

## 12.0 Estimation Theory

### 12.1 Conditional Operations

Say  $\mathbf{X}$  and  $\mathbf{Y}$  have joint density  $f_{XY}(\mathbf{x}, \mathbf{y})$ . Then

$$f_{Y|X}(\mathbf{y}|\mathbf{x}) = \frac{f_{XY}(\mathbf{x}, \mathbf{y})}{f_X(\mathbf{x})}.$$

$$E[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \int \cdots \int \mathbf{y} f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{y}$$

in the continuous case and

$$E[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \sum_{\mathbf{y}} \mathbf{y} f_{Y|X}(\mathbf{y}|\mathbf{x})$$

in the discrete case. In this latter case

$$f_{XY}(\mathbf{x}, \mathbf{y}) = P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}).$$

### Properties of Conditional Expectation

1. Usual expectation is a special case. Let  $\mathbf{X} = \mathbf{c}$  (constant vector).

$$E[\mathbf{Y}|\mathbf{X}] = E[\mathbf{Y}].$$

2. Linearity.

$$E[a\mathbf{Y}_1 + b\mathbf{Y}_2|\mathbf{X}] = aE[\mathbf{Y}_1|\mathbf{X}] + bE[\mathbf{Y}_2|\mathbf{X}].$$

3. If  $X$  is independent of  $Y$  then

$$E[\mathbf{Y}|\mathbf{X}] = E[\mathbf{Y}].$$

**Proof:**

$$f_{Y|X}(\mathbf{y}|\mathbf{x}) = \frac{f_{XY}(\mathbf{x}, \mathbf{y})}{f_X(\mathbf{x})} = \frac{f_Y(\mathbf{y})f_X(\mathbf{x})}{f_X(\mathbf{x})} = f_Y(\mathbf{y}).$$

4.  $E[h(\mathbf{X})|\mathbf{X}] = h(\mathbf{X})$ .

5.  $E[g(\mathbf{X}, \mathbf{Y})|\mathbf{X} = \mathbf{x}] = E[g(\mathbf{x}, \mathbf{Y})|\mathbf{X} = \mathbf{x}]$ . Note that  $\mathbf{Y}$  may depend on  $\mathbf{X}$  so we still need the conditional  $\mathbf{X} = \mathbf{x}$ .
6.  $E[r(\mathbf{X})h(\mathbf{Y})|\mathbf{X}] = r(\mathbf{X})E[h(\mathbf{Y})|\mathbf{X}]$ .
7. Double Expectation Formula:

$$E[E(\mathbf{Y}|\mathbf{X})] = E(\mathbf{Y}).$$

**Proof:**

$$\begin{aligned} E[E(\mathbf{Y}|\mathbf{X})] &= \int f_X(\mathbf{x}) \left[ \int \mathbf{y} f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right] d\mathbf{x} \\ &= \int \int \mathbf{y} f_{Y|X}(\mathbf{y}|\mathbf{x}) f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= \int \int \mathbf{y} f_{XY}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= \int \mathbf{y} f_Y(\mathbf{y}) d\mathbf{y} \\ &= E[\mathbf{Y}]. \end{aligned}$$

8. Product Expectation Formula:

$$E[r(\mathbf{X})h(\mathbf{Y})] = E[E(r(\mathbf{X})h(\mathbf{Y})|\mathbf{X})] = E[r(\mathbf{X})E(h(\mathbf{Y})|\mathbf{X})].$$

**12.2 Stein's Equation** Consider the standard normal density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbf{R}$$

that corresponds to the normally distributed random variable  $Z \sim N(0, 1)$ . Suppose we have a differentiable function  $f$  such that  $E[f'(z)]$  exists. Then

$$E[f'(z)] = E[zf(z)].$$

The above equation is known as Stein's equation which we will now develop. First note that

$$\phi'(z) = -z\phi(z).$$

Then

$$E[f'(z)] = \int_{-\infty}^{\infty} f'(y)\phi(y)dy$$

$$\begin{aligned}
&= \int_0^\infty f'(y) \int_y^\infty z\phi(z)dzdy - \int_{-\infty}^0 f'(y) \int_{-\infty}^y z\phi(z)dzdy \\
&= \int_0^\infty z\phi(z) \left[ \int_0^z f'(y)dy \right] dz - \int_{-\infty}^0 z\phi(z) \left[ \int_z^0 f'(y)dy \right] dz \\
&= \int_0^\infty \phi(z)z [f(z) - f(0)] dz + \int_{-\infty}^0 \phi(z)z [f(z) - f(0)] dz \\
&= \int_{-\infty}^\infty zf(z)\phi(z)dz - f(0) \int_{-\infty}^\infty z\phi(z)dz \\
&= \int_{-\infty}^\infty zf(z)\phi(z)dz \\
&= E[zf(z)].
\end{aligned}$$

The reference for the above is Charles Stein, *Approximate Computation of Expectations* (Monograph, 1986).

### 12.3 Minimum Mean Square Error Estimation

Say we observe  $X$  and predict or estimate  $Y$ . Predict  $Y$  by  $g(X)$ . Say  $E[Y] = \mu$  and we use the best constant for prediction, i.e.,  $g(X) = c$ .

$$\begin{aligned}
E[(Y - c)^2] &= E[(Y - \mu + \mu - c)^2] \\
&= E[(Y - \mu)^2] + E[(\mu - c)^2] + 2E[(Y - \mu)(\mu - c)] \\
&= E[(Y - \mu)^2] + (\mu - c)^2
\end{aligned}$$

so the best we can do is to set  $c = \mu$ . Thus, we predict  $Y$  by its mean  $\mu$ .

More generally,

$$\begin{aligned}
E[(Y - g(X))^2] &= E \left[ (Y - E[Y|X] + E[Y|X] - g(X))^2 \right] \\
&= E \left[ (Y - E[Y|X])^2 \right] + E \left[ (E[Y|X] - g(X))^2 \right] + \text{crossterm}
\end{aligned}$$

where

$$\text{crossterm} = 2E \left[ (Y - E[Y|X])(E[Y|X] - g(X)) \right].$$

Now

$$\begin{aligned}
\text{crossterm} &= 2E \left\{ E \left[ (Y - E[Y|X])(E[Y|X] - g(X)) \mid X \right] \right\} \\
&= 2E \left\{ (E[Y|X] - g(X)) E \left[ (Y - E[Y|X]) \mid X \right] \right\}.
\end{aligned}$$

But

$$\begin{aligned} E[(Y - E[Y|X])|X] &= E[Y|X] - E[E[Y|X]|X] \\ &= E[Y|X] - E[Y|X] = 0 \end{aligned}$$

so

$$\begin{aligned} MSE &= E[(Y - g(X))^2] = E[(Y - E[Y|X])^2] + E[(E[Y|X] - g(X))^2] \\ &\geq E[(Y - E[Y|X])^2] \end{aligned}$$

thus the best we can do is make LHS=RHS by setting  $g(X) = E[Y|X]$ . Hence the best MSE estimator for  $Y$  given  $X$  is

$$\hat{Y} = E[Y|X].$$

Similarly, for vectors we compute

$$\min \|\mathbf{Y} - g(\mathbf{X})\|^2$$

to get

$$\hat{\mathbf{Y}} = E[\mathbf{Y}|\mathbf{X}].$$

**Definition:** An estimate  $\hat{\theta}$  of  $\theta$  is said to be *unbiased* if

$$E[\hat{\theta}] = \theta.$$

Let

$$\begin{aligned} \hat{\mathbf{Y}} &= E[\mathbf{Y}|\mathbf{X}] = E_{Y|\mathbf{X}}[\mathbf{Y}|\mathbf{X}]. \\ E[\hat{\mathbf{Y}}] &= E_{\mathbf{X}} [E_{Y|\mathbf{X}}[\mathbf{Y}|\mathbf{X}]] = E[\mathbf{Y}] \end{aligned}$$

so  $E[\mathbf{Y}|\mathbf{X}]$  is unbiased.

### Orthogonality Principle

Here

$$E[h^T(\mathbf{X})(\mathbf{Y} - \hat{\mathbf{Y}})^*] = 0$$

or

$$E[h^T(\mathbf{X})(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))^*] = 0.$$

### MMSE $\rightarrow$ Orthogonality

$$E[h^T(\mathbf{X})(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))^*] = E[E[(h^T(\mathbf{X})(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))^*)^*|\mathbf{X}]]$$

$$= E \left[ (h^T(\mathbf{X})[E(\mathbf{Y}|\mathbf{X})^* - E(\mathbf{Y}|\mathbf{X})^*]) \right] = 0.$$

**Orthogonality**  $\longrightarrow$  **MMSE**

Now assume orthogonality holds. The MMSE is

$$\epsilon_{min} = E[(\mathbf{Y} - g(\mathbf{X}))^T(\mathbf{Y} - g(\mathbf{X}))^*]$$

where

$$g(\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}).$$

Consider an alternative estimate  $h(\mathbf{X}) \neq g(\mathbf{X})$ . Then using this estimate we have

$$\begin{aligned} \epsilon &= E[(\mathbf{Y} - g(\mathbf{X}) + g(\mathbf{X}) - h(\mathbf{X}))^T(\mathbf{Y} - g(\mathbf{X}) + g(\mathbf{X}) - h(\mathbf{X}))^*] \\ &= \epsilon_{min} + E[(g(\mathbf{X}) - h(\mathbf{X}))^T(g(\mathbf{X}) - h(\mathbf{X}))^*] \\ &\quad + E[(g(\mathbf{X}) - h(\mathbf{X}))^T(\mathbf{Y} - g(\mathbf{X}))^*] + E[(\mathbf{Y} - g(\mathbf{X}))^T(g(\mathbf{X}) - h(\mathbf{X}))^*]. \end{aligned}$$

Now the last two terms are zero by orthogonality and

$$E[(g(\mathbf{X}) - h(\mathbf{X}))^T(g(\mathbf{X}) - h(\mathbf{X}))^*] > 0 \quad \forall h(\mathbf{X}) \neq g(\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}).$$

Hence  $\epsilon > \epsilon_{min}$  if  $h(\mathbf{X}) \neq E(\mathbf{Y}|\mathbf{X})$ . So the best we can do is let  $h(\mathbf{X}) = E(\mathbf{Y}|\mathbf{X})$ .

**Linear Functions**

Consider a subspace of linear functions of  $x$ . We want to find the best MSE linear predictor of real  $Y$ . Look at

$$\hat{Y} = a(X - \mu) + b$$

where  $E[X] = \mu$ . Find  $a, b$ . We have

$$E[\hat{Y}] = b.$$

Consider the subspace:  $\alpha(x - \mu) + \beta$ . Then

$$(Y - \hat{Y}) \perp \alpha(x - \mu) + \beta, \quad \forall \alpha, \beta.$$

Let  $\alpha = 0, \beta = 1$ . Then

$$(Y - \hat{Y}) \perp 1.$$

So,

$$E[1 \cdot (Y - \hat{Y})] = 0 \Rightarrow E[\hat{Y}] = E[Y].$$

Therefore,

$$\hat{Y} = a(X - \mu) + E[Y] \Rightarrow \hat{Y} - E[Y] = a(X - \mu).$$

Now let  $\alpha = 1$ ,  $\beta = 0$ . Then

$$(Y - \hat{Y}) \perp (X - \mu).$$

Hence,

$$E[(X - \mu)(Y - \hat{Y})] = 0.$$

So,

$$E[(X - \mu)Y] = E[(X - \mu)\hat{Y}].$$

Thus,

$$E[(X - \mu)(Y - E[Y])] = E[(X - \mu)(\hat{Y} - E[Y])] = E[(X - \mu)a(X - \mu)].$$

Therefore,

$$Cov(X, Y) = aE[(X - \mu)^2].$$

So,

$$a = \frac{Cov(X, Y)}{Var(X)}.$$

Thus,

$$\hat{Y} = \frac{Cov(X, Y)}{Var(X)}(X - \mu) + E[Y].$$

Let

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

Then

$$\hat{Y} = \rho \frac{\sigma_Y}{\sigma_X}(X - \mu) + \mu_Y.$$

## Consistency

Let

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

where the  $X_i$  are i.i.d.

**Definition:** An estimate  $\hat{\theta}$  of  $\theta$  is said to be *consistent* if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0 \quad \forall \epsilon > 0.$$

By Chebyshev

$$P(|\hat{\mu} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

So,  $\hat{\mu} \rightarrow \mu$  in probability and is consistent.

## 12.4 Linear Minimum Mean Square Error Estimation

### Orthogonality Principle

$$E[h^t(\mathbf{X})(\mathbf{Y} - \hat{\mathbf{Y}})^*] = 0.$$

In particular, with

$$h^t(\mathbf{X}) = (0, 0, \dots, X_i, 0, \dots, 0)$$

we see that the error  $\mathbf{e} = (\mathbf{Y} - \hat{\mathbf{Y}})$  is orthogonal to each component of the data. So

$$E[X_i(\mathbf{Y} - \hat{\mathbf{Y}})_j^*] = 0, \quad 1 \leq i, j \leq n.$$

$$\Rightarrow E[(\mathbf{Y} - \hat{\mathbf{Y}})_j X_i^*] = 0$$

$$\Rightarrow E[(\mathbf{Y} - \hat{\mathbf{Y}})X^\dagger] = \mathbf{0}.$$

Let

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{X}.$$

Then

$$E[(\mathbf{Y} - \mathbf{A}\mathbf{X})X^\dagger] = \mathbf{0}.$$

$$\Rightarrow E[\mathbf{Y}X^\dagger] = \mathbf{A}E[\mathbf{X}X^\dagger] = \mathbf{A}\mathbf{R}_\mathbf{X}.$$

If  $\mathbf{R}_X$  is invertible then

$$\mathbf{A} = \mathbf{R}_{YX}\mathbf{R}_X^{-1}$$

so

$$\hat{\mathbf{Y}} = \mathbf{R}_{YX}\mathbf{R}_X^{-1}\mathbf{X}.$$

This is in fact the LMMSE estimator of  $\mathbf{Y}$ .

### Another Way

Next we derive the same estimator using an alternative approach.

We want to find

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{X}.$$

To minimize mean square error we consider

$$\min \left\{ tr E[(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})^\dagger] \right\}$$

i.e.,

$$\min tr [\mathbf{K}_e]$$

where

$$\mathbf{K}_e = E[(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})^\dagger].$$

We now force the estimator to be linear and consider

$$\begin{aligned} \mathbf{K}_e &= E[(\mathbf{Y} - \mathbf{A}\mathbf{X})(\mathbf{Y} - \mathbf{A}\mathbf{X})^\dagger] \\ &= \mathbf{A}\mathbf{R}_X\mathbf{A}^\dagger - \mathbf{R}_{YX}\mathbf{A}^\dagger - \mathbf{A}\mathbf{R}_{XY} + \mathbf{R}_Y. \end{aligned}$$

This last result may be written

$$\mathbf{K}_e = \mathbf{R}_Y + (\mathbf{A} - \mathbf{R}_{YX}\mathbf{R}_X^{-1})\mathbf{R}_X(\mathbf{A} - \mathbf{R}_{YX}\mathbf{R}_X^{-1})^\dagger - \mathbf{R}_{YX}\mathbf{R}_X^{-1}\mathbf{R}_{XY}.$$

Write

$$\mathbf{K}_e = \alpha + \beta + \gamma$$

where

$$\alpha = \mathbf{R}_Y, \quad \beta = (\mathbf{A} - \mathbf{R}_{YX}\mathbf{R}_X^{-1})\mathbf{R}_X(\mathbf{A} - \mathbf{R}_{YX}\mathbf{R}_X^{-1})^\dagger, \quad \gamma = -\mathbf{R}_{YX}\mathbf{R}_X^{-1}\mathbf{R}_{XY}.$$

Now

$$tr [\mathbf{K}_e] = tr (\alpha + \beta + \gamma) = tr (\alpha) + tr (\beta) + tr (\gamma).$$



Note we have no control over  $tr(\alpha)$  and  $tr(\gamma)$ . Observe  $tr(\beta) \geq 0$  and  $tr(\beta) = 0$  if  $\mathbf{A} = \mathbf{R}_{\mathbf{YX}}\mathbf{R}_{\mathbf{X}}^{-1}$ . Hence, the LMMSE estimator of  $Y$  is

$$\hat{\mathbf{Y}} = \mathbf{R}_{\mathbf{YX}}\mathbf{R}_{\mathbf{X}}^{-1}\mathbf{X}.$$

The resulting mean square error is

$$MSE = tr(\mathbf{R}_{\mathbf{Y}} - \mathbf{R}_{\mathbf{YX}}\mathbf{R}_{\mathbf{X}}^{-1}\mathbf{R}_{\mathbf{XY}}).$$

## 12.5 Affine Linear Minimum Mean Square Error Estimation

Here

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{X} + \mathbf{m}.$$

Let

$$\tilde{\mathbf{Y}} = \mathbf{Y} - \mu_{\mathbf{Y}}, \quad \tilde{\mathbf{X}} = \mathbf{X} - \mu_{\mathbf{X}}.$$

So,

$$E[\tilde{\mathbf{Y}}] = E[\tilde{\mathbf{X}}] = \mathbf{0},$$

$$\mathbf{R}_{\tilde{\mathbf{Y}}} = \mathbf{K}_{\tilde{\mathbf{Y}}} = \mathbf{K}_{\mathbf{Y}},$$

$$\mathbf{R}_{\tilde{\mathbf{X}}} = \mathbf{K}_{\tilde{\mathbf{X}}} = \mathbf{K}_{\mathbf{X}}.$$

We know

$$\hat{\tilde{\mathbf{Y}}} = \mathbf{A}\tilde{\mathbf{X}}$$

or

$$\hat{\tilde{\mathbf{Y}}} = \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\tilde{\mathbf{X}}.$$

Now

$$\begin{aligned} \hat{\mathbf{Y}} &= \hat{\tilde{\mathbf{Y}}} + \mu_{\mathbf{Y}} = \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\tilde{\mathbf{X}} + \mu_{\mathbf{Y}} \\ &= \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}}) + \mu_{\mathbf{Y}} \\ &= \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mathbf{X} - \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mu_{\mathbf{X}} + \mu_{\mathbf{Y}}. \end{aligned}$$

So

$$\mathbf{A} = \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1},$$

$$\mathbf{m} = -\mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mu_{\mathbf{X}} + \mu_{\mathbf{Y}}.$$

## 12.6 Gaussian Examples

Say  $\mathbf{X}$  and  $\mathbf{Y}$  are jointly normal. We want the conditional distribution of  $\mathbf{Y}$  given  $\mathbf{X}$  in order to find  $E[\mathbf{Y}|\mathbf{X}]$ .

Let

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix}.$$

Then

$$\mathbf{K}_Z = \begin{bmatrix} \mathbf{K}_Y & \mathbf{K}_{YX} \\ \mathbf{K}_{XY} & \mathbf{K}_X \end{bmatrix}.$$

Let

$$\mathbf{W} = (\mathbf{Y} - \mu_Y) - \mathbf{K}_{YX}\mathbf{K}_X^{-1}(\mathbf{X} - \mu_X).$$

Then

$$\begin{pmatrix} \mathbf{W} \\ \mathbf{X} \end{pmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{K}_{YX}\mathbf{K}_X^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} + \begin{pmatrix} -\mu_Y + \mathbf{K}_{YX}\mathbf{K}_X^{-1}\mu_X \\ \mathbf{0} \end{pmatrix}.$$

Note  $\begin{pmatrix} \mathbf{W} \\ \mathbf{X} \end{pmatrix}$  is written as a linear transformation of a multivariate normal and is hence multivariate normal.

We get

$$E[\mathbf{W}] = \mathbf{0}.$$

$$\begin{aligned} \text{Cov}(\mathbf{W}, \mathbf{X}) &= \text{Cov}(\mathbf{W}, \mathbf{X} - \mu_X) = E[\mathbf{W}(\mathbf{X} - \mu_X)^\dagger] \\ &= E[(\mathbf{Y} - \mu_Y)(\mathbf{X} - \mu_X)^\dagger - \mathbf{K}_{YX}\mathbf{K}_X^{-1}(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^\dagger] \\ &= \mathbf{K}_{YX} - \mathbf{K}_{YX}\mathbf{K}_X^{-1}\mathbf{K}_X = \mathbf{0}. \end{aligned}$$

Now write

$$\mathbf{Y} = (\mathbf{Y} - \mathbf{W}) + \mathbf{W} = (\mu_Y + \mathbf{K}_{YX}\mathbf{K}_X^{-1}(\mathbf{X} - \mu_X)) + \mathbf{W}.$$

We wish to find the conditional distribution of  $\mathbf{Y}$  given  $\mathbf{X} = \mathbf{x}$ . Now

$$\mathcal{L}(\mathbf{W}|\mathbf{X} = \mathbf{x}) = \mathcal{L}(\mathbf{W})$$

since  $Cov(\mathbf{W}, \mathbf{X}) = 0$ . Also,

$$\begin{aligned}\mathcal{L}(\mu_{\mathbf{Y}} + \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}}) | \mathbf{X} = \mathbf{x}) &= \text{constant} \\ &= \mu_{\mathbf{Y}} + \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{x} - \mu_{\mathbf{X}}).\end{aligned}$$

So

$$\mathcal{L}(\mathbf{Y} | \mathbf{X} = \mathbf{x}) = \mathcal{L}(\text{constant} + \mathbf{W} | \mathbf{X} = \mathbf{x})$$

which is normally distributed as

$$\sim \mathcal{N}(\mu_{\mathbf{Y}} + \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{x} - \mu_{\mathbf{X}}), Cov(\mathbf{W})).$$

Now

$$\begin{aligned}Cov(\mathbf{W}) &= E[\mathbf{W}\mathbf{W}^\dagger] \\ &= E\left[\left[(\mathbf{Y} - \mu_{\mathbf{Y}}) - \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}})\right]\left[(\mathbf{Y} - \mu_{\mathbf{Y}}) - \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}})\right]^\dagger\right] \\ &= \mathbf{K}_{\mathbf{Y}} - \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mathbf{K}_{\mathbf{XY}} - \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mathbf{K}_{\mathbf{XY}} + \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mathbf{K}_{\mathbf{X}}\mathbf{K}_{\mathbf{X}}^{-1}\mathbf{K}_{\mathbf{XY}} \\ &= \mathbf{K}_{\mathbf{Y}} - \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}\mathbf{K}_{\mathbf{XY}}.\end{aligned}$$

We have derived the MMSE estimate of  $\mathbf{Y}$  given  $\mathbf{X}$  as

$$\hat{\mathbf{Y}} = E[\mathbf{Y} | \mathbf{X}] = \mu_{\mathbf{Y}} + \mathbf{K}_{\mathbf{YX}}\mathbf{K}_{\mathbf{X}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}})$$

which is affine linear in  $\mathbf{X}$ .

## 12.7 Least Squares Estimation

Let

$$z_i = h_i(\theta) + \epsilon_i, \quad \theta = (\theta_1, \dots, \theta_n)^T$$

where  $h_i(\theta)$  is some known function of  $\theta$ . Assume

$$E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2, \quad E(\epsilon_i\epsilon_j) = 0 \text{ for } i \neq j.$$

**Linear Regression**

$$h_i(\theta) = \theta_0 + \theta_1 X_i.$$

Consider

$$\sum_{i=1}^n (Z_i - h_i(\theta))^2.$$

We want to choose  $\theta$  to minimize the above expression.

**Example:** Let  $h_i(\theta) = \theta$ . Then

$$Z_i = \theta + \epsilon_i.$$

Consider

$$\min \left\{ \sum_{i=1}^n (Z_i - \theta)^2 \right\}.$$

Compute

$$\frac{\partial}{\partial \theta} \Rightarrow \sum_{i=1}^n (Z_i - \theta) = 0 \Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n Z_i = \bar{Z}.$$

**Example:** Let  $h_i(\theta) = \theta_1 + \theta_2 X_i$ . Then

$$Z_i = \theta_1 + \theta_2 X_i + \epsilon_i.$$

We want to

$$\min \left\{ \sum_{i=1}^n (Z_i - \theta_1 - \theta_2 X_i)^2 \right\}.$$

Compute

$$\frac{\partial}{\partial \theta_2} \Rightarrow \sum_{i=1}^n (Z_i - \theta_1 - \theta_2 X_i) X_i = 0.$$

$$\frac{\partial}{\partial \theta_1} \Rightarrow \sum_{i=1}^n (Z_i - \theta_1 - \theta_2 X_i) = 0.$$

We get

$$\bar{Z} - \hat{\theta}_1 - \hat{\theta}_2 \bar{X} = 0 \Rightarrow \hat{\theta}_1 = \bar{Z} - \hat{\theta}_2 \bar{X}.$$

Substitute

$$\sum_{i=1}^n (Z_i - \bar{Z} + \hat{\theta}_2 \bar{X} - \hat{\theta}_2 X_i) X_i = 0.$$

$$\hat{\theta}_2 \sum_{i=1}^n X_i (\bar{X} - X_i) = \sum_{i=1}^n X_i (\bar{Z} - Z_i)$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})(\bar{Z} - Z_i)}{\sum_{i=1}^n (X_i - \bar{X})(\bar{X} - X_i)} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}.$$

Note:

$$\sum_{i=1}^n \bar{X}(\bar{Z} - Z_i) = \bar{X} \sum_{i=1}^n (\bar{Z} - Z_i) = 0$$

and

$$\sum_{i=1}^n \bar{X}(\bar{X} - X_i) = 0.$$

Let

$$\rho = \frac{E[(X - \mu_X)(Z - \mu_Z)]}{\sqrt{\text{Var}(X)\text{Var}(Z)}}.$$

Then

$$\hat{\rho} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Z_i - \bar{Z})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})^2}}.$$

Thus,

$$\hat{\theta}_2 = \hat{\rho} \frac{\hat{\sigma}_Z}{\hat{\sigma}_X}.$$

$$\hat{Z} = \hat{\theta}_1 + \hat{\theta}_2 X = \bar{Z} - \hat{\theta}_2 \bar{X} + \hat{\theta}_2 X = \hat{\theta}_2 (X - \bar{X}) + \bar{Z}$$

so

$$\hat{Z} = \hat{\rho} \frac{\hat{\sigma}_Z}{\hat{\sigma}_X} (X - \bar{X}) + \bar{Z}.$$

This looks like the linear predictor of Section 12.3.

Now consider

$$\mathbf{Z} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$$

where

$$E(\mathbf{n}) = \mathbf{0}, \quad \mathbf{K}_n = \sigma^2 \mathbf{I}.$$

Consider derivatives

$$\frac{\partial(\mathbf{X}^t \mathbf{A} \mathbf{X})}{\partial \mathbf{X}} = 2\mathbf{A} \mathbf{X}$$

for  $\mathbf{A}$  symmetric.

$$\frac{\partial(\mathbf{a}^t \mathbf{X})}{\partial \mathbf{X}} = \mathbf{a}$$

$$\frac{\partial(\mathbf{Y}^t \mathbf{A} \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}^t \mathbf{Y}$$

$$\frac{\partial(\mathbf{X}^t \mathbf{A} \mathbf{Y})}{\partial \mathbf{X}} = \mathbf{A} \mathbf{Y}.$$

Now

$$\min_{\theta} S = [(\mathbf{Z} - \mathbf{H}\theta)^t(\mathbf{Z} - \mathbf{H}\theta)].$$

Observe

$$S = \mathbf{Z}^t\mathbf{Z} + \theta^t\mathbf{H}^t\mathbf{H}\theta - \theta^t\mathbf{H}^t\mathbf{Z} - \mathbf{Z}^t\mathbf{H}\theta.$$

$$\frac{\partial S}{\partial \theta} = 0 = 2\mathbf{H}^t\mathbf{H}\hat{\theta} - \mathbf{H}^t\mathbf{Z} - \mathbf{H}^t\mathbf{Z}$$

$$(\mathbf{H}^t\mathbf{H})\hat{\theta} = \mathbf{H}^t\mathbf{H}\mathbf{Z}$$

or

$$\hat{\theta} = (\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t\mathbf{Z}.$$

## 12.8 Estimation of Parameters of PDFs

### Maximum Likelihood Estimation (MLE)

Here  $\mathbf{x}$  is observed and we estimate  $\theta$ . We compute

$$\max_{\theta} f(\mathbf{x}|\theta) \text{ or } \max_{\theta} [\ln f(\mathbf{x}|\theta)].$$

**Example:**  $\theta$  is a scalar. Here we have  $n$  i.i.d. observations  $x_i$ ,  $i = 1, 2, \dots, n$ .

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta).$$

Say

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi}\theta} \exp\left\{-\frac{(x_i - \theta)^2}{2\sigma^2}\right\}.$$

Then

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

so

$$\ln f(\mathbf{x}|\theta) = \text{constant} + \sum_{i=1}^n \left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right).$$

Set

$$\frac{\partial}{\partial \theta} \ln f(\mathbf{x}|\theta) = 0$$

to get

$$\sum_{i=1}^n \frac{(x_i - \theta)}{\sigma^2} = 0 \Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i.$$

### Maximum a Posteriori (MAP) Estimation

$$\max_{\theta} f(\theta|\mathbf{x}) = \max_{\theta} \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})}.$$

But  $f(\mathbf{x})$  is a constant since  $\mathbf{x}$  is observed. So we compute

$$\max_{\theta} f(\mathbf{x}|\theta)f(\theta)$$

where  $f(\theta)$  is a priori density for vector  $\theta$ .