Lecture 24: LINEAR MMSE ESTIMATION THEORY

05 April 2016

In the previous lecture, we dealt with *Kalman-Bucy* filter which is the recursive/sequential algorithm to output the optimal MMSE state of a linear dynamical system with Gaussian statistics. Recursive relations for the above algorithm were derived. In this lecture, we will discuss the general theory of linear estimation which reduces computational complexity. Orthogonality principle, it's alternative version will be proved and *Levinson-Durbin* filter will be introduced.

1 Linear MMSE Estimation Theory

Suppose we have two real valued discrete time random processes $\{X_n\}_{n=0}^{\infty}$ & $\{Y_n\}_{n=0}^{\infty}$ and we want to estimate X_t using the observations $(Y_0, Y_1, ..., Y_s)$ to minimize mean square error(MSE) i.e., $\mathbb{E}[(X_t - \hat{X}_t)^2]$. The optimum estimator in the MMSE sense is the conditional mean of X_t

$$\hat{X}_t = \mathbb{E}[X_t | Y_1, Y_2, ..., Y_s]. \tag{1}$$

However, if the number of observations are large then computation of conditional mean is quite cumbersome, unless the problem exhibits special structure (as in *Kalman-Bucy* model).

Computational complexity of these problems can be reduced by constrained estimators. One such class of constraints is linear constraint, where the estimate is linear function of $(Y_0, Y_1, ..., Y_s)$, i.e., functions of the form

$$\hat{X}_t = \sum_{n=1}^{s} h_n Y_n + h_0. \tag{2}$$

Let $\mathcal{H}^s$ denote the set of all linear estimators, such that

$$\{\hat{X}_t : \hat{X}_t = \sum_{n=1}^{s} h_n Y_n + h_0\}. \tag{3}$$

Consider the best linear estimator problem

$$\text{minimize } \mathbb{E}[(X_t - \hat{X}_t)^2] \;\; \text{s.t} \;\; \hat{X}_t \in \mathcal{H}^s. \tag{4}$$

Theorem 1.1. *(Orthogonality Principle)*
Consider the best linear estimation problem as in eqn. (4). $\hat{X}_t \in \mathcal{H}^s$ *solves the problem if and only if* $\mathbb{E}[(\hat{X}_t - X_t)Z] = 0, \;\; \forall \;\; Z \in \mathcal{H}^s$.

Proof. Assume that $\mathbb{E}[(\hat{X}_t - X_t)Z] = 0 \; \forall Z \in \mathcal{H}^s$. Let $\tilde{X}_t \in \mathcal{H}^s$ be any linear estimator.

$$\begin{aligned}
\mathbb{E}[(\tilde{X}_t - X_t)^2] &= \mathbb{E}[(\tilde{X}_t - \hat{X}_t + \hat{X}_t - X_t)^2], \\
&= \mathbb{E}[(\tilde{X}_t - \hat{X}_t)^2] + \mathbb{E}[(\hat{X}_t - X_t)^2] + 2\mathbb{E}[(\tilde{X}_t - \hat{X}_t)(\hat{X}_t - X_t)], \\
&= \underbrace{\mathbb{E}[(\tilde{X}_t - \hat{X}_t)^2]}_{\geqslant 0} + \underbrace{\mathbb{E}[(\hat{X}_t - X_t)^2]}_{\geqslant 0} + \underbrace{2\mathbb{E}[(\tilde{X}_t - \hat{X}_t)(\hat{X}_t - X_t)]}_{=0}, \\
&\geq \mathbb{E}[(\hat{X}_t - X_t)^2].
\end{aligned} \tag{5}$$

The third term in the equation (5) is zero since $\mathbb{E}[(\hat{X}_t - X_t)Z] = 0$, where $Z = \tilde{X}_t - \hat{X}_t$ (difference of two linear estimators) is also a linear estimator. Assume that $\tilde{X}_t \in \mathcal{H}^s$, $Z \in \mathcal{H}^s$ such that $\mathbb{E}[(\tilde{X}_t - X_t)Z] \neq 0$. We will construct a better estimator than $\tilde{X}_t$. Define

$$\hat{X}_t = \tilde{X}_t - \left[\frac{\mathbb{E}[(\tilde{X}_t - X_t)Z]Z}{\mathbb{E}[Z^2]}\right] = \tilde{X}_t - \left[\frac{aZ}{b}\right],$$

where $a = \mathbb{E}[(\tilde{X}_t - X_t)Z]$ and $b = \mathbb{E}[Z^2]$.

$$\begin{aligned}
\mathbb{E}[(\tilde{X}_t - X_t)^2] &= \mathbb{E}[(\tilde{X}_t - \frac{aZ}{b} - X_t)^2], \\
&= \mathbb{E}[(\tilde{X}_t - X_t)^2] + \frac{a^2}{b^2}\mathbb{E}[Z^2] - 2\frac{a}{b}\mathbb{E}[(\tilde{X}_t - X_t)Z)], \\
&= \mathbb{E}[(\tilde{X}_t - X_t)^2] + \frac{a^2}{b} - \frac{2a^2}{b}, \\
&= \mathbb{E}[(\tilde{X}_t - X_t)^2] - \frac{a^2}{b}, \\
&\leq \mathbb{E}[(\tilde{X}_t - X_t)^2].
\end{aligned}$$

Hence orthogonality principle is proved. □

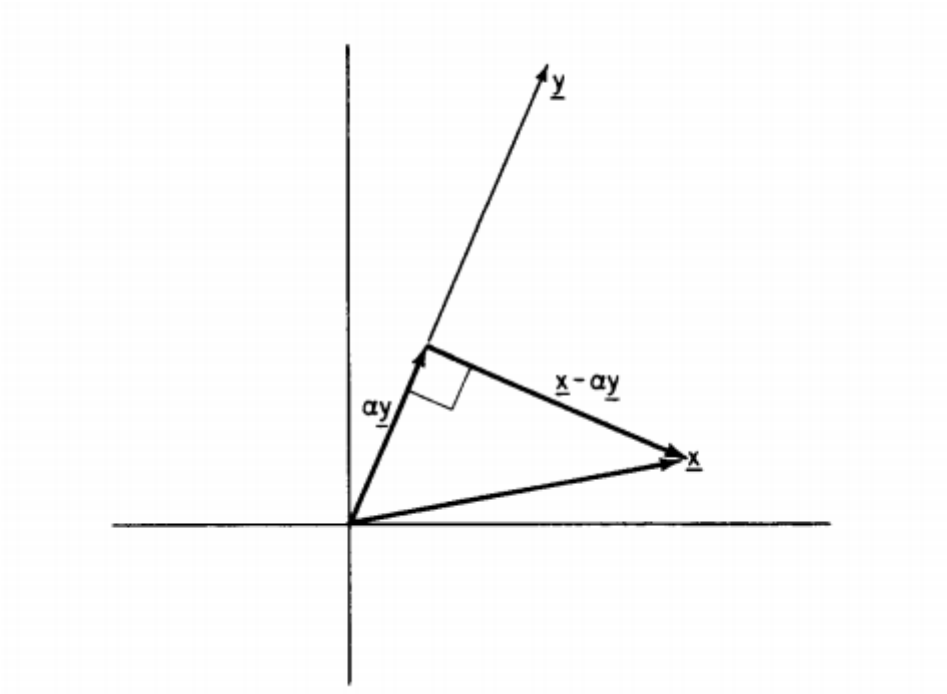

Figure 1: Illustration of Orthogonality Principle for 1-dimension

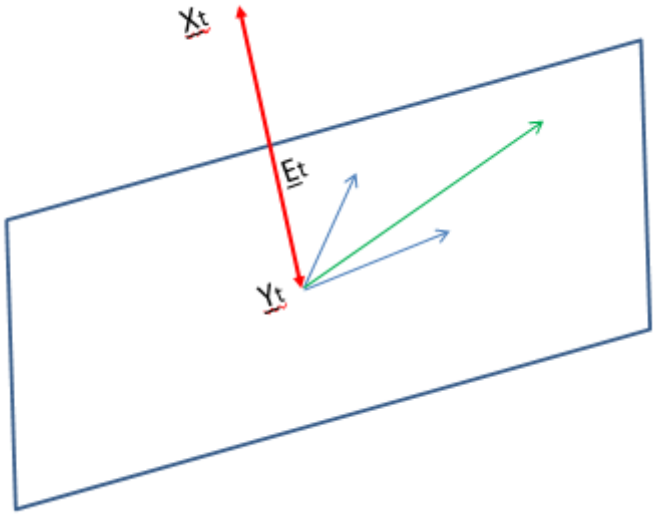

Figure 2: Illustration of Orthogonality Principle for 2-dimensions

Suppose that $\underline{x}$ and $\underline{y}$ are two vectors of same dimension, and suppose that we would like to approximate $\underline{x}$ by a constant, say α, times $\underline{y}$ such that length of error vector $\underline{x} - \alpha\underline{y}$ is as small as possible. It is is easy to see that α minimizes this length if and only if error vector is perpendicular to the line that is along $\underline{y}$ (see Fig. 1). In Fig. 2, the plane shows the subspace of $\mathcal{H}^s$, $\underline{Y}_t$ is the LMMSE estimate of $\underline{X}_t$, since error vector $\underline{E}_t = (\underline{Y}_t - \underline{X}_t)$ is perpendicular to the plane.

Theorem 1.2. *Another Orthogonality Principle:* $\hat{X}_t \in \mathcal{H}^s$ *solves eqn.* (4) *iff* $\mathbb{E}[\hat{X}_t] = \mathbb{E}[X_t]$ *and* $\mathbb{E}[(\hat{X}_t - X_t)Y_l] = 0$, $\forall\, l = 1, 2, 3, ..., s$.

This is an alternative orthogonality condition.

1.1 Explicit Solution for the Optimal LMMSE Estimator

Assume that the best linear estimator of X_t given $(Y_1, ..., Y_s)$ is

$$\hat{X}_t = \sum_{n=1}^{s} h_n Y_n + h_0. \tag{6}$$

From theorem 1.2, we have

$$\mathbb{E}[\hat{X}_t] = \sum_{n=1}^{s} h_n \mathbb{E}[Y_n] + h_0 = \mathbb{E}[X_t]. \tag{7}$$

This gives,

$$h_0 = \mathbb{E}[X_t] - \sum_{n=1}^{s} h_n \mathbb{E}[Y_n]. \tag{8}$$

Also, for all $l = 1, \ldots, s$,

$$\mathbb{E}\left\{ \left(X_t - \underbrace{\left[\sum_{n=1}^{s} h_n Y_n + h_0 \right]}_{\hat{X}_t} \right) Y_l \right\} = 0 \tag{9}$$

Substituting eqn. (8) in eqn. (9), we get

$$\mathbb{E}\left\{ \left((X_t - \mathbb{E}[X_t]) - \sum_{n=1}^{s} h_n (Y_n - \mathbb{E}\{Y_n\}) \right) Y_l \right\} = 0, \tag{10}$$

$$\mathbb{E}[(X_t - X_t)Y_l] = \sum_{n=1}^{s} h_n \mathbb{E}\left[(Y_n - \mathbb{E}\{Y_n\})\, Y_l\right],$$

$$\mathrm{Cov}(X_t, Y_l) = \sum_{n=1}^{s} h_n \mathrm{Cov}(Y_n, Y_l),\ \forall\ 1 \leq l \leq s,$$

$$C_{XY}(t,l) = \sum_{n=1}^{s} h_n C_Y(n,l),\ \forall\ 1 \leq l \leq s. \tag{11}$$

the above equations are called *Yule-Walker Equations (or) Weiner-Hopf Equations*, where $C_{XY}(t,l) \triangleq \mathrm{Cov}(X_t, Y_l)$ is the cross covariance function of $\{X_n\}_{n=0}^{\infty}$ and $\{Y_n\}_{n=0}^{\infty}$ and $C_Y(n,l) \triangleq \mathrm{Cov}(Y_n, Y_l)$ is the auto covariance function of the sequence $\{Y_n\}_{n=0}^{\infty}$.

Note 1. First and second order statistics of X and Y completely determine the optimal Linear Estimator.

In Matrix form,

$$\underline{\sigma}_{XY} = \Sigma_Y \underline{h},$$

where

$$\begin{aligned} \underline{\sigma}_{XY} &\triangleq [C_{XY}(t,l), ..., C_{XY}(t,s)]^T, \\ \underline{h} &\triangleq [h_{t,1}, ..., h_{t,s}]^T. \end{aligned}$$

If Σ_Y is positive definite, then the optimum estimator coefficients are given by

$$\underline{h} = \Sigma_Y^{-1} \underline{\sigma}_{XY}.$$

Note 2. Inverting Σ_Y of size $s \times s$ is expensive, if s is large (in general, time required for inversion of Σ_Y is $O(s^3)$. In our case s is the number of observations, which grows linearly with time for many signal estimation applications. So, the computation of optimum coefficients from above equation cannot be accomplished in real time.

However, with more structure in the problem, one can hope to solve it faster. The following models provide more efficient computation of these coefficients.

1. Kalman filter for linear dynamical systems.
2. Levinson-Durbin filter for wide sense stationary processes (WSSP).

2 Levinson-Durbin Filter

Suppose $\{Y_n\}_{n=0}^{\infty}$ is a wide sense stationary random process with zero mean and auto-covariance function $C_Y(n,l) = C_Y(n-l,0) = C_Y(n-l)$. At each time $t = 1, 2, ...$, we want to output best linear estimate of Y_{t+1} using $(Y_0, Y_1, Y_2, ..., Y_t)$. In the previous notation $X_t \leftrightarrow Y_{t+1}$ and $s \leftrightarrow t$.

$$\begin{aligned} C_{XY} &= \text{Cov}(X_t, Y_l), \\ &= \text{Cov}(Y_{t+1}, Y_l), \\ &= C_Y(t+1-l). \end{aligned}$$

2.1 Yule-Walker Equations

Let the optimal estimator be $\hat{Y}_{t+1} = \sum_0^t h_{t,n} Y_n$.

$$\begin{bmatrix} C_Y(t+1) \\ C_Y(t) \\ . \\ . \\ C_Y(1) \end{bmatrix} = \underbrace{\begin{bmatrix} C_Y(0) & C_Y(1) & . & . & . & C_Y(t) \\ C_Y(1) & C_Y(2) & . & . & . & C_Y(t-1) \\ . & . & & & & . \\ . & . & & & & . \\ C_Y(t) & C_Y(t-1) & . & . & . & C_Y(0) \end{bmatrix}}_{\Sigma_Y} \begin{bmatrix} h_{t,0} \\ h_{t,1} \\ . \\ . \\ h_{t,t} \end{bmatrix} \tag{12}$$

The matrix Σ_Y in the eqn. (12) is a *Toeplitz* matrix, which means that its entries are constant along the diagonals (since $C_Y(n,l) = C_Y(n-l)$).