TAMS32 STOKASTISKA PROCESSER
Komplettering 4

Torkel Erhardsson

24 augusti 2015

- LINEAR MINIMAL MEAN SQUARE ESTIMATION (LMMSE).
- LINEAR PREDICTION & FILTERING.
- YULE-WALKER EQUATIONS & TOEPLITZ MATRICES.

1 LMMSE

Assume that we are interested in the value of a random variable X, and that we cannot observe this directly. However, we can observe the values of an n-dimensional random variable $Y = (Y_1, \ldots, Y_n)^T$. (Note: we will use vector and matrix notation, so we think of Y as a stochastic column vector of dimension n.) We can then try to **predict (estimate)** X using a linear **predictor (estimator)**

$$\widehat{X} = a^T Y + b = a_1 Y_1 + \ldots + a_n Y_n + b,$$

where $a = (a_1, \ldots, a_n)^T$ is a real valued column vector of dimension n, and b is a real number.

It is common procedure to choose $a = (a_1, \ldots, a_n)^T$ and b so that the quantity

$$E((X - a^T Y - b)^2), \tag{1.1}$$

called the **mean square prediction error**, is minimized. The resulting predictor (estimator) $\widehat{X} = \widehat{a}^T Y + \widehat{b}$ is called the **linear minimal mean square estimator**, or **LMMSE**, of X based on Y, and is sometimes denoted $\widehat{X}_{LMMSE}$.

We shall derive some theorems for LMMSE.

Theorem 1.1 *Let X be a random variable, and $Y = (Y_1, \ldots, Y_n)^T$ be an n-dimensional random variable with mean (column) vector μ_Y and covariance matrix C_Y, where $\det C_Y > 0$. Then, the minimal mean square estimator (LMMSE) of X based on Y is given by $\widehat{X} = \widehat{a}^T Y + \widehat{b}$, where*

$$\widehat{a} = C_Y^{-1} C_{X,Y}, \qquad \widehat{b} = E(X) - \widehat{a}^T \mu_Y,$$

and $C_{X,Y}$ is the n-dimensional column vector with elements

$$(C_{X,Y})_i = C(X, Y_i) \qquad \forall i = 1, \ldots, n.$$

Proof: We rewrite the mean square prediction error by adding and subtracting $E(X) - a^T \mu_Y$ inside the square, expanding the square, and taking the expectation of each term in the expansion:

$$E((X - a^T Y - b)^2) = E((X - E(X) - a^T(Y - \mu_Y) + E(X) - a^T \mu_Y - b)^2)$$

$$= E((X - E(X))^2 + a^T(Y - \mu_Y)(Y - \mu_Y)^T a$$

$$+(E(X) - a^T \mu_Y - b)^2 - 2a^T(Y - \mu_Y)(X - E(X))$$

$$-2a^T(Y-\mu_Y)(E(X)-a^T\mu_Y-b)+(X-E(X))(E(X)-a^T\mu_Y-b))$$
$$=E((X-E(X))^2)+a^TE((Y-\mu_Y)(Y-\mu_Y)^T)a$$
$$+(E(X)-a^T\mu_Y-b)^2-2a^TE((Y-\mu_Y)(X-E(X)))$$
$$-2a^TE(Y-\mu_Y)(E(X)-a^T\mu_Y-b)+E(X-E(X))(E(X)-a^T\mu_Y-b))$$
$$=V(X)+a^TC_Ya+(E(X)-a^T\mu_Y-b)^2-2a^TC_{X,Y}-0+0.$$

In this expression, the third term is the only one that depends on b, and it is greater than or equal to 0. If a is chosen so that the sum of the other terms is minimized, then the third term can be set to 0 by choosing

$$b=\widehat{b}=E(X)-\widehat{a}^T\mu_Y.$$

It now remains to choose a so that the function $f(a)=a^TC_Ya-2a^TC_{X,Y}$ is minimized. This function is a polynomial in $a_1,\ldots,a_n$ of degree 2. A necessary condition for $a=(a_1,\ldots,a_n)^T$ to be a minimum is that

$$\frac{\partial f}{\partial a_i}=2(C_Ya)_i-2(C_{X,Y})_i=0 \qquad \forall i=1,\ldots,n,$$

or, in vector and matrix notation, $C_Ya=C_{X,Y}$. The solution to this system of linear equations is $a=\widehat{a}=C_Y^{-1}C_{X,Y}$. Furthermore, since

$$\frac{\partial^2 f}{\partial a_i \partial a_j}=2(C_Y)_{i,j} \qquad \forall i,j=1,\ldots,n,$$

the matrix of second derivatives is $2C_Y$, which is positive definite by assumption (since $\det C_Y>0$). Hence, the function $f(a)=a^TC_Ya-2a^TC_{X,Y}$ is strictly convex, which implies that $\widehat{a}$ is in fact the global minimum of the function f (see a basic course in multivariate analysis). ∎

Theorem 1.2 *Let X be a random variable, and $Y=(Y_1,\ldots,Y_n)^T$ be an n-dimensional random variable with mean (column) vector μ_Y and covariance matrix C_Y, where $\det C_Y>0$. Then, the mean square prediction error corresponding to the minimal mean square estimator (LMMSE) of X based on Y is:*

$$E((X-\widehat{a}^TY-\widehat{b})^2)=V(X)-\widehat{a}^TC_{X,Y}.$$

Proof: From the proof of Theorem 1.1,

$$E((X-\widehat{a}^TY-\widehat{b})^2)=V(X)+\widehat{a}^TC_Y\widehat{a}+(E(X)-\widehat{a}^T\mu_Y-\widehat{b})^2-2\widehat{a}^TC_{X,Y}$$
$$=V(X)+\widehat{a}^TC_YC_Y^{-1}C_{X,Y}+0-2\widehat{a}^TC_{X,Y}=V(X)-\widehat{a}^TC_{X,Y}.$$

∎

2 Examples

Example 2.1 Let both X and Y be (one-dimensional) random variables. Then, the LMMSE $\widehat{X} = \widehat{a}Y + \widehat{b}$ of X based on Y is given by:

$$\widehat{a} = \frac{C(X,Y)}{V(Y)}, \qquad \widehat{b} = E(X) - \frac{C(X,Y)}{V(Y)}E(Y).$$

The corresponding variance of the prediction error is:

$$E((X - \widehat{a}Y - \widehat{b})^2) = V(X) - \frac{C(X,Y)}{V(Y)}C(X,Y)$$

$$= V(X)(1 - \frac{C(X,Y)^2}{V(Y)V(X)}) = V(X)(1 - \rho(X,Y)^2),$$

where $\rho(X,Y)$ as usual denotes the correlation coefficient.

Example 2.2 Let (X,Y) have a two-dimensional normal distribution. Then, we know that the conditional distribution of X given $Y = y$ is the normal distribution

$$N\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho^2)\right),$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = V(X)$, $\sigma_Y^2 = V(Y)$ and $\rho = \rho(X,Y)$. In this case, we see that for **the conditional expectation** of X given Y,

$$E(X|Y) = E(X|Y = y)\big|_{y=Y} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y)$$

$$= \mu_X + \frac{C(X,Y)}{V(Y)}(Y - \mu_Y) = \widehat{a}Y + \widehat{b}.$$

In words, $E(X|Y)$ equals the LMMSE of X based on Y. This is rather unusual; in most situations it does not hold. Furthermore, the conditional variance of X given Y is $\sigma_X^2(1 - \rho^2)$, which equals the variance of the prediction error for the LMMSE. This is also quite unusual.

Example 2.3 Let the process $\{X_n; n \in \mathbb{Z}\}$ be wide sense stationary with mean $E(X_n) = 0$ and autocovariance (autocorrelation) function

$$C_X(\tau) = c^{|\tau|} \qquad \forall \tau \in \mathbb{Z}, \tag{2.1}$$

where $|c| < 1$. We wish to use the current value of the process to predict (estimate) future values. We formulate this in terms of the LMMSE theory above. We set $X = X_{n+k}$ and $Y = X_n$. The LMMSE of X based on Y is given by

$$\widehat{X}_{n+k} = \frac{C(X_{n+k}, X_n)}{V(X_n)} X_n = \frac{C_X(k)}{C_X(0)} X_n = c^k X_n,$$

and the variance of the prediction error for the LMMSE is

$$E((X_{n+k} - \widehat{X}_{n+k})^2) = V(X_{n+k}) - \frac{C(X_{n+k}, X_n)^2}{V(X_n)} = 1 - c^{2k}.$$

From these expressions we see that if k is large, the predictor $c^k X_n$ should often be small, close to $E(X_{n+k}) = 0$ and the error variance close to its maximum value. This expresses the reasonable idea that the reliability of the prediction should decrease for prediction large steps ahead. If c is close to zero, then again the prediction is close to $E(X_{n+k}) = 0$, the a priori prediction. If $|c|$ is close to 1, then the predictor is highly correlated with the predicted variable.

Example 2.4 Let the process $\{X_n; n \in \mathbb{Z}\}$ be wide sense stationary with mean $E(X_n) = 0$ and autocovariance (autocorrelation) function

$$C_X(\tau) = c^{|\tau|} \qquad \forall \tau \in \mathbb{Z}, \tag{2.2}$$

where $|c| < 1$. Define the process $\{Y_n; n \in \mathbb{Z}\}$ by

$$Y_n = X_n + Z_n \qquad \forall n \in \mathbb{Z},$$

where $\{Z_n; n \in \mathbb{Z}\}$ is I.I.D. white noise with $E(Z_n) = 0$ and $V(Z_n) = \sigma_Z^2 < \infty$, and $\{Z_n; n \in \mathbb{Z}\}$ is independent of $\{X_n; n \in \mathbb{Z}\}$. We can think of Y_n as a noisy measurement of X_n, and we wish to predict (estimate) X_n based on Y_n. This is known as *filtering*. The LMMSE formulation is to let $X = X_n$ and $Y = Y_n$. The LMMSE of X based on Y is given by

$$\widehat{X}_n = \frac{C(X_n, Y_n)}{V(Y_n)} Y_n = \frac{V(X_n)}{V(X_n) + V(Z_n)} Y_n = \frac{1}{1 + \sigma_Z^2} Y_n,$$

and the variance of the prediction error for the LMMSE is

$$E((X_n - \widehat{X}_n)^2) = V(X_n) - \frac{C(X_n, Y_n)^2}{V(Y_n)} = 1 - \frac{1}{1 + \sigma_Z^2}.$$

Example 2.5 Let $\{Z_n; n \in \mathbb{Z}\}$ be I.I.D. white noise, with $E(Z_n) = 0$ and $V(Z_n) = 1$, and define the MA(1) process $\{X_n; n \in \mathbb{Z}\}$ by

$$X_n = Z_n + \frac{1}{2} Z_{n-1} \qquad \forall n \in \mathbb{Z}.$$

The process $\{X_n; n \in \mathbb{Z}\}$ has the autocovariance (autocorrelation) function

$$C_X(\tau) = R_X(\tau) = \begin{cases} \frac{5}{4}, & \text{if } \tau = 0; \\ \frac{1}{2}, & \text{if } |\tau| = 1; \\ 0, & \text{otherwise.} \end{cases}$$

(Either check this yourself, or see Kompletteringshäfte 3.) The LMMSE of X_n based on $Y = (X_{n-1}, X_{n-2})^T$ can be obtained using Theorem 1.1. We first note that

$$C_Y = \begin{pmatrix} \frac{5}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{5}{4} \end{pmatrix}$$

and

$$C_{X_n,Y} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

The LMMSE is given by $\widehat{X}_n = \widehat{a}^T Y = \widehat{a}_1 X_{n-1} + \widehat{a}_2 X_{n-2}$, where

$$\begin{aligned} \widehat{a} &= C_Y^{-1} C_{X_n,Y} = \begin{pmatrix} 0.9524 & -0.3810 \\ -0.3810 & 0.9524 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} = \\ &= \begin{pmatrix} 0.4762 \\ -0.1905 \end{pmatrix}. \end{aligned} \tag{2.3}$$

3 One-step prediction of a wide sense stationary process

Let $\{X_n; n \in \mathbb{Z}\}$ be a wide sense stationary random sequence with mean function equal to zero and with autocorrelation function $R_X(\tau)$. We consider the problem of predicting (estimating) X_n based on the p most recent values of the process, or (equivalently) on the p-dimensional random variable $Y =$

$(X_{n-1}, \ldots, X_{n-p})$. The covariance (correlation) matrix of Y is

$$C_Y = \begin{pmatrix} R_X(0) & R_X(1) & \ldots & R_X(p-2) & R_X(p-1) \\ R_X(1) & R_X(0) & R_X(1) & \ldots & R_X(p-2) \\ R_X(2) & R_X(1) & R_X(0) & \ldots & R_X(p-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R_X(p-2) & R_X(p-3) & \ldots & \ddots & R_X(1) \\ R_X(p-1) & R_X(p-2) & \ldots & R_X(1) & R_X(0) \end{pmatrix}. \tag{3.1}$$

Using Theorem 1.1, we see that the LMMSE of X_n based on Y exists if $\det C_Y > 0$ (that is, if C_Y has full rank). In this case, the LMMSE is

$$\widehat{X}_n = \widehat{a}^T Y + b = \widehat{a}_1 X_{n-1} + \ldots + \widehat{a}_p X_{n-p}, \tag{3.2}$$

where $\widehat{a}$ satisfies the system of linear equations

$$C_Y \widehat{a} = C_{X_n,Y}, \tag{3.3}$$

known as the **Yule-Walker equations**, where

$$C_{X_n,Y} = (R_X(1), \ldots, R_X(p))^T .$$

It should be noted that neither C_Y nor $C_{X_n,Y}$ depend on n, so by (3.2), the **process of one-step predictors (estimators)** $\{\widehat{X}_n; n \in \mathbb{Z}\}$ is the output of a linear time-invariant filter (a LTI), with finite impulse response $\{\widehat{a}_{k+1}; k = 0, \ldots, p-1\}$ (a FIR filter). This kind of FIR filter is called a **prediction filter**.

We also note that C_Y has the property that the elements along every diagonal are identical. Matrices with this property are known as **Toeplitz** matrices. There are a number of algorithms for inverting C_Y that take advantage of the Toeplitz property, e.g. Levinson - Durbin or Berlekamp-Massey algorithms.