FSAN-815/ELEG-815: Foundations of Statistical Learning

Gonzalo R. Arce

Chapter 2: Stationary Processes and Models

Fall 2014

Course Objectives & Structure

The course provides an introduction to the mathematics of data analysis and a detailed overview of statistical models for inference and prediction.

Course Structure:

- Weekly lectures [notes: www.ece.udel.edu/∼arce/Courses]
- Homework & computer assignments [30%]
- Midterm & Final examinations [70%]

Textbooks:

- Papoulis and Pillai, Probability, random variables, and stochastic processes.
- Hastie, Tibshirani and Friedman, The elements of statistical learning.
- Haykin, Adaptive Filter Theory.

Stationary Process and Models

- Statistic process describes the time evolution of statistical phenomena
- A stochastic process is not a single function of time but an infinite number of possible realizations
- A single realization is called a time series
- A full joint distribution function of an arbitrary stochastic process is difficult to obtain or estimate
- Settle for a partial characterization

Consider a discrete-time stochastic process

$$x(n), x(n-1), \ldots, x(n-M)$$

which may be complex.

Definitions (Mean, Auto-Correlation, and Auto-Covariance)

The mean process is given by

$$\mu(n) = E\{x(n)\}$$

The auto–correlation is defined as

$$r(n, n-k) = E\{x(n)x^*(n-k)\}$$

The auto–covariance is given by

$$\begin{aligned} c(n, n-k) &= E\{[x(n)-\mu(n)][x(n-k)-\mu(n-k)]^*\} \\ &= r(n, n-k) - \mu(n)\mu^*(n-k) \end{aligned}$$

Definition (Wide-Sense Stationary)

A discrete-time stochastic process is wide–sense stationary (WSS) if

$$\begin{aligned} \mu(n) &= \mu \quad \text{for all n} \\ r(n, n-k) &= r(k) \quad \text{and} \\ c(n, n-k) &= c(k) \quad k = 0, \pm 1, \pm 2, \ldots \end{aligned}$$

Let $\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-M+1)]^T$ be a $M \times 1$ observation vector. Then for $\{x(n)\}$ WSS, the correlation matrix is

$$\mathbf{R} = E\{\mathbf{x}(n)\mathbf{x}^H(n)\} = \begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r(-1) & r(0) & \cdots & r(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(-M+1) & r(-M+2) & \cdots & r(0) \end{bmatrix}$$

Properties of the correlation matrices

For a stationary discrete time process: $\mathbf{R}^H = \mathbf{R}$ (Hermetian)

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r(-1) & r(0) & \cdots & r(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(-M+1) & r(-M+2) & \cdots & r(0) \end{bmatrix} =$$

$$\begin{bmatrix} r(0) & r^*(-1) & \cdots & r^*(-M+1) \\ r^*(1) & r(0) & \cdots & r^*(-M+2) \\ \vdots & \vdots & \ddots & \vdots \\ r^*(M-1) & r^*(M-2) & \cdots & r(0) \end{bmatrix}$$

Consequence: $\Rightarrow r(-k) = r^*(k)$

The correlation matrix is Toeplitz

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(M-1) \\ r^*(1) & r(0) & r(1) & \cdots & r(M-2) \\ r^*(2) & r^*(1) & r(0) & \cdots & r(M-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r^*(M-1) & r^*(M-2) & r^*(M-3) & \cdots & r(0) \end{bmatrix}$$

For any non-zero vector **a**

$$\mathbf{a}\mathbf{R}\mathbf{a}^H \geq 0 \quad \text{(positive semi-definite)}$$

and usually

$$\mathbf{a}\mathbf{R}\mathbf{a}^H > 0 \quad \text{(positive definite)}$$

Result: **R** is positive definite if the samples in **x** are not linearly dependent. In this case $\mathbf{R}^{-1}$ exists.

Historical Note: Diagonal–constant matrices are named after the mathematician Otto Toeplitz (1881–1940)

Stochastic Models

- A model is used to describe the hidden laws governing the generation of physical data observed
- We assume that $x(n), x(n-1), \cdots$ have statistical dependencies that can be modeled as

  where $v(n)$ is a purely random process
- Linear model types:
  1. Auto Regressive – no past model input samples used
  2. Moving Average – no past model output samples used
  3. Auto Regressive Moving Average – both past input and output used

General Stochastic Model:

$$\left(\begin{array}{c}\text{Model}\\ \text{output}\end{array}\right) + \underbrace{\left(\begin{array}{c}\text{Linear combination}\\ \text{of past outputs}\end{array}\right)}_{\text{AR part}} = \underbrace{\left(\begin{array}{c}\text{Linear combination of}\\ \text{present \& past inputs}\end{array}\right)}_{\text{MA part}}$$

Three model possibilities:

1. AR – auto regressive
2. MA – moving average
3. ARMA – mixed AR and MA

Model Input: assumed to be an i.i.d. zero mean Gaussian process:

$$E\{v(n)\} = 0 \quad \text{for all } n$$

$$E\{v(n)v^*(k)\} = \begin{cases} \sigma_v^2 & k = n \\ 0 & \text{otherwise} \end{cases}$$

Auto-Regressive Models

Definition (Auto-Regressive)

The time series $\{x(n)\}$ is said to be generated by an AR model if

$$x(n) + a_1^* x(n-1) + \cdots + a_M^* x(n-M) = v(n)$$

or

$$x(n) = w_1^* x(n-1) + \cdots + w_M^* x(n-M) + v(n)$$

where $w_k = -a_k$.

- This is an order M model and $v(n)$ is referred to as the noise term
- Note that we can set $a_0 = 1$ and write

$$\sum_{k=0}^{M} a_k^* x(n-k) = v(n)$$

which is a convolution sum

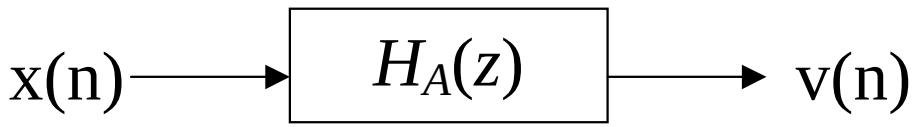
Thus taking Z-transforms

$$\begin{aligned} Z\{a_n^*\} &= A(z) = \sum_{n=0}^{M} a_n^* z^{-n} \\ Z\{x(n)\} &= X(z) = \sum_{n=0}^{\infty} x(n) z^{-n} \\ Z\{v(n)\} &= V(z) = \sum_{n=0}^{\infty} v(n) z^{-n} \end{aligned}$$

and

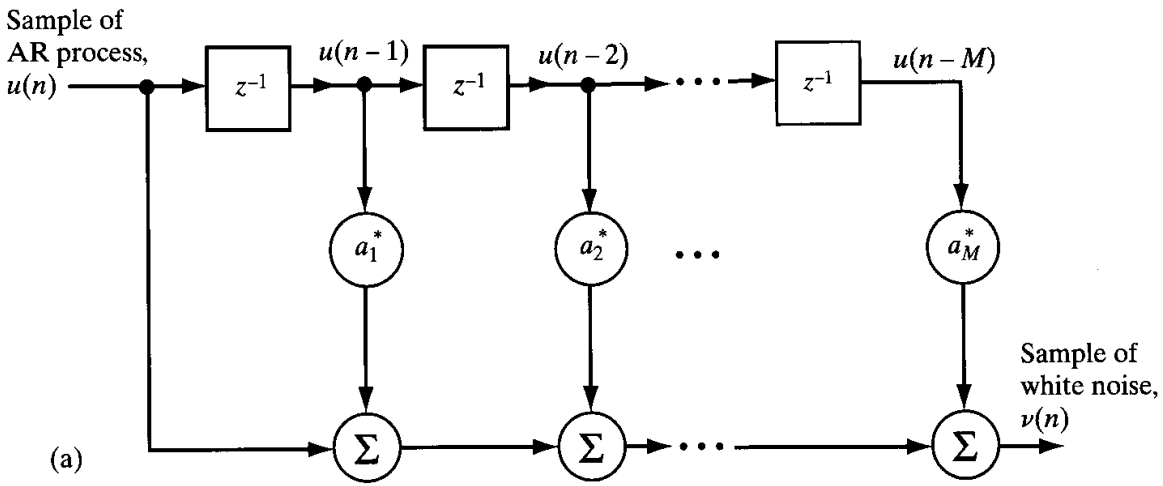
$$\sum_{k=0}^{M} a_k^* x(n-k) = v(n) \qquad \Rightarrow \qquad A(z)X(z) = V(z)$$

If we regard $v(n)$ as the output, then



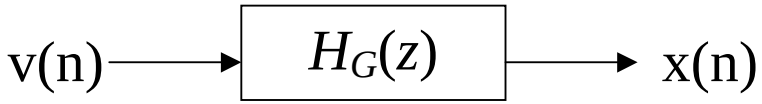
where $H_A(z) = \frac{V(z)}{X(z)} = A(z)$

[Notation note: figure uses $u(n)$ as input, i.e., $u(n) = v(n)$]



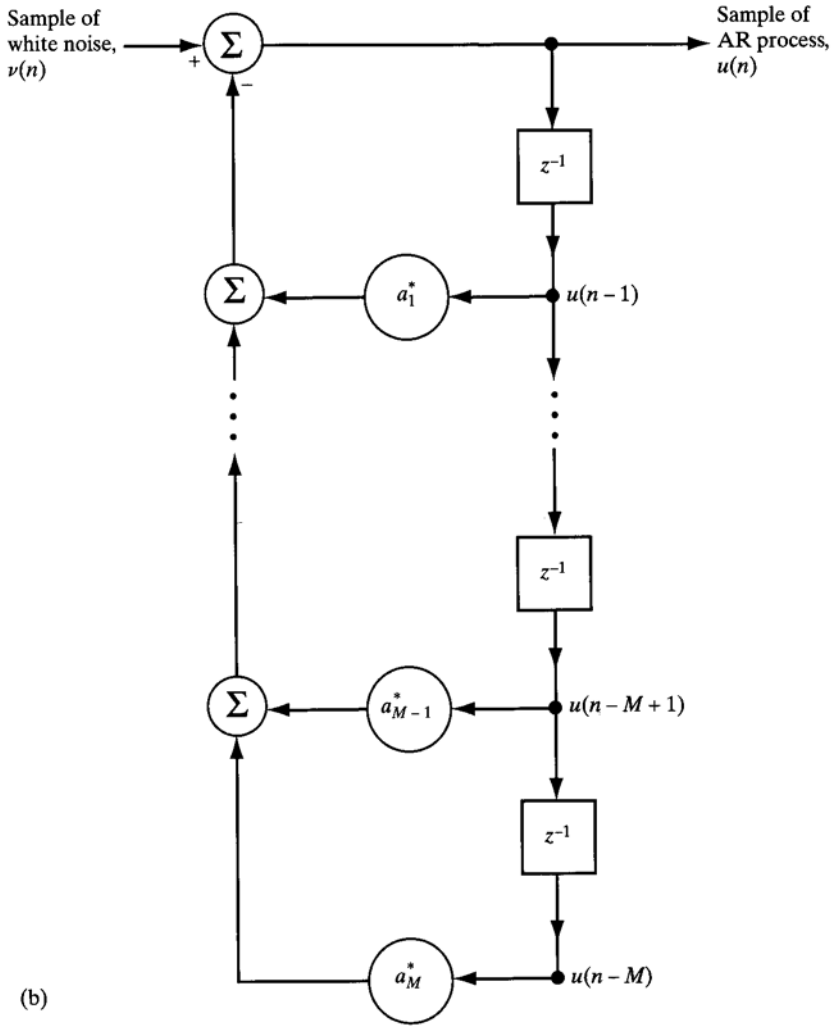
- This is called the process analyzer
- Analyzer is an all zero system
  - Impulse response is finite (FIR)
  - System is BIBO stable

If we view $v(n)$ as the input, then we have the process generator



$$H_G(z) = \frac{X(z)}{V(z)} = \frac{1}{A(z)}$$

- The process generator is an all pole system
  - Impulse response is infinite (IIR)
  - System stability is an issue



Note

$$H_G(z) = \frac{1}{A(z)} = \frac{1}{\sum_{n=0}^{M} a_n^* z^{-n}}$$

- Factor the denominator and represent $H_G(z)$ in terms of its poles

$$H_G(z) = \frac{1}{(1-p_1 z^{-1})(1-p_2 z^{-1})\cdots(1-p_M z^{-1})}$$

- $p_1, p_2, \ldots, p_M$ are the poles of $H_G(z)$ defined as the roots of the characteristic equation

$$1 + a_1^* z^{-1} + a_2^* z^{-2} + \cdots + a_M^* z^{-M} = 0$$

- $H_G(z)$ is all pole (IIR) and BIBO stable only if all poles are in the unit circle, i.e.,

$$|p_n| < 1 \quad n = 1, 2, \cdots, M$$

Moving Average Model

Definition (Moving Average)

The time series $\{x(n)\}$ is said to be generated by a Moving Average (MA) model if

$$x(n) = v(n) + b_1^* v(n-1) + \cdots + b_K^* v(n-K)$$

where $b_1, b_2, \cdots, b_k$ are the parameters of the order K MA model

- $v(n)$ is zero mean white Gaussian noise
- The process generation model is all zero (FIR)

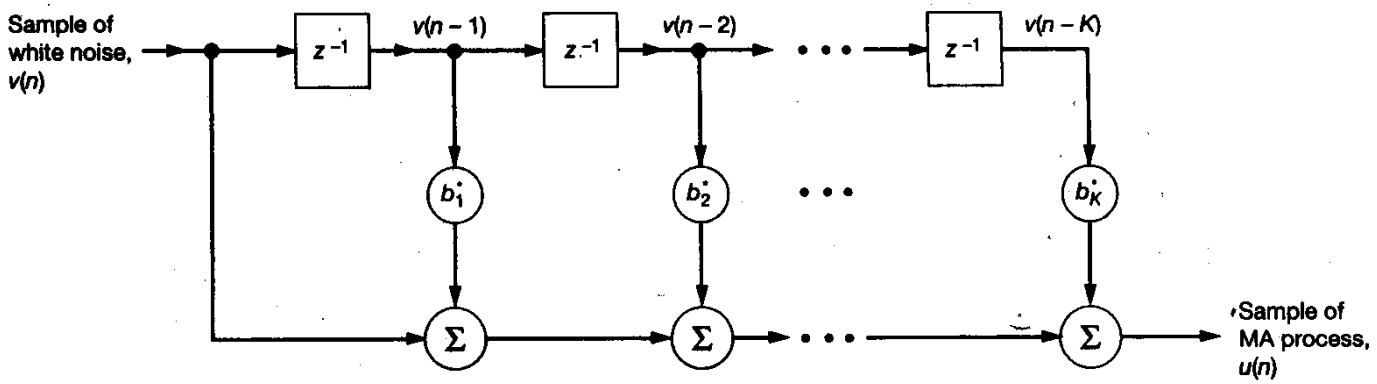


Figure 2.3 Moving average model (process generator).

Auto-Regressive Moving Average Model

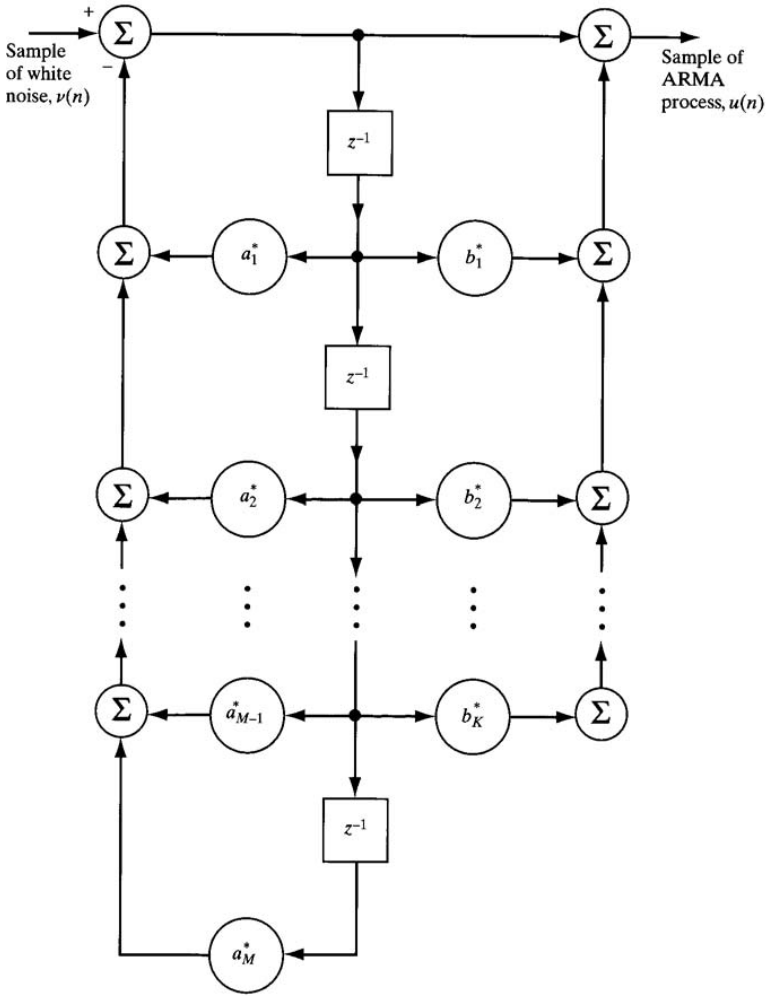
Definition (Auto-Regressive Moving Average)

In this case, $\{x(n)\}$ is a mixed process where the output is a function of past outputs and current/past inputs

$$x(n) + a_1^* x(n-1) + \cdots + a_M^*(n-M) = v(n) + b_1^* v(n-1) + \cdots + b_K^* v(n-K)$$

The order is (M, K).

- $v(n)$ is zero mean white Gaussian noise
- The process model has zeros and poles (IIR)



Wold Decomposition (after Herman Wold (1908–92))

Any WSS discrete time stochastic process $y(n)$ can be expressed as

$$y(n) = x(n) + s(n)$$

where:

- $x(n)$ and $s(n)$ are uncorrelated
- $x(n)$ can be expressed by the MA model

$$x(n) = \sum_{k=0}^{\infty} b_k^* v(n-k)$$

  - $b_0 = 1$ and $\sum_{k=0}^{\infty} |b_k| < \infty$
  - $v(n)$ is white noise uncorrelated with $s(n)$
- $s(n)$ is perfectly predictable

Note: If $B(z)$ is minimum phase, then it can be represented by an all pole (AR) system.

- AR models are widely used because they are tractable

Asymptotic statistics of AR processes

Recall that $\{x(n)\}$ is generated by

$$x(n) + a_1^* x(n-1) + a_2^* x(n-2) + \cdots + a_M^* x(n-M) = v(n)$$

or

$$x(n) = w_1^* x(n-1) + w_2^* x(n-2) + \cdots + w_M^* x(n-M) + v(n)$$

- Linear constant coefficient difference equation of order M driven by $v(n)$.
- Z-transform representation:

$$X(z) = \frac{V(z)}{1 + \sum_{k=1}^{M} a_k^* z^{-k}}$$

Inverse transforming $X(z) = \frac{V(z)}{1+\sum_{k=1}^{M} a_k^* z^{-k}}$ yields

$$x(n) = \underbrace{x_c(n)}_{\text{Homogeneous Solution}} + \underbrace{x_p(n)}_{\text{Particular Solution}}$$

- The particular solution is the result of driving $H_G(z)$ with $v(n)$

$$x_p(n) = H_G(z)v(n),$$

  where z^{-1} is taken as the delay operator.
- The particular solution has stationary statistics

The homogeneous solution is of the form

$$x_c(n) = B_1 p_1^n + B_2 p_2^n + \cdots + B_M p_M^n$$

where $p_1, p_2, \cdots, p_M$ are the roots of

$$1 + a_1^* z^{-1} + a_2^* z^{-2} + \cdots + a_M^* z^{-M} = 0$$

- The B values depend on the initial conditions
- The homogeneous solution is not stationary
- The process is asymptotically stationary if $|p_n| < 1$

Correlation of a stationary AR process

Recall that an AR process can be written as

$$\sum_{k=0}^{M} a_k^* x(n-k) = v(n)$$

where $a_0 = 1$.
Multiply both sides by $x^*(n-l)$ and take $E\{\ \}$.

$$E\left\{ \sum_{k=0}^{M} a_k^* x(n-k)x^*(n-l) \right\} = E\{v(n)x^*(n-l)\}$$

Note that

$$\begin{aligned} E\{x(n-k)x^*(n-l)\} &= r(l-k) \\ E\{v(n)x^*(n-l)\} &= 0 \quad \text{for } l > 0 \end{aligned}$$

Thus

$$E\left\{\sum_{k=0}^{M} a_k^* x(n-k)x^*(n-l)\right\} = E\{v(n)x^*(n-l)\}$$

$$\Rightarrow \sum_{k=0}^{M} a_k^* r(l-k) = 0 \quad \text{for } l>0$$

Accordingly, the auto-correlation of the AR process satisfies

$$r(l) = w_1^* r(l-1) + w_2^* r(l-2) + \cdots + w_M^* r(l-M)$$

where $w_k = -a_k$. Note that this also has the solution

$$r(m) = \sum_{k=1}^{M} c_k p_k^m$$

where p_k is the kth root of

$$1 - w_1^* z^{-1} - w_2^* z^{-2} - \cdots - w_M^* z^{-M} = 0$$

Why? Diff. equation (no driving function; homogeneous solution only)

Recall that the AR characteristic equation is

$$1 + a_1^* z^{-1} + a_2^* z^{-2} + \cdots + a_M^* z^{-M} = 0$$

This is identical to the auto-correlation characteristic equation

$$1 - w_1^* z^{-1} - w_2^* z^{-2} - \cdots - w_M^* z^{-M} = 0$$

$\Rightarrow$ the roots are equal
Result: A stable AR process $\Rightarrow |p_k| < 1$ and

$$\lim_{m\to\infty} r(m) = \lim_{m\to\infty} \sum_{k=1}^{M} c_k p_k^m = 0$$

(asymptotically uncorrelated)

Yule-Walker Equations

An AR model of order M is completely specified by

- AR coefficients: $a_1, a_2, \ldots, a_M$
- Variance of $v(n)$: σ_v^2

Proposition: These parameters can be determined by the auto-correlation values: $r(0), r(1), \ldots, r(M)$.

Recall

$$r(l) = w_1^* r(l-1) + w_2^* r(l-2) + \cdots + w_M^* r(l-M)$$

Case 1: Let $l = 1$

$$r(1) = w_1^* r(0) + w_2^* r(-1) + \cdots + w_M^* r(1-M)$$

Using the fact $r(-k) = r^*(k)$

$$r(1) = w_1^* r(0) + w_2^* r^*(1) + \cdots + w_M^* r^*(M-1)$$

Taking the complex conjugate

$$\begin{aligned} r^*(1) &= w_1 r(0) + w_2 r(1) + \cdots + w_M r(M-1) \\ &= \mathbf{w}^T [r(0), r(1), \cdots, r(M-1)]^T \end{aligned}$$

where $\mathbf{w}^T = [w_1, w_2, \cdots, w_M]$

Case 2: Now let $l = 2$

$$\begin{aligned} r(2) &= w_1^* r(1) + w_2^* r(0) + w_3^* r(-1) + \cdots + w_M^* r(2-M) \\ \Rightarrow r^*(2) &= w_1 r^*(1) + w_2 r(0) + w_3 r(1) \cdots + w_M r(M-2) \\ &= \mathbf{w}^T [r^*(1), r(0), r(1), \cdots, r(M-2)]^T \end{aligned}$$

Case 3: Similarly, for $l = 3$

$$\begin{aligned} r(3) &= w_1^* r(2) + w_2^* r(1) + w_3^* r(0) + w_4^* r(-1) \cdots + w_M^* r(3-M) \\ \Rightarrow r^*(3) &= w_1 r^*(2) + w_2 r^*(1) + w_3 r(0) + w_4 r(1) \cdots + w_M r(M-3) \\ &= \mathbf{w}^T [r^*(2), r^*(1), r(0), r(1), \cdots, r(M-3)]^T \end{aligned}$$

Repeating the process & combining results in matrix form

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r^*(1) & r(0) & \cdots & r(M-2) \\ r^*(2) & r^*(1) & \cdots & r(M-3) \\ \vdots & \vdots & \ddots & \vdots \\ r^*(M-1) & r^*(M-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_M \end{bmatrix} = \begin{bmatrix} r^*(1) \\ r^*(2) \\ r^*(3) \\ \vdots \\ r^*(M) \end{bmatrix}$$

or more compactly

$$\mathbf{Rw} = \mathbf{r} \qquad \text{where} \quad \mathbf{r} = [r(1), r(2), r(3), \ldots, r(M)]^H$$

Result: Given the auto-correlation values, the AR coefficients we can be uniquely determine

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{r}$$

where

$$a_k = -w_k \quad k = 1, 2, \cdots, M$$

Assumption: **R** is nonsingular

Still to be determined: variance of the driving sequence $v(n)$

Recall: $E\left\{\sum_{k=0}^{M} a_k^* x(n-k)x^*(n-l)\right\} = E\{v(n)x^*(n-l)\}$

$$\Rightarrow \sum_{k=0}^{M} a_k^* r(l-k) = E\{v(n)x^*(n-l)\} \qquad (*)$$

Note that

$$x^*(n) = [w_1 x^*(n-1) + w_2 x^*(n-2) + \cdots + w_M x^*(n-M) + v^*(n)] \qquad (**)$$

Let $l = 0$ in $(*)$ and use $(**)$ on the RHS

$$\begin{aligned}\sum_{k=0}^{M} a_k^* r(-k) &= E\{v(n)x^*(n)\} \\ &= E\{v(n)[w_1 x^*(n-1) + w_2 x^*(n-2) + \cdots \\ &\quad + w_M x^*(n-M) + v^*(n)]\}\end{aligned}$$

Note that

$$E\{v(n)w_k x^*(n-k)\} = 0, \quad k = 1, 2, \cdots, M$$

Thus $E\{v(n)w_k x^*(n-k)\} = 0, k = 1,2,\cdots,M$ gives

$$\begin{aligned}\sum_{k=0}^{M} a_k^* r(-k) &= E\{v(n)x^*(n)\} \\ &= E\{v(n)[w_1 x^*(n-1) + w_2 x^*(n-2) + \cdots \\ &\quad + w_M x^*(n-M) + v^*(n)]\} \\ &= E\{v(n)v^*(n)\}\end{aligned}$$

or conjugating

$$\sigma_v^2 = \sum_{k=0}^{M} a_k r(k)$$

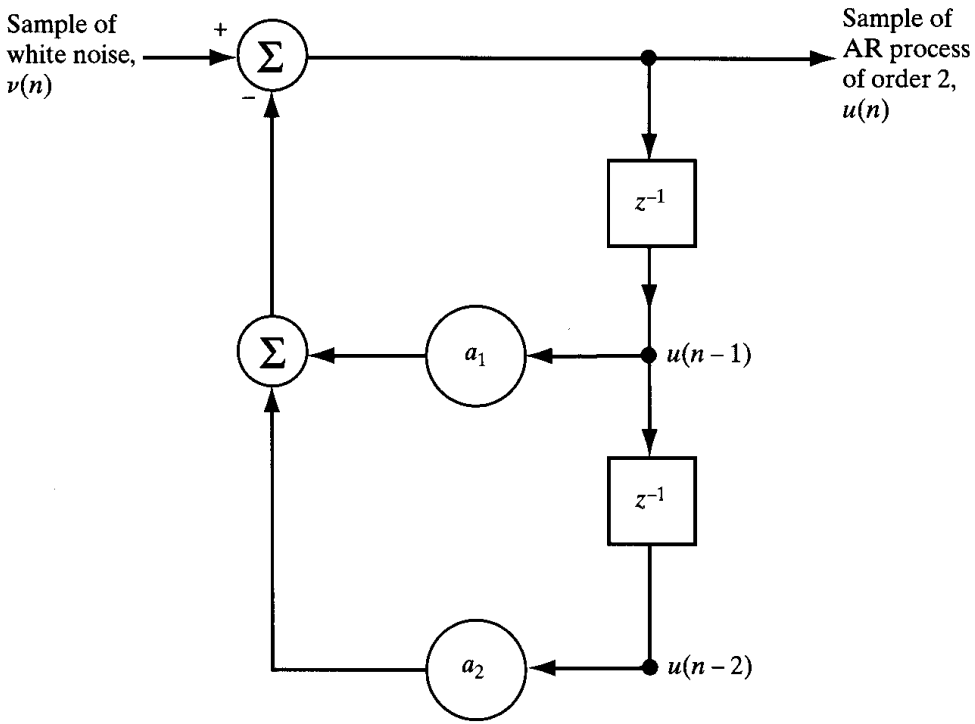
Final Yule–Walker Result

$$\mathbf{Rw} = \mathbf{r} \qquad \text{and} \qquad \sigma_v^2 = \sum_{k=0}^{M} a_k r(k)$$

Example (AR Order-2 Process)

Consider the process defined by

$$x(n) + a_1 x(n-1) + a_2 x(n-2) = v(n)$$



The process

$$x(n)+a_1x(n-1)+a_2x(n-2)=v(n)$$

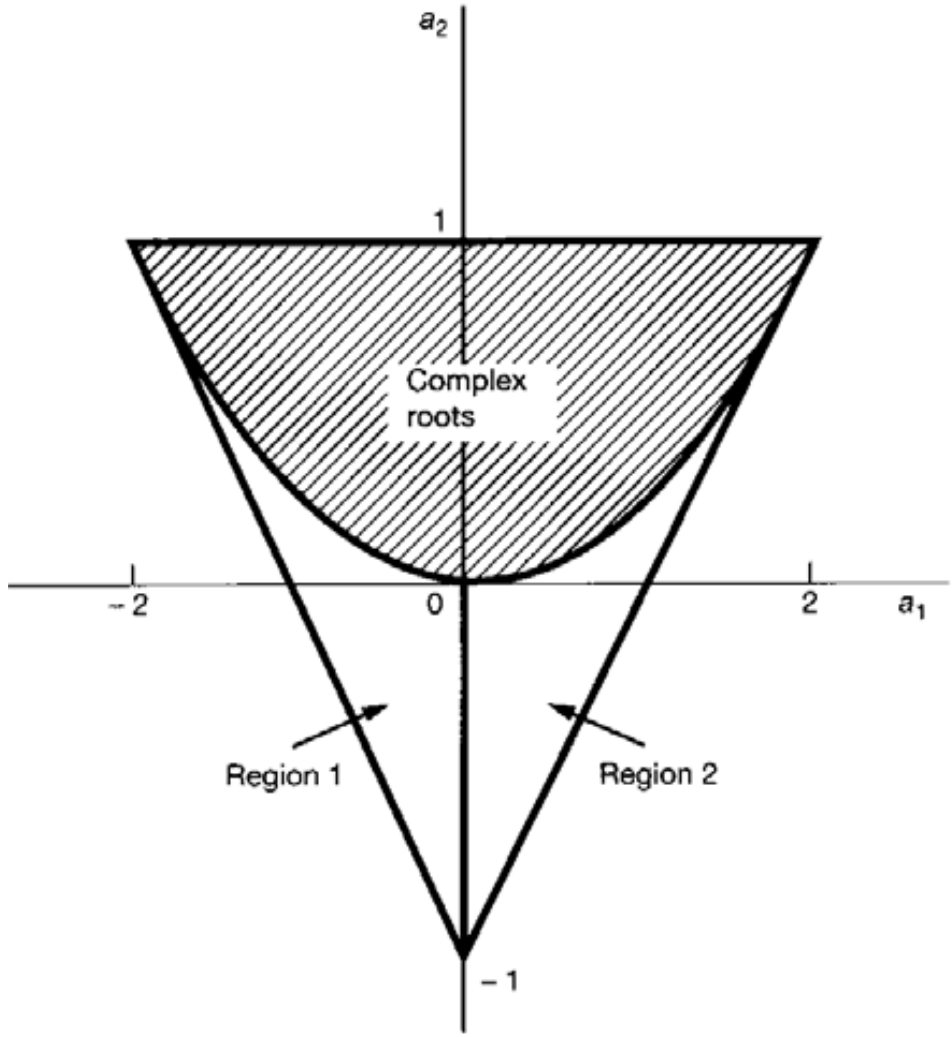
has characteristic equation

$$1+a_1z^{-1}+a_2z^{-2}=0$$

$$\Rightarrow p_1,p_2=\frac{-a_1\pm\sqrt{a_1^2-4a_2}}{2}$$

Stability enforces the constraints

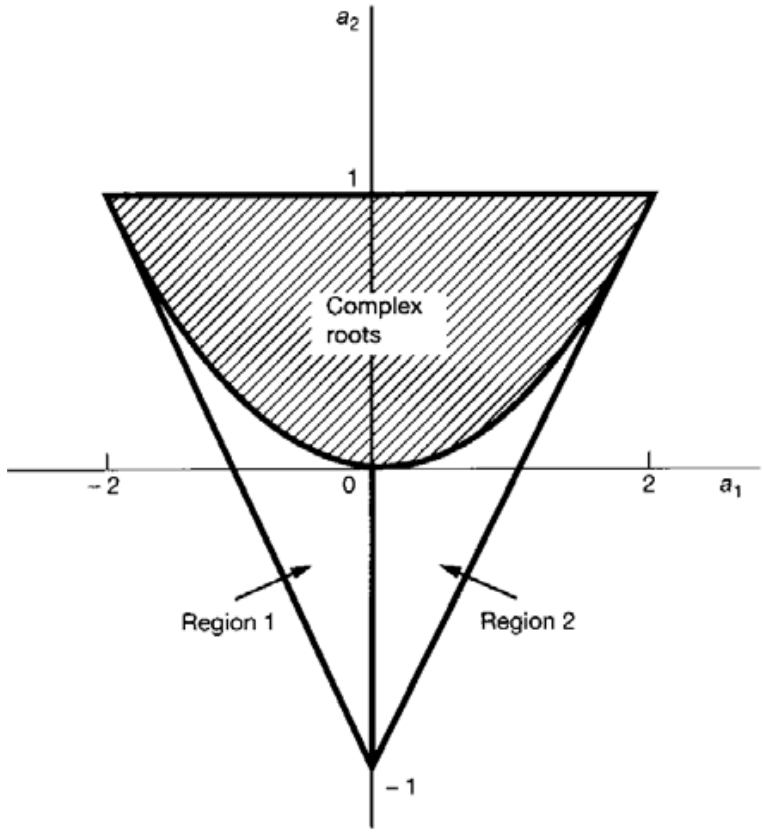
$$|p_k|<1\Rightarrow\begin{cases}-1\leq a_2+a_1\\-1\leq a_2-a_1\\-1\leq a_2\leq 1\end{cases}$$

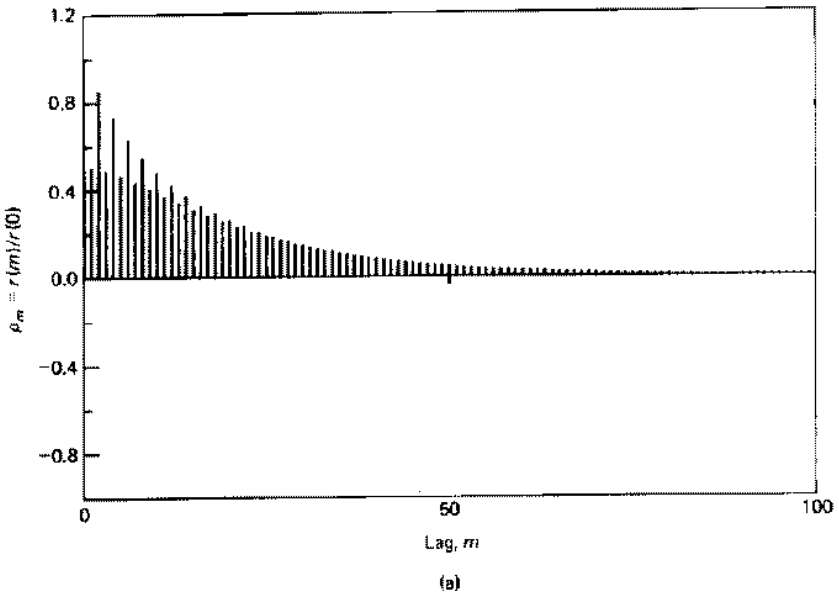


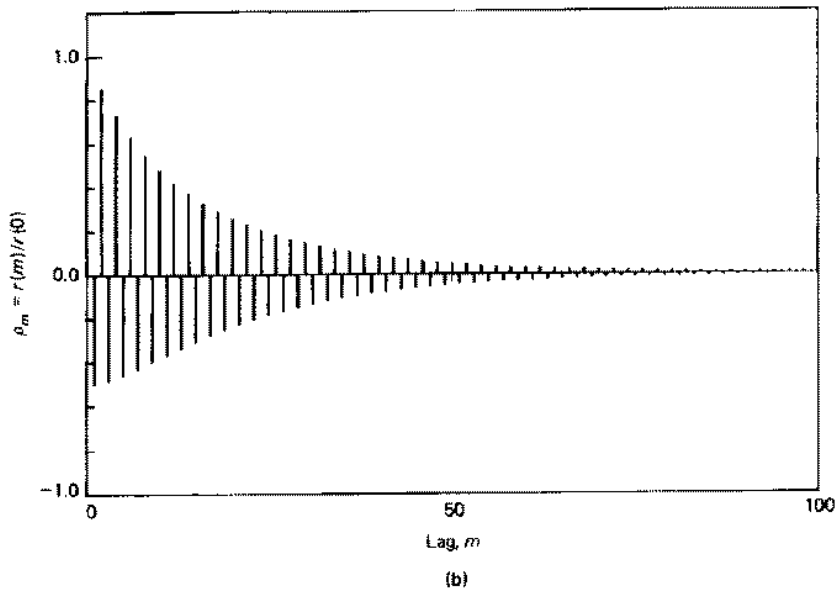
Recall that the auto-correlation can be expressed as

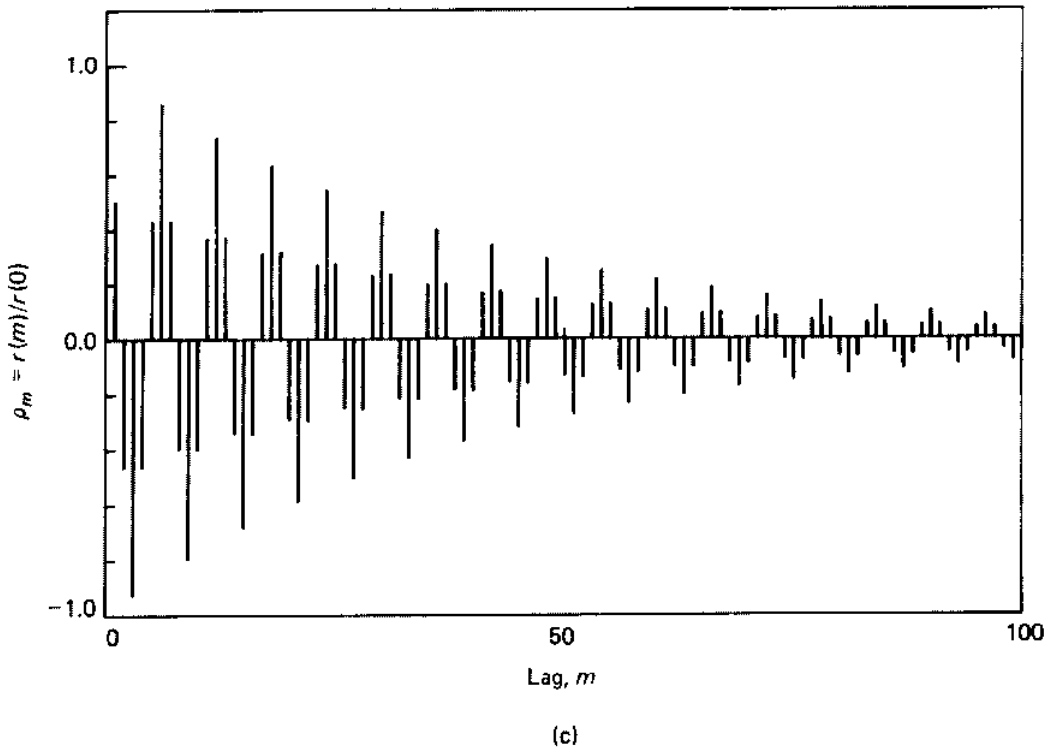
$$r(m) = \sum_{k=1}^{M} c_k p_k^m = c_1 p_1^m + c_2 p_2^m$$

- p_1, p_2 real, positive $\Rightarrow$ $r(m)$ positive decaying exponential
- p_1, p_2 real, negative $\Rightarrow$ $r(m)$ alternate sign decaying exponential
- p_1, p_2 complex conjugate $\Rightarrow$ $r(m)$ exponentially decaying sinusoid

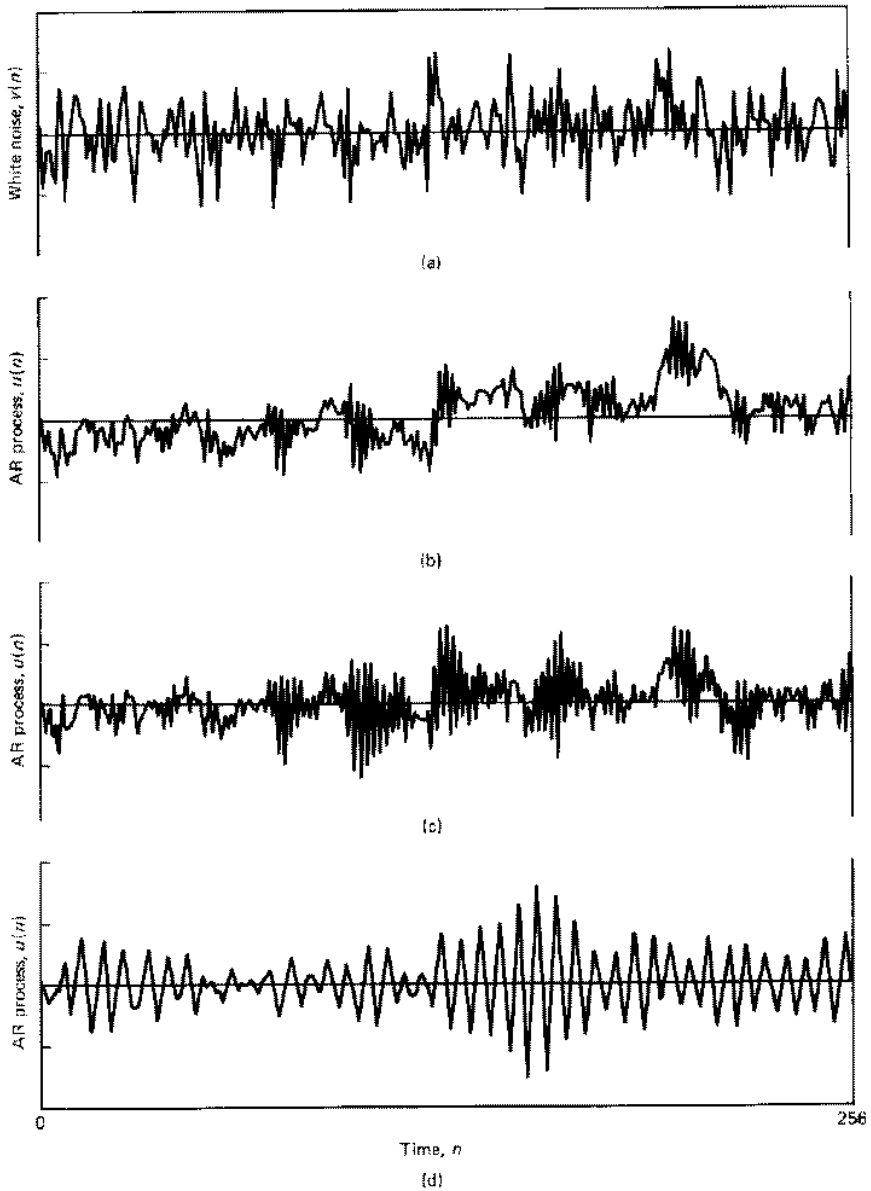



1.2
0.8
0.4
0.0
−0.4
−0.8
0
50
100
Lag, m
(a)


1.0
0.0
−1.0
0
50
100
Lag, m
(b)



The characteristics of the AR process vary in a related fashion to the pole placements.


White noise, v(n)
(a)
AR process, u(n)
(b)
AR process, u(n)
(c)
AR process, u(n)
0
256
Time, n
(d)

Model Order Selection

Model Order Selection

A model is typically estimated from a finite set of observation data.

- Result: Use Yule-Walker equations to estimate model parameters
- Open Question: How do we estimate the model order?
- Solution: Use information theoretic criteria.
  - Akaike's information criterion, developed by Hirotsugu Akaike under the name of "an information criterion" (AIC) in 1971

- Take $x_i = x(i), i = 1, 2, \cdots, N$ to be N observations of a stationary discrete time process.
- Let $\hat{\theta}$ be the estimated model (AR/MA/ARMA) order m parameters

$$\hat{\theta}_m = [\hat{\theta}_{1m}, \hat{\theta}_{2m}, \cdots, \hat{\theta}_{Mm}]^T$$

- Let $f_x(x_i|\hat{\theta}_m)$ be the conditional pdf of x_i given the estimated model defined by $\hat{\theta}_m$.
- Set $L(\hat{\theta}_m) = \max_{\hat{\theta}_m} \sum_{i=1}^{N} \ln f_x(x_i|\hat{\theta}_m)$.
  - The likelihood function (log of conditional pdf evaluated at the maximum likelihood estimates of the model parameters, $\hat{\theta}_m$).
- Then the AIC model order is given by m that minimizes

$$\text{AIC}(m) = \underbrace{-2L(\hat{\theta}_m)}_{\text{Always decreasing}} + \underbrace{2m}_{\text{Parameter cost function}}$$

- The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters.