Chapter 1

ESTIMATION THEORY

1.1 Estimation of Random Variables

Suppose $X, Y_1, Y_2, ..., Y_n$ are random variables defined on the same probability space $(\Omega, \mathcal{S}, P)$. We consider $Y_1, ..., Y_n$ to be the observed random variables and X to be the random variable to be estimated. An *estimator* of X in terms of $Y_1, ..., Y_n$ is a random variable which is a function of $Y_1, ..., Y_n$, say $g(Y_1, Y_2, ..., Y_n)$. If the observed values of $Y_1, ..., Y_n$ are $y_1, y_2, ..., y_n$ respectively, then an *estimate* of X is given by $g(y_1, y_2, ..., y_n)$.

In order to decide how good an estimator is compared to another, we need some criterion which measures the closeness of an estimator to the true random variable X. The criterion we shall use is the mean square error criterion, formulated as follows.

Let $\varepsilon_g = X - g(Y_1, Y_2, ..., Y_n)$. ε_g is the *estimation error* and is itself a random variable. Thus $E\varepsilon_g^2$ is the mean square error and is a number which depends on the choice of the estimator g. The *minimum mean square error estimator g_0*, or *least squares estimator* for short, is that estimator satisfying the property

$$E\varepsilon_{g_0}^2 \leq E\varepsilon_g^2$$

for any estimator g. The least squares estimator is also called the *optimal estimator* in the least squares sense, or simply the optimal estimator if the estimation criterion is understood to be least squares. Note that the least squares criterion is meaningful only in cases where the random variables involved have finite second moments. We shall always make that implicit assumption.

Theorem 1.1.1: The least squares estimator of X in terms of $Y_1, ..., Y_n$ is given by $E\{X|Y_1, Y_2, ..., Y_n\}$, the conditional expectation of X given $Y_1, Y_2, ..., Y_n$.

Proof: For convenience, write $\hat{X}_g = g(Y_1, Y_2, ..., Y_n)$. Then

$$E\varepsilon_g^2 = E\{X - \hat{X}_g + \hat{X}_{g_0} - \hat{X}_{g_0}\}^2$$

where

$$\hat{X}_{g_0} = E\{X|Y_1, Y_2, ..., Y_n\}$$

Expanding the square, we get

$$E\varepsilon_g^2 = E(X - \hat{X}_{g_0})^2 + 2E\{(X - \hat{X}_{g_0})(\hat{X}_{g_0} - \hat{X}_g)\} + E(\hat{X}_{g_0} - \hat{X}_g)^2$$

Now

$$
\begin{aligned}
E\{(X - \hat{X}_{g_0})(\hat{X}_{g_0} - \hat{X}_g)\} &= E\{E[(X - \hat{X}_{g_0})(\hat{X}_{g_0} - \hat{X}_g)|Y_1, ..., Y_n]\} \\
&= E\{(\hat{X}_{g_0} - \hat{X}_g)E[(X - \hat{X}_{g_0})|Y_1, Y_2, ..., Y_n]\} = 0
\end{aligned}
$$

Thus

$$E\varepsilon_g^2 = E(X - \hat{X}_{g_0})^2 + E(\hat{X}_{g_0} - \hat{X}_g)^2 \geq E(X - \hat{X}_{g_0})^2$$

which proves the theorem.

Example 1.1.1

Consider 2 random variables X, Y with a joint density function given by

$$
\begin{aligned}
f_{X,Y}(x,y) &= \lambda^2 e^{-\lambda x} \quad 0 \leq y \leq x \\
&= 0 \quad otherwise
\end{aligned}
$$

To find $E(X|Y)$, we proceed as follows.

(a) Find the marginal density function of Y:

$$
\begin{aligned}
f_Y(y) &= \int_y^\infty \lambda^2 e^{-\lambda x} dx \\
&= -\lambda e^{-\lambda x} \Big|_y^\infty \\
&= \lambda e^{-\lambda y} \quad 0 \leq y < \infty
\end{aligned}
$$

(b) Find the conditional density function of X given Y:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \lambda e^{-\lambda(x-y)} \quad 0 \leq y \leq x$$

(c) Determine $E(X|Y = y)$ using

$$
\begin{aligned}
E(X|Y = y) &= \int x f_{X|Y}(x|y) dx \\
&= \int_y^\infty x \lambda e^{-\lambda(x-y)} dx \\
&= e^{\lambda y} \int_y^\infty x \lambda e^{-\lambda x} dx \\
&= e^{\lambda y} [-xe^{-\lambda x}\Big|_y^\infty + \int_y^\infty e^{-\lambda x} dx] \\
&= e^{\lambda y} [ye^{-\lambda y} + (-\frac{1}{\lambda}e^{-\lambda x})\Big|_y^\infty] \\
&= y + \frac{1}{\lambda}
\end{aligned}
$$

(d) Finally, write down $E(X|Y)$ by replacing the variable y in the expression for $E(X|Y = y)$ with the random variable Y:

$$E(X|Y) = Y + \frac{1}{\lambda}$$

The above example has a conditional expectation $E(X|Y)$ which is an affine (linear plus constant) function of Y. In general, the conditional expectation can be any nonlinear function of Y. Here are some further examples.

Example 1.1.2

Consider 2 random variables X, Y with a joint density function

$$f_{X,Y}(x,y) = xe^{-x(y+1)} \quad x,y \geq 0$$

We would like to determine $E(X|Y)$. Calculations involving integration of exponential functions can often be easily carried out using the following function

$$g(\lambda) = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}$$

Note that

$$-\frac{d}{d\lambda}g(\lambda) = \int_0^\infty xe^{-\lambda x} dx = \frac{1}{\lambda^2}$$

Also,

$$\frac{d^2}{d\lambda^2}g(\lambda) = \int_0^\infty x^2 e^{-\lambda x} dx = \frac{2}{\lambda^3}$$

Using these results, we obtain

$$f_Y(y) = \frac{1}{(y+1)^2}$$

Hence

$$f_{X|Y}(x|y) = (y+1)^2 xe^{-x(y+1)} \quad x,y \geq 0$$

This gives

$$E(X|Y = y) = \int_0^\infty (y+1)^2 x^2 e^{-x(y+1)} dx = \frac{2}{y+1}$$

Finally

$$E(X|Y) = \frac{2}{Y+1}$$

Example 1.1.3:

Let X be a random variable uniformly distributed on $[0\ 1]$. The density function of X is given by

$$f_X(x) = 1, \quad 0 \leq x \leq 1$$

Let V be an exponential distributed random variable with parameter 1, i.e., its density function is given by

$$f_V(v) = e^{-v}, \quad v \geq 0$$

Assume X and V are independent. Let the observed random variable Y be given by

$$Y = \log\frac{1}{X} + V$$

We would like to find $E(X|Y)$.

First, we need to find the joint density of (X, Y), starting from the joint density of (X, V). By independence, we have

$$f_{X,V}(x, v) = e^{-v}, \quad 0 \le x \le 1, \ v \ge 0$$

Define the one-to-one transformation T mapping (X, V) to (X, Y) by

$$(X, Y) = T(X, V) = (X, \log \frac{1}{X} + V)$$

Clearly, the inverse of T is given by

$$(X, V) = T^{-1}(X, Y) = (X, Y - \log \frac{1}{X}) = (X, Y + \log X)$$

The Jacobian matrix is given by

$$\frac{\partial(x, v)}{\partial(x, y)} = \begin{bmatrix} 1 & 0 \\ \frac{1}{x} & 1 \end{bmatrix}$$

Hence, the absolute value of the determinant of the Jacobian matrix is 1. We therefore get

$$f_{X,Y}(x, y) = f_{X,V}(x, y + \log(x)) = e^{-(y + \log x)} = \frac{1}{x} e^{-y}, \quad 0 \le x \le 1, \ y \ge -\log x$$

The constraints on the values of x and y can be combined to give

$$
\begin{aligned}
f_{X,Y}(x, y) &= \frac{1}{x} e^{-y}, \quad e^{-y} \le x \le 1, \ y \ge 0 \\
&= 0, \quad otherwise
\end{aligned}
$$

From here, we can get the marginal density of Y:

$$
\begin{aligned}
f_Y(y) &= \int_{e^{-y}}^{1} \frac{1}{x} e^{-y} dx \\
&= y e^{-y}, \quad y \ge 0
\end{aligned}
$$

Hence the conditional density of X given Y is given by

$$f_{X|Y}(x|y) = \frac{\frac{1}{x} e^{-y}}{y e^{-y}} = \frac{1}{xy}, \quad e^{-y} \le x \le 1, \ y \ge 0$$

The conditional expectation is then given by

$$E(X|Y = y) = \int_{e^{-y}}^{1} x f_{X|Y}(x|y) dx = \int_{e^{-y}}^{1} \frac{1}{y} dx = \frac{1 - e^{-y}}{y}$$

Finally, we obtain

$$E(X|Y) = \frac{1 - e^{-Y}}{Y}$$

In principle, then, the problem of optimal estimation in the least squares sense is solved. All we need to do is to compute the conditional expectation $E\{X|Y_1, Y_2, ..., Y_n\}$. There are, however, a number of difficulties:

(i) The calculation of $E\{X|Y_1, Y_2, ..., Y_n\}$ requires the knowledge of the joint distribution function of $X, Y_1, Y_2, ..., Y_n$. This may not be available as a priori knowledge.

(ii) Even if the joint distribution is known, the conditional expectation is in general a complicated nonlinear function of the observations $Y_1, ..., Y_n$ and may have no analytical formula.

Example 1.1.4:

Let X be a Gaussian random variable with mean m and variance σ^2. Let N be a uniformly distributed random variable with density function

$$f_N(n) = \frac{1}{2b} \quad -b \le n \le b$$

Furthermore X and N are independent. Set $Y = X + N$. We want to determine $E(X|Y)$. We find

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} f_N(y-x) dx \\
&= \frac{1}{2b} \int_{y-b}^{y+b} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx
\end{aligned}
$$

We know from mathematical tables that the integral

$$\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

is the normalized Gaussian distribution function and does not have a closed form expression. Hence there is no closed form expression for $E(X|Y)$ as well.

Other estimators can be considered. For example, the maximum aposteriori estimate, or map estimate for short, is the value of x which maximizes the conditional density $f_{X|Y}(x|y)$.

$$\hat{x}_{map} = arg\max_x f_{X|Y}(x|y)$$

The motivation for the map estimate is that it is the most likely value associated with the conditional density. Since the value of $\hat{x}_{map}$ varies as a function of y, this gives rise to a function of the form $\hat{x}_{map} = \eta(y)$. The map estimator is then defined to be

$$\hat{X}_{map} = \eta(Y)$$

In practice, when the maximimum occurs in the interior of an interval, the maximization is often carried out by finding the roots of the equation

$$\frac{\partial \log f_{X|Y}(x|y)}{\partial x} = 0$$

For example 1.1.3, the maximum of the conditional density $f_{X|Y}(x|y)$ occurs at the boundary, when $x = e^{-y}$. Hence the map estimator is given by

$$\hat{X}_{map} = e^{-Y}$$

Note that the map estimator is also generally a nonlinear function of Y and has similar drawbacks to the conditional mean. In the map estimator case, it also requires determination of the conditional density. As well, the solution of the maximization problem may not result in an analytical expression for $\hat{x}_{map}$.

It is therefore of interest to find estimators which do not require as much prior knowledge about the random variables, which are easy and simple to implement, and which are still good (though not optimal) in some way. This leads us to the study of linear least squares estimators.

1.2   Linear Least Squares Estimation

Let us assume $E|X|^2 < \infty$, $E|Y_i|^2 < \infty$, all i, and that we know or can calculate the mean and covariances of $X, Y_1, ..., Y_n$. There is then no loss of generality in assuming that $EX = EY_i = 0$ for all i, and we shall do so in this section. The general formula for nonzero random vectors is given in Section 1.6.

A linear estimator is simply one which is linear in $Y_1, Y_2, ..., Y_n$, i.e.

$$g(Y_1, Y_2, ..., Y_n) = \sum_{i=1}^{n} \alpha_i Y_i$$

for some scalars $\{\alpha_i\}_{i=1}^n$. The estimation error ε_ℓ is now given by

$$\varepsilon_\ell = X - \sum_{i=1}^{n} \alpha_i Y_i$$

The *linear least squares estimator* is defined to be that linear estimator such that $E\varepsilon_\ell^2$ is minimized.

Let us introduce some notation. The subspace spanned by $Y_1, Y_2, ..., Y_n$, denoted by $\mathcal{L}(Y_1, Y_2, ..., Y_n)$, is defined by

$$\mathcal{L}(Y_1, Y_2, ..., Y_n) = \left\{ \sum_{i=1}^{n} \beta_i Y_i | \beta_i \in \mathbb{R}, \quad i = 1, 2, ..., n \right\}$$

A linear estimator is then an element of $\mathcal{L}(Y_1, Y_2, ..., Y_n)$. A linear least squares estimator is an element of $\mathcal{L}(Y_1, Y_2, ..., Y_n)$ which minimizes $E\varepsilon_\ell^2$ over all elements in $\mathcal{L}(Y_1, Y_2, ..., Y_n)$.

We can now characterize linear least squares estimators.

Theorem 1.2.1: Let $\hat{X} \in \mathcal{L}(Y_1, Y_2, ..., Y_n)$. Then $\hat{X}$ is a linear least squares estimator of X if and only if

$$E(X - \hat{X})Y_i = 0 \qquad \text{for} \quad i = 1, 2, ..., n \tag{1.2.1}$$

(or equivalently, $E(X - \hat{X})Z = 0$ for $Z \in \mathcal{L}(Y_1, Y_2, ..., Y_n)$).

Proof: Suppose (1.2.1) is satisfied. Let Z be any other linear estimator. Then $E(X - Z)^2 = E(X - \hat{X} + \hat{X} - Z)^2$. Since $Z \in \mathcal{L}(Y_1, Y_2, ..., Y_n)$, $\hat{X} - Z$ is also. Thus

$$E(X - \hat{X} + \hat{X} - Z)^2 = E(X - \hat{X})^2 + E(\hat{X} - Z)^2 \geq E(X - \hat{X})^2$$

so that $\hat{X}$ is a linear least squares estimator.

Conversely, suppose $\hat{X}$ is a linear least squares estimator. We need to show that (1.2.1) holds. We proceed by contradiction. Suppose for some i, (1.2.1) is not satisfied. Set

$$Z = \hat{X} + \frac{E(X - \hat{X})Y_i}{EY_i^2} Y_i$$

Since $\hat{X}$ is linear least squares estimator,

$$E(X - Z)^2 \geq E(X - \hat{X})^2$$

But by direct computation,

$$
\begin{aligned}
E(X-Z)^2 &= E\left(X - \hat{X} - \frac{E(X-\hat{X})Y_i}{EY_i^2}Y_i\right)^2 \\
&= E(X-\hat{X})^2 - \frac{[E(X-\hat{X})Y_i]^2}{EY_i^2} < E(X-\hat{X})^2, \quad \text{a contradiction.}
\end{aligned}
$$

This shows that (1.2.1) must be satisfied.

1.3 Geometric Interpretation of Linear Least Squares Estimators

The characterization of the linear least squares estimators in Theorem 1.2.1 can be given a geometric interpretation.

Consider the class of random variables X defined on the probability space $(\Omega, \mathcal{S}, P)$ such that $EX = 0$, $EX^2 < \infty$. This class of random variables are called *second order*. It is easily verified (Ex.) that the space of second order random variables is a vector space. We endow this vector space, denoted $\tilde{\mathcal{H}}$, with the following inner product:

$$\langle X, Y \rangle_{\tilde{\mathcal{H}}} = EXY$$

The inner product induces a notion of length and distance in $\tilde{\mathcal{H}}$. We define the norm of $X \in \tilde{\mathcal{H}}$ by

$$\|X\|_{\tilde{\mathcal{H}}} = [\langle X, X \rangle_{\tilde{\mathcal{H}}}]^{1/2}$$

and the distance between two elements X and Y by

$$d(X, Y) = \|X - Y\|_{\tilde{\mathcal{H}}}$$

In this language, the problem of finding a linear least squares estimator boils down to finding an element $\hat{X} \in \mathcal{L}(Y_1, ..., Y_n)$, which is a subspace of $\tilde{\mathcal{H}}$, such that among the elements of $\mathcal{L}(Y_1, ..., Y_n)$, $\hat{X}$ is closest to X in terms of the distance defined on $\tilde{\mathcal{H}}$. Theorem 1.2.1 then tells us that a linear least squares estimator is characterized by

(i) $\hat{X}$ belongs to $\mathcal{L}(Y_1, Y_2, ..., Y_n)$.

(ii) $X - \hat{X}$ is orthogonal to $\mathcal{L}(Y_1, ..., Y_n)$ in the sense that $\langle X - \hat{X}, Z \rangle_{\tilde{\mathcal{H}}} = 0$ for $Z \in \mathcal{L}(Y_1, ..., Y_n)$.

Any $\hat{X}$ satisfying (i) and (ii) is called the *orthogonal projection* of X onto $\mathcal{L}(Y_1, Y_2, ..., Y_n)$.

Pictorially, we can visualize the situation as follows (taking $n = 1$)
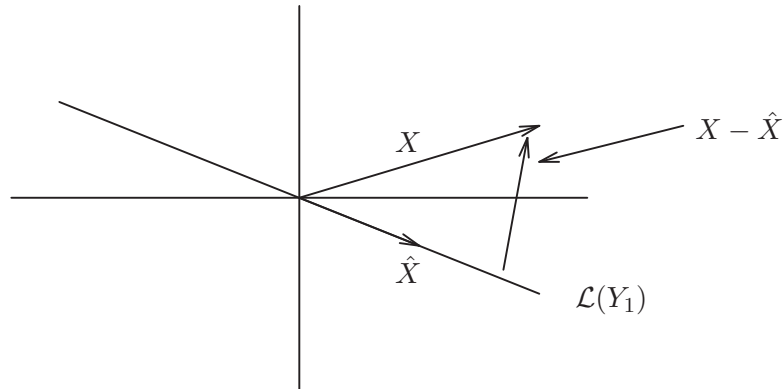
We obtain $\hat{X}$ by dropping the perpendicular from X onto $\mathcal{L}(Y_1)$. Hence the name orthogonal projection.

1.4 The Normal Equation

To determine the linear least squares estimator explicitly, we apply Theorem 1.2.1 as follows:

Let $\alpha^T = [\alpha_1 ... \alpha_n]$, $Y^T = [Y_1 ... Y_n]$. Then $\hat{X} = \alpha^T Y$. Equation (1.2.1) can now be written as

$$E(X - \alpha^T Y)Y^T = 0 \tag{1.4.1}$$

Equation (1.4.1) is referred to as the normal equation.

If $E(YY^T) > 0$, we get

$$\begin{aligned} \alpha^T &= E(XY^T)E(YY^T)^{-1} \\ &= \text{cov}(X,Y)\text{cov}(Y)^{-1} \end{aligned}$$

The l.l.s. estimate is thus given by

$$\hat{X} = \text{cov}(X,Y)\text{cov}(Y)^{-1}Y \ .$$

It can be shown that the normal equation

$$\alpha^T E(YY^T) = EXY^T$$

always has a solution even if $E(YY^T)$ is not positive definite.

1.5 Estimation of One Random Vector in Terms of Another

We shall now generalize that situation to one of estimating one random vector in terms of another, not necessarily of the same dimension. Let X be an n-dimensional random vector, the one to be estimated, and Y be an m-dimensional random vector, the one observed. We wish to construct a linear estimator $\hat{X}$ such that

$$E\|X - \hat{X}\|^2 = E\left\{ \sum_{i=1}^{n} (X_i - \hat{X}_i)^2 \right\}$$

is minimized. It is easy to see that this problem is really that of n sub-problems of estimating the various components X_i in terms of Y. We have seen that the solutions of these sub-problems are given by the Projection Theorem. We now explicitly characterize the linear estimator $\hat{X}$.

We shall assume, as before, that $EX = EY = 0$ and $E\|X\|^2 < \infty$, $E\|Y\|^2 < \infty$. Then we can write

$$\hat{X}_i = \sum_{j=1}^{m} a_{ij} Y_j$$

By the projection theorem, we must have

$$E(X_i - \hat{X}_i)Y_j = 0 \qquad\qquad j = 1, ..., m$$

Using the inner product introduced in Section 1.3 and dropping the subscript $\tilde{\mathcal{H}}$ for convenience, we get

$$\langle X_i, Y_j \rangle = \sum_{k=1}^{m} a_{ik} \langle Y_k, Y_j \rangle \qquad\qquad j = 1, ..., m$$

This gives

$$[\langle X_i, Y_1\rangle \langle X_i, Y_2\rangle, ..., \langle X_i, Y_m\rangle] = [a_{i1}, a_{i2}, ..., a_{im}] \begin{bmatrix} \langle Y_1, Y_1\rangle & \langle Y_1, Y_2\rangle & \cdots & \langle Y_1, Y_m\rangle \\ \langle Y_2, Y_1\rangle & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \langle Y_m, Y_1\rangle & \cdots & \cdots & \langle Y_m, Y_m\rangle \end{bmatrix}$$

$$= [a_{i1}, ..., a_{im}]E(YY^T)$$

If we now write the equations in matrix form for the various values of i, we have

$$\begin{bmatrix} \langle X_1, Y_1\rangle & \cdots & \langle X_1, Y_m\rangle \\ \vdots & & \vdots \\ \langle X_n, Y_1\rangle & \cdots & \langle X_n, Y_m\rangle \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & & \cdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} E(YY^T)$$

or

$$AE(YY^T) = E(XY^T)$$

If we assume that $E(YY^T)$ is invertible, then the unique A is given by

$$A = E(XY^T)E(YY^T)^{-1}$$

Since

$$\hat{X} = AY$$

we get

$$\hat{X} = E(XY^T)E(YY^T)^{-1}Y \tag{1.5.1}$$

1.6 Extensions

(i) If the random variables involved are not of zero mean, we can add the trivial random variable $Y_0 = 1$. Applying the Projection Theorem with Y_0 included, we find

$$\hat{X} = m_X + \text{cov}(X, Y)\text{cov}(Y)^{-1}(Y - m_Y) \tag{1.6.1}$$

where by $\text{cov}(X, Y)$ we mean $E\{(X - m_X)(Y - m_Y)^T\}$. Although $\hat{X}$ is an affine (linear plus constant) function of Y, we still call it the **linear least squares** estimator.

(ii) If X and Y are jointly Gaussian, the minimum mean square error estimator of X in terms of Y is in fact linear in Y. Since (1.6.1) gives the best linear estimator, in the Gaussian case, it is also the minimum mean square error estimator.

(iii) The linear least squares (l.l.s.) estimate also has the following property:

If $\hat{X}$ is the l.l.s. estimator of X, then $T\hat{X}$ is the l.l.s. estimator of TX. To prove this, assume for simplicity that EX=0, EY=0. Suppose KY is the l.l.s. estimator of TX. By the application of the Projection Theorem, we obtain

$$E(TX - KY)Y^T = 0$$
$$\therefore \quad K = TE(XY^T)E(YY^T)^{-1}$$

so that

$$KY = T\hat{X} .$$

(Exercise: Extend this to the nonzero mean case)

Example 1.6.1:

Suppose X and Y are jointly Gaussian, each with zero mean and variance 1, and $EXY = \rho$. The l.l.s. estimate of X in terms of Y is then given by

$$\hat{X} = \rho Y$$

which is the same as the conditional mean $E(X|Y)$.

Example 1.6.2:

Consider 2 random variable X, Y with a joint density function

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda x} \quad 0 \le y \le x$$

In Example 1.1.1, we computed $E(X|Y)$. Here, we determine the l.l.s. estimate of X based on Y.

(a) Determine m_X:

$$f_X(x) = \int_0^x \lambda^2 e^{-\lambda x} dy = \lambda^2 x e^{-\lambda x}$$

Thus,

$$
\begin{aligned}
m_X &= \int_0^\infty \lambda^2 x^2 e^{-\lambda x} dx \\
&= \lambda^2 \frac{2}{\lambda^3} = \frac{2}{\lambda}
\end{aligned}
$$

(b) Determine m_Y: We know from Example 1.1.1 that

$$f_Y(y) = \lambda e^{-\lambda y} \quad 0 \le y < \infty$$

Hence $m_Y = \frac{1}{\lambda}$.

(c) Determine $cov(Y) = EY^2 - m_Y^2$:

$$EY^2 = \int_0^\infty \lambda y^2 e^{-\lambda y} dy = \frac{2}{\lambda^2}$$

Hence

$$cov(Y) = \frac{1}{\lambda^2}$$

(d) Determine $cov(X,Y) = EXY - m_X m_Y$:

$$
\begin{aligned}
EXY &= \int_0^\infty \int_0^x \lambda^2 xy e^{-\lambda x} dx dy \\
&= \int_0^\infty \lambda^2 \frac{x^3}{2} e^{-\lambda x} dx \\
&= \frac{3}{\lambda^2}
\end{aligned}
$$

Hence

$$cov(X,Y) = \frac{3}{\lambda^2} - \frac{2}{\lambda^2} = \frac{1}{\lambda^2}$$

(e) Finally,

$$\hat{X} = \frac{2}{\lambda} + \frac{1}{\lambda^2}(\frac{1}{\lambda^2})^{-1}(Y - \frac{1}{\lambda}) = Y + \frac{1}{\lambda}$$

We see that the l.l.s. estimator is the same as $E(X|Y)$, which is to be expected since $E(X|Y)$ was seen to be an affine function of Y.

1.7 Adding One Observation to Improve Estimation

The previous sections study the problem of estimating one random vector X in terms of another random vector Y. Both X and Y do not change. In this section, we re-examine the estimation problem from a different point of view. We consider the information available for estimation as increasing, with $Y_1, Y_2, \cdots, Y_n$, the components of Y, being observed sequentially. We shall show that this viewpoint leads naturally to recursive updates of the estimate of X.

We first consider the following simple situation: Assume all random variables are zero mean, and we make two observations, Y_1, and Y_2, about the random variable X. By the previous results, the l.l.s. estimator is given by

$$\hat{X} = a_1 Y_1 + a_2 Y_2$$

where a_1 and a_2 satisfy the equations

$$\begin{bmatrix} \langle X, Y_1 \rangle \\ \langle X, Y_2 \rangle \end{bmatrix} = \begin{bmatrix} \langle Y_1, Y_1 \rangle & \langle Y_2, Y_1 \rangle \\ \langle Y_1, Y_2 \rangle & \langle Y_2, Y_2 \rangle \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{1.7.1}$$

The solution of (1.7.1) would be trivial if Y_1 and Y_2 were uncorrelated or orthogonal. For then $\langle Y_1, Y_2 \rangle = 0$ and

$$\begin{aligned} a_1 &= \frac{\langle X, Y_1 \rangle}{\langle Y_1, Y_1 \rangle} \\ a_2 &= \frac{\langle X, Y_2 \rangle}{\langle Y_2, Y_2 \rangle} \end{aligned}$$

Recalling the interpretation that $\hat{X}$ is simply an element in $\mathcal{L}(Y_1, Y_2)$, this suggests that we seek $\tilde{Y}_1$ and $\tilde{Y}_2$ such that $\mathcal{L}(Y_1, Y_2) = \mathcal{L}(\tilde{Y}_1, \tilde{Y}_2)$ and that $\tilde{Y}_1 \perp \tilde{Y}_2$. Geometrically, this corresponds to two coordinate axes which are orthogonal.
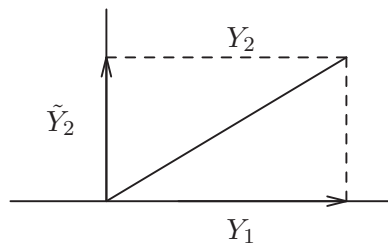
One way of doing this is the following: we take $\tilde{Y}_1 = Y_1$. Now the orthogonal projection of Y_2 onto Y_1 is given by

$$\frac{\langle Y_2, Y_1 \rangle}{\langle Y_1, Y_1 \rangle} Y_1$$

and we know from the Projection Theorem that

$$Y_2 - \frac{\langle Y_2, Y_1 \rangle}{\langle Y_1, Y_1 \rangle} Y_1$$

is orthogonal to Y_1. Now clearly if we define

$$\tilde{Y}_2 = Y_2 - \frac{\langle Y_2, Y_1 \rangle}{\langle Y_1, Y_1 \rangle} Y_1$$

then $\mathcal{L}(Y_1, Y_2) = \mathcal{L}(Y_1, \tilde{Y}_2)$ and $Y_1 \perp \tilde{Y}_2$. We may therefore seek

$$\hat{X} = b_1 Y_1 + b_2 \tilde{Y}_2$$

By the Projection Theorem

$$
\left[\begin{array}{c} \langle X, Y_1 \rangle \\ \langle X, \tilde{Y}_2 \rangle \end{array} \right]
= \left[\begin{array}{cc} \langle Y_1, Y_1 \rangle & \langle \tilde{Y}_2, Y_1 \rangle \\ \langle Y_1, \tilde{Y}_2 \rangle & \langle \tilde{Y}_2, \tilde{Y}_2 \rangle \end{array} \right]
\left[\begin{array}{c} b_1 \\ b_2 \end{array} \right]
$$

$$
= \left[\begin{array}{cc} \langle Y_1, Y_1 \rangle & 0 \\ 0 & \langle \tilde{Y}_2, \tilde{Y}_2 \rangle \end{array} \right]
\left[\begin{array}{c} b_1 \\ b_2 \end{array} \right] \qquad \text{by orthogonality}
$$

So

$$b_1 = \frac{\langle X, Y_1 \rangle}{\langle Y_1, Y_1 \rangle}$$

$$b_2 = \frac{\langle X, \tilde{Y}_2 \rangle}{\langle \tilde{Y}_2, \tilde{Y}_2 \rangle}$$

An alternative characterization of $\hat{X}$ is then given by

$$\hat{X} = \frac{\langle X, Y_1 \rangle}{\langle Y_1, Y_1 \rangle} Y_1 + \frac{\langle X, \tilde{Y}_2 \rangle}{\langle \tilde{Y}_2, \tilde{Y}_2 \rangle} \tilde{Y}_2 \tag{1.7.2}$$

The first term on the right hand side of (1.7.2) is recognized as the orthogonal projection of X onto Y_1 alone. We can thus interpret the above result as follows: If we first make the observation Y_1, then the best estimate of X is given by $\hat{X}_1 = \frac{\langle X, Y_1 \rangle}{\langle Y_1, Y_1 \rangle} Y_1$. If we add a second observation Y_2, the new best estimate is given as the sum of the old estimate $\hat{X}_1$ and the best estimate of X in terms of $\tilde{Y}_2$ alone. $\tilde{Y}_2$ may therefore be considered as the new information contained in the additional observation Y_2.

We now extend these results to zero mean random vectors. The nonzero mean case requires only minor modifications. If we define, in the case where X, Y_1, Y_2 are vectors, $\langle X, Y \rangle = EXY^T$, then we can straightforwardly generalize (1.7.2) to the vector case

$$\hat{X} = \langle X, Y_1 \rangle \langle Y_1, Y_1 \rangle^{-1} Y_1 + \langle X, \tilde{Y}_2 \rangle \langle \tilde{Y}_2, \tilde{Y}_2 \rangle^{-1} \tilde{Y}_2 \tag{1.7.3}$$

The change in the error covariance can also be easily evaluated. Writing

$$\hat{X} = \hat{X}_1 + \hat{X}_2 \qquad \text{where} \qquad \hat{X}_2 = \langle X, \tilde{Y}_2 \rangle \langle \tilde{Y}_2, \tilde{Y}_2 \rangle^{-1} \tilde{Y}_2$$

we have

$$
\begin{aligned}
E(X - \hat{X})(X - \hat{X})^T &= E\{(X - \hat{X}_1 - \hat{X}_2)(X - \hat{X}_1 - \hat{X}_2)^T\} \\
&= E(X - \hat{X}_1)(X - \hat{X}_1)^T - EX\hat{X}_2^T - E\hat{X}_2 X^T + E\hat{X}_2 \hat{X}_2^T
\end{aligned}
$$

using the orthogonality of $\hat{X}_1$ and $\hat{X}_2$.

Now

$$
\begin{aligned}
E\{X\hat{X}_2^T\} &= E(X\tilde{Y}_2^T)\langle \tilde{Y}_2, \tilde{Y}_2 \rangle^{-1} E(\tilde{Y}_2 X^T) \\
&= E\{\hat{X}_2 X^T\} = E\hat{X}_2 \hat{X}_2^T
\end{aligned}
$$

So

$$E(X - \hat{X})(X - \hat{X})^T = E(X - \hat{X}_1)(X - \hat{X}_1)^T - E(X\tilde{Y}_2^T)E(\tilde{Y}_2\tilde{Y}_2^T)^{-1}E(\tilde{Y}_2 X^T) \tag{1.7.4}$$

Let the error covariances associated with $\hat{X}$ and $\hat{X}_1$ be P and P_1 respectively. Then

$$P_1 - P = E(X\tilde{Y}_2^T)E(\tilde{Y}_2\tilde{Y}_2^T)^{-1}E(\tilde{Y}_2 X^T) \tag{1.7.5}$$

which is a positive semidefinite matrix.

This means that by adding a measurement, we would improve our estimate in the sense that the error covariance would be reduced. This is of course intuitively reasonable.

We specialize the above results now to the following situation:

Suppose from previous observations, we have formed the estimate $\hat{X}_1$, whose error covariance is given by P_1. The additional measurement Y is related to X in the form

$$Y = CX + V$$

where X and V are orthogonal, with V zero mean, $EVV^T = R > 0$ (i.e. R is positive definite), and also orthogonal to the past observations. To find the updated estimate, we first construct the innovation of Y. Let the orthogonal projection of Y onto the past observations be given by $\mathcal{P}Y$. Then

$$\mathcal{P}Y = \mathcal{P}(CX + V) = \mathcal{P}CX + \mathcal{P}V = C\hat{X}_1$$

Thus the innovation $\tilde{Y}$ is given by

$$\tilde{Y} = Y - C\hat{X}_1$$

From the previous analysis,

$$\hat{X} = \hat{X}_1 + E(X\tilde{Y}^T)E(\tilde{Y}\tilde{Y}^T)^{-1}\tilde{Y}$$

But

$$\begin{aligned} EX\tilde{Y}^T &= E[X(Y - C\hat{X}_1)^T] = E\{X[(X - \hat{X}_1)^T C^T + V^T]\} \\ &= P_1 C^T \\ E(\tilde{Y}\tilde{Y}^T) &= E\{C(X - \hat{X}_1)(X - \hat{X}_1)^T C^T + V(X - \hat{X}_1)^T C^T + C(X - \hat{X}_1)V^T + VV^T\} \end{aligned}$$

By the orthogonality between V and X, and V and past observations, we get, on letting $EVV^T = R$

$$E(\tilde{Y}\tilde{Y}^T) = CP_1 C^T + R$$

Thus the updated estimate $\hat{X}$ is given by

$$\hat{X} = \hat{X}_1 + P_1 C^T (CP_1 C^T + R)^{-1}(Y - C\hat{X}_1) \tag{1.7.6}$$

From (1.7.5), we also see that the updated error covariance P is given by

$$P = P_1 - P_1 C^T (CP_1 C^T + R)^{-1}CP_1 \tag{1.7.7}$$

We shall see in Chapter 3 that by combining the above results with linear system dynamics will give us the Kalman filter.

1.8   Least Squares Parameter Estimation

The Projection Theorem has many applications. It can be used to solve the least squares parameter estimation problem. The problem can be formulated as follows.

Let Y be a given m-vector, possibly random. Let Φ be an $m \times n$ matrix, with possibly random entries. The columns of Φ are referred to as regression vectors. The least sqaures parameter estimation problem is to determine a *nonrandom* n-vector θ, such that the least squares criterion $\|Y - \Phi\theta\|^2 = (Y - \Phi\theta)^T(Y - \Phi\theta)$ is minimized. The problem can be interpreted as finding the best approximation, in the sense of shortest Euclidean distance, of Y by linear combinations of the columns of Φ (which is $\Phi\theta$). This optimal parameter which minimizes the criterion is called the least squares parameter estimate and is denoted by $\hat{\theta}$.

We can use the Projection Theorem to characterize the optimal choice for θ. By the Projection Theorem, the optimal error $Y - \Phi\hat{\theta}$ must be orthogonal to the columns of Φ. This implies that the following equation holds:

$$\Phi^T(Y - \Phi\hat{\theta}) = 0$$

The least squares estimate $\hat{\theta}$ therefores satisfies the normal equation

$$\Phi^T\Phi\hat{\theta} = \Phi^TY \tag{1.8.1}$$

If $\Phi^T\Phi$ is invertible, we can solve for $\hat{\theta}$ explicitly to get

$$\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^TY \tag{1.8.2}$$

This result does not depend on the way Y is defined. If we assume more knowledge on how Y is generated, we can get more detailed results on the properties of the least squares estimate. Suppose Y satisfies

$$Y = \Phi\theta + V$$

where V is a zero mean "noise" vector independent of the regressor Φ. Here, the interpretation is that there is a "true" parameter θ which, together with the additive noise V, gives the observation Y. Assume that $\Phi^T\Phi$ is invertible. We then have

$$\hat{\theta} = \theta + (\Phi^T\Phi)^{-1}\Phi^TV$$

Since V is zero mean and independent of Φ,

$$E\hat{\theta} = \theta \tag{1.8.3}$$

Note that the property given in (1.8.3) is true, regardless of what the parameter value for θ is.

A parameter estimator is called unbiased if its expectation is equal to the true value of the parameter. Thus the least squares parameter estimate, under the assumption of independence of Φ and V, is unbiased.

The least squares estimate can be easily generalized to a weighted least squares criterion. Let Q be a positive semidefinite matrix. Define the weighted least sqaures criterion to be

$$J(\theta) = (Y - \Phi\theta)^TQ(Y - \Phi\theta) \tag{1.8.4}$$

By factoring $Q = Q^{\frac{1}{2}}Q^{\frac{1}{2}}$, (1.8.4) can be expressed as the least squares criterion $(Q^{\frac{1}{2}}Y - Q^{\frac{1}{2}}\Phi\theta)^T(Q^{\frac{1}{2}}Y - Q^{\frac{1}{2}}\Phi\theta)$. Using previous results, we find that the θ which minimizes $J(\theta)$ is given by the solution of the normal equation

$$\Phi^TQ\Phi\hat{\theta} = \Phi^TQY \tag{1.8.5}$$

If $\Phi^TQ\Phi$ is nonsingular, then the unique weighted least squares estimate is given by

$$\hat{\theta} = (\Phi^TQ\Phi)^{-1}\Phi^TQY \tag{1.8.6}$$

Example 1.8.1:

Let
$$y_i = b + v_i \quad i = 1, \cdots, N$$
where b is a constant, and $Ev_i = 0$, all i. Thus, b is the mean of Y_i. Put
$$Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T$$
$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_N \end{bmatrix}^T$$
and
$$\Phi = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T$$
Then
$$\Phi^T \Phi = N$$
$$\Phi^T Y = \sum_{i=1}^{N} y_i$$
so that the least squares estimate is given by
$$\hat{b} = \frac{1}{N} \sum_{i=1}^{N} y_i$$
which is the arithmetic mean of the y_i's. We can also express
$$\hat{b} = b + \frac{1}{N} \sum_{i=1}^{N} v_i$$
so that $\hat{b}$ is unbiased. If we assume further that $Ev_iv_j = 0$ for $i \neq j$, and $Ev_i^2 = \sigma^2$, all i, then
$$E(\hat{b} - b)^2 = \frac{\sigma^2}{N}$$
which converges to 0 as $N \to \infty$.

Exercises

1. Let X and Y be independent random variables with densities

$$
\begin{aligned}
f_X(x) &= \alpha e^{-\alpha x} & x \geq 0 & \quad \alpha > 0 \\
&= 0 & x < 0 & \\
f_Y(y) &= \beta e^{-\beta y} & y \geq 0 & \quad \beta > 0 \\
&= 0 & y < 0 &
\end{aligned}
$$

   Let $Z = X + Y$.

   (i) For $\alpha \neq \beta$, find the conditional density function $f_{X|Z}(x|z)$ (Hint: Use the transformation of densities formula to determine first the joint density $f_{X,Z}(x,z)$). Verify that it is a probability density function. Determine $E(X|Z)$. Verify that $E[E(X|Z)] = E(X)$.

   (ii) Suppose $\alpha = \beta$. Repeat the calculations of (i). Verify that the same results can be obtained from (i) by taking the limit as $\alpha \to \beta$.

2. Consider the random variables X, Y, and Z as described in problem 1.

   (i) For $\alpha = 1$, $\beta = 2$, find the linear least squares (l.l.s.) estimator of X given Z (Remember that if the random variables are not zero mean, the l.l.s. estimator takes the form $aZ + b$).

   (ii) Suppose $\alpha = \beta$. Again find the l.l.s. estimator of X given Z and compare it to $E(X/Z)$ determined in problem 2. Can you generalize this result for the sum of 2 independent exponentially distributed random variables to the situation involving $Z = \sum_{i=1}^{n} X_i$, X_i's independent, identically distributed, but not necessarily exponential?

3. Let X and Y have a joint density

$$
f_{X,Y}(x,y) = \frac{1}{x} \quad 0 \leq y \leq x \leq 1
$$

   Determine the linear least square estimate of X based on Y.

4. Let Y and V be independent zero mean Gaussian random variables with variance 1. Let $X = Y^2 + V$.

   (i) Determine the optimal least squares estimator $\hat{X} = E(X|Y)$, and find the mean square estimation error $E(X - \hat{X})^2$.

   (ii) Determine the linear least squares estimator $\hat{X}_l$ in the form $aY + b$, and find the mean square estimation error $E(X - \hat{X}_l)^2$.

   (iii) Compare the results of (i) and (ii) and discuss.

5. In Example 1.1.3, we found $E(X|Y)$ using the probabilistic information provided about (X, V) and how Y is related to (X, V). For the same probabilistic information, find the linear least squares estimate of X based on Y. Sketch $E(X|Y)$ and the l.l.s. estimator for large Y, and comment on the differences in behaviour, if any, between the two estimators.

6. Consider the $(n + p)$-vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$ where X and Y are jointly Gaussian random vectors of dimension n and p respectively. The joint density of X and Y is given by

$$
f_{X,Y}(x,y) = \frac{1}{(2\pi)^{\frac{n+p}{2}} |\det \Sigma|^{1/2}} e^{-\frac{1}{2}(z-m)^T \Sigma^{-1}(z-m)} \tag{ex3.1}
$$

where $m = EZ = \begin{bmatrix} m_X \\ m_Y \end{bmatrix}$ and $\Sigma = E(Z - m)(Z - m)^T = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}$ and $z = \begin{bmatrix} x \\ y \end{bmatrix}$. The density of Y is therefore given by

$$f_Y(y) = \frac{1}{(2\pi)^{\frac{p}{2}}|\det \Sigma_Y|^{1/2}} e^{-\frac{1}{2}(y - m_Y)^T \Sigma_Y^{-1}(y - m_Y)} \qquad (ex3.2)$$

(i) Write down the conditional density $f_{X|Y}(x|y)$ using (ex3.1) and (ex3.2). Denote the exponent of $f_{X|Y}(x|y)$ by $-\frac{1}{2}J_{X|Y}$.

(ii) Verify that for any invertible matrix P,

$$J_{X/Y} = (z - m)^T P^T (P\Sigma P^T)^{-1} P(z - m) - (y - m_Y)^T \Sigma_Y^{-1}(y - m_Y)$$

find $P(z - m)$ and $P\Sigma P^T$ when P is chosen to be

$$\begin{bmatrix} I & -\Sigma_{XY}\Sigma_Y^{-1} \\ 0 & I \end{bmatrix}$$

(iii) Show from here that $f_{X|Y}(x|y)$ has the form of a Gaussian density. Find $E(X|Y)$ and the conditional covariance $E\{[X - E(X|Y)][X - E(X|Y)]^T|_Y\}$. Does the conditional covariance depend on Y?

7. (a) You are given 2 random variables, Y_1 and Y_2, each with zero mean and with second order statistics $EY_1^2 = 1$, $EY_1Y_2 = 2$, and $EY_2^2 = 4$. You are asked to find the linear least squares estimate of a zero mean random variable X with $EXY_1 = 1$ and $EXY_2 = 5$. Explain why this is not a meaningful problem.

(Hint: Examine the covariance matrix of the random vector $Y = [Y_1 \ Y_2]^T$.)

(b) You are given 2 random variables, Y_1 and Y_2, each with zero mean and with second order statistics $EY_1^2 = 1$, $EY_1Y_2 = 2$, and $EY_2^2 = 4$. You are asked to find the linear least squares estimate of a zero mean random variable X with $EXY_1 = 1$ and $EXY_2 = 2$. Note that with $Y = [Y_1 \ Y_2]^T$, $E(YY^T)$ is singular, and that the situation is very similar to that of part (a), except that the present problem is meaningful and consistent.

  (i) Even though $E(YY^T)$ is singular, it can be shown that a solution to the equation

$$\alpha^T E(YY^T) = E(XY^T)$$

  always exists. Find the general solution for α for the particular X, Y given above. This should be a one parameter family of solutions.

  (ii) Determine the linear least squares estimate of X in terms of Y. Show that the l.l.s. estimate $\hat{X}$ does not depend on the specific value of the free parameter in part (a).

  (iii) Now consider determining the l.l.s. estimate $\hat{X}$ by sequentially processing Y_1, and then Y_2. What is $\hat{X}_1$, the l.l.s. estimate of X based on Y_1 alone? Update $\hat{X}_1$ to $\hat{X}_2$ by processing Y_2 also. Compare $\hat{X}_2$ to $\hat{X}_1$ and to $\hat{X}$ in part (b). Explain your results.

8. We have developed 2 formulas for the l.l.s. estimator, one using the complete observation vector all at once, as in Section 1.5, the other one using the observation vector sequentially, as in Section 1.7. Consider 3 zero mean random variables X, Y_1, and Y_2. Show explicitly the 2 formulas for computing the l.l.s of X based on Y_1 and Y_2 give the same answer.

9. Consider the signal in noise problem

$$Y_i = X + V_i \qquad i = 1, ..., n$$

with $EX = EV_i = 0$, all i, $EX^2 = \sigma_X^2$, $EV_i^2 = \sigma_V^2$, $EXV_i = 0$, all i, and $EV_iV_j = 0$, for $i \neq j$. The problem is to find the linear least squares estimate of X based on $Y = [Y_1 \cdots Y_n]^T$. Let the optimal estimate be given by

$$\hat{X} = \sum_{i=1}^{n} \alpha_i Y_i$$

Determine and solve the normal equation satisfied by $\alpha_1 \cdots \alpha_n$. What is the form of $\hat{X}$ as $n \to \infty$? (Hint: Examine the form of the normal equation and guess its solution.)

10. In this problem, we establish some additional useful results for the weighted least squares parameter estimate. Let

$$J(\theta) = (Y - \Phi\theta)^T Q (Y - \Phi\theta)$$

(a) Show that for any θ and $\hat{\theta}$, the following equation holds:

$$J(\theta) - J(\hat{\theta}) = 2(\hat{\theta} - \theta)^T (\Phi^T QY - \Phi^T Q\Phi\hat{\theta}) + (\theta - \hat{\theta})^T \Phi^T Q\Phi(\theta - \hat{\theta}) \qquad (ex8.1)$$

From the above equation, we see that if θ satisfies the normal equation, i.e., it corresponds to the least squares estimate, we obtain

$$J(\theta) - J(\hat{\theta}) = (\theta - \hat{\theta})^T \Phi^T Q\Phi(\theta - \hat{\theta}) \geq 0, \quad \text{for any } \theta$$

(b) Conversely, if $\hat{\theta}$ minimizes $J(\theta)$, it must satisfy the normal equation. To prove this directly from (ex8.1), assume that

$$\Phi^T QY - \Phi^T Q\Phi\hat{\theta} \neq 0$$

Choose θ in the form

$$\hat{\theta} - \theta = -\alpha(\Phi^T QY - \Phi^T Q\Phi\hat{\theta})$$

Show that for α sufficiently small, we get $J(\theta) - J(\hat{\theta}) < 0$, so that $\hat{\theta}$ cannot be optimal.