12.0 Estimation Theory

12.1 Conditional Operations

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

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

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

in the continuous case and

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

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).

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

2. Linearity.

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

3. If X is independent of Y then

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

   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

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

$$= \int_0^\infty f'(y) \int_y^\infty z\phi(z)dzdy - \int_{-\infty}^0 f'(y) \int_{-\infty}^y z\phi(z)dzdy$$

$$= \int_0^\infty z\phi(z) \left[\int_0^z f'(y)dy\right] dz - \int_{-\infty}^0 z\phi(z) \left[\int_z^0 f'(y)dy\right] dz$$

$$= \int_0^\infty \phi(z)z\left[f(z) - f(0)\right] dz + \int_{-\infty}^0 \phi(z)z\left[f(z) - f(0)\right] dz$$

$$= \int_{-\infty}^\infty zf(z)\phi(z)dz - f(0)\int_{-\infty}^\infty z\phi(z)dz$$

$$= \int_{-\infty}^\infty zf(z)\phi(z)dz$$

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

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., $g(X) = c$.

$$E[(Y-c)^2] = E[(Y-\mu+\mu-c)^2]$$

$$= E[(Y-\mu)^2] + E[(\mu-c)^2] + 2E[(Y-\mu)(\mu-c)]$$

$$= E[(Y-\mu)^2] + (\mu-c)^2$$

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

More generally,

$$E[(Y-g(X))^2] = E\left[(Y - E[Y|X] + E[Y|X] - g(X))^2\right]$$

$$= E\left[(Y-E[Y|X])^2\right] + E\left[(E[Y|X]-g(X))^2\right] + crossterm$$

where

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

Now

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

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

But

$$E\left[(Y - E[Y|X])|X\right] = E[Y|X] - E\left[E[Y|X]|X\right]$$
$$= E[Y|X] - E[Y|X] = 0$$

so

$$MSE = E[(Y - g(X))^2] = E\left[(Y - E[Y|X])^2\right] + E\left[(E[Y|X] - g(X))^2\right]$$
$$\geq E\left[(Y - E[Y|X])^2\right]$$

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 θ is said to be *unbiased* if

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

Let

$$\hat{\mathbf{Y}} = E[\mathbf{Y}|\mathbf{X}] = E_{Y|X}[\mathbf{Y}|\mathbf{X}].$$
$$E[\hat{\mathbf{Y}}] = E_X\left[E_{Y|X}[\mathbf{Y}|\mathbf{X}]\right] = E[\mathbf{Y}]$$

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 $\longrightarrow$ Orthogonality

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

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

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

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

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

$$+E[(g(\mathbf{X}) - h(\mathbf{X}))^T(\mathbf{Y} - g(\mathbf{X}))^*] + E[(\mathbf{Y} - g(\mathbf{X}))^T(g(\mathbf{X}) - h(\mathbf{X}))^*].$$

Now the last two terms are zero by orthogonality and

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

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 θ is said to be *consistent* if

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

By Chebyshev

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

So, $\hat{\mu} \to \mu$ 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

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

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

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

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

$$\hat{\mathbf{Y}} = \mathbf{AX}.$$

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

$$\mathbf{K_e} = E[(\mathbf{Y} - \mathbf{AX})(\mathbf{Y} - \mathbf{AX})^\dagger]$$

$$= \mathbf{A}\mathbf{R_X}\mathbf{A}^\dagger - \mathbf{R_{YX}}\mathbf{A}^\dagger - \mathbf{A}\mathbf{R_{XY}} + \mathbf{R_Y}.$$

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 $\mathbf{A} = \mathbf{R_{YX}}\mathbf{R_X^{-1}}$. Hence, the LMMSE estimator of Y is

$$\hat{\mathbf{Y}} = \mathbf{R_{YX}}\mathbf{R_X^{-1}}\mathbf{X}.$$

The resulting mean square error is

$$MSE = tr\ (\mathbf{R_Y} - \mathbf{R_{YX}}\mathbf{R_X^{-1}}\mathbf{R_{XY}}).$$

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,

$$E[\tilde{\mathbf{Y}}] = E[\tilde{\mathbf{X}}] = \mathbf{0},$$
$$\mathbf{R}_{\tilde{\mathbf{Y}}} = \mathbf{K}_{\tilde{\mathbf{Y}}} = \mathbf{K_Y},$$
$$\mathbf{R}_{\tilde{\mathbf{X}}} = \mathbf{K}_{\tilde{\mathbf{X}}} = \mathbf{K_X}.$$

We know

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

or

$$\hat{\tilde{\mathbf{Y}}} = \mathbf{K_{YX}}\mathbf{K_X^{-1}}\tilde{\mathbf{X}}.$$

Now

$$\hat{\mathbf{Y}} = \hat{\tilde{\mathbf{Y}}} + \mu_{\mathbf{Y}} = \mathbf{K_{YX}}\mathbf{K_X^{-1}}\tilde{\mathbf{X}} + \mu_{\mathbf{Y}}$$
$$= \mathbf{K_{YX}}\mathbf{K_X^{-1}}(\mathbf{X} - \mu_{\mathbf{X}}) + \mu_{\mathbf{Y}}$$
$$= \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mathbf{X} - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mu_{\mathbf{X}} + \mu_{\mathbf{Y}}.$$

So

$$\mathbf{A} = \mathbf{K_{YX}}\mathbf{K_X^{-1}},$$
$$\mathbf{m} = -\mathbf{K_{YX}}\mathbf{K_X^{-1}}\mu_{\mathbf{X}} + \mu_{\mathbf{Y}}.$$

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

$$\mathbf{W} = (\mathbf{Y} - \mu_\mathbf{Y}) - \mathbf{K_{YX}}\mathbf{K_X^{-1}}(\mathbf{X} - \mu_\mathbf{X}).$$

Then

$$\begin{pmatrix} \mathbf{W} \\ \mathbf{X} \end{pmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{K_{YX}}\mathbf{K_X^{-1}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} + \begin{pmatrix} -\mu_\mathbf{Y} + \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mu_\mathbf{X} \\ \mathbf{0} \end{pmatrix}.$$

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

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

$$Cov(\mathbf{W},\ \mathbf{X}) = Cov(\mathbf{W},\ \mathbf{X} - \mu_\mathbf{X}) = E\left[\mathbf{W}(\mathbf{X} - \mu_\mathbf{X})^\dagger\right]$$

$$= E\left[(\mathbf{Y} - \mu_\mathbf{Y})(\mathbf{X} - \mu_\mathbf{X})^\dagger - \mathbf{K_{YX}}\mathbf{K_X^{-1}}(\mathbf{X} - \mu_\mathbf{X})(\mathbf{X} - \mu_\mathbf{X})^\dagger\right]$$

$$= \mathbf{K_{YX}} - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mathbf{K_X} = \mathbf{0}.$$

Now write

$$\mathbf{Y} = (\mathbf{Y} - \mathbf{W}) + \mathbf{W} = \left(\mu_\mathbf{Y} + \mathbf{K_{YX}}\mathbf{K_X^{-1}}(\mathbf{X} - \mu_\mathbf{X})\right) + \mathbf{W}.$$

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,

$$\mathcal{L}\left(\mu_{\mathbf{Y}} + \mathbf{K_{YX}}\mathbf{K_X^{-1}}\left(\mathbf{X} - \mu_{\mathbf{X}}\right)|\mathbf{X} = \mathbf{x}\right) = \text{constant}$$

$$= \mu_{\mathbf{Y}} + \mathbf{K_{YX}}\mathbf{K_X^{-1}}\left(\mathbf{x} - \mu_{\mathbf{X}}\right).$$

So

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

which is normally distributed as

$$\sim \mathcal{N}\left(\mu_{\mathbf{Y}} + \mathbf{K_{YX}}\mathbf{K_X^{-1}}\left(\mathbf{x} - \mu_{\mathbf{X}}\right),\ \text{Cov}(\mathbf{W})\right).$$

Now

$$Cov(\mathbf{W}) = E\left[\mathbf{W}\mathbf{W}^{\dagger}\right]$$

$$= E\left[\left[\left(\mathbf{Y} - \mu_{\mathbf{Y}}\right) - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\left(\mathbf{X} - \mu_{\mathbf{X}}\right)\right]\left[\left(\mathbf{Y} - \mu_{\mathbf{Y}}\right) - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\left(\mathbf{X} - \mu_{\mathbf{X}}\right)\right]^{\dagger}\right]$$

$$= \mathbf{K_Y} - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mathbf{K_{XY}} - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mathbf{K_{XY}} + \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mathbf{K_X}\mathbf{K_X^{-1}}\mathbf{K_{XY}}$$

$$= \mathbf{K_Y} - \mathbf{K_{YX}}\mathbf{K_X^{-1}}\mathbf{K_{XY}}.$$

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

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

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

12.7 Least Squares Estimation

Let

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

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

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

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

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

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

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

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 Likelihhod Estimation (MLE)

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

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

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

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

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 θ.