Module 6.3
Optimal linear forecast:
Wiener's Approach

S. Lakshmivarahan

School of Computer Science
University of Oklahoma
Norman, OK, 73071
USA

- Let $w = (w_1, w_2, \ldots, w_n)^T \in \mathbb{R}^n$ be vector of correlated random variables with $\mu_i = \mathbf{E}[w_i]$ for $1 \leq i \leq n$ as the mean.
- Let $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T \in \mathbb{R}^n$ be the vector of the mean of w
- Let $\Gamma = [\Gamma_{ij}] \in \mathbb{R}^{n \times n}$ be the $n \times n$ covariance matrix of w where $\Gamma_{ij} = \mathbf{E}[(w_i - \mu_i)(w_j - \mu_j)]$, a Toeplitz matrix
- It is assumed that Γ is known and is positive definite (ie) Γ is a symmetric positive definite (SPD) matrix

Statement of the problem

- Let $y \in \mathbb{R}$ be an unknown random variable with known mean $\mu_y = \mathbf{E}[y]$ and finite variance: $\mathbf{Var}(y) = \mathbf{E}[y - \mu_y]^2 < \infty$
- The unknown y is correlated with each component of w with known (cross) correlation vector $c_{yw} \in \mathbb{R}^n$ with $c_{yw}(i) = \mathbf{E}[(y - \mu_y)(w_i - \mu_i)]$ for $1 \leq i \leq n$
- Problem: Knowing the second-order properties Γ and c_{yw} of w and y, find the best linear estimate of the unknown y given w

Estimation Problem

- Let
$$\hat{y} = a_0 + a_1 w_1 + a_2 w_2 + \cdots + a_n w_n \tag{1}$$
be the linear estimator of the unknown y in terms of the known w, where a_i's are to be determined
- Define the error e in the estimate as
$$e = y - \hat{y} \tag{2}$$
-
$$\text{MSE} = \mathbf{E}\left[y - \hat{y}\right]^2 \tag{3}$$
- The optimal estimate is obtained by selecting the coefficients a_i's such that MSE is a minimum

Minimization of MSE

- As a first step, set the $(n+1)$ derivatives of MSE with respect to $a_i, 0 \leq i \leq n$ to zero and solve for a_i's
- Setting $w_0 = 1$, we get

$$
\begin{aligned}
0 = \frac{\partial \ \text{MSE}}{\partial a_j} &= \frac{\partial}{\partial a_j} \mathbf{E}[y - \sum_{i=0}^{n} a_i w_i]^2 \\
&= \mathbf{E}\left[\frac{\partial}{\partial a_j} \left(y - \sum_{i=0}^{n} a_i w_i \right)^2 \right] \\
&= 2\mathbf{E}\left[\left(y - \sum_{i=0}^{n} a_i w_i \right) (-w_j) \right]
\end{aligned} \tag{4}
$$

Minimization of MSE

- Simplifying (4) becomes:

$$\mathbf{E}\left[yw_j\right] = \mathbf{E}\left[\sum_{i=0}^{n} a_i w_i w_j\right] = \sum_{i=0}^{n} a_i \mathbf{E}\left[w_i w_j\right] \tag{5}$$

- First set $j = 0$ in (5): Since $w_0 = 1$ we get

$$\mathbf{E}[y] = \mu_y = \sum_{i=0}^{n} a_i \mathbf{E}\left[w_i\right] = a_0 + \sum_{i=1}^{n} a_i \mu_i$$

- That is

$$a_0 = \mu_y - \sum_{i=1}^{n} a_i \mu_i \tag{6}$$

Minimization of MSE

- Recall

$$\begin{aligned}\Gamma_{ij} &= \mathbf{E}[(w_i - \mu_i)(w_j - \mu_j)] \\ &= \mathbf{E}[w_i w_j - w_i \mu_j - \mu_i w_j + \mu_i \mu_j] \\ &= \mathbf{E}[w_i w_j] - \mu_i \mu_j\end{aligned} \tag{7}$$

- Likewise:

$$\begin{aligned}c_{yw_j} &= \mathbf{E}[(y - \mu_y)(w_j - \mu_j)] \\ &= \mathbf{E}[yw_j] - \mu_y \mu_j\end{aligned} \tag{8}$$

Minimization of MSE

- Rewrite (5) for $1 \leq j \leq n$ as:

$$\mathbf{E}\left[yw_j\right] = a_0 \mathbf{E}\left[w_j\right] + \sum_{i=1}^{n} a_i \mathbf{E}\left[w_i w_j\right] \tag{9}$$

- Substituting (6),(7) and (8) in (9) and simplifying

$$c_{yw_j} + \mu_y \mu_j = (\mu_y - \sum_{i=1}^{n} a_i \mu_i)\mu_j + \sum_{i=1}^{n} a_i \left[\Gamma_{ij} + \mu_i \mu_j\right] \tag{10}$$

- That is

$$c_{yw_j} = \sum_{i=1}^{n} a_i \Gamma_{ij} \tag{11}$$

Optimal vector $a = (a_1, a_2, \ldots, a_n)$

- The n equations one for each j in (11) can be collectively written in the matrix form as (Γ is SPD)

$$\Gamma a = c_{yw} \tag{12}$$

-

$$a^* = \Gamma^{-1} c_{yw} \tag{13}$$

Observation 1 - a Centered Formulation

- Substituting (6) in (1) :

$$\hat{y} = \mu_y + \sum_{i=1}^{n} a_i(w_i - \mu_i) \tag{14}$$

- Hence $\mathbf{E}(\hat{y}) = \mu_y$ and $\hat{y}$ is unbiased
- Defining $\overline{y} = y - \mu_y$ and $\overline{w}_i = w_i - \mu_i$, we could have started with a linear estimator as

$$\hat{\overline{y}} = \sum_{i=0}^{n} a_i \overline{w}_i \tag{15}$$

- Where we express the estimate of the centered y using the centered w_i. Then $\hat{y} = \mu_y + \hat{\overline{y}}$
- Verify that a is given by: $\Gamma a = c_{yw}$

Observation 2 - Minimum Value of MSE

- Recall, using (6), with $w_0 = 1$

$$\begin{aligned} \sum_{i=0}^{n} a_i w_i &= a_0 + \sum_{i=1}^{n} a_i w_i \\ &= \mu_y + \sum_{i=1}^{n} a_i (w_i - \mu_i) \\ &= \mu_y + a^T \overline{w}, \end{aligned} \tag{16}$$

where $a = (a_1, a_2, \ldots, a_n)^T$ and $\overline{w} = (\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_n)^T$

- Hence

$$
\begin{aligned}
\text{MSE} &= \mathbf{E}\left[y - \sum_{i=0}^{n} a_i w_i\right]^2 = \mathbf{E}\left[(y - \mu_y) - a^T \overline{w}\right]^2 \\
&= \mathbf{E}\left[\overline{y} - a^T \overline{w}\right]^2 \quad \text{(transpose of a scalar is a scalar)} \\
&= \mathbf{E}\left[(\overline{y} - a^T \overline{w})^T (\overline{y} - a^T \overline{w})\right] \\
&= \mathbf{E}\left[(\overline{y}^2 - \overline{y} a^T \overline{w} - (a^T \overline{w})^T \overline{y} + (a^T \overline{w})^T (a^T \overline{w})\right]
\end{aligned}
\tag{17}
$$

Facts from matrix algebra

$$
\left.\begin{aligned}
&\overline{y} a^T \overline{w} = \overline{y}\, \overline{w}^T a = \left(\overline{y}\, \overline{w}^T\right) a = a^T \left(\overline{y}\, \overline{w}\right) \\
&\left(a^T \overline{w}\right) \overline{y} = \left(\overline{w}^T a\right) \overline{y} = a^T \left(\overline{w}^T\, \overline{y}\right) = a^T \left(\overline{y}\, \overline{w}\right) \\
&\left(a^T \overline{w}\right)^T \left(a^T \overline{w}\right) = \left(\overline{w}^T a\right)^T \left(\overline{w}^T a\right) = a^T \left(\overline{w}\, \overline{w}^T\right) a
\end{aligned}\right\} \qquad (18)
$$

- Using (18) in (17) and simplifying:

$$\begin{aligned} \text{MSE} &= \mathbf{E}\left[\overline{y}\right]^2 - 2a^T\mathbf{E}\left[\overline{y}\,\overline{w}\right] + a^T\mathbf{E}\left[\overline{w}\,\overline{w}^T\right]a \\ &= \mathbf{Var}\left[\overline{y}\right] - 2a^Tc_{yw} + a^T\Gamma a \end{aligned} \tag{19}$$

- From (13), optimal

$$a^* = \Gamma^{-1}c_{yw} \tag{20}$$

Minimum MSE

- Substituting (20) in (19) and simplifying:

$$
\begin{aligned}
\min_{a}(\text{MSE}) &= \mathbf{Var}\,[y] - 2c_{yw}^{T}\Gamma^{-1}c_{yw} + \left(c_{yw}^{T}\Gamma^{-1}\right)\Gamma\left(\Gamma^{-1}c_{yw}\right) \\
&= \mathbf{Var}\,[y] - 2c_{yw}^{T}\Gamma^{-1}c_{yw} + c_{yw}^{T}\Gamma^{-1}c_{yw} \\
&= \mathbf{Var}\,[y] - c_{yw}^{T}\Gamma^{-1}c_{yw} \\
&< \mathbf{Var}\,[y] \ \text{ since } \ \Gamma \text{ and } \ \Gamma^{-1} \text{ are SPD}
\end{aligned}
\tag{21}
$$

Observation 3 - orthogonal projection

- From (4):

$$\mathbf{E}\left[\left(y - \sum_{i=0}^{n} a_i w_i\right) w_j\right] = 0$$

- That is, error $e = y - \hat{y} = (y - \sum_{i=0}^{n} a_i w_i)$ is orthogonal to each w_j. Stated in other words, the prediction error is orthogonal to the linear space generated by $\{1, w_1, w_2, \ldots, w_n\}$

Example

- Let $y = x^2 + v$, where $x \sim N(0,1), V \sim N(0,1)$ and x and v are independent.
- v is the observation noise in measuring x^2
- Consider a linear estimator $\hat{y} = ax + b$
- Clearly, the error $e = y - (ax + b) \perp x$ and $e \perp 1$

Example

- That is

$$0 = \mathbf{E}\left[\left(y - (ax + b)\right)x\right] = \mathbf{E}\left[yx - ax^2 + bx\right]$$
$$\therefore \mathbf{E}\left[yx\right] = a\mathbf{E}([x^2] + b\mathbf{E}[x] = a$$

- But

$$y = x^2 + v \Rightarrow \mathbf{E}\left[yx\right] = \mathbf{E}[x^3] + \mathbf{E}[xv] = 0 \quad \text{(Why?)}$$

- Hence $a = 0$ (since odd moments of x are zero)

Example

- Again

$$0 = \mathbf{E}\left[(y - (ax + b)) \cdot 1\right] = \mathbf{E}\left[y\right] - a\mathbf{E}\left[x\right] - b$$

- Therefore

$$\mathbf{E}\left[y\right] = a\mathbf{E}\left[x\right] + b = b$$

- But $\mathbf{E}\left[y\right] = \mathbf{E}\left[x^2 + v\right] = \mathbf{E}\left[x^2\right] + \mathbf{E}\left[v\right] = 1$ (Why?)
- Hence, $b = 1$

Example

- The best linear estimation $\hat{y} = 1$

$$\begin{aligned}\text{MSE: } \mathbf{E}[y-1]^2 &= \mathbf{E}\left[x^2 + v - 1\right]^2 \\ &= \mathbf{E}\left[x^4\right] + \mathbf{E}\left[v^2\right] \\ &+ 1 + 2\mathbf{E}\left[x^2\right]\mathbf{E}[v] - 2\mathbf{E}[v]\cdot 1 - 2\mathbf{E}\left[x^2\right]\cdot 1\end{aligned}$$

- Recall $\mathbf{E}\left[x^4\right] = 3, \mathbf{E}\left[v^2\right] = 1, \mathbf{E}\left[x^2\right] = 1, \mathbf{E}[v] = 0$
- Hence, MSE $= \mathbf{E}[y-1] = 3 + 1 + 1 - 2 = 3$
- This MSE is much larger than the MSE using conditional expectation as the predictor (Refer to Module 6.2, Slide 12)

Application to stationary Time Series prediction:

- Let $\{x_k\}$ be a second-order stationary time series
- By identifying w with $x(1:n) = (x_1, x_2, \ldots, x_n)$ a vector with the first-n members of the time series and y with x_{n+1} given $x(1:n) = (x_1, x_2, \ldots, x_n)$
- Without loss of generality, we assume that the $\{x_k\}$ series has been centered and has mean zero

Required covariance matrices and vectors

- Identifying $w = x(1:n)$ or $w_i = x_i$ for $1 \le i \le n$, it is immediate that (since $x_i{}'s$ are centered) we have $\Gamma_n = [\Gamma_{ij}] \in \mathbb{R}^{n \times n}$, where

$$\Gamma_{ij} = \mathbf{E}\left[x_i x_j\right] = \gamma(|j-i|) \tag{22}$$

  where $\gamma(k)$ is the ACF for the given series
- Similarly

$$c_{yw} = c_{x_{n+1},x(1:n)} \in \mathbb{R}^n, \tag{23}$$

  where i^{th} element is

$$c_{x_{n+1},x(1:n)}(i) = \mathbf{E}\left[x_{n+1} x_i\right] = \gamma(n+1-i),\ 1 \le i \le n \tag{24}$$

- Henceforth

$$c_{x_{n+1},x(1:n)} = \gamma(n:1) = (\gamma(n), \gamma(n-1), \ldots, \gamma(1))^T \tag{25}$$

Structure of the Optimal Estimate

- Following the above development,

$$\hat{x}_{n+1} = a^T x(1:n) = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \qquad (26)$$

where $a \in \mathbb{R}^n$ is the solution of

$$\Gamma_n a = \gamma(n:1) \text{ or } a^* = \Gamma^{-1}\gamma(n:1)$$

Example: $n = 3$

- Given $\{x_1, x_2, x_3\}$, find the best linear predictor for x_4
- Following (26), $\Gamma \in \mathbb{R}^3$ and $\gamma(3:1) \in \mathbb{R}^3$ are given by

$$\Gamma_3 = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) \\ \gamma(1) & \gamma(0) & \gamma(1) \\ \gamma(2) & \gamma(1) & \gamma(0) \end{bmatrix} \text{ and } \gamma(3:1) = \begin{bmatrix} \gamma(3) \\ \gamma(2) \\ \gamma(1) \end{bmatrix}$$

- Equation

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) \\ \gamma(1) & \gamma(0) & \gamma(1) \\ \gamma(2) & \gamma(1) & \gamma(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \gamma(3) \\ \gamma(2) \\ \gamma(1) \end{bmatrix} \tag{27}$$

is called Yule-Walker equation.

Optimal Linear Prediction of x_{n+s} Given $\{x_1, x_2, \ldots, x_n\}$ for $s \geq 1$

- In this case $\Gamma \in \mathbb{R}^{n\times n}$ is the same as above
- By y is x_{n+s}. Hence

$$c_{yw} = c_{x_{n+s},x(1:n)} \in \mathbb{R}^n \tag{28}$$

where i^{th} element

$$c_{x_{n+s},x_{(1:n)}} = \mathbf{E}\left[x_{n+s}x_i\right] = \gamma(n+s-1) \tag{29}$$

- Hence for the

$$c_{x_{n+s},x(1:n)} = \gamma(n+s-1:s) \in \mathbb{R}^n \tag{30}$$

Example: $n = 3$ and $s = 3$

- Given $x(1:3) = (x_1, x_2, x_3)^T$, for structure of the optimal predictor for x_6 is given by

$$\hat{x}_6 = a_1 x_1 + a_2 x_2 + a_3 x_3$$

where a is the solution of

$$\Gamma_3 a = \gamma(5:3) \tag{31}$$

- $\Gamma_3 \in \mathbb{R}^{3\times 3}$ is as given in (27) and

$$\gamma(5:3) = (\gamma(5), \gamma(4), \gamma(3))^T \tag{32}$$

Additional Fact

- Let $\{x_k\}$ be a zero mean second-order stationary time series
- Consider a vector $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ of the first three elements of the series
- Let $y = (x_3, x_2, x_1)^T \in \mathbb{R}^3$ be the reversal of the vector x
- Since $\mathbf{E}(x_i x_j) = \gamma(|j - i|)$ it follows that

$$\mathbf{E}[xx^t] = \mathbf{E}\begin{bmatrix} x_1^2 & x_1x_2 & x_1x_3 \\ x_2x_1 & x_2^2 & x_2x_3 \\ x_3x_1 & x_3x_2 & x_3^2 \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_2 \end{bmatrix} = \mathbf{E}[yy^T] \tag{33}$$

where $\gamma_i = \gamma(i)$ for simplicity of notation, that is x and y share the same covariance matrix

Two Equivalent Formulations

- Let $x(1:n) = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ and $y(1:n) = (x_n, x_{n-1}, \ldots, x_2, x_1)^T \in \mathbb{R}^n$, the reversal of x
- Let $a = (a_1, a_2, \ldots, a_n)^T \in \mathbb{R}^n$ and $b = (b_1, b_2, \ldots, b_n)^T \in \mathbb{R}^n$
- Then we can express $\hat{x}_{n+s}$, for $s \geq 1$ in two ways:

$$\left.\begin{aligned} \hat{x}_{n+s} &= a^T x(1:n) = \textstyle\sum_{i=1}^n a_i x_i \\ &= b^T y(1:n) = \textstyle\sum_{i=1}^n b_i x_{n-i+1} \end{aligned}\right\} \tag{34}$$

Relation between a and b in (34)

- By (33), recall that x_n and y_n have the same covariance matrix $\Gamma_n = [\Gamma_{ij}]$, where $\Gamma_{ij} = \gamma(|i-j|)$
- Also

$$\left.\begin{aligned} c_{x_{n+s},x(1:n)} &= \gamma(n+s-1:s) \\ c_{x_{n+s},y(1:n)} &= \gamma(s:n-i+1) \end{aligned}\right\} \tag{35}$$

- Hence the solutions are related by

$$a_i = b_{n-i+1}, \text{ for } 1 \leq i \leq n \tag{36}$$

(ie) b is a reversal of a where $\gamma(s:n+s-1)$ is a reversal of $\gamma(n+s-1:s)$

- Given $\{x_k\}$, first compute $\gamma(k)$, the ACF
- Then we can form the matrix Γ and the vector c for a given n and s
- Solving $\Gamma a = c$ is the time consuming part
- We will discuss two types of recursive algorithm for solving $\Gamma a = c$ by exploiting the structure Γ in the subsequent modules

Homework

- Let

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \text{ and } \overline{f} = \begin{bmatrix} f_2 \\ f_1 \end{bmatrix}$$

Solve explicitly $Ax = f$ and $Ay = \overline{f}$ where $x = (x_1, x_2)^T$ and $y = (y_1, y_2)^T$. Verify $x_1 = y_2$ and $x_2 = y_1$

References

- Since Γ is a Toeplitz matrix, the linear system $\Gamma a = \gamma$ can be solved by a special class of methods called Durbin-Levinson algorithms requiring $\mathbb{O}(n^2)$ operations.
- Refer to Chapter 4 (Section 4.7) in the following book for details:

Golub, G.H., & Van Load, C.F. (1989). *Matrix Computations* Johns Hopkins University Press (Second Edition)