients $\hat{c}_1, \hat{c}_2, \hat{c}_3, \ldots, \hat{c}_L$.

- from the estimated coefficients $\hat{c}_1, \hat{c}_2, \hat{c}_3, \ldots, \hat{c}_L$ compute the estimated values of the residuals of the model, $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_N$,

$$\hat{\varepsilon}_t = y_t - \hat{c}_1 y_{t-1} - \hat{c}_2 y_{t-2} - \ldots - \hat{c}_L y_{t-L}$$

Estimation of ARMA models: Two-stage LS method

- We return to estimating the unknown coefficients in the ARMA(n,m), using now the equations for all $t = 1, \ldots, N$

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + \hat{\varepsilon}_t + b_1 \hat{\varepsilon}_{t-1} - + \ldots + b_m \hat{\varepsilon}_{t-m}$$

from which we obtain a system of equations, for all $t = 1, \ldots, N$

$$y_t - \hat{\varepsilon}_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_n y_{t-n} + + b_1 \hat{\varepsilon}_{t-1} - + \ldots + b_m \hat{\varepsilon}_{t-m}$$

with the unknowns $\underline{\theta} = [a_1 \ a_2 \ \ldots a_n \ b_1 \ b_2 \ \ldots b_m]^T$. Denoting the data vector $\underline{u}_t = [y_{t-1} \ y_{t-2} \ \cdots y_{t-n} \ \hat{\varepsilon}_{t-1} \ \hat{\varepsilon}_{t-2} \ \ldots \hat{\varepsilon}_{t-m}]^T$ and the "desired" values $d_t = y_t - \hat{\varepsilon}_t$ the overdetermined system of equations becomes

$$\begin{cases} d_1 = \underline{u}_1^T \underline{\theta} \\ \ldots \\ d_N = \underline{u}_N^T \underline{\theta} \end{cases}$$

having the LS solution

$$\underline{\hat{\theta}} = \left(\sum_{i=1}^{N} \underline{u}_i \underline{u}_i^T \right)^{-1} \sum_{i=1}^{N} \underline{u}_i d_i \qquad (2)$$

Selecting the order of ARMA models

- A set of candidate orders is given, most usually $k \in \{1, 2, \ldots, K\}$. An estimated method is used for obtaining for each order k the estimates of the parameters $\underline{\hat{\theta}}_k$ and the power $\hat{\sigma}_k^2$ of the residuals computed with the estimated vector $\underline{\hat{\theta}}_k$.
- For each candidate order k a certain "criterion" is computed and the selected order is the order k^* minimizing the criterion.
- There are many criteria which are in use, derived based on various considerations, more often "information theoretic" considerations. None of them is good in all situations. A good policy is to check the optimal value given by each criterion and take a voting decision, taking into account the dangers of overfitting or underfitting in the current application.
- All criteria are penalizing a too large number of parameters and contain a term which rewards very good fits. The selection is a trade-off between overfitting and underfitting.
- Overfitting occurs when having a very good fit to data, obtained for very large k, but in which case the noise in the data is modeled as well. Such a model will not perform well on other data, different than the one where it was fitted.
- Under-fitting occurs when we have too few parameters, and the fit to the data is poor.

Criteria for order selection

- Final prediction error (FPE) criterion

$$FPE(k) = \frac{N+k}{N-k}\hat{\sigma}_k^2$$

FPE tends to underestimate the order.

- Akaike Information Criterion (AIC)

$$AIC(k) = N \ln(\hat{\sigma}_k^2) + 2k$$

For large N, the AIC tends to overestimate the order (is not penalizing enough). FPE never exceeds the selection of AIC, hence is better for large N.

- Minimum Description Length (MDL) criterion, also known as Bayesian Information Criterion (BIC)

$$MDL(k) = N \ln(\hat{\sigma}_k^2) + k \ln N$$

For large N, MDL estimates correctly the order (is penalizing correctly the number of parameters).

- However, for small N the performance of AIC and MDL is similar. There are no definitive ranking of the three methods.