Chapter 4

Estimation

4.1 Introduction

Estimation, as considered here, involves a probabilistic experiment with two random vectors (r$\vec{v}$'s) $\vec{X}$ and $\vec{Y}$. The experiment is performed, but only the resulting sample value $\vec{y}$ of the r$\vec{v}$ $\vec{Y}$ is observed. The observer then "estimates" the unknown sample value $\vec{x}$ of $\vec{X}$ from the observation $\vec{y}$. For simplicity, we assume that the r$\vec{v}$'s have continuous distribution functions with finite probability densities and conditional densities. The marginal density $p_{\vec{X}}(\vec{x})$ is called the *a priori density* of $\vec{X}$ and the conditional density $p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})$ is called the *a posteriori density*. The a posteriori density $p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})$ gives us the conditional distribution of $\vec{X}$, given $\vec{Y} = \vec{y}$. The observer, given the observed $\vec{y}$ and knowing the a posteriori density, chooses an estimate, $\widehat{X}(\vec{y})$ of the unknown sample value $\vec{x}$. That estimate might be the conditional mean, the conditional median, or the conditional mode of the a posteriori distribution, or might be any other function of $\vec{y}$.

Estimation problems occur in an amazing variety of situations, and are often referred to as measurement problems or recovery problems. For example, in communication systems, the timing and the phase of the transmitted signals must be *recovered* at the receiver. Often it is necessary to *measure* the channel, and finally, for analog data, the receiver must *estimate* the transmitted waveform at finely spaced sampling times. In control systems, the state of the system must be *estimated* in order to generate appropriate control signals. In statistics, we try to *estimate* parameters for a probabilistic model from sample values of the system being modeled. In any experimental science, one is always concerned with *measuring* quantities in the presence of noise and experimental error.

The problem of estimation is very similar to that of detection. With detection, we must decide between a finite set of alternatives on the basis of an observation, whereas here we make a selection from a vector continuum of choices. Although this does not appear to be a fundamental difference, it leads to a surprising set of differences in approach. In many typical detection problems, the cost of different kinds of errors is secondary and we are concerned primarily with maximizing the probability of correct choice. In typical estimation problems, with a continuum of alternatives, the probability of selecting the exact correct value is zero; thus we can not avoid questioning what types of errors are most costly.

A fundamental approach to estimation is to use a cost function, $C(\vec{x}\,', \vec{x})$, to quantify the cost associated with an estimate $\vec{x}\,'$ when the true sample value is $\vec{x}$. This cost function, $C(\vec{x}\,', \vec{x})$, is analogous to the cost C_{ij}, defined in Chapter 3, of making decision i when j is the correct hypothesis. The *minimum cost criterion* or *Bayes criterion* for estimation is to choose $\widehat{X}(\vec{y})$, for the observed $\vec{y}$, to minimize the expected cost conditional on $\vec{Y} = \vec{y}$. Specifically, for each observation $\vec{y}$, $\widehat{X}(\vec{y})$ is chosen to minimize

$$\int C\left(\widehat{X}(\vec{y}), \vec{x}\right) p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})\, d\vec{x} = E\left[C[\widehat{X}(\vec{y}), \vec{X}] \mid \vec{Y}{=}\vec{y}\right]$$

Thus, this minimum cost estimate $\widehat{X}(\vec{y})$ has the property that, for any function $g(\vec{y})$,

$$E\left[C\left(\widehat{X}(\vec{y}), \vec{X}\right) \Big| \vec{Y}{=}\vec{y}\right] \le E\left[C[g(\vec{y}), \vec{X}] \Big| Y{=}\vec{y}\right] \qquad \text{for all } \vec{y} \tag{4.1}$$

Another way to express $\widehat{X}(\vec{y})$, for each $\vec{y}$, is as follows:

$$\widehat{X}(\vec{y}) = \arg\min_{\vec{x}\,'} \int_{\vec{x}} C\left[\vec{x}\,', \vec{x}\right] p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})\, d\vec{x} = \arg\min_{\vec{x}\,'} E\left[C[\vec{x}\,', \vec{X}] \Big| \vec{Y}{=}\vec{y}\right] \tag{4.2}$$

The arg min of a function is, by definition, the value of an argument that minimizes the function. Now, suppose we multiply both sides of (4.1) by $p_{\vec{Y}}(\vec{y})$ and integrate over $\vec{y}$. We obtain

$$E\left[C\left(\widehat{X}(\vec{Y}), \vec{X}\right)\right] \le E\left[C\left(g(\vec{Y}), \vec{X}\right)\right] \tag{4.3}$$

Thus the minimum cost estimate not only minimizes the expected cost for each sample value $\vec{y}$ but also minimizes the overall expected cost.

In (4.3), the minimum cost estimation problem is expressed as minimizing the expected cost of the r$\vec{v}$ $\widehat{X}(\vec{Y})$ relative to the r$\vec{v}$ $\vec{X}$. Viewed this way, the mathematical problem of finding the minimum cost estimate is formulated strictly in terms of the probability model and does not depend on any notion of actually performing an experiment. Because of this we often refer to estimating the r$\vec{v}$ $\vec{X}$ from $\vec{Y}$. This point of view can sometimes be confusing, since the r$\vec{v}$ $\vec{X}$ is a mapping from the underlying sample space to real vectors $\vec{x}$, and this mapping is assumed to be known. What is unknown is the sample value $\vec{x}$ of $\vec{X}$ corresponding to the observation $\vec{y}$.

There are other kinds of estimation problems, briefly discussed in section 4.7, where we do not assume a full probabilistic description of the r$\vec{v}$s. There we assume that there is an unknown underlying probability model, and that an estimate must be chosen that does not depend on the unknown parts of the model.

4.1.1 The Squared Cost Function

In practice, the most important cost function is the squared cost function,

$$C[\vec{x}\,', \vec{x})] \overset{\text{def}}{=} \sum_i [x_i' - x_i]^2 \tag{4.4}$$

The estimate $\widehat{X}(\vec{y})$ that satisfies (4.1) or, equivalently, (4.2) for the squared cost function is called the *Minimum Mean Square Error Estimate (MMSE Estimate)*. It is also often called the Bayes least squares estimate. In order to minimize $E\left[\sum_i [\widehat{X}_i(\vec{y}) - X_i]^2 \mid \vec{Y}=\vec{y}\right]$ over $\widehat{X}(\vec{y})$, it is sufficient to choose $\widehat{X}_i(\vec{y})$, for each i, to minimize $E\left[[\widehat{X}_i(\vec{y}) - X_i]^2 \mid \vec{Y}=\vec{y}\right]$. Note that $E\left[(\widehat{X}_i(y) - X_i)^2 \mid \vec{Y}=\vec{y}\right]$ is simply the second moment of X_i around $\widehat{X}_i(\vec{y})$, conditioned on $\vec{Y} = \vec{y}$. Recall that the second moment of an arbitrary random variable $U = E[U] + \widetilde{U}$ around an arbitrary number α satisfies

$$E[(U-\alpha)^2] = E[(\widetilde{U} + E[U] - \alpha)^2] = E[\widetilde{U}^2] + (E[U] - \alpha)^2 \tag{4.5}$$

This is minimized over α when $\alpha = E[U]$, i.e., the second moment of a random variable around α is minimized by choosing α equal to the mean. Here we are dealing with the second moment of the distribution *conditional* on a given observed sample value $\vec{Y} = \vec{y}$, so the second moment is minimized when $\widehat{X}_i(\vec{y})$ is the conditional mean[1] of X_i, conditional on $\vec{Y} = \vec{y}$. Thus, the MMSE of X_i, given $\vec{Y} = \vec{y}$, is given by $\widehat{X}_i(\vec{y}) = E[X_i \mid \vec{Y}=\vec{y}]$. Since the MMSE estimate of $\vec{X}$ is simply the vector of MMSE estimates of the components X_i, $\widehat{X}(\vec{y})$ is simply the conditional mean of $\vec{X}$, conditional on $\vec{Y} = \vec{y}$. Simple though this result is, it is a central result of estimation theory, and we state it as a theorem.

Theorem 4.1 *The MMSE estimate $\widehat{X}(\vec{y})$, as a function of the observation $\vec{Y} = \vec{y}$, is given by*

$$\widehat{X}(\vec{y}) = E\left[\vec{X} \mid \vec{Y}=\vec{y}\right] = \int \vec{x}\, p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})\, d\vec{x} \tag{4.6}$$

Define the *estimation error* $\vec{\Xi}$ as $\widehat{X}(\vec{Y}) - \vec{X}$. We shall be interested not only in the MMSE estimate, but also in $E[\vec{\Xi}\vec{\Xi}^T]$, which is the correlation matrix of $\vec{\Xi}$. Note that the mean square error, $E[\vec{\Xi}^T\vec{\Xi}]$ is the trace of the correlation matrix. Since $\widehat{X}(\vec{y})$ is the conditional mean of $\vec{X}$ conditional on $\vec{Y} = \vec{y}$, we see that $E[\vec{\Xi} \mid \vec{Y}=\vec{y}] = 0$ for all $\vec{y}$, and thus[2] $E[\vec{\Xi}] = 0$. The correlation matrix and covariance matrix of $\vec{\Xi}$ are thus the same and given by

$$K_{\vec{\Xi}} = E\left[\vec{\Xi}\,\vec{\Xi}^T\right] = E\left[\left(\widehat{X}(\vec{Y}) - \vec{X}\right)\left(\widehat{X}(\vec{Y}) - \vec{X}\right)^T\right] \tag{4.7}$$

This can be simplified by recalling that $E\left[\vec{\Xi} \mid \vec{Y}=\vec{y}\right] = 0$ for all $\vec{y}$. Thus if $g(\vec{y})$ is any vector valued function of the same dimension as $\vec{x}$ and $\vec{\Xi}$, then

$$E\left[\vec{\Xi} \mid \vec{Y}=\vec{y}\right] g^T(\vec{y}) = E\left[\vec{\Xi}\, g^T(\vec{Y}) \mid \vec{Y}=\vec{y}\right] = 0$$

[1] There is a very important, albeit simple, principle involved here. When we condition on something ($\vec{Y} = \vec{y}$), we are restricting the sample space to those points for which $\vec{Y} = \vec{y}$, and renormalizing the probabilities to sum to 1. This leaves us with a new probability space, but everything we know about probabilities and random variables is still true in this new space. Mathematical purists will (and rightfully should) be concerned about measurability in the new space, but we will ignore this, since it is usually unimportant in sensibly modeled situations.

[2] As we will explain in section 4.7, $E[\vec{\Xi}]$ is the expected value of the estimate bias. The bias of an estimate $\widehat{X}(\vec{Y})$ is defined there as $E[\widehat{X}(\vec{Y}) - \vec{x} \mid \vec{X} = \vec{x}]$. This is a function of the sample value $\vec{x}$. Bias and expected bias are often confused in the literature.

Averaging over $\vec{y}$, we get the important relation that the MMSE estimation error and any $g(\vec{y})$ satisfies

$$E\left[\vec{\Xi}\, g^T(\vec{Y})\right] = 0 \tag{4.8}$$

In subsection 4.3.4, we will interpret this equation as an orthogonality principle. For now, since $g(\vec{y})$ is an arbitrary function (of the right dimension), we can replace it with $\widehat{X}(\vec{y})$, getting $E\left[\vec{\Xi}\,\widehat{X}^T(\vec{Y})\right] = 0$. Substituting this into (4.7), we finally get a useful simplified expression for the covariance of the estimation error for MMSE estimation,

$$K_{\vec{\Xi}} = -E\left[\vec{\Xi}\,\vec{X}^T\right] = K_{\vec{X}} - E\left[\widehat{X}(\vec{Y})\,\vec{X}^T\right] \tag{4.9}$$

4.1.2 Other Cost Functions

Subsequent sections of this chapter focus primarily on the squared cost function. First, however, we briefly discuss several other cost functions. One is the absolute value cost function, $C(\vec{x}\,', \vec{x}) = \sum_i |x_i' - x_i|$; this expected cost is minimized, for each $\vec{y}$, by choosing $\widehat{X}_i(\vec{y})$ to be the conditional median of $p_{X_i|\vec{Y}}(x_i \mid \vec{y})$ (see Exercise 1.1). The absolute value cost function weighs large estimation errors more lightly than the squared cost function. The reason for the greater importance of the squared cost function, however, has less to do with the relative importance of large errors than with the conceptual richness and computational ease of working with means and variances rather than medians and absolute errors.

Another cost function is the maximum error cost function. Here, for some given number ϵ, the cost is 1 if the magnitude of any component of the error exceeds ϵ and is 0 otherwise. That is, $C(\vec{x}\,', \vec{x}) = 1$ if $|x_i' - x_i| > \epsilon$ for any i and $C(\vec{x}\,', \vec{x}) = 0$ otherwise. For any observation $\vec{y}$, this expected cost is minimized by that value $\widehat{X}(\vec{y})$ that maximizes the conditional probability that $\vec{X}$ lies in a cube with sides of length 2ϵ centered on $\widehat{X}(\vec{y})$.

For ϵ very small, this maximum error cost function is approximated by choosing $\widehat{X}(\vec{y})$ to be that $\vec{x}$ that maximizes $p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})$ (i.e., for given $\vec{y}$, one chooses the mode of $p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})$). This estimation rule is called the *Maximum A posteriori Probability (MAP) estimate.* Thus, the MAP estimate, denoted $\widehat{X}_{MAP}(\vec{Y})$, is given by[3]

$$\widehat{X}_{MAP}(\vec{Y}) = \arg\max_{\vec{x}} p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y}) \tag{4.10}$$

One awkward property of MAP estimation is that, in some cases, the MAP estimate of x_i is different from the i^{th} component of the MAP estimate of $\vec{x} = (x_1, \ldots, x_m)$ (see Exercise 4.2).

In studying detection, we focused on the MAP rule, whereas here we focus on MMSE. To see the reason for this, note that in the typical case where $p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})$ is finite, the probability of choosing the correct sample value $\vec{x}$ exactly is equal to zero. Thus, in principle, attempting to maximize the probability of correct choice is foolish (since that probability will be zero

[3] In what follows, we use a subscript on an estimate, such as $\widehat{X}_{MAP}$, to be explicit about the type of estimate beng used. We usually omit the subscript for MMSE estimates unless necessary for explicitness.

anyway). The reason the MAP rule is often used is, first, analytical convenience, and second, some confidence, for the particular problem at hand, that the conditional mode is a reasonable choice. For many situations, particularly those with Gaussian statistics, MMSE, MAP, and minimum absolute error estimates are all equivalent.

A more fundamental question in estimation theory is whether it is reasonable to assume a complete probabilistic model. A popular kind of model is that in which $p_{\vec{Y}|\vec{X}}(\vec{y} \mid \vec{x})$ is specified, but one is unwilling or unable to specify $p_{\vec{X}}(\vec{x})$. For example, $\vec{X}$ might model a set of parameters about which almost nothing is known, whereas $\vec{Y}$ might be the sum of $\vec{X}$ plus statistically well characterized noise. In these situations, we shall still regard $\vec{X}$ as a random vector, but look at estimates that do not depend on $p_{\vec{X}}(\vec{x})$. This will allow us both to compare the two approaches easily and to employ the methodologies of probabilistic models. The most common estimate that does not depend on $p_{\vec{X}}(\vec{x})$ is called the *maximum likelihood (ML) estimate*; the maximum likelihood estimate, $\widehat{X}_{ML}(\vec{y})$, is defined by

$$\widehat{X}_{ML}(\vec{y}) = \arg\max_{\vec{x}} p_{\vec{Y}|\vec{X}}(\vec{y} \mid \vec{x}) \tag{4.11}$$

We can view the ML estimate as a limit of MAP estimates, taking the limit as the a priori density on $\vec{X}$ approaches a constant. For example, we could model $\vec{X} \sim \mathcal{N}(0, \sigma^2 I)$ in the limit as $\sigma^2 \to \infty$. This limiting density does not exist, since the density approaches 0 everywhere, but the MAP estimate typically will approach a limit. All of the previous comments about MAP estimates clearly carry over to ML.

4.2 MMSE Estimation for Jointly Gaussian Random Vectors

Minimum mean square error estimation turns out to be particularly simple when the observed rv $\vec{Y}$ and the rv $\vec{X}$ to be estimated are jointly Gaussian. One of the simplifications is that the estimate and its error depend only on the means and joint covariances of $\vec{X}$ and $\vec{Y}$. We say that $\vec{X}$ and $\vec{Y}$ are *jointly non-singular* if the covariance matrix $K_{\vec{Z}}$ of $\vec{Z}^T = (\vec{X}^T, \vec{Y}^T)$ is non-singular. *Throughout this chapter, except where explicitly stated otherwise, we assume that $\vec{X}$ and $\vec{Y}$ are jointly non-singular.* For $\vec{X}$ and $\vec{Y}$ jointly non-singular, jointly Gaussian, and zero mean, we saw in Theorem 2.3 that $\vec{X}$ can always be represented as $\vec{X} = G\vec{Y} + \vec{V}$ where $\vec{V}$ is a zero mean Gaussian rv independent of $\vec{Y}$ and where

$$G = K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1}; \quad K_{\vec{V}} = K_{\vec{X}} - K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} K_{\vec{X}\vec{Y}}^T \tag{4.12}$$

This means that $\vec{X}$, conditional on $\vec{Y}=\vec{y}$, is $\mathcal{N}(G\vec{y}, K_{\vec{V}})$, and therefore the conditional mean, $E[\vec{X} \mid \vec{Y}=\vec{y}]$, is equal to $G\vec{y}$. The MMSE estimate, $\widehat{X}(\vec{y})$, given $\vec{Y}=\vec{y}$, is equal to the conditional mean, so

$$\widehat{X}(\vec{y}) = G\vec{y} = K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} \vec{y} \tag{4.13}$$

The estimation error, $\vec{\Xi} = \widehat{X}(\vec{Y}) - \vec{X}$ is then $G\vec{Y} - \vec{X}$, which is $-\vec{V}$. Since $\vec{\Xi} = -\vec{V}$, $\vec{\Xi}$ and $\vec{V}$ have the same covariance, so from (4.12),

$$K_{\vec{\Xi}} = K_{\vec{X}} - K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} K_{\vec{X}\vec{Y}}^T \tag{4.14}$$

As a check, this equation also results from substituting (4.13) into (4.9). The estimate in (4.13) is linear in the observation sample value $\vec{y}$. Also, since $\vec{Y}$ and $\vec{V}$ are independent r$\vec{\text{v}}$'s, the covariance of the estimation error does not depend on the observation $\vec{y}$. That is,

$$K_{\vec{\Xi}} = E\left[(\widehat{X}(\vec{y}) - \vec{X})(\widehat{X}(\vec{y}) - \vec{X})^T \mid \vec{Y}{=}\vec{y}\right] \quad \text{for each } \vec{y} \tag{4.15}$$

These are major simplifications that arise because the variables are jointly Gaussian.

More generally, let $\vec{X}$ and $\vec{Y}$ be jointly Gaussian random vectors with arbitrary means and define the fluctuations, $\vec{U} = \vec{X} - E[\vec{X}]$ and $\vec{Z} = \vec{Y} - E[\vec{Y}]$. The observation $\vec{y}$ of $\vec{Y}$ is equivalent to an observation $\vec{z} = \vec{y} - E[\vec{Y}]$ of $\vec{Z}$. Since $\vec{U}$ and $\vec{Z}$ are zero mean, the MMSE estimate $\widehat{U}$, given $\vec{Z} = \vec{z} = \vec{y} - E[\vec{Y}]$, can be found from (4.13), i.e.,

$$\widehat{U} = K_{\vec{U}\vec{Z}} K_{\vec{Z}}^{-1}(\vec{y} - E[\vec{Y}])$$

Since $K_{\vec{X}\vec{Y}} = E\left[(\vec{X} - E[\vec{X}])(\vec{Y} - E[\vec{Y}])^T\right] = K_{\vec{U}\vec{Z}}$ and, similarly, $K_{\vec{Y}} = K_{\vec{Z}}$, this can be rewritten as

$$\widehat{U} = K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1}(\vec{y} - E[\vec{Y}])$$

Finally, since $\vec{X} = \vec{U} + E[\vec{X}]$, the corresponding estimate of $\vec{X}$, given $\vec{Y} = \vec{y}$, is

$$\widehat{X}(\vec{y}) = E[\vec{X}] + K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1}(\vec{y} - E[\vec{Y}]) \tag{4.16}$$

This can be rewritten as

$$\widehat{X}(\vec{y}) = \vec{b} + K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1}\vec{y} \quad \text{where} \quad \vec{b} = E[\vec{X}] - K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} E[\vec{Y}] \tag{4.17}$$

The error covariance in estimating $\vec{X}$ from $\vec{Y}$ is clearly the same as that in estimating $\vec{U}$ from $\vec{Z}$, so that there is no lack of generality in restricting our attention to zero mean r$\vec{\text{v}}$'s. Note, however, that the estimate in (4.17) is a constant plus a linear function of $\vec{y}$. This is properly called an affine transformation. These results are summarized in the following theorem:

Theorem 4.2 *If $\vec{X}$ and $\vec{Y}$ are jointly non-singular Gaussian random vectors, the MMSE estimate of $\vec{X}$ from the observation $\vec{Y}{=}\vec{y}$ is given by (4.16) which, for the zero mean case, is (4.13). The covariance of the estimation error is given by (4.14), which gives both the covariance conditional on $\vec{Y}{=}\vec{y}$ and the unconditional covariance.*

In the remainder of this section, we look at a number of simple examples of MMSE estimation for jointly Gaussian variables. We start with the simplest scalar cases and work our way up to recursive and Kalman estimation. More general vector examples are given in sections 4.4 and 4.5.

Example 4.1 (Scalar Signal plus Noise) The simplest estimation problem one can imagine is to estimate X from the observation of $Y = X + Z$ where X and Z are zero mean independent Gaussian rv's with variances σ_X^2 and σ_Z^2 (i.e., $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Z \sim \mathcal{N}(0, \sigma_X^2)$).

Since X and Y are zero mean and jointly Gaussian, X, conditional on an observation $Y = y$, is given by (4.13) and (4.14) as $\mathcal{N}(K_{XY}K_Y^{-1}y, K_X - K_{XY}K_Y^{-1}K_{XY}^T)$. Since $Y = X + Z$, we have $K_{XY} = \sigma_X^2$ and $K_Y = \sigma_X^2 + \sigma_Z^2$. From (4.13), the MMSE estimate is

$$\widehat{X}(y) = \frac{K_{XY}}{\sigma_Y^2}y = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}y \tag{4.18}$$

From the symmetry between X and Z,

$$\widehat{Z}(y) = \frac{\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}y \tag{4.19}$$

$\widehat{X}(y) + \widehat{Z}(y) = y$, so the estimation simply splits the observation between signal and noise according to their variances. From an intuitive standpoint, recall that both X and Z are zero mean, so if one has a larger variance than the other, it is reasonable to attribute the major part of Y to the variable with the larger variance, as is done in (4.19). The estimation error, $\Xi = \widehat{X}(Y) - X = Z - \widehat{Z}(Y)$, conditional on $Y = y$, is $\mathcal{N}(0, K_X - K_{XY}K_Y^{-1}K_{XY}^T)$. This does not depend on y, so Ξ is statistically independent of Y. Because of this, the variance of the estimation error, conditional on $Y = y$, i.e., $E[\Xi^2 \mid Y{=}y]$ does not depend on y and is the same as the unconditional variance of the error, which we call σ_Ξ^2. Thus, for all y

$$\begin{aligned}\sigma_\Xi^2 &= E\left[\left(\widehat{X}(y) - X\right)^2 \mid Y{=}y\right] = \sigma_X^2 - \frac{K_{XY}^2}{\sigma_Y^2} \\ &= \sigma_X^2 - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2} = \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}\end{aligned} \tag{4.20}$$

Rearranging (4.19) and (4.20) we get the following forms:

$$\widehat{X}(y) = \frac{\sigma_\Xi^2}{\sigma_Z^2}y \quad ; \qquad \frac{1}{\sigma_\Xi^2} = \frac{1}{\sigma_X^2} + \frac{1}{\sigma_Z^2} \tag{4.21}$$

This is a special case of the alternative forms for conditional means and covariances in (2.50) and (2.46). This expression for σ_Ξ^2 makes it clear that it is smaller than the original variances of X and Z, and that σ_Ξ^2 decreases as either σ_X^2 and σ_Z^2 decrease.

Example 4.2 (Attenuated Scalar Signal plus Noise) Now generalize the above problem to $Y = hX + Z\cdot$where h is a scale factor. X and Z are still zero mean, independent, and Gaussian. The conditional probability $p_{X|Y}(x \mid y)$ for given y is again $\mathcal{N}(K_{XY}K_Y^{-1}y, K_X - K_{XY}K_Y^{-1}K_{XY}^T)$. Now $K_{XY} = h\sigma_X^2$ and $\sigma_Y^2 = h^2\sigma_X^2 + \sigma_Z^2$, so (4.13) and (4.14) become

$$\widehat{X}(y) = \frac{h\sigma_X^2 y}{h^2\sigma_X^2 + \sigma_Z^2} \quad ; \qquad \sigma_\Xi^2 = \sigma_X^2 - \frac{h^2\sigma_X^4}{h^2\sigma_X^2 + \sigma_Z^2} = \frac{\sigma_X^2\sigma_Z^2}{h^2\sigma_X^2 + \sigma_Z^2} \tag{4.22}$$

These expressions can be rearranged to the forms:

$$\widehat{X}(y) = \frac{\sigma_\Xi^2}{\sigma_Z^2}hy \quad ; \qquad \frac{1}{\sigma_\Xi^2} = \frac{1}{\sigma_X^2} + \frac{h^2}{\sigma_Z^2} \tag{4.23}$$

These forms are also special cases of Equations (2.50) and (2.46). It is insightful to view the observation as $Y/h = X + Z/h$. Thus the variance of this scaled noise is σ_Z^2/h^2, which shows that (4.22) and (4.23) follow directly from (4.18), (4.20), and (4.21).

As a final extension, suppose $Y = hX + Z$ and X has a mean $\overline{X}$; i.e., $X \sim \mathcal{N}(\overline{X}, \sigma_X^2)$. Z is still zero mean. Then $E[Y] = h\overline{X}$. Using (4.22) for the fluctuations in X and Y, and using (4.16),

$$\widehat{X}(y) = \overline{X} + \frac{h\sigma_X^2[y - h\overline{X}]}{h^2\sigma_X^2 + \sigma_Z^2} \quad ; \qquad \sigma_\Xi^2 = \frac{\sigma_X^2\sigma_Z^2}{h^2\sigma_X^2 + \sigma_Z^2} \tag{4.24}$$

This can be rearranged as follows:

$$\widehat{X}(y) = \frac{\sigma_\Xi^2}{\sigma_X^2}\overline{X} + \frac{\sigma_\Xi^2}{\sigma_Z^2}\,hy \quad ; \qquad \frac{1}{\sigma_\Xi^2} = \frac{1}{\sigma_X^2} + \frac{h^2}{\sigma_Z^2} \tag{4.25}$$

Example 4.3 (Scalar Recursive Extimation) Suppose that we make multiple noisy observations, $y_1, y_2, \dots$ of a single random variable $X \sim \mathcal{N}(\bar{X}, \sigma_X^2)$. After the i^{th} observation, for each i, we want to make the best estimate of X based on $y_1, \dots, y_i$. We shall see that this can be done recursively, using the estimate based on $y_1, \dots, y_{i-1}$ to help form the estimate based on $y_1, \dots, y_i$. The observation random variables are related to X by

$$Y_i = h_iX + Z_i$$

where $\{Z_i; i \geq 1\}$ and X are independent Gaussian rv's, $Z_i \sim \mathcal{N}(0, \sigma_{Z_i}^2)$. Let $Y_{1;i}$ denote the first i observation random variables, $Y_1, \dots, Y_i$, and let $y_{1;i}$ denote the corresponding sample values $y_1, \dots, y_i$. Let $\widehat{X}(y_{1;i})$ be the MMSE estimate of X based on the observation of $Y_{1;i} = y_{1;i}$ and let $\sigma_{\Xi_i}^2$ be the variance of the estimation error, $\Xi_i = \widehat{X}(Y_{1;i}) - X$.

For $i = 1$, we solved this problem in (4.24); given $Y_1 = y_1$,

$$\widehat{X}(y_1) = \overline{X} + \frac{h_1\sigma_X^2[y_1 - h_1\overline{X}]}{h_1^2\sigma_X^2 + \sigma_{Z_1}^2} \quad ; \qquad \sigma_{\Xi_1}^2 = \frac{\sigma_X^2\sigma_{Z_1}^2}{h_1^2\sigma_X^2 + \sigma_{Z_1}^2} \tag{4.26}$$

The alternate forms, from (4.25) are

$$\widehat{X}(y_1) = \frac{\sigma_{\Xi_1}^2}{\sigma_X^2}\overline{X} + \frac{\sigma_{\Xi_1}^2}{\sigma_{Z_1}^2}\,h_1y_2 \quad ; \qquad \frac{1}{\sigma_{\Xi_1}^2} = \frac{1}{\sigma_X^2} + \frac{h_1^2}{\sigma_{Z_1}^2} \tag{4.27}$$

Conditional on $Y_1 = y_1$, X is Gaussian with mean $\widehat{X}(y_1)$ and variance $\sigma_{\Xi_1}^2$; it is also independent of Z_2, which, conditional on $Y_1 = y_1$ is still $\mathcal{N}(0, \sigma_{Z_2}^2)$. Thus, in the portion of the sample space restricted to $Y_1 = y_1$, and with probabilities conditional on $Y_1 = y_1$, we want to estimate X from $Y_2 = h_2X + Z_2$; this is an instance of (4.24) (here using conditional probabilities in the space $Y_1 = y_1$), with $\widehat{X}(y_1)$ in place of $\overline{X}$, $\sigma_{\Xi_1}^2$ in place of σ_X^2, and h_2 in place of h, so

$$\widehat{X}(y_{1;2}) = \widehat{X}(y_1) + \frac{h_2\sigma_{\Xi_1}^2[y_2 - h_2\widehat{X}(y_1)]}{h_2^2\sigma_{\Xi_1}^2 + \sigma_{Z_2}^2} \quad ; \qquad \sigma_{\Xi_2}^2 = \frac{\sigma_{\Xi_1}^2\sigma_{Z_2}^2}{h_2^2\sigma_{\Xi_1}^2 + \sigma_{Z_2}^2} \tag{4.28}$$

The alternate forms are

$$\widehat{X}(y_{1;2}) = \frac{\sigma^2_{\Xi_2}}{\sigma^2_{\Xi_1}}\widehat{X}(y_1) + \frac{\sigma^2_{\Xi_2}}{\sigma^2_{Z_2}} h_2 y \quad ; \qquad \frac{1}{\sigma^2_{\Xi_2}} = \frac{1}{\sigma^2_{\Xi_1}} + \frac{h_2^2}{\sigma^2_{Z_2}} \tag{4.29}$$

We see that, conditional on $Y_1 = y_1$ and $Y_2 = y_2$, X is Gaussian with mean $\widehat{X}(y_{1;2})$ and variance $\sigma^2_{\Xi_2}$. What we are doing here is first conditioning on $Y_1{=}y_1$, and then, in that conditional space, conditioning on $Y_2{=}y_2$. This is the same as directly conditioning on $Y_1{=}y_1, Y_2{=}y_2$. To see this more concretely, let $q_{XY_2}(xy_2)$, $q_{X|Y_2}(x|y_2)$, and $q_{Y_2}(y_2)$ represent conditional probability densities in the space restricted to $Y_1 = y_1$. Then

$$q_{X|Y_2}(x|y_2) = \frac{q_{XY_2}(xy_2)}{q_{Y_2}(y_2)} = \frac{p_{XY_2|Y_1}(xy_2 \mid y_1)}{p_{Y_2|Y_1}(y_2 \mid y_1)} = p_{X|Y_2Y_1}(x \mid y_2y_1) \tag{4.30}$$

Now we can iterate the argument in (4.28), (4.29) to arbitrary $i > 1$, getting

$$\widehat{X}(y_{1;i}) = \widehat{X}(y_{1;i-1}) + \frac{h_i\sigma^2_{\Xi_{i-1}}[y_i - h_i\widehat{X}(y_{1;i-1})]}{h_i^2\sigma^2_{\Xi_{i-1}} + \sigma^2_{Z_i}} \quad ; \qquad \sigma^2_{\Xi_i} = \frac{\sigma^2_{\Xi_{i-1}}\sigma^2_{Z_i}}{h_i^2\sigma^2_{\Xi_{i-1}} + \sigma^2_{Z_i}} \tag{4.31}$$

Alternately,

$$\widehat{X}(y_{1;i}) = \frac{\sigma^2_{\Xi_i}}{\sigma^2_{\Xi_{i-1}}}\widehat{X}(y_{1;i-1}) + \frac{\sigma^2_{\Xi_i}}{\sigma^2_{Z_i}} h_i y_i \quad ; \qquad \frac{1}{\sigma^2_{\Xi_i}} = \frac{1}{\sigma^2_{\Xi_{i-1}}} + \frac{h_i^2}{\sigma^2_{Z_i}} \tag{4.32}$$

These equations can be combined (see Exercise 4.3) over subsequent observations from 1 to i to yield

$$\widehat{X}(y_{1;i}) = \sigma^2_{\Xi_i}\left[\frac{\overline{X}}{\sigma^2_X} + \sum_{j=1}^{i}\frac{h_jy_j}{\sigma^2_{Z_j}}\right] ; \qquad \frac{1}{\sigma^2_{\Xi_i}} = \frac{1}{\sigma^2_X} + \sum_{j=1}^{i}\frac{h_j^2}{\sigma^2_{Z_j}} \tag{4.33}$$

This is derived in a different way in (4.112) and (4.113).

Assuming that h_j^2 and $\sigma^2_{Z_j}$ are bounded away from 0 and ∞, we see from (4.33) that $\sigma^2_{\Xi_i}$ approaches 0 with increasing i, and thus $\widehat{X}(Y_{i;j})$ converges in mean square to the true x. Also, as i increases, the effect of the a priori mean and variance becomes increasingly small.

Example 4.4 (Scalar Kalman Filter) We now extend the recursive estimation above to the case where X evolves with time. In particular, assume that $X_1 \sim \mathcal{N}(\overline{X}_1, \sigma^2_{X_1})$. For each $i \geq 1$, X_{i+1} evolves from X_i as

$$X_{i+1} = \alpha_i X_i + W_i\,; \qquad W_i \sim \mathcal{N}(0, \sigma^2_{W_i}) \tag{4.34}$$

For each $i \geq 1$, α_i and $\sigma^2_{W_i}$ are known scalars. Noisy observations Y_i are made satisfying

$$Y_i = h_i X_i + Z_i\,; \qquad i \geq 1; \quad Z_i \sim \mathcal{N}(0, \sigma^2_{Z_i}) \tag{4.35}$$

where, for each $i \geq 1$, h_i is a known scalar. Finally, assume that $\{Z_i; i \geq 1\}$, $\{W_i; i > 1\}$, and X_1 are all independent.

For each $i \geq 1$, we want to find the MMSE estimate of both X_i and X_{i+1} conditional on the first i observations, $Y_{1;i} = y_{1;i} = y_1, \dots, y_i$. We denote those estimates as $\widehat{X}_i(y_{1;i})$ and $\widehat{X}_{i+1}(y_{1;i})$. We denote the errors in these estimates as $\Xi_i = \widehat{X}_i(Y_{1;i}) - X_i$ and $\zeta_{i+1} = \widehat{X}_{i+1}(Y_{1;i}) - X_{i+1}$. For $i = 1$, conditional on $Y_1 = y_1$, the solution for $\widehat{X}_1(y_1)$ and the variance $\sigma^2_{\Xi_i}$ of the estimation error is given by (4.26) and (4.27),

$$\widehat{X}_1(y_1) = \overline{X}_1 + \frac{h_1\sigma^2_{X_1}[y_1 - h_1\overline{X}_1]}{h_1^2\sigma^2_{X_1} + \sigma^2_{Z_1}} \quad ; \qquad \sigma^2_{\Xi_1} = \frac{\sigma^2_{X_1}\sigma^2_{Z_1}}{h_1^2\sigma^2_{X_1} + \sigma^2_{Z_1}} \tag{4.36}$$

$$\widehat{X}_1(y_1) = \frac{\sigma^2_{\Xi_1}}{\sigma^2_{X_1}}\overline{X}_1 + \frac{\sigma^2_{\Xi_1}}{\sigma^2_{Z_1}}h_1 y_1 \quad ; \qquad \frac{1}{\sigma^2_{\Xi_1}} = \frac{1}{\sigma^2_{X_1}} + \frac{h_1^2}{\sigma^2_{Z_1}} \tag{4.37}$$

This means that, conditional on $Y_1{=}y_1$, X_1 is Gaussian with mean $\widehat{X}_1(y_1)$ and variance $\sigma^2_{\Xi_1}$. Thus, conditional on $Y_1{=}y_1$, $X_2 = \alpha_1 X_1 + W_1$ is Gaussian with mean $\alpha_1\widehat{X}_1(y_1)$ and variance $\alpha_1^2\sigma^2_{\Xi_1} + \sigma^2_{W_1}$. It follows that the MMSE estimate $\widehat{X}_2(y_1)$ and the error variance $\sigma^2_{\zeta_2}$ for X_2, conditional on $Y_1{=}y_1$, are given by

$$\widehat{X}_2(y_1) = \alpha_1\widehat{X}_1(y_1) \quad ; \qquad \sigma^2_{\zeta_2} = \alpha_1^2\sigma^2_{\Xi_1} + \sigma^2_{W_1} \tag{4.38}$$

In the restricted space $Y_1 = y_1$, we now want to estimate X_2 from the additional observation $Y_2 = y_2$. We can use (4.24) and (4.25) again, with $\widehat{X}_2(y_1)$ in place of $\overline{X}$ and σ_{ζ_2} in place of σ_X

$$\widehat{X}_2(y_{1;2}) = \widehat{X}_2(y_1) + \frac{h_2\sigma^2_{\zeta_2}[y_2 - h_2\widehat{X}_2(y_1)]}{h_2^2\sigma^2_{\zeta_2} + \sigma^2_{Z_2}} \quad ; \qquad \sigma^2_{\Xi_2} = \frac{\sigma^4_{\zeta_2}\sigma^2_{Z_2}}{h_2^2\sigma^2_{\zeta_2} + \sigma^2_{Z_2}} \tag{4.39}$$

$$\widehat{X}_2(y_{1;2}) = \frac{\sigma^2_{\Xi_2}}{\sigma^2_{\zeta_2}}\widehat{X}_2(y_1) + \frac{\sigma^2_{\Xi_2}}{\sigma^2_{Z_2}}h_2 y_2 \quad ; \qquad \frac{1}{\sigma^2_{\Xi_2}} = \frac{1}{\sigma^2_{\zeta_2}} + \frac{h_2^2}{\sigma^2_{Z_2}} \tag{4.40}$$

In general, we can repeat the arguments leading to (4.38, 4.39, 4.40) for each i, leading to

$$\widehat{X}_i(y_{1;i-1}) = \alpha_{i-1}\widehat{X}_{i-1}(y_{1;i-1}) \quad ; \qquad \sigma^2_{\zeta_i} = \alpha^2_{i-1}\sigma^2_{\Xi_{i-1}} + \sigma^2_{W_{i-1}} \tag{4.41}$$

$$\widehat{X}_i(y_{1;i}) = \widehat{X}_i(y_{1;i-1}) + \frac{h_i\sigma^2_{\zeta_i}[y_i - h_i\widehat{X}_i(y_{1;i-1})]}{h_i^2\sigma^2_{\zeta_i} + \sigma^2_{Z_i}} \; ; \qquad \sigma^2_{\Xi_i} = \frac{\sigma^2_{\zeta_i}\sigma^2_{Z_i}}{\sigma^2_{\zeta_i} + \sigma^2_{Z_i}} \tag{4.42}$$

An alternate form to (4.42) is

$$\widehat{X}_i(y_{1;i}) = \frac{\sigma^2_{\Xi_i}}{\sigma^2_{\zeta_i}}\widehat{X}_i(y_{1;i-1}) + \frac{\sigma^2_{\Xi_i}}{\sigma^2_{Z_i}}h_i y_i \quad ; \qquad \frac{1}{\sigma^2_{\Xi_i}} = \frac{1}{\sigma^2_{\zeta_i}} + \frac{h_i^2}{\sigma^2_{Z_i}} \tag{4.43}$$

These are the scalar Kalman filter equations. The idea is that one "filters" the observed values $Y_1, Y_2, \dots$ to generate the MMSE estimate of the "signal" $X_1, X_2, \dots$. The variance terms, $\sigma^2_{\zeta_i}$ and $\sigma^2_{\Xi_i}$ do not depend on the observations, and thus can be precomputed. It can then be seen that the estimates are affine in the observations.

An important special case of this result occurs where h_i, α_i, $\sigma^2_{Z_i}$, and $\sigma^2_{W_i}$ are all independent of i, with $0 < \alpha < 1$. As we will see later, the discrete process $\{X_n; n \geq 1\}$ is a Markov process as well as a Gaussian process, and is usually referred to as Gauss Markov. As n

becomes large, the mean of X_n approaches 0 and the variance approaches $\sigma_W^2/(1-\alpha^2)$. We would expect that if α is very close to 1, we would be able to estimate X_n quite accurately since an independent measurement is taken each unit of time and the process changes very slowly. To investigate the steady state behavior under these circumstances, we substitute the second half of (4.41) into the second half of (4.43), getting

$$\frac{1}{\sigma_{\Xi_i}^2} = \frac{1}{\alpha^2\sigma_{\Xi_{i-1}}^2 + \sigma_W^2} + \frac{h^2}{\sigma_Z^2} \tag{4.44}$$

It is not hard to see that $\sigma_{\Xi_i}^2$ monotonically approaches a limiting value λ, and that λ must satisfy the quadratic equation

$$\alpha^2 h^2 \sigma_Z^{-2} \lambda^2 + \left[h^2 \sigma_W^2 \sigma_Z^{-2} + 1 - \alpha^2\right] \lambda - \sigma_W^2 = 0 \tag{4.45}$$

The simplicity of the results in these last two examples came from two features. The first is that the conditional distribution of X_i, conditional on $Y_{1;i}$ contains all the relevant information about $Y_{1;i}$ for estimation of $X_{i+1}, \dots$. The second is that the conditional distribution is Gaussian and thus specified by mean and variance.

4.3 Linear Least Squares Error Estimation and the Projection Principle

In section 4.2, we saw that, for zero mean jointly Gaussian r$\vec{\text{v}}$s, the MMSE estimate $\widehat{X}(\vec{y})$ is a linear function of the observation $\vec{y}$. For non-Gaussian cases, the MMSE estimate is sometimes difficult to find and messy to compute. For this and several other reasons, we sometimes want to constrain an estimate to be a linear function of the observation, and to minimize the mean square error subject to this constraint. Such an estimate is called a *linear least squares error estimate (LLSE estimate)*. We start by calculating these LLSE estimates strictly as an optimization problem. We then view the problem in terms of a particular abstract linear vector space and view the solution as an orthogonal projection in that space.

4.3.1 LLSE Estimation; zero mean case

As usual, we first look at the zero mean case and then generalize the problem slightly to look at non-zero means. We also look only at the estimation of a single rv from the observation of an n dimensional r$\vec{\text{v}}$ (a random vector can always be estimated by estimating each component separately). In particular, then, let X be a zero mean rv and let $\vec{Y}$ be a zero mean n dimensional r$\vec{\text{v}}$ with a non-singular covariance matrix $K_{\vec{Y}}$. The LLSE estimate of X from $\vec{Y}$ is an estimate $\widehat{X}(\vec{Y})$ of the form

$$\widehat{X}(\vec{Y}) = \vec{\alpha}^T\vec{Y} = \sum_{i=1}^{n} \alpha_i Y_i \tag{4.46}$$

where the vector $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)^T$ is chosen to minimize $E[(\vec{\alpha}^T\vec{Y} - X)^2]$, i.e., to minimize

$$E\left[(\vec{\alpha}^T\vec{Y} - X)(\vec{\alpha}^T\vec{Y} - X)^T\right] = E\left[(\vec{\alpha}^T\vec{Y}\vec{Y}^T\vec{\alpha} - 2X\vec{Y}^T\vec{\alpha} + X^2\right] \tag{4.47}$$

$$= \vec{\alpha}^T K_{\vec{Y}}\vec{\alpha} - 2K_{X\vec{Y}}\vec{\alpha} + \sigma_X^2 \tag{4.48}$$

To minimize this, we take the derivative of $\vec{\alpha}^T K_{\vec{Y}}\vec{\alpha} - 2K_{X\vec{Y}}\vec{\alpha} + \sigma_X^2$ with respect to $\vec{\alpha}$, getting $2\vec{\alpha}^T K_{\vec{Y}} - 2K_{X\vec{Y}}$ (Exercise 4.4 explains this for readers who are rusty on vector calculus). Setting this equal to zero, we get

$$\vec{\alpha}^T K_{\vec{Y}} = K_{X\vec{Y}}\,; \qquad \vec{\alpha}^T = K_{X\vec{Y}}K_{\vec{Y}}^{-1} \tag{4.49}$$

Assuming that this stationary point is actually the minimum[4], the LLSE estimate is

$$\widehat{X}_{LLSE}(\vec{Y}) = K_{X\vec{Y}}K_{\vec{Y}}^{-1}\vec{Y} \tag{4.50}$$

Next suppose that $\vec{X}$ is an m dimensional r$\vec{\text{v}}$. Then (4.50) can be applied to each component of $\vec{X}$, and the LLSE estimate is given by

$$\widehat{X}_{LLSE}(\vec{Y}) = K_{\vec{X}\vec{Y}}K_{\vec{Y}}^{-1}\vec{Y} \tag{4.51}$$

Finally, let $\vec{\Xi}_L = \widehat{X}_{LLSE}(\vec{Y}) - \vec{X}$ be the error in the LLSE estimate. Since $\vec{X}$ and $\vec{Y}$ are zero mean, $E[\vec{\Xi}_L] = 0$, so the correlation matrix and covariance matrix of this error are the same and given by

$$K_{\vec{\Xi}_L} = E\left[(K_{\vec{X}\vec{Y}}K_{\vec{Y}}^{-1}\vec{Y} - \vec{X})(K_{\vec{X}\vec{Y}}K_{\vec{Y}}^{-1}\vec{Y} - \vec{X})^T\right] = -K_{\vec{X}\vec{Y}}K_{\vec{Y}}^{-1}K_{\vec{X}\vec{Y}}^T + K_{\vec{X}} \tag{4.52}$$

Note that the estimation rule in (4.51) is the same as the MMSE Gaussian estimation rule in (4.13) and the error covariance in (4.52) is the same as the error covariance for Gaussian MMSE in (4.14). This says that the LLSE estimate for a zero mean r$\vec{\text{v}}$ $\vec{X}$ from a zero mean r$\vec{\text{v}}$ $\vec{Y}$ is the same as the MMSE estimate of a Gaussian zero mean r$\vec{\text{v}}$ from a jointly Gaussian zero mean r$\vec{\text{v}}$ with the same covariances.

To understand why MMSE estimation for the Gaussian case yields the same result as LLSE estimation for the arbitrary case, first note that MMSE and LLSE are equivalent for the Gaussian case. That is, the MMSE estimate for the Gaussian case is (from (4.13)) a linear estimate, and thus the linear constraint for LLSE is not really a constraint.

Next note,from (4.48), that the minimization problem solved to find the LLSE estimate involves only joint covariances of X and $\vec{Y}$. Thus the LLSE estimation rule, as generalized in (4.51), depends only on covariances and thus must be the same for any non-Gaussian case and Gaussian case with the same covariances. Combining these two facts, we see that the MMSE estimation rule for the Gaussian case and the LLSE estimation rule for the arbitrary case must be the same. Similarly, the error covariance matrices must be the same. Thus,

[4] For those familiar with vector calculus, the second derivitive (Hessian) matrix of (4.47) with respect to $\vec{\alpha}$ is $K_{\vec{Y}}$. The fact that this is positive definite guarantees that the stationary point is actually the minimum. We shall soon give a geometric argument that also shows that (4.49) achieves the minimum.

aside from the appealing simplicity of the optimization above, it wasn't really necessary to carry it out.

For an arbitrary zero mean rv X and r$\vec{v}$ $\vec{Y}$, the LLSE estimate minimizes the mean square estimation error, $E[(\widehat{X}(\vec{Y})-X)^2]$ subject to the constraint that $\widehat{X}(\vec{Y})$ is linear in $\vec{Y}$, and the MMSE minimizes the same quantity with no constraint. Thus the mean square estimation error in the LLSE case is greater than or equal to mean square estimation error for the MMSE estimate. In fact, letting $\Xi_L = (\widehat{X}_{LLSE}(\vec{Y}) - X$ and $\Xi = (\widehat{X}_{MMSE}(\vec{Y}) - X$, exercise 4.5 shows that

$$E[\Xi_L^2] = E[\Xi^2] + E\left[(\widehat{X}_{LLSE}(\vec{Y}) - \widehat{X}_{MMSE}(\vec{Y}))^2\right] \tag{4.53}$$

For the Gaussian case, $E[\Xi_L^2] = E[\Xi^2]$, i.e., the mean square errors for MMSE and LLSE are the same. Thus, for estimating X from $\vec{Y}$, the mean square error $E[\Xi^2]$ for MMSE estimation in the non-Gaussian case is less than or equal to the mean square error for the Gaussian case with the same covariances. In this sense, the Gaussian case is the case that yields the largest mean square estimation errors. This same result clearly extends to the estimation of a r$\vec{v}$ $\vec{X}$ instead of a rv X. What is more, an estimator, constructed under the assumption of Gaussian r$\vec{v}$'s, will work just as well for non-Gaussian r$\vec{v}$'s with the same covariances. Naturally, an estimator constructed to take advantage of the non-Gaussian statistics might do even better.

4.3.2 LLSE Estimation; arbitrary means

Next let X be a rv with arbitrary mean and let $\vec{Y}$ be a r$\vec{v}$ with an arbitrary mean. The LLSE estimate of X is defined to be an estimate $\widehat{X}(\vec{Y})$ of the form

$$\widehat{X}(\vec{Y}) = \beta + \vec{\alpha}^T\vec{Y} = \beta + \sum_{i=1}^{n} \alpha_i Y_i \tag{4.54}$$

where the constant β and the vector $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)^T$ are chosen to minimize $E[(\beta+\alpha^T\vec{Y} - X)^2]$. Because of the constant β, this estimate should be called an affine estimate rather than linear, but the terminology is too standard to be changed. Note that this definition of LLSE estimate differs from that in the zero mean case because of the constant β, but we shall soon see that the optimizing constant β turns out to be zero in the zero mean case.

Let $\overline{X}$ and $\widetilde{X}$ be the mean and fluctuation of X, and similarly for $\vec{Y}$. We then want to choose β and $\vec{\alpha}$ to minimize

$$E\left[(\beta + \vec{\alpha}^T\overline{Y} - \overline{X} + \vec{\alpha}^T\widetilde{Y} - \widetilde{X})^2\right] = (\beta + \vec{\alpha}^T\overline{Y} - \overline{X})^2 + E\left[(\vec{\alpha}^T\widetilde{Y} - \widetilde{X})^2\right] \tag{4.55}$$

$$= (\beta + \vec{\alpha}^T\overline{Y} - \overline{X})^2 + \vec{\alpha}^T K_{\vec{Y}}\vec{\alpha} - 2K_{X\vec{Y}}\vec{\alpha} + \sigma_X^2 \tag{4.56}$$

We used the fact that the last two terms on the left of (4.55) are zero mean, and the first three terms are constants, so, in squaring, the cross terms between them are zero. For any choice of $\vec{\alpha}$, the first term in (4.56) is minimized, at the value 0, by choosing

$$\beta = \overline{X} - \vec{\alpha}^T\overline{Y} \tag{4.57}$$

Note that β is zero for the zero mean case, thus justifying the slight difference in definition of LLSE in this section and the last. Note also that the remaining terms are exactly the same as the quantity we minimized in the previous subsection. Thus the minimum occurs at

$$\vec{\alpha}^T K_{\vec{Y}} = K_{X\vec{Y}}\,; \qquad \vec{\alpha}^T = K_{X\vec{Y}} K_{\vec{Y}}^{-1} \tag{4.58}$$

Assuming again that this stationary point is actually the minimum, the LLSE estimate is

$$\widehat{X}_{LLSE}(\vec{Y}) = \overline{X} + K_{X\vec{Y}} K_{\vec{Y}}^{-1}(\vec{Y} - \overline{Y}) \tag{4.59}$$

As before, if $\vec{X}$ is an m dimensional r$\vec{v}$, then (4.59) can be applied to each component of $\vec{X}$, and the LLSE estimate is given by

$$\widehat{X}_{LLSE}(\vec{Y}) = E[\vec{X}] + K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1}(\vec{Y} - E[\vec{Y}]) \tag{4.60}$$

This estimate is the same as that in (4.16). The LLSE estimate for an arbitrary r$\vec{v}$ $\vec{X}$ from an arbitrary r$\vec{v}$ $\vec{Y}$ is the same as the MMSE estimate of a Gaussian r$\vec{v}$ from another jointly Gaussian r$\vec{v}$ with the same means and covariances.

Also as before, let $\vec{\Xi}_L = \widehat{X}_{LLSE}(\vec{Y}) - \vec{X}$ be the error in the LLSE estimate. We see from (4.60) that $E[\vec{\Xi}_L] = 0$, so the correlation matrix and covariance matrix of $\vec{\Xi}_L$ are the same and given by (4.52).

The argument why LLSE estimation for the non-Gaussian case yields the same result as MMSE estimation for the Gaussian case is the same for non-zero means as it was for zero means. Also, as in the zero mean case, the mean square error for the MMSE estimate must be less than or equal to the mean square error for the LLSE estimate. In this sense, the jointly Gaussian case is the case that yields the largest mean square estimation errors. The results on LLSE estimation are summarized in the following theorem:

Theorem 4.3 *If $\vec{X}$ and $\vec{Y}$ are jointly non-singular random vectors, the LLSE estimate of $\vec{X}$ from the observation $\vec{Y} = \vec{y}$ is given in (4.60) and is equal to the MMSE estimate for the Gaussian case as given by (4.16) and, for zero means, by (4.13). The error covariance, averaged over $\vec{Y}$, i.e., $K_{\vec{\Xi}_L} = E[\vec{\Xi}_L \, \vec{\Xi}_L^T]$ is given by (4.52) and is equal to the MMSE covariance for the Gaussian case in (4.14).*

4.3.3 The Projection Principle; zero mean case

In order to get some added insight into the linear optimization we have just done, and to obtain some later generalizations, we now view linear variables as elements of an abstract linear vector space. Up until now, we have used vectors simply as n-tuples of real numbers, and done so primarily as a notational tool to make expressions look simpler. We are now about to view random variables as vectors in their own right. To review the definition of a vector space, a *real vector space* $\mathcal{V}$ is a set of elements, called vectors, along with two operations, addition and scalar multiplication. Under the addition operation, any two vectors, $X \in \mathcal{V}$, $Y \in \mathcal{V}$ can be added to produce another vector, denoted $X + Y \in \mathcal{V}$. Under scalar multiplication, any real number α (called a scalar) can be multiplied by any vector $X \in \mathcal{V}$ to produce another vector, denoted $\alpha X \in \mathcal{V}$. A real vector space must satisfy the following axioms for all vectors X, Y, Z and all real numbers α, β:

1. Addition is commutative and associative, i.e. $X + Y = Y + X$ and $(X + Y) + Z = X + (Y + Z)$.
2. scalar multiplication is associative, i.e., $(\alpha\beta)X = \alpha(\beta X)$.
3. Scalar multiplication by 1 satisfies $1X = X$.
4. The distributive laws hold: $\alpha(X + Y) = (\alpha X) + (\alpha Y)$; $(\alpha + \beta)X = (\alpha X) + (\beta X)$.
5. There is a zero vector 0 such that $X + 0 = X$ for all $X \in \mathcal{V}$.
6. For each $X \in \mathcal{V}$, there is a unique $-X \in \mathcal{V}$ such that $X + (-X) = 0$.

The reason for all this formalism is that vector space results can be applied to many situations other than n-tuples of real numbers. Some common examples are polynomials and other sets of functions. All that is necessary to apply all the known results about vector spaces to the given situation is to check whether the given set of elements satisfy the axioms above. The particular set of elements of concern to us here is the set of zero mean random variables defined in some given probability space. We can add zero mean random variables to get other zero mean random variables, we can scale random variables by scalars, and it is easy to verify that the above axioms are satisfied.

It is important to understand that viewing a random variable as a vector is very different from the random vectors that we have been considering. A random vector $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is a string of n random variables, and is a function from the sample space to the space of n-dimensional real vectors. It is primarily a notational artifice to refer compactly to several random variables as a unit. Here, each random variable Y_i; $1 \le i \le n$ is viewed as a vector in its own right, and the sample values of these random variables do not live in this vector space at all.

An abstract real vector space, as defined above, contains no notion of length or orthogonality. To achieve these notions, we must define an *inner product*, $\langle X, Y\rangle$ as an additional operation on a real vector space, mapping pairs of vectors into real numbers. A *real inner product vector space* is a real vector space with an inner product operation that satisfies the following axioms for all vectors X, Y, Z and all scalars α:

1. $\langle X, Y\rangle = \langle Y, X\rangle$
2. $\alpha\langle X, Y\rangle = \langle \alpha X, Y\rangle$
3. $\langle X + Y, Z\rangle = \langle X, Z\rangle + \langle Y, Z\rangle$
4. $\langle X, X\rangle \ge 0$ with equality iff X = 0.

In a real inner product vector space, two vectors, X and Y, are defined to be *orthogonal* if $\langle X, Y\rangle = 0$. The *length* of a vector X is defined to be $\| X \| = \sqrt{\langle X, X\rangle}$. Similarly, the *distance* between two vectors X and Y is $\| X - Y \| = \sqrt{\langle X - Y, X - Y\rangle}$. For the conventional vector space in which a vector $\vec{x}$ is an n-tuple of real numbers, $\vec{x} = (x_1, \dots, x_n)^T$, the inner product $\langle \vec{x}, \vec{y}\rangle$ is conventionally taken to be $\langle \vec{x}, \vec{y}\rangle = x_1y_1 + \cdots + x_ny_n$. With this convention, length and orthogonality take on their conventional geometric significance. For the vector space of zero mean random variables, the natural definition for an inner product is the covariance,

$$\langle X, Y\rangle = E[XY] \tag{4.61}$$

Note that the covariance satisfies all the axioms for an inner product above. In this vector space, two zero mean random variables are orthogonal if they are uncorrelated. Also, the length of a random variable is its standard deviation.

We now use this vector space to interpret the LLSE estimate of a zero mean random variable X from the observation of n zero mean random variables $Y_1, \dots, Y_n$. The estimate here is a random variable that is required to be a linear combination of the observed variables,

$$\widehat{X}(Y_1, \dots, Y_n) = \sum_{i=1}^{n} \alpha_i Y_i \tag{4.62}$$

Viewing the random variables $Y_1, \dots, Y_n$ as vectors in the real inner product vector space discussed above, a linear estimate $\widehat{X}(Y_1, \dots, Y_n)$ must then be in the subspace[5] $\mathcal{S}$ spanned by $Y_1, \dots, Y_n$. The estimation error is the vector (random variable) $\Xi = \widehat{X}(Y_1, \dots, Y_n) - X$. Thus, in terms of this inner product space, the LLSE estimate $\widehat{X}(Y_1, \dots, Y_n)$ is that point $P \in \mathcal{S}$ that is closest to X.

Finding the closest point in a subspace to a given vector is a fundamental and central problem for inner product vector spaces. As indicated in Figure 4.1, the closest point is found by dropping a "perpendicular" from the point X to the subspace. The point where this perpendicular intersects the subspace is called the projection P of X onto the subspace. Formally, P is defined to be the *projection* of X onto $\mathcal{S}$ if $\langle P - X\, , Y\rangle = 0$ for all $Y \in \mathcal{S}$. The following well known result summarizes this:

Theorem 4.4 (Projection Principle) *Let X be a vector in a real inner product space $\mathcal{V}$, and let $\mathcal{S}$ be a finite dimensional[6] subspace of $\mathcal{V}$. There is a unique vector $P \in \mathcal{S}$ such that $\langle X - P, Y\rangle = 0$ for all $Y \in \mathcal{S}$. That vector P is the closest point in $\mathcal{S}$ to X, i.e., $\| X - P \| < \| X - Y \|$ for all $Y \in \mathcal{S}$, $Y \neq P$.*

Proof: Initially assume $X \notin \mathcal{S}$ and assume that a vector $P \in \mathcal{S}$ exists such that $\langle X - P\, , Y\rangle = 0$ for all $Y \in \mathcal{S}$. Let $Y \in \mathcal{S}, Y \neq P$ be arbitrary. The figure suggests that the three points

[5] A subspace of a vector space $\mathcal{V}$ is a subset of elements of $\mathcal{V}$ that constitutes a vector space in its own right. In particular, the linear combination of any two vectors in the subspace is also in the subspace; for this reason, subspaces are often called linear subspaces.

[6] The theorem also holds for infinite dimensional vector spaces if they satisfy a condition called completeness; this will be discussed later.

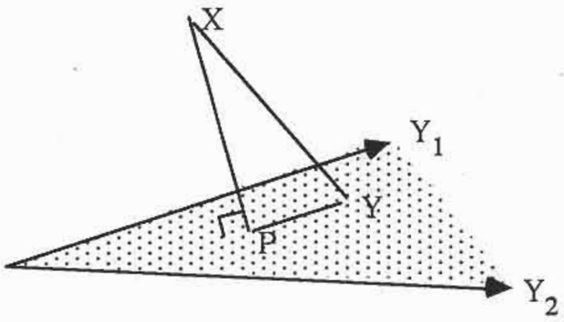

Figure 4.1: P is the projection of the vector X onto the subspace spanned by Y_1 and Y_2. That is, $X-P$ is orthogonal to all points in the subspace, and, as seen geometrically, the distance from X to P is smaller than the distance from X to any other point Y in the subspace.

X, P, Y form a right triangle with sides $U = X-P$, $V = P-Y$, and hypoteneuse $W = P-Y$. We first show that the squared length of the hypoteneuse, $\langle W, W\rangle$, is equal to the sum of the squared lengths of the sides, $\langle U, U\rangle + \langle V, V\rangle$[7] . W, U, V are vectors and $W = U + V$. i.e., $X-Y = (X-P) + (P-Y)$. Expanding the inner product,

$$\begin{aligned}\langle W, W\rangle &= \langle (U+V), (U+V)\rangle \\ &= \langle U, U\rangle + 2\langle U, V\rangle + \langle V, V\rangle\end{aligned}$$

Since $P \in \mathcal{S}$ and $Y \in \mathcal{S}$, $V = P - Y \in \mathcal{S}$, so $\langle U, V\rangle = 0$. Thus

$$\langle W, W\rangle = \langle U, U\rangle + \langle V, V\rangle \tag{4.63}$$

Since $\langle V, V\rangle$ is positive for all $Y \neq P$, $\langle W, W\rangle > \langle U, U\rangle$, so P is the unique point in $\mathcal{S}$ closest to X, and thus is the unique $P \in \mathcal{S}$ for which $\langle P - Y, X\rangle = 0$ for all $\vec{Y} \in \mathcal{S}$. We assumed above that $X \notin \mathcal{S}$, but if $X \in \mathcal{S}$, we simply choose $P = X$, and the result follows immediately.

The remaining question in completing the proof is whether a P exists such that $\langle X-P\,, Y\rangle = 0$ for all $Y \in \mathcal{S}$. Since $\mathcal{S}$ is finite dimensional, we can choose a basis $Y_1, Y_2, \dots, Y_n$ for $\mathcal{S}$. Then $\langle X-P\,, Y\rangle = 0$ for all $Y \in \mathcal{S}$ iff $\langle X-P\,, Y_i\rangle = 0$, or equivalently $\langle X\,, Y_i\rangle = \langle P\,, Y_i\rangle$ for all i, $1 \le i \le n$. Representing P as $\alpha_1 Y_1 + \alpha_2 Y_2 + \ldots + \alpha_n Y_n$, the question reduces to whether the set of linear equations

$$\langle X, Y_i\rangle = \sum_{j=1}^{n} \alpha_j \langle Y_j, Y_i\rangle = 0 \quad ; \qquad 1 \le i \le n \tag{4.64}$$

has a solution. But the vectors $Y_1, \dots, Y_n$ are linearly independent, so it follows (see Exercise 4.7) that the matrix with elements $\langle Y_j, Y_i\rangle$ must be non-singular. Thus (4.64) has a unique solution and the proof is complete.

[7] The issue here is not whether the familiar Pythagorean theorem of plane geometry is correct, but rather whether, in an abstract vector space, the definition of orthogonality corresponds to the plane geometry notion of right angle.

Returning to LLSE estimation, we have seen that the LLSE estimate $\widehat{X}_{LLSE}(Y_1, \dots, Y_n)$ is the projection P of X onto the subspace spanned by $Y_1, \dots, Y_n$. Thus,

$$\widehat{X}_{LLSE}(Y_1, \dots, Y_n) = P = \sum_{i=1}^{n} \alpha_i Y_i \tag{4.65}$$

where $\alpha_1, \dots, \alpha_n$ is the unique solution to (4.64). Since the inner products here correspond to covariances, (4.64) becomes

$$E[XY_i] = \sum_{j=1}^{n} \alpha_j E[Y_j Y_i] \quad ; \qquad 1 \leq i \leq n \tag{4.66}$$

Equations (4.65) and (4.66) agree (as they must) with (4.50). This also gives us a simple, and very geometric, proof of the fact that the stationary point that we found in (4.49) is actually a minimum. It turns out that (4.66) must have a solution whether or not $K_{\widetilde{Y}}$ is non-singular, and according to the theorem, $P = \widehat{X}_{LLSE}(Y_1, \dots, Y_n) = \sum_{j=1}^{n} \alpha_j Y_j$ is uniquely specified. The subtlety in the case where $K_{\widetilde{Y}}$ is singular is that, although P is uniquely specified, $(\alpha_1, \dots, \alpha_n)$ is not. As usual, however, the singular case is best handled by eliminating the dependent observations from consideration.

4.3.4 Projection Principle; non-zero mean

The above discussion was restricted to viewing *zero mean* random variables as vectors. It is often useful to remove this restriction and to consider the entire set of random variables in a given probability space as forming a real inner product vector space. In this generalization, the inner product is again defined as $\langle X, Y \rangle = E[XY]$, but the correlation now includes the mean values, i.e., $E[XY] = E[X]E[Y] + E[\widetilde{X}\widetilde{Y}]$. Note that for any random variable X with a non-zero mean, the difference between X and its fluctuation $\widetilde{X}$ is a constant random variable equal to $E[X]$. This constant random variable maps all sample points into the constant $E[X]$. This constant rv must also be interpreted as a vector in the real inner product space under consideration (since the difference of two vectors must again be a vector). Let D be the random variable (i.e., vector) that maps all sample points into unity. Thus $X - \widetilde{X} = E[X]D$ is a constant rv with value $E[X]$. As in subsection 4.3.2, a linear estimate involving rv's with non-zero means is invariably taken to mean an affine estimate $\widehat{X}(Y_1, \dots, Y_n) = \beta + \sum_i \alpha_i Y_i$. Writing this in terms of the unit constant rv D,

$$\widehat{X}(Y_1, \dots, Y_n) = \beta D + \sum_{i=1}^{n} \alpha_i Y_i \tag{4.67}$$

Interpreting these random variables as vectors, the LLSE estimate is the projection P of X onto the subspace $\mathcal{S}$ spanned by $D, Y_1, Y_2, \dots, Y_n$,

$$\widehat{X}_{LLSE}(Y_1, \dots, Y_n) = P = \beta D + \sum_{i=1}^{n} \alpha_i Y_i \tag{4.68}$$

where β and α_i ; $1 \le i \le n$, are chosen so that $\langle X - P, Y\rangle = 0$, or equivalently $\langle X, Y\rangle = \langle P, Y\rangle$ for all $Y \in \mathcal{S}$. This condition is satisfied iff both $\langle X, D\rangle = \langle P, D\rangle$ and $\langle X, Y_i\rangle = \langle P, Y_i\rangle$ for $1 \le i \le n$. Now for any vector Y we have $\langle Y, D\rangle = E[YD] = E[Y]$, so $\langle X, D\rangle = \langle P, D\rangle$ can be rewritten as

$$E[X] = E[P] = \beta + \sum_{i=1}^{n} \alpha_i E[Y_i] \tag{4.69}$$

Also, $\langle X, Y_i\rangle = \langle P, Y_i\rangle$ can be rewritten as

$$E[XY_i] = E[PY_i] = \beta E[Y_i] + \sum_{j=1}^{n} \alpha_j E[Y_j Y_i] \tag{4.70}$$

Exercise 4.8 shows that (4.68-4.70) are equivalent to (4.59). Note that in the world of random variables, $\widehat{X}(Y_1, \dots, Y_n) = \beta + \sum_i \alpha_i Y_i$ is an *affine* function of $Y_1, \dots, Y_n$, whereas it can also be viewed as a *linear* function of D, $Y_1, \dots, Y_n$. In the vector space world, $\widehat{X}(Y_1, \dots, Y_n) = \beta D + \sum_i \alpha_i Y_i$ must be viewed as a vector in the linear subspace spanned by $D, Y_1, \dots, Y_n$. The notion of linearity here is dependent on the context.

Next consider a situation in which we want to allow a limited amount of nonlinearity into an estimate. For example, suppose we consider estimating X as a linear combination of the constant random variable D, the observation variables $Y_1, \dots, Y_n$, and the squares of the observation variables $Y_1^2, \dots, Y_n^2$. Note that a sample value y_i of Y_i also specifies the sample value y_i^2 of Y_i^2, so such an estimate can be formed from the observation $\vec{y}$. Our estimate is then

$$\widehat{X}(Y_1, \dots, Y_n) = \beta D + \sum_{j=1}^{n} (\alpha_j Y_j + \gamma_j Y_j^2) \tag{4.71}$$

We wish to choose the scalars β, α_j, and γ_j, $1 \le j \le n$, to minimize the expected squared error between X and $\widehat{X}(Y_1, \dots, Y_n)$. In terms of the inner product space of random variables, we want to estimate X as the projection P of X onto the subspace spanned by the vectors $D, Y_1, \dots, Y_n, Y_1^2, \dots, Y_n^2$. It is sufficient for $X - P$ to be orthogonal to D, Y_i, and Y_i^2 for $1 \le i \le n$, so that $\langle X, D\rangle = \langle P, D\rangle$, $\langle X, Y_i\rangle = \langle P, Y_i\rangle$ and $\langle X, Y_i^2\rangle = \langle P, Y_i^2\rangle$ for $1 \le i \le n$. Using (4.71), the coefficients β, α_j, and γ_j must satisfy

$$E[X] = \beta + \sum_{j=1}^{n} (\alpha_j E[Y_j] + \gamma_j E[Y_j^2]) \tag{4.72}$$

$$E[XY_i] = \beta E[Y_i] + \sum_{j=1}^{n} (\alpha_j E[Y_i Y_j] + \gamma_j E[Y_i Y_j^2]) \tag{4.73}$$

$$E[XY_i^2] = \beta E[Y_i^2] + \sum_{j=1}^{n} (\alpha_j E[Y_i^2 Y_j] + \gamma_j E[Y_i^2 Y_j^2]) \tag{4.74}$$

This example should make one even more careful about the word "linear." The projection P is a linear function of the vectors D, Y_i, Y_i^2, $1 \le i \le n$, but in the underlying probability space with random variables, Y_i^2 is a nonlinear function of Y_i.

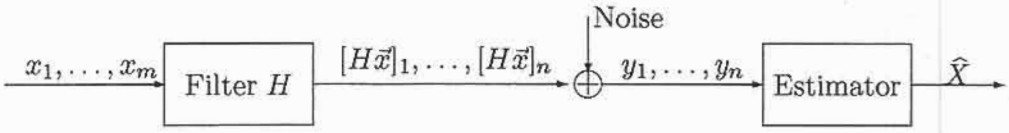

Figure 4.2: The input, over some interval of time, is the m dimensional sample sequence $\vec{x} = x_1, \ldots, x_m$. This is filtered into the n dimensional sample sequence $H\vec{x}$. Noise is added to get the n dimensional sample output $\vec{y} = y_1, \ldots, y_n$.

This example can be extended with cross terms, cubic terms, and so forth. Ultimately, we can consider the subspace $\mathcal{S}$ of all random variables $g(Y_1, \ldots, Y_n)$ that are functions of $Y_1, \ldots, Y_n$. The projection P of X onto $\mathcal{S}$ by definition satisfies $\langle X - P, g(Y_1, \ldots, Y_n)\rangle = 0$ for all $g(Y_1, \ldots, Y_n)$. Theorem 4.4 does not quite assure us that this projection exists, since $\mathcal{S}$ is infinite dimensional, but we already know that the MMSE estimate $\widehat{X}(Y_1, \ldots, Y_n)$ exists and, from (4.8), that the estimation error $\Xi = \widehat{X}(Y_1, \ldots, Y_n) - X$ satisfies $\langle \Xi, g(Y_1, \ldots, Y_n)\rangle = 0$ for all $g(Y_1, \ldots, Y_n)$. Thus the projection P does exist, and is equal to the MMSE estimate. In principle, then, the MMSE estimate can be found from (4.8), although usually it is more simply determined as the conditional mean of X, conditional on the observed sample values of $Y_1, \ldots, Y_n$.

As we progress from linear (or affine) estimates to estimates with square terms to estimates with yet more terms, the mean square error must be non-increasing, since the simpler estimates are non-optimal versions of the more complex estimates with some of the terms set to zero. Viewing this geometrically, in terms of projecting from X onto a subspace $\mathcal{S}$, the distance from X to the subspace must be non-increasing as the subspace is enlarged.

4.4 Filtered Vector Signal Plus Noise

Consider estimating a Gaussian m dimensional r$\vec{v}$ $\vec{X}$ from an observed sample of a Gaussian n dimensional r$\vec{v}$ $\vec{Y}$ where

$$\vec{Y} = H\vec{X} + \vec{Z} \tag{4.75}$$

Assume that $\vec{X}$ and $\vec{Z}$ are independent with non-singular covariance matrices and that H is an arbitrary n by m matrix. It follows from this that $\vec{X}$ and $\vec{Y}$ are jointly non-singular (see exercise 4.9). Initially, we also assume that $\vec{X}$ and $\vec{Z}$ are zero mean.

In a very real sense, this is the canonic estimation problem, estimating $\vec{X}$ from noisy observations of linear functions of $\vec{X}$. In communication terms, we can view $\vec{X}$ as a discrete time input to a communication system (see figure 4.2). H then represents the filter (perhaps time varying) that the input passes through, and $\vec{Z}$ represents discrete time noise that is added to the filtered signal. From (4.13) and 4.14), the MMSE estimate, $\widehat{X}(\vec{Y})$, and the covariance $K_{\vec{\Xi}}$ of the error, $\vec{\Xi} = \widehat{X}(\vec{Y}) - \vec{X}$, are given by

$$\widehat{X}(\vec{y}) = K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} \vec{y} \tag{4.76}$$

$$K_{\vec{\Xi}} = K_{\vec{X}} - K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} K_{\vec{X}\vec{Y}}^{T} \tag{4.77}$$

Also, we have seen that the error $\vec{\Xi}$ is Gaussian and independent of $\vec{X}$, and the conditional probability $p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y})$, for any given $\vec{y}$, is the Gaussian density $\mathcal{N}(\widehat{X}(\vec{y}), K_{\vec{\Xi}}$, $K_{\vec{X}} - K_{\vec{X}\vec{Y}}K_{\vec{Y}}^{-1}K_{\vec{X}\vec{Y}}^{T})$. To express this in terms of H, $K_{\vec{X}}$, and $K_{\vec{Z}}$, we have

$$K_{\vec{X}\vec{Y}} = E\left[\vec{X}(H\vec{X} + \vec{Z})^T\right] = K_{\vec{X}}H^T \tag{4.78}$$

$$K_{\vec{Y}} = E\left[(H\vec{X} + \vec{Z})(H\vec{X} + \vec{Z})^T\right] = HK_{\vec{X}}H^T + K_{\vec{Z}} \tag{4.79}$$

Substituting these into (4.76) and (4.77),

$$\widehat{X}(\vec{y}) \;=\; K_{\vec{X}}H^T[HK_{\vec{X}}H^T + K_{\vec{Z}}]^{-1}\vec{y} \tag{4.80}$$

$$K_{\vec{\Xi}} \;=\; K_{\vec{X}} - K_{\vec{X}}H^T[HK_{\vec{X}}H^T + K_{\vec{Z}}]^{-1}HK_{\vec{X}} \tag{4.81}$$

From Equations (2.46) and (2.50), alternative forms are

$$\widehat{X}(\vec{y}) \;=\; K_{\vec{\Xi}}H^TK_{\vec{Z}}^{-1}\vec{y} \tag{4.82}$$

$$K_{\vec{\Xi}} \;=\; [K_{\vec{X}}^{-1} + H^TK_{\vec{Z}}^{-1}H]^{-1} \tag{4.83}$$

As we have seen in the previous section, all of these results are also valid for LLSE estimation whether or not the r$\vec{v}$'s are Gaussian.

Example 4.5 (Gaussian Signal plus Noise) It is easier to interpret these results in the simpler case where H is an identity matrix I_n, i.e. where $\vec{Y} = \vec{X} + \vec{Z}$. Equations (4.80) and (4.81) then become

$$\widehat{X}(y) = K_{\vec{X}}(K_{\vec{X}} + K_{\vec{Z}})^{-1}\vec{y} \tag{4.84}$$

$$K_{\vec{\Xi}} = K_{\vec{X}} - K_{\vec{X}\vec{Y}}K_{\vec{Y}}^{-1}K_{\vec{X}\vec{Y}}^T = K_{\vec{X}} - K_{\vec{X}}K_{\vec{Y}}^{-1}K_{\vec{X}} \tag{4.85}$$

We now show how the alternative forms in (4.82) and (4.83) follow directly from these equations. Factoring $K_{\vec{X}}K_{\vec{Y}}^{-1}$ out of the right side of (4.85), we get

$$K_{\vec{\Xi}} = K_{\vec{X}}K_{\vec{Y}}^{-1}[K_{\vec{Y}} - K_{\vec{X}}] = K_{\vec{X}}K_{\vec{Y}}^{-1}K_{\vec{Z}} = K_{\vec{X}}[K_{\vec{X}} + K_{\vec{Z}}]^{-1}K_{\vec{Z}} \tag{4.86}$$

Inverting both sides of this equation,

$$K_{\vec{\Xi}}^{-1} = K_{\vec{Z}}^{-1}[K_{\vec{X}} + K_{\vec{Z}}]K_{\vec{X}}^{-1} = K_{\vec{X}}^{-1} + K_{\vec{Z}}^{-1} \tag{4.87}$$

This is (4.83) in the special case $H = I_n$. Finally, substituting (4.86) into (4.84), we get

$$\widehat{X}(\vec{y}) = K_{\vec{\Xi}}K_{\vec{Z}}^{-1}\vec{y} \tag{4.88}$$

in agreement with (4.82). This example is simply the vector form of example 4.1 and is interpreted in the same way. In particular, because of the symmetry between $\vec{X}$ and $\vec{Z}$, we can also estimate $\vec{Z}$ in the same way, getting $\widehat{Z}(\vec{y}) = K_{\vec{\Xi}}K_{\vec{X}}^{-1}\vec{y}$. Note that $\widehat{X}(\vec{y})$ +$\widehat{Z}(\vec{y})$= $\vec{y}$, so that the observation is split between the estimate of $\vec{X}$ and that of $\vec{Z}$ according to the covariances.

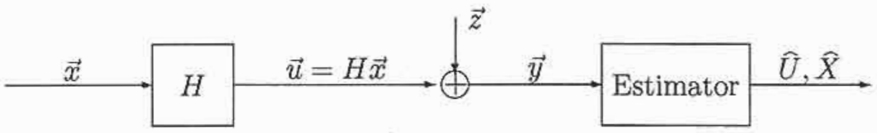

Figure 4.3: Viewing the input to a filter followed by noise as a transformed version of the filter output followed by noise.

4.4.1 Interpretation in terms of Input transformations

If we transform the input in figure 4.2, then we should be able to determine the estimate and estimation error of the transformed quantities from the untransformed quantities. To see how this works, let H be an invertible n by n matrix and let $\vec{Y} = H\vec{X} + \vec{Z}$ where $\vec{X}$ and $\vec{Y}$ are zero mean Gaussian (see figure 4.3). Define $\vec{U}$ as the filter output, so that

$$\vec{U} = H\vec{X} \quad ; \qquad \vec{X} = H^{-1}\vec{U} \tag{4.89}$$

It follows that $E[\vec{U} \mid \vec{Y}{=}\vec{y}] = H\,E[\vec{X} \mid \vec{Y}{=}\vec{y}]$ and $E[\vec{X} \mid \vec{Y}{=}\vec{y}] = H^{-1}E[\vec{U} \mid \vec{Y}{=}\vec{y}]$. Since these conditional means are the MMSE estimates, $\widehat{X}(\vec{y})$ and $\widehat{U}(\vec{y})$, we have

$$\widehat{U}(\vec{y}) = H\widehat{X}(\vec{y}) \quad ; \qquad \widehat{X}(\vec{y}) = H^{-1}\widehat{U}(\vec{y}) \tag{4.90}$$

From (4.89), the covariances of X and U are related by

$$K_{\vec{U}} = HK_{\vec{X}}H^T \quad ; \qquad K_{\vec{X}} = H^{-1}K_{\vec{U}}(H^{-1})^T \tag{4.91}$$

Letting $\vec{\Xi}_X = \widehat{X}(\vec{y}) - \vec{X}$ and $\vec{\Xi}_U = \widehat{U}(\vec{y}) - \vec{U}$ denote the estimation errors, we have $\vec{\Xi}_U = H\vec{\Xi}_X$ and $\vec{\Xi}_X = H^{-1}\vec{\Xi}_U$, so that

$$K_{\vec{\Xi}_U} = HK_{\vec{\Xi}_X}H^T \quad ; \qquad K_{\vec{\Xi}_X} = H^{-1}K_{\vec{\Xi}_U}(H^{-1})^T \tag{4.92}$$

We can now use these equations to find $\widehat{X}(\vec{y})$ and $K_{\vec{\Xi}_X}$ in terms of $\widehat{U}(\vec{y})$ and $K_{\vec{\Xi}_U}$, which, in turn, have been found in example 4.5. Using (4.90) for $\widehat{X}(\vec{y})$ and (4.88) for $\widehat{U}(\vec{y})$, we get

$$\widehat{X}(\vec{y}) = H^{-1}K_{\vec{\Xi}_U}K_{\vec{Z}}^{-1}\vec{y} = K_{\vec{\Xi}_X}H^TK_{\vec{Z}}^{-1}\vec{y} \tag{4.93}$$

where we have used (4.92) for the second equality. Similarly, using (4.92) for $K_{\vec{\Xi}_X}$ and (4.87) for $K_{\vec{\Xi}_U}$, we get

$$K_{\vec{\Xi}_X}^{-1} = H^T\left[K_{\vec{U}}^{-1} + K_{\vec{Z}}^{-1}\right]H = K_{\vec{X}}^{-1} + H^TK_{\vec{Z}}^{-1}H \tag{4.94}$$

Equations (4.93) and (4.94) agree with (4.82) and (4.83), but the derivation here is considerably more insightful than the strictly algebraic derivation of these equations from chapter 2. Equations (4.80) and (4.81) can also be derived in the same way (see Exercise 4.10) Note, however, that these derivations apply only to the case where H is invertible. We look at the non-invertible case in subsection 4.4.3.

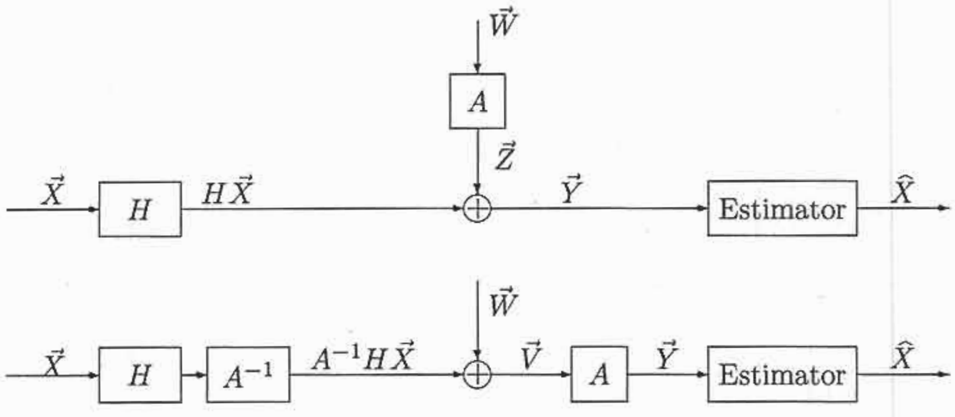

Figure 4.4: Transforming the output to $\vec{V} = A^{-1}\vec{Y}$ has the effect of making the noise IID and modifying the filter H to $A^{-1}H$. Note that $\vec{X}$ is operated on by H and then A^{-1}, the result is $A^{-1}H\vec{X}$, i.e., a left to right sequence of operations in a block diagram corresponds to a right to left ordering in an equation.

4.4.2 Interpretation in terms of Noise Transformations

An equally useful transformation comes from viewing the noise $\vec{Z}$ as filtered IID noise. To do this, let A be a matrix such that $AA^T = K_{\vec{Z}}$ (for example A could be $\sqrt{K_{\vec{Z}}}$ as discussed in section 2.5). We can then represent $\vec{Z}$ as $A\vec{W}$ where $\vec{W}$ is $\mathcal{N}(\vec{0}, I_n)$. Define

$$\vec{V} = A^{-1}\vec{Y} = A^{-1}H\vec{X} + \vec{W} \tag{4.95}$$

This can be viewed as the output filtered to make the noise components IID, and is represented in block diagram form in Figure 4.4. $\vec{V}$ is an invertible transformation of the actual observation $\vec{Y}$, so we can form the MMSE estimate of $\vec{X}$ from $\vec{V}$ as well as from $\vec{Y}$. However, defining $G = A^{-1}H$, we see that $\vec{V} = G\vec{X} + \vec{W}$. Thus, any such Gaussian problem with an an arbitrary non-singular covariance can be converted into a problem with IID noise by modifying the input filter and the output.

In terms of this transformation, the estimate and estimation error, in the form of (4.82) and (4.83), are given by

$$\widehat{X}(\vec{v}) = K_{\vec{\Xi}}G^TK_{\vec{W}}^{-1}\vec{v} \tag{4.96}$$

$$K_{\vec{\Xi}} = [K_{\vec{X}}^{-1} + G^TK_{\vec{W}}^{-1}G]^{-1} \tag{4.97}$$

As a sanity check, we verify that if we substitute $G = A^{-1}H$, $\vec{v} = A^{-1}\vec{y}$, $K_{\vec{W}} = I_n$, and $AA^T = K_{\vec{Z}}$ into these equations, we come back to (4.82) and (4.83) in their untransformed form.

$$K_{\vec{\Xi}}G^TK_{\vec{W}}^{-1}\vec{v} = K_{\vec{\Xi}}[A^{-1}H]^TA^{-1}\vec{y} = K_{\vec{\Xi}}H^TK_{\vec{Z}}^{-1}\vec{y} \tag{4.98}$$

$$[K_{\vec{X}}^{-1} + G^TK_{\vec{W}}^{-1}G]^{-1} = [K_{\vec{X}}^{-1} + \left[A^{-1}H]^TA^{-1}H\right]^{-1} = \left[K_{\vec{X}}^{-1} + H^TK_{\vec{Z}}^{-1}H\right]^{-1}] \tag{4.99}$$

This transformation essentially shows that if we solve problems with IID noise, we can then easily modify the solution to deal with arbitrary non-singular noise covariances. We shall use this idea several times in what follows.

4.4.3 Interpretation in terms of Sufficient Statistics

In many applications, the dimension m of $\vec{X}$ is much smaller than the dimension n of $\vec{Y}$ and $\vec{Z}$. Also the components of $\vec{Z}$ are often independent, so that $K_{\vec{Z}}$ is easy to invert. In these applications, (4.82) and (4.83) are easier to work with and more insightful than (4.80) and (4.81). In this section, we re-derive and interpret (4.82) and (4.83) from the viewpoint of likelihood ratios and sufficient statistics.

Starting from first principles, $\vec{Y}$, conditional on a given $\vec{X} = \vec{x}$, is a Gaussian vector with covariance $K_{\vec{Z}}$ and mean $H\vec{x}$. The density, or likelihood, is

$$p_{\vec{Y}|\vec{X}}(\vec{y} \mid \vec{x}) = \frac{\exp\left(-\frac{1}{2}(\vec{y} - H\vec{x})^T K_{\vec{Z}}^{-1}(\vec{y} - H\vec{x})\right)}{(2\pi)^{n/2}\sqrt{\det K_{\vec{Z}}}} \tag{4.100}$$

$$= \frac{\exp\left(-\frac{1}{2}\left[\vec{y}^T K_{\vec{Z}}^{-1}\vec{y} - 2(H\vec{x})^T K_{\vec{Z}}^{-1}\vec{y} + (H\vec{x})^T K_{\vec{Z}}^{-1}H\vec{x}\right]\right)}{(2\pi)^{n/2}\sqrt{\det K_{\vec{Z}}}} \tag{4.101}$$

In going from (4.100) to (4.101 we used the fact that $(H\vec{x})^T K_{\vec{Z}}^{-1}\vec{y}$ and $\vec{y}^T K_{\vec{Z}}^{-1} H\vec{x}$ are transposes of each other and are one dimensional; thus they must be the same. The *likelihood ratio* for estimating $\vec{X}$ from $\vec{Y}$ is defined as

$$\Lambda(\vec{y}, \vec{x}) = \frac{p_{\vec{Y}|\vec{X}}(\vec{y} \mid \vec{x})}{p_{\vec{Y}|\vec{X}}(\vec{y} \mid \vec{x}_o)} \tag{4.102}$$

where $\vec{x}_o$ is some fixed vector that we take to be 0. Recall that for hypothesis testing with M hypotheses, the likelihood ratio was a function of $\vec{y}$ and was indexed by the hypothesis, whereas here, it is a function of both $\vec{y}$ and $\vec{x}$. Thus, the likelihood ratio here is the extension of that for the detection case in which $\vec{x}$ takes values from the uncountably infinite set of real vectors rather than from a finite set of M values. From (4.101), with $\vec{x}_o = 0$, the likelihood ratio is given by

$$\Lambda(\vec{y}, \vec{x}) = \exp\left(\vec{x}^T H^T K_{\vec{Z}}^{-1}\vec{y} - \frac{1}{2}(H\vec{x})^T K_{\vec{Z}}^{-1} H\vec{x}\right) \tag{4.103}$$

For estimation problems, as with detection problems, a *sufficient statistic* $\mathcal{T}(\vec{y})$ is a function of the observation $\vec{y}$ from which the likelihood ratio can be calculated for all $\vec{x}$. Thus, from (4.103), we see that

$$\mathcal{T}(\vec{y}) = H^T K_{\vec{Z}}^{-1}\vec{y} \tag{4.104}$$

is a sufficient statistic. Intuitively, a sufficient statistic is a function of the observation that includes all the statistical information about $\vec{X}$ contained in the observation. More precisely, $\mathcal{T}(\vec{y})$ specifies $\Lambda(\vec{y},\vec{x})$ for all $\vec{x}$. To see the significance of a sufficient statistic, note that the a posteriori density can be calculated as

$$p_{\vec{X}|\vec{Y}}(\vec{x} \mid \vec{y}) = \frac{\Lambda(\vec{y},\vec{x})p_{\vec{X}}(\vec{x})}{\int \Lambda(\vec{y},\vec{x}_1)p_{\vec{X}}(\vec{x}_1)d\vec{x}_1} \tag{4.105}$$

The MMSE estimate, and any other minimum expected cost estimate, can be found from the a posteriori probability density, and thus from a sufficient statistic. Also, the ML estimate can be calculated by maximizing $\Lambda(\vec{y},\vec{x})$ over $\vec{x}$, and the MAP estimate can be calculated by maximizing $\Lambda(\vec{y},\vec{x})p_{\vec{X}}(\vec{x})$ over $\vec{x}$.

Note that $\mathcal{T}(\vec{y})$ above is an m dimensional quantity, whereas $\vec{y}$ has dimension n. Thus, for $n > m$, replacing the observation $\vec{y}$ with the sufficient statistic $\mathcal{T}(\vec{y})$ has greatly simplified the problem. It is important to recognize that nothing is lost if $\mathcal{T}(\vec{y})$ is found from the observation $\vec{y}$ and then $\vec{y}$ is discarded; all of the relevant information has been extracted from $\vec{y}$. There are usually many choices of sufficient statistics. For example, the observation $\vec{y}$ itself is a sufficient statistic, and any invertible transformation of $\vec{y}$ is also sufficient. Also, any invertible transformation of a sufficient statistic $\mathcal{T}(\vec{y})$ is another sufficient statistic.

What is important about $\mathcal{T}(\vec{y}) = HK_{\vec{Z}}^{-1}\vec{y}$ is, first, that the dimensionality has been reduced from n (the dimension of the observation $\vec{y}$) to m. Second, the operation $\mathcal{T}(\vec{y})$ is the same operation that we studied for detection; this will be discussed shortly. Third, sufficient statistics do not depend on the probability distribution of the rv $\vec{X}$ to be estimated, and thus, that distribution can be changed without changing the sufficient statistic.

Let us now view our estimation problem as first finding the sufficient statistic $\mathcal{T}(\vec{y}) = HK_{\vec{Z}}^{-1}\vec{y}$ and then considering the m dimensional estimation problem for estimating $\vec{X}$ from $\mathcal{T}(\vec{y})$. Let $\mathcal{T}(\vec{Y})$ be the rv form of the sufficient statistic so that

$$\mathcal{T}(\vec{Y}) = H^T K_{\vec{Z}}^{-1}\left[H\vec{X}+\vec{Z}\right] = A\vec{X}+\vec{U} \tag{4.106}$$

where the matrix A and rv $\vec{U}$ are given by

$$A = H^T K_{\vec{Z}}^{-1} H \quad ; \qquad \vec{U} = H^T K_{\vec{Z}}^{-1}\vec{Z} \tag{4.107}$$

We then have

$$K_{\vec{U}} = E\left[H^T K_{\vec{Z}}^{-1} Z Z^T K_{\vec{Z}}^{-1} H\right] = H^T K_{\vec{Z}}^{-1} H \tag{4.108}$$

Note that A is the same as $K_{\vec{U}}$. Assuming[8] that the columns of H are linearly independent, A (and thus $K_{\vec{U}}$ are invertible. Thus, estimating $\vec{X}$ from $\mathcal{T} = A\vec{X}+\vec{U}$ is the problem solved in subsection 4.4.1. From (4.94), the estimation error in $\vec{X}$ for the problem $\mathcal{T} = A\vec{X}+\vec{U}$ is given by

$$K_{\vec{\Xi}}^{-1} = K_{\vec{X}}^{-1} + A^T K_{\vec{U}}^{-1} A = K_{\vec{X}}^{-1} + A = K_{\vec{X}}^{-1} + H^T K_{\vec{Z}}^{-1} H \tag{4.109}$$

[8] If the columns of H are not independent, then a sufficient statistic of lower dimension than m can be found, and after creating such an estimate, the procedure above can be followed.

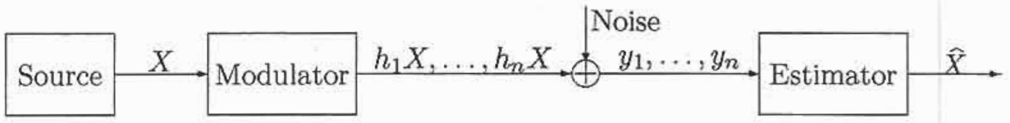

Figure 4.5: The source, in some given interval of time, attempts to transmit a number, modeled as the rv X. The transmitter is abstracted into a modulator which maps X into the signal $(h_1X,\dots,h_nX)$. An IID Gaussian noise vector, independent of the source output, is added to the modulated signal. The estimator receives the signal plus noise and estimates the source output.

Thus we have derived and interpreted the estimation error in (4.83) in terms of a sufficient statistic. For the estimate itself, (4.93), applied to estimating $\vec{X}$ from $\mathcal{T} = A\vec{X} + \vec{U}$, yields

$$\widehat{X}(\mathcal{T}) = K_{\vec{\Xi}}A^TK_{\vec{U}}^{-1}\mathcal{T} = K_{\vec{\Xi}}\mathcal{T} \tag{4.110}$$

where we have used the fact that $A = K_{\vec{U}}$. Since $\mathcal{T}(\vec{y}) = HK_{\vec{Z}}^{-1}\vec{y}$, this is equivalent to (4.82).

Example 4.6 (Scalar Signal plus IID Vector Noise) Consider the special case of (4.109) and (4.110) where X is a scalar rv, H is an n by 1 matrix denoted as $\vec{h}$, and the components $\{Z_i; 1 \le i \le n\}$ of $\vec{Z}$ are independent with variances $\sigma^2_{Z_i}$. We can view this as a communication problem in which a single real number, modeled as the Gaussian rv X, is first modulated into the signal $\vec{h}X$ and then transmitted in the presence of independent noise (see Figure 4.5). This problem is the same as Example 4.3 (specialized here to $E[X] = 0$).

Using (4.104), we see that a sufficient statistic is given by

$$\mathcal{T}(\vec{y}) = \vec{h}^TK_{\vec{Z}}\vec{y} = \sum_{i=1}^{n}\frac{h_iy_i}{\sigma^2_{Z_i}} \tag{4.111}$$

Note that $\mathcal{T}(\vec{y})$ above is a one dimensional quantity, whereas $\vec{y}$ has dimension n. Thus, in this case, replacing the observation $\vec{y}$ with the sufficient statistic reduces the problem to a single dimension. In the special case where $\sigma^2_{Z_i}$ is the same for all i (i.e., where the Z_i are IID), the operation $\mathcal{T}(\vec{y})$ is the same matched filter operation that we studied for detection. When $\sigma^2_{Z_i}$ varies with i. the operation of forming $\mathcal{T}(\vec{y})$ is an almost trivial variation on the matched filter operation. When the observed values y_i come from different locations, as opposed to being observed at different times, $\mathcal{T}(\vec{y})$ is called *maximal ratio combining*. As we see, however, all of this terminology simply refers to this one dimensional sufficient statistic. Finally note that (4.111) yields a sufficient statistic whether or not X is Gaussian.

Finally, to complete the example, we find the MMSE estimate and the estimation error. From (4.110),

$$\widehat{X}(\mathcal{T}(\vec{y})) = \sigma^2_{\Xi}\mathcal{T}(\vec{y}) = \sigma^2_{\Xi}\sum_{i=1}^{n}\frac{h_iy_i}{\sigma^2_{Z_i}} \tag{4.112}$$

Figure 4.6: Illustration of one dimensional signal x and 2 dimensional observation $\vec{y}$ where $y_1 = 2x + z_1$ and $y_2 = x + z_2$. The noise terms, z_1 and z_2, are sample values of IID Gaussian rv's. The concentric circles are regions where $p_{\vec{Y}|X}(\vec{y} \mid x)$ is constant for given x. The straight line through the origin is the locus of points $(2x, x)$ as x varies; this is the received signal vector , as a function of the signal x, in the absence of noise. The other straight line is perpendicular to the first line; it is the locus of points where $2y_1 + y_2$ is constant. Thus the sufficient statistic $2y_1 + y_2$ can be viewed as projecting (y_1, y_2) onto the straight line $(2x, x)$ and ignoring the noise in the perpendicular direction.

$$\frac{1}{\sigma^2_{\Xi}} = \frac{1}{\sigma^2_X} + \sum_{i=1}^{n} \frac{h_i^2}{\sigma^2_{Z_i}} \tag{4.113}$$

Note that this is the same as the solution to the scalar recursive estimation problem in (4.33). Figure 4.6 illustrates this for $n = 2$ and illustrates how the sufficient statistic focuses on the received vector in the direction of the modulated signal, ignoring the irrelevant directions.

Look again now at the case where $\vec{X}$ is m dimsional and $\vec{Z}$ and $\vec{Y}$ are n dimensional, $n > m$. The m dimensional sufficient statistic given in (4.104) can be broken down into m one dimensional equations,

$$\mathcal{T}_i(\vec{y}) = \sum_{i=1}^{n} \frac{H_{ij} y_i}{\sigma^2_{Z_i}} \tag{4.114}$$

Thus each component of $\mathcal{T}$ is a matched filter type operation as in the previous example, and does not depend on the covariance of the input. The problem of finding the estimate of each component of $\vec{X}$ in the presence of "intersymbol interference" from the other components is then handled by the matrix K_{Ξ}

Example 4.7 (Scalar Signal plus Arbitrary Vector Noise) As a slightly more general scalar signal example , let $\vec{Y} = \vec{h}X + \vec{Z}$ again, but let $\vec{Z}$ have an arbitrary non-singular covariance matrix $K_{\vec{Z}}$. Let A be a matrix such that $AA^T = K_{\vec{Z}}$. We can then represent $\vec{Z}$ as $A\vec{W}$ where $\vec{W}$ is $\mathcal{N}(\vec{0}, I_n)$. Using the transformation of figure 4.4, let $\vec{V} = A^{-1}\vec{Y}$ be the transformed output and let $\vec{g} = A^{-1}\vec{h}$ be the transformed filter. Then $\vec{V} = \vec{g}^T X + \vec{W}$. This is then the problem of the previous example. The sufficient statistic is then $\mathcal{T} = \vec{g}^T \vec{v} = \vec{g}^T A^{-1} \vec{y} = \vec{h}^T K_{\vec{Z}} \vec{y}$.

4.4.4 Orthornomal Expansions of Input and Output

Next consider the situation $\vec{Y} = H\vec{X} + \vec{Z}$ where $\vec{X} = (X_1, \dots, X_m)^T$ has m IID components each of variance σ^2 and $\vec{Z} = (Z_1, \dots, Z_n)^T$ has $n \geq m$ IID components each of variance 1. We could view each component X_i of $\vec{X}$ as passing through a filter $\vec{h}_i = (h_{1,i}, \dots, h_{n,i})$, where $\vec{h}_i$ is the i^{th} column of H, and the solution in (4.82) passes the observation $\vec{y}$ through n matched filters $\vec{h}_i^T \vec{y}$ in forming the MMSE estimate of $\vec{X}$. However, in (4.82) the matched filter outputs are then operated on by $K_{\vec{\Xi}}$, which somehow takes care of scaling and of the interference between the different components of $\vec{X}$.

To get another viewpoint of why $K_{\vec{\Xi}}$ is the right operation in (4.82), we rederive (4.82) in yet another way, through viewing the operator H in terms of orthonormal expansions of input and output. Note that $H^T H$ is a symmetric, non-negative definite m by m matrix, and it has m eigenvalue, eigenvector pairs,

$$H^T H \phi_i = \lambda_i \phi_i; \quad 1 \leq i \leq m \tag{4.115}$$

The eigenvalues need not be distinct, and need not be non-zero, but they are all non-negative, and we can always choose the set of ϕ_i to be orthonormal, which we henceforth assume. The matrix F with columns $\phi_1, \phi_2, \dots, \phi_m$ has the property that $F^T F = I$ (since $\phi_i^T \phi_j = \delta_{ij}$) so $F^T = F^{-1}$. Assume that $\lambda_i > 0$ for $1 \leq i \leq m$ and define the n dimensional vectors θ_i to be $\lambda_i^{-1/2} H \phi_i$. These vectors must be orthonormal (see Exercise 4.6) since the vectors ϕ_i are orthonormal, and each θ_i is an eigenvector, with eigenvalue λ_i, of HH^T (which is an n by n matrix). Let Q be the n by m matrix made up of the columns $\theta_1, \dots, \theta_m$. Then $Q = HF\Lambda^{-1/2}$ where Λ is the diagonal matrix with elements $\lambda_1, \dots, \lambda_m$. We then have

$$H = Q\Lambda^{1/2} F^T = \sum_i \theta_i \lambda_i^{1/2} \phi_i^T \tag{4.116}$$

$$\vec{Y} = Q\Lambda^{1/2} F^T \vec{X} + \vec{Z} \tag{4.117}$$

Now note that, because of the orthonormality of $\{\theta_i\}$, $Q^T Q = I$ (even though Q is n by m), and thus

$$Q^T \vec{Y} = \Lambda^{1/2} F^T \vec{X} + Q^T \vec{Z} \tag{4.118}$$

Now define $F^T \vec{X}$ as $\vec{X}_\phi$. That is, $\vec{X}_\phi$ is the random vector $\vec{X}$ in the basis $\phi_1, \dots, \phi_n$, i.e., the random vector with components $\{\phi_i^T \vec{X};\ 1 \leq i \leq m\}$. Note that $E[\vec{X}_\phi \vec{X}_\phi^T] = E[F^T \vec{X} \vec{X}^T F] = \sigma^2 F^T I F = \sigma^2 I$. Similarly, define $Q^T \vec{Z}$ as $\vec{Z}_\theta$. By the same argument, $E[\vec{Z}_\theta \vec{Z}_\theta^T] = I$. Finally define $Q^T \vec{Y}$ as $\vec{Y}_\theta$. We then have $\vec{Y}_\theta = \Lambda^{1/2} \vec{X}_\phi + \vec{Z}_\theta$. Since the components of both $\vec{X}_\phi$ and $\vec{Z}_\theta$ are independent, we have m independent one dimensional estimation problems. Note that $\vec{Z}_\theta$ does not fully represent the noise $\vec{Z}$ (see Exercise 4.6), but the remaining $n - m$ components of the noise are independent of $\vec{X}$ and of $\vec{Z}_\theta$, so they are irrelevant and can be ignored. From (4.22) and (4.23), the MMSE estimate of each component of $\vec{X}_\phi$ is given by

$$(\widehat{X}_\phi(\vec{y}_\theta))_i = \frac{\sqrt{\lambda_i}\,\sigma^2 (\vec{y}_\theta)_i}{\lambda_i \sigma^2 + 1}; \quad E\left[\left|(\widehat{X}_\phi(\vec{y}_\theta))_i - (\vec{X}_\phi)_i\right|^2\right] = [\sigma^{-2} + \lambda_i]^{-1} \tag{4.119}$$

In vector notation, define the error, in the basis $\phi_1, \ldots, \phi_n$, as $\vec{\Xi}_\phi = \widehat{X}_\phi(\vec{y}_\theta) - \vec{X}_\phi$. The covariance matrix of this, from 4.119, is a diagonal matrix given by

$$K_{\vec{\Xi}_\phi} = [\sigma^{-2} I + \Lambda]^{-1} \tag{4.120}$$

The estimate, from 4.119 then becomes

$$\widehat{X}(\vec{y}_\theta) = K_{\vec{\Xi}_\phi} \sqrt{\Lambda} \vec{y}_\phi \tag{4.121}$$

Finally we change back to the original basis. Since $\vec{X} = F\vec{X}_\phi$, we have $\widehat{X}(\vec{y}) = F\widehat{X}_\phi(\vec{y}_\theta)$. Also $\vec{y}_\theta = Q^T \vec{y}$, so

$$\widehat{X}(\vec{y}) = FK_{\vec{\Xi}_\phi} \sqrt{\Lambda}\, Q^T \vec{y} \;=\; (FK_{\vec{\Xi}_\phi} F^T)(F\sqrt{\Lambda} Q^T)\vec{y} \tag{4.122}$$

As shown below, the the first term above is $K_{\vec{\Xi}}$.

$$K_{\vec{\Xi}} = E\left[\mid \vec{X} - \widehat{X}(\vec{y}) \mid^2\right] = E\left[F(\vec{X}_\phi - \widehat{X}_\phi)(\vec{X}_\phi^T - \widehat{X}_\phi^T)F^T\right] = FK_{\vec{\Xi}_\phi} F^T \tag{4.123}$$

Also, from (4.116), $H^T = F\Lambda^{1/2} Q^T$. Substituting this and (4.123) into (4.122)

$$\widehat{X}(\vec{y}) = K_{\vec{X}i} H^T \vec{y} \tag{4.124}$$

This agrees with (4.82), since $K_{\vec{Z}} = I$. Finally, from (4.123) and then (4.120),

$$K_{\vec{\Xi}}^{-1} = F(K_{\vec{\Xi}_\phi})^{-1} F^T = F(\sigma^{-2} I + \Lambda) F^T = \sigma^{-2} I + H^T H$$

which agrees with (4.83).

For the more general case when $\vec{X}$ is not IID, one can transform $\vec{X}$ into IID variables, $\vec{X}' = A\vec{X}$, and then apply the above approach to $\vec{X}'$. Similarly, if $\vec{Z}$ is not IID, one can first whiten it into IID variables $\vec{W} = B\vec{Z}$. Not much would be gained by this, since we already know the answer from (4.80) and (4.82), and the purpose of these examples was to make those results look less like linear algebra magic.

As a final generalization, assume $\vec{Y} = H\vec{X} + \vec{Z}$ as before, but now assume $\vec{X}$ has a mean, $\overline{X}$. Thus $E[\vec{Y}] = H\overline{X}$. Applying (4.80) and (4.82) to the fluctuations, $X - \bar{X}$ and $\vec{Y} - H\overline{X}$, we have the alternative forms

$$\widehat{X}(\vec{y}) \;=\; \overline{X} + K_{\vec{X}} H^T \left[HK_{\vec{X}} H^T + K_{\vec{Z}}\right]^{-1} (\vec{y} - H\overline{X}) \tag{4.125}$$

$$\widehat{X}(\vec{y}) \;=\; \overline{X} + K_{\vec{\Xi}} H^T K_{\vec{Z}}^{-1} (\vec{y} - H\overline{X}) \tag{4.126}$$

The error covariance is still given by (4.81) and (4.83).

4.5 Vector Recursive and Kalman Estimation

We now make multiple noisy vector observations, $\vec{y}_1, \vec{y}_2, \dots$, of a single m dimensional Gaussian rv⃗ $\vec{X} \sim \mathcal{N}(\overline{X}, K_{\vec{X}})$. For each i, we assume the observation random vectors have the form

$$\vec{Y}_i = H_i\vec{X} + \vec{Z}_i; \quad \vec{Z}_i \sim \mathcal{N}(\vec{0}, K_{\vec{Z}_i}); \quad i \geq 1 \tag{4.127}$$

We assume that $\vec{X}, \vec{Z}_1, \vec{Z}_2, \dots$ are mutually independent and that $H_1, H_2, \dots$ are known matrices. For each i, we want to find the MMSE estimate of $\vec{X}$ based on $\{\vec{y}_j;\, 1 \leq j \leq i\}$. We will do this recursively, using the estimate based on $\vec{y}_1, \dots, \vec{y}_{i-1}$ to help in finding the estimate based on $\vec{y}_1, \dots, \vec{y}_i$. Let $\vec{Y}_{1;i}$ denote the first i observation rv⃗'s and let $\vec{y}_{1;i}$ denote the corresponding i sample observations. Let $\widehat{X}(\vec{y}_{1;i})$ denote the MMSE estimate, $\vec{\Xi}_i = \widehat{X}(\vec{y}_{1;i}) - \vec{X}$ denote the estimation error, and $K_{\vec{\Xi}_i}$ denote the error covariance, all based on the observation $\vec{y}_{1;i} = \vec{y}_1, \vec{y}_2, \dots, \vec{y}_i$. For $i = 1$, the result is the same as that in (4.125) and (4.126), yielding

$$\widehat{X}(\vec{y}_1) = \overline{X} + K_{\vec{X}}H_1^T\left[H_1K_{\vec{X}}H_1^T + K_{\vec{Z}_1}\right]^{-1}(\vec{y}_1 - H_1\overline{X}) \tag{4.128}$$

$$\widehat{X}(\vec{y}_1) = \overline{X} + K_{\vec{\Xi}_1}H_1^T K_{\vec{Z}_1}^{-1}(\vec{y}_1 - H_1\overline{X}) \tag{4.129}$$

From (4.81) and (4.83), the error covariance is

$$K_{\vec{\Xi}_1} = K_{\vec{X}} - K_{\vec{X}}H_1^T[H_1K_{\vec{X}}H_1^T + K_{\vec{Z}_1}]^{-1}H_1K_{\vec{X}} \tag{4.130}$$

$$K_{\vec{\Xi}_1}^{-1} = K_{\vec{X}}^{-1} + H_1^TK_{\vec{Z}_1}^{-1}H_1 \tag{4.131}$$

Using the same argument that we used for the scalar case, we see that for each $i > 1$, the conditional mean of $\vec{X}$, conditional on the observation $\vec{y}_{1;i-1}$, is by definition $\widehat{X}(\vec{y}_{1;i-1})$ and the conditional covariance is $K_{\vec{\Xi}_{i-1}}$. Using these quantities in place of $\overline{X}$ and $K_{\vec{X}}$ in (4.125) and (4.126) then yields

$$\widehat{X}(\vec{y}_{1;i}) = \widehat{X}(\vec{y}_{1;i-1}) + K_{\vec{\Xi}_{i-1}}H_i^T\left[H_iK_{\vec{\Xi}_{i-1}}H_i^T + K_{\vec{Z}_i}\right]^{-1}(\vec{y}_i - H_i\widehat{X}(\vec{y}_{1;i-1})) \tag{4.132}$$

$$\widehat{X}(\vec{y}_{1;i}) = \widehat{X}(\vec{y}_{1;i-1}) + K_{\vec{\Xi}_{i-1}}H_i^TK_{\vec{Z}_i}^{-1}(\vec{y}_i - H_i\widehat{X}(\vec{y}_{1;i-1})) \tag{4.133}$$

From (4.81) and (4.83), the error covariance is given by

$$K_{\vec{\Xi}_i} = K_{\vec{\Xi}_{i-1}} - K_{\vec{\Xi}_{i-1}}H_i^T\left[H_iK_{\vec{\Xi}_{i-1}}H_i^T + K_{\vec{Z}_i}\right]^{-1}H_iK_{\vec{\Xi}_{i-1}} \tag{4.134}$$

$$K_{\vec{\Xi}_i}^{-1} = K_{\vec{\Xi}_{i-1}}^{-1} + H_i^TK_{\vec{Z}_i}^{-1}H_i \tag{4.135}$$

4.5.1 Vector Kalman Filter

Finally, consider the vector case of recursive estimation on a sequence of m dimensional time varying Gaussian rv⃗'s, say $\vec{X}_1, \vec{X}_2, \dots$. Assume that $\vec{X}_1 \sim \mathcal{N}(\overline{X}_1, K_{\vec{X}_1})$, and that $\vec{X}_{i+1}$

evolves from $\vec{X}_i$ for $i \geq 1$ according to the equation

$$\vec{X}_{i+1} = A_i\vec{X}_i + \vec{W}_i\,; \quad \vec{W}_i \sim \mathcal{N}(0, K_{\vec{W}_i})\,; \quad i \geq 1 \tag{4.136}$$

Here, for each $i \geq 1$, A_i is a given matrix and $K_{\vec{W}_i}$ is a given invertible covariance matrix.. Noisy n dimensional observations $\vec{Y}_i$ are made satisfying

$$\vec{Y}_i = H_i\vec{X}_i + \vec{Z}_i\,; \quad \vec{Z}_i \sim \mathcal{N}(0, K_{\vec{Z}_i})\,; \quad i \geq 1 \tag{4.137}$$

where for each $i \geq 1$, H_i is a known matrix, and $K_{\vec{Z}_i}$ is an invertible covariance matrix.

Assume that $\vec{X}_1, \{\vec{W}_i; i \geq 1\}$ and $\{\vec{Z}_i; i \geq 1\}$ are all independent. As in the scalar case, we want to find the MMSE estimate of both $\vec{X}_i$ and $\vec{X}_{i+1}$ conditional on $\vec{Y}_1, \vec{Y}_2, \dots, \vec{Y}_i$. We denote these estimates as $\widehat{X}_i(\vec{y}_{1;i})$ and $\widehat{X}_{i+1}(\vec{y}_{1;i})$ respectively. We denote the errors in these estimates as $\vec{\Xi}_i = \widehat{X}_i(\vec{y}_{1;i}) - \vec{X}_i$ and $\vec{\zeta}_{i+1} = \widehat{X}_{i+1}(\vec{y}_{1;i}) - \vec{X}_{i+1}$ respectively and we denote the covariance of these estimation errors as $K_{\vec{\Xi}_i}$ and $K_{\vec{\zeta}_{i+1}}$ respectively. For $i = 1$, the problem is the same as in (4.128-4.131). Alternate forms for the estimate and error covariance are

$$\widehat{X}_1(\vec{y}_1) = \overline{X}_1 + K_{\vec{X}_1}H_1^T\left[H_1K_{\vec{X}_1}H_1^T + K_{\vec{Z}_1}\right]^{-1}(\vec{y}_1 - H_1\overline{X}_1) \tag{4.138}$$

$$\widehat{X}_1(\vec{y}_1) = \overline{X}_1 + K_{\vec{\Xi}_1}H_1^TK_{\vec{Z}_1}^{-1}(\vec{y}_1 - H_1\overline{X}_1) \tag{4.139}$$

$$K_{\vec{\Xi}_1} = K_{\vec{X}_1} - K_{\vec{X}_1}H_1^T[H_1K_{\vec{X}_1}H_1^T + K_{\vec{Z}_1}]^{-1}H_1K_{\vec{X}_1} \tag{4.140}$$

$$K_{\vec{\Xi}_1}^{-1} = K_{\vec{X}_1}^{-1} + H_1^TK_{\vec{Z}_1}^{-1}H_1 \tag{4.141}$$

This means that, conditional on $\vec{Y}_1 = \vec{y}_1$, $\vec{X}_1 \sim \mathcal{N}(\widehat{X}_1(\vec{y}_1), K_{\vec{\Xi}_1})$. Thus, conditional on $\vec{Y}_1 = \vec{y}_1$, $\vec{X}_2 = A_1\vec{X}_1 + \vec{W}_1$ is Gaussian with mean $A_1\widehat{X}_1(\vec{y}_1)$ and with covariance $A_1K_{\vec{\Xi}_1}A_1^T + K_{\vec{W}_1}$. Thus

$$\widehat{X}_2(\vec{y}_1) = A_1\widehat{X}_1(\vec{y}_1)\,; \quad K_{\vec{\zeta}_2} = A_1K_{\vec{\Xi}_1}A_1^T + K_{\vec{W}_1} \tag{4.142}$$

Conditional on $\vec{Y}_1 = \vec{y}_1$, (4.142) gives the mean and covariance of $\vec{X}_2$. Thus, given the additional observation $\vec{Y}_2 = \vec{y}_2$, where $\vec{Y}_2 = H_2\vec{X}_2 + \vec{Z}_2$, we can use (4.125) and (4.126) with this mean and covariance to get the alternative forms

$$\widehat{X}_2(\vec{y}_{1;2}) = \widehat{X}_2(\vec{y}_1) + K_{\vec{\zeta}_2}H_2^T\left[H_2K_{\vec{\zeta}_2}H_2^T + K_{Z_2}\right]^{-1}(\vec{y}_2 - H_2\widehat{X}_2(\vec{y}_1)) \tag{4.143}$$

$$\widehat{X}_2(\vec{y}_{1;2}) = \widehat{X}_2(\vec{y}_1) + K_{\vec{\Xi}_2}H_2^TK_{\vec{Z}_2}^{-1}(\vec{y}_2 - H_2\vec{X}_2(\vec{y}_1)) \tag{4.144}$$

The covariance of the error is given by the alternative forms

$$K_{\vec{\Xi}_2} = K_{\vec{\zeta}_2} - K_{\vec{\zeta}_2}H_2^T\left[H_2K_{\vec{\zeta}_2}H_2^T + K_{\vec{Z}_2}\right]^{-1}H_2K_{\vec{\zeta}_2} \tag{4.145}$$

$$K_{\vec{\Xi}_2}^{-1} = K_{\vec{\zeta}_2}^{-1} + H_2^TK_{\vec{Z}_2}^{-1}H_2 \tag{4.146}$$

Continuing this same argument recursively for all $i > 1$, we obtain the Kalman filter equations,

$$\widehat{X}_i(\vec{y}_{1;i}) = \widehat{X}_i(\vec{y}_{1;i-1}) + K_{\vec{\zeta}_i} H_i^T \Big[H_i K_{\vec{\zeta}_1} H_i^T + K_{\vec{Z}_i}\Big]^{-1} (\vec{y}_i - H_i \widehat{X}_i(\vec{y}_{1;i-1})) \qquad (4.147)$$

$$\widehat{X}_i(\vec{y}_{1;i}) = \widehat{X}_i(\vec{y}_{1;i-1}) + K_{\vec{\Xi}_i} H_i^T K_{\vec{Z}_i}^{-1} (\vec{y}_i - H_i \widehat{X}_i(\vec{y}_{1;i-1})) \qquad (4.148)$$

$$K_{\vec{\Xi}_i} = K_{\vec{\zeta}_i} - K_{\vec{\zeta}_i} H_i^T [H_i K_{\vec{\zeta}_1} H_i^T + K_{\vec{Z}_i}]^{-1} H_i K_{\vec{\zeta}_i} \qquad (4.149)$$

$$K_{\vec{\Xi}_i}^{-1} = K_{\vec{\zeta}_i}^{-1} + H_i^T K_{\vec{Z}_i}^{-1} H_i \qquad (4.150)$$

$$\widehat{X}_{i+1}(\vec{y}_{1;i}) = A_i \widehat{X}_i(\vec{y}_{1;i}) \quad ; \qquad K_{\vec{\zeta}_{i+1}} = A_i K_{\vec{\Xi}_i} A_i^T + K_{\vec{W}_i} \qquad (4.151)$$

The alternative forms above are equivalent and differ in the size and type of matrix inversions required; these matrix inversions do not depend on the data, however, and thus can be precomputed.

4.6 Estimation for Complex Random Variables

A complex random variable (crv) is a mapping from the underlying sample space onto the complex numbers. A crv X can be viewed as a pair of real rv's, X_r and X_{im}, where for each sample point ω, $X_r(\omega) = \Re[X(\omega)]$ and $X_{im}(\omega) = \Im[X(\omega)]$. Thus $X = X_r + jX_{im}$ where $j = \sqrt{-1}$. The complex conjugate X^* of a crv X is $X_r - jX_{im}$. It is always possible to simply represent each crv as a pair of real rv's, but this often obscures insights. In what follows, we first look at vectors of crv's and their covariance matrices. We then look at crv's as vectors in their own right and extend the orthogonality principle to the corresponding complex inner product space.

Let $X_1, \dots, X_n$ be an n-tuple of crv's. We refer to $\vec{X} = (X_1, \dots, X_n)^T$ as a complex r$\vec{\text{v}}$. It is a mapping from sample points into n-vectors of complex numbers. The complex conjugate, $\vec{X}^*$, of a complex r$\vec{\text{v}}$ $\vec{X}$ is the vector of complex conjugates, $(X_1^*, \dots, X_n^*)^T$. The *covariance matrix* of a zero mean complex r$\vec{\text{v}}$ $\vec{X}$ is defined as

$$K_{\vec{X}} = E\left[\vec{X}\vec{X}^{*T}\right] \qquad (4.152)$$

The covariance matrix K of a complex r$\vec{\text{v}}$ has the special property that $K = K^{*T}$. Such matrices are called Hermetian. A complex matrix is said to be positive definite (positive semi-definite) if it is Hermetian and if, for all complex non-zero vectors $\vec{b}$, $\vec{b}^T K \vec{b}^* > 0 (\geq 0)$. A matrix K is a covariance matrix of a complex r$\vec{\text{v}}$ iff K is positive semi-definite.

All the eigenvalues of a Hermitian matrix are real, and the eigenvectors $\vec{q}_i$ (which in general are complex) can be chosen to span the space and to be orthogonal (in the sense that $\vec{q}_i{}^T \vec{q}_j{}^* = 0, i \neq j$). These properties of Hermetian matrices can be derived in much the same way as the properties of symmetric matrices were derived in section 2.4.

There is a very important difference between covariance matrices of complex r$\vec{v}$'s and real r$\vec{v}$'s. The covariance matrix of a complex r$\vec{v}$ $\vec{X}$ does not contain all the second moment information about the $2n$ real rv's $X_{1,r}, X_{1,im}, \ldots, X_{n,r}, X_{n,im}$. For example, in the one dimensional case, we have

$$E[XX^*] = E[(X_r + jX_{im})(X_r - jX_{im})] = E[X_r^2 + X_{im}^2] \tag{4.153}$$

This only specifies the sum of the variances of X_r and X_{im}, but does not specify the individal variances or the covariance. As shown in Exercise 4.11, all of the second moment information can be extracted from the combination of $K_{\vec{X}}$ and a matrix $E\left[\vec{X}\vec{X}^T\right]$ sometimes known as the pseudo-covariance matrix.

In many of the situations in which complex random variables appear naturally, there is a type of circular symmetry which causes $E\left[\vec{X}\vec{X}^T\right]$ to be identically zero. To be more specific, a complex r$\vec{v}$ $\vec{X}$ is *circularly symmetric* if $\vec{X}$ and $e^{j\phi}\vec{X}$ are statistically the same (i.e., have the same joint distribution function) for all ϕ. This implies that each complex random variable X_i, $1 \leq i \leq n$, is circularly symmetric in the sense that if the joint density for $X_{i,r}, X_{i,im}$ is expressed in polar form, it will be independent of the angle. It also implies that if two complex variables, X_i and X_k, are each rotated by the same angle, then the joint density will again be unchanged. If $\vec{X}$ is circularly symmetric, then

$$E\left[\vec{X}\vec{X}^T\right] = E\left[(e^{j\phi}\vec{X})(e^{j\phi}\vec{X})^T\right] = e^{2j\phi}E\left[\vec{X}\vec{X}^T\right] \tag{4.154}$$

This equation can only be satisfied for all ϕ if $E\left[\vec{X}\vec{X}^T\right] = 0$. Thus for circularly symmetric complex r$\vec{v}$'s, the pseudo-covariance matrix is zero and the covariance matrix specifies all the second moments. We also define $\vec{X}$ and $\vec{Y}$ to be jointly circularly symmetric if $(\vec{X}^T, \vec{Y}^T)^T$ is circularly symmetric.

Next, we extend the definition of real vector spaces to complex vector spaces. This is particularly easy since the definition of a *complex vector space* is the same as that of a real vector space except that the scalars are now complex numbers rather than real numbers. We must also extend inner products. A *complex inner product space* is a complex vector space $\mathcal{V}$ with an inner product $\langle X, Y\rangle$ that maps pairs of vectors X, Y into complex numbers. The axioms are the same as for real inner product spaces except that for all $X, Y \in \mathcal{V}$,

$$\langle X, Y\rangle = \langle Y, X\rangle^* \tag{4.155}$$

As with real inner product spaces, two vectors, X, Y are orthogonal if $\langle X, Y\rangle = 0$. Also, the *length* of X is defined to be $\| X \| = \sqrt{\langle X, Y\rangle}$, and the *distance* between X and Y is $\| X - Y \|$. For the conventional complex vector space in which $\vec{x}$ is an n-tuple of complex numbers, the inner product $\langle \vec{x}, \vec{y}\rangle$ is conventionally taken to be $\langle \vec{x}, \vec{y}\rangle = x_1y_1^* + \cdots + x_ny_n^*$. Note that this choice makes $\langle x, x\rangle \geq 0$, whereas without the complex conjugates, this would not be true. The complex vector space of interest here is the set of zero mean crv's defined in some given probability space. The inner product for zero mean crv's is defined to be the covariance of the crv's, i.e.,

$$\langle X, Y\rangle = E[XY^*] \tag{4.156}$$

It can be verified directly that this definition satisfies all the axioms of an inner product.

The orthogonality principle can also be applied to complex inner product vector spaces. The proof is a minor modification of the proof in the real case and is left as Exercise 4.12.

Theorem 4.5 (ORTHOGONALITY PRINCIPLE, COMPLEX CASE) *Let X be a vector in a complex inner product space $\mathcal{V}$, and let $\mathcal{S}$ be a finite dimensional subspace of $\mathcal{V}$. There is a unique vector $P \in \mathcal{S}$ such that $\langle X - P, Y\rangle = 0$ for all $Y \in \mathcal{S}$. That vector P is the closest point in $\mathcal{S}$ to X, i.e., $\| X - P \| < \| X - Y \|$ for all $Y \in \mathcal{S}$, $Y \neq P$.*

Consider LLSE estimation of a crv X from the crv's $Y_1, \dots, Y_n$. The LLSE estimate for crv's is defined as

$$\widehat{X}(Y_1, \dots, Y_n) = \sum_i \alpha_i Y_i \tag{4.157}$$

where $\alpha_1, \dots, \alpha_n$ are complex numbers chosen to minimize $E\left[\left|\widehat{X}(Y_1, \dots, Y_n) - X\right|^2\right]$. In terms of the complex vector space of zero mean crv's, the quantity to be minimized is $\| \widehat{X}(Y_1, \dots, Y_n) - X \|$. Thus, from the theorem, $\widehat{X}(Y_1, \dots, Y_n)$ is the projection P of X on the space spanned by $Y_1, \dots, Y_n$. Since $\langle X - P, Y\rangle = 0$ for all Y in the subspace spanned by $Y_1, \dots, Y_n$, it suffices to find that P for which $\langle X - P, Y_i\rangle = 0$ for $1 \leq i \leq n$. Equivalently, P must satisfy $\langle X, Y_i\rangle = \langle P, Y_i\rangle$ for $1 \leq i \leq n$. Since $P = \widehat{X}(Y_1, \dots, Y_n) = \sum_{j=1}^n \alpha_j Y_j$, this means that $\alpha_1, \dots, \alpha_n$ must satisfy (4.57). Using (4.156), we have

$$E[XY_i^*] = \sum_{j=1}^{n} \alpha_j E[Y_j Y_i^*] \quad ; \qquad ; 1 \leq i \leq n \tag{4.158}$$

If we define $K_{X\vec{Y}}$ as $E[X\vec{Y}^{*T}]$, this can be written in vector form as

$$\vec{\alpha}^T K_{\vec{Y}} = K_{X\vec{Y}}\,; \qquad \vec{\alpha}^T = K_{X\vec{Y}} K_{\vec{Y}}^{-1} \tag{4.159}$$

Thus the LLSE estimate $\widehat{X}(\vec{Y})$ is given by

$$\widehat{X}(\vec{Y}) = K_{X\vec{Y}} K_{\vec{Y}}^{-1} \vec{Y}, \tag{4.160}$$

This is valid whether or not the variables are circularly symmetric. This is somewhat surprising since the covariance matrices do not contain all the information about second moments in this case. We explain this shortly, but first, we look at LLSE estimation of a vector of complex variables, $\vec{X}$, from the observation of $\vec{Y}$. Since (4.160) can be used for the LLSE estimate of each component of $\vec{X}$, the LLSE estimate of the vector $\vec{X}$ is given by

$$\widehat{X}(\vec{Y}) = K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} \vec{Y}, \tag{4.161}$$

The estimation error matrix, $K_{\vec{X}|\vec{Y}} = E\left[\left(\widehat{X}(\vec{Y}) - \vec{X}\right)\left(\widehat{X}(\vec{Y}) - \vec{X}\right)^{*T}\right]$ is then given by

$$K_{\vec{X}|\vec{Y}} = K_{\vec{X}} - K_{\vec{X}\vec{Y}} K_{\vec{Y}}^{-1} K_{\vec{X}\vec{Y}}^{*T} \tag{4.162}$$

The surprising nature of these results is resolved by being careful about what linearity means. What we have found in (4.160) is the linear least squares estimate of the crv X based on the crv's $Y_1, \dots Y_n$. This is **not** the same as the linear least squares estimate of X_r and X_{im} based on $Y_{i,r}, Y_{i,im}\,;\; 1 \le i \le n$. This can be seen by noting that an arbitrary linear transformation from two variables Y_r, Y_{im} to $\widehat{X}_r, \widehat{X}_{im}$ is specified by a two by two matrix, which contains four arbitrary real numbers. An arbitrary linear transformation from a complex variable Y to a complex variable $\widehat{X}$ is specified by a single complex variable, i.e., by two real variables.

To further clarify the difference between LLSE estimates on complex variables and LLSE estimates on the real and imaginary parts of the complex variables, consider estimating the crv X as a linear function of both $\vec{Y}$ and $\vec{Y}^*$, i.e.,

$$\widehat{X}(\vec{Y}) = \sum_i [\alpha_i Y_i + \beta_i Y_i^*] \tag{4.163}$$

where α_i and β_i are chosen, for each i, to minimize the mean squared error. From the orthogonality theorem, 4.4,

$$\langle X\,,\,Y_i\rangle \;=\; \sum_j \alpha_j \langle Y_j\,,\,Y_i\rangle + \sum_j \beta_j \langle Y_j^*\,,\,Y_i\rangle \tag{4.164}$$

$$\langle X\,,\,Y_i^*\rangle \;=\; \sum_j \alpha_j \langle Y_j\,,\,Y_i^*\rangle + \sum_j \beta_j \langle Y_j^*\,,\,Y_i^*\rangle \tag{4.165}$$

It can be seen from (4.163) that this estimate must have a real part and imaginary part that are linear functions of the real and imaginary parts of $\vec{Y}$. It can also be seen (see Exercise 4.11) that any such linear function of the real and imaginary parts can be represented in this way. Thus the LLSE estimate based on complex variables using both the variables and complex conjugates is equivalent to the LLSE estimate for real and imaginary parts.

Now consider the case where X and $\vec{Y}$ are jointly circularly symmetric. Then $\langle X\,,\,Y_i^*\rangle = 0$ and $\langle Y_j\,,\,Y_i^*\rangle = \langle Y_j^*\,,\,Y_i\rangle = 0$. In this case, (4.165) implies that $\beta_j = 0$ for each j, and (4.163) is then the same as (4.159). This means that the complex LLSE estimate of (4.160) is the same as the LLSE estimate using real and imaginary values in the circularly symmetric case. Similarly if $\vec{X}$ and $\vec{Y}$ are jointly circularly symmetric, the complex LLSE estimate in (4.161) is the same as the LLSE estimate using real and imaginary values, and the covariance error is also the same.

The next question is when the circularly symmetric case arises. Consider the case $\vec{Y} = H\vec{X} + \vec{Z}$ and suppose that $\vec{X}$ and $\vec{Z}$ are independent and each circularly symmetric. The first thing to observe is that for any ϕ, $e^{j\phi}\vec{Y} = H[e^{j\phi}\vec{X}] + e^{j\phi}\vec{Z}$. Thus, since $\vec{X}$ and $e^{j\phi}\vec{X}$ have the same distribution and $\vec{Z}$ and $e^{j\phi}\vec{Z}$ have the same distribution, $\vec{Y}$ and $e^{j\phi}\vec{Y}$ also have the same distribution. Thus $\vec{Y}$ is circularly symmetric. In the same way, we see that $\vec{X}$ and $\vec{Y}$ are jointly circularly symmetric. Thus, in this case, the complex LLSE estimate is again the same as the real LLSE estimate.

Finally, define $\vec{X}$ to be a complex Gaussian r$\vec{v}$ if $\Re(\vec{X})$ and $\Im(\vec{X})$ are jointly Gaussian. Similarly $\vec{X}$ and $\vec{Y}$ are jointly complex Gaussian if the real and imaginary parts are collectively jointly Gaussian. For this jointly Gaussian case, the MMSE estimate of $\vec{X}$ from observation of $\vec{Y}$ is given by the MMSE estimate of the real and imaginary parts of $\vec{X}$ from the observation of the real and imaginary parts of $\vec{Y}$. This is the same as the LLSE estimate of $\vec{X}$ as a linear function of both $\vec{Y}$ and $\vec{Y}^*$. Finally, if $\vec{X}$ and $\vec{Y}$ are jointly circularly symmetric, the MMSE estimate is also given by (4.161).

For the circularly symmetric Gaussian case, it is also nice to express the density of $\vec{X}$ in terms of the covariance matrix, $K_{\vec{X}}$. As shown in Exercises 4.14 and 4.15, this is given by

$$p_{\vec{X}}(\vec{x}) = \frac{\exp\left[-\vec{x}^* K_{\vec{X}}^{-1}\vec{x}\right]}{\pi^n det(K_{\vec{X}})} \tag{4.166}$$

4.7 Parameter Estimation and the Cramer-Rao Bound

We now focus on estimation problems in which there is no appropriate model for a priori probabilities on the quantity x to be estimated. We view x as a parameter which is known to lie in some interval on the real line. We can view the parameter x as a sample value of a random variable X whose distribution is unknown to us, or we can view it simply as an unknown value. When x is viewed simply as an unknown value, then we don't have an overall probability space to work with—we only have a probability space for each individual value of x. We can not take overall expected values of random variables, but can only take expected values given particular values of x. Strictly speaking, we can not even regard the observation as a random vector, but can only view the observation as a distinct random vector for each value of x. By viewing x as a sample value of a rv X whose distribution is unknown, we avoid these notational problems. We denote the expected value of $\vec{Y}$, given $X = x$, by $E_x(\vec{Y})$, but the overall expected value of $\vec{Y}$ can not be found, since the distribution of X is unknown.

Consider an estimation problem in which we want to estimate the parameter x, and we observe a sample value $\vec{y}$ of the observation $\vec{Y}$. Let $f(\vec{y} \mid x)$ denote the probability density of $\vec{Y}$ at sample value $\vec{y}$, given that the parameter has the value x. This is not a conditional probability in the usual sense, since x is only a parameter. We assume that small changes in x correspond to small changes in the density $f(\vec{y} \mid x)$, and in fact we assume that $f(\vec{y} \mid x)$ and $\ln[f(\vec{y} \mid x)]$ are differentiable with respect to x. Define $V_x(\vec{y})$ as

$$V_x(\vec{y}) = \frac{\partial \ln(f(\vec{y} \mid x))}{\partial x} = \frac{1}{f(\vec{y} \mid x)} \frac{\partial (f(\vec{y} \mid x)}{\partial x} \tag{4.167}$$

Recall that the maximum likelihood estimate $\widehat{X}_{ML}(\vec{y})$ is that value of x that maximizes $f(\vec{y} \mid x)$ for the given observation $\vec{y}$. $\widehat{X}_{ML}(\vec{y})$ also is the x that maximizes $\ln[f(\vec{y} \mid x)]$, and if the maximum occurs at a stationary point, it occurs where $V_x(\vec{y}) = 0$. Note that for each x, $V_x(\vec{Y})$ is a random variable, and as shown in the next equation, $V_x(\vec{Y})$ has zero mean for each x.

$$E_x[V_x(\vec{Y})] \quad = \quad \int f(\vec{y} \mid x) V_x(\vec{y})\, d\vec{y} = \int f(\vec{y} \mid x) \frac{1}{f(\vec{y} \mid x)} \frac{\partial f(\vec{y} \mid x)}{\partial x}\, d\vec{y}$$

$$= \frac{\partial \int f(\vec{y} \mid x)\, d\vec{y}}{\partial x} = \frac{\partial 1}{\partial x} = 0 \tag{4.168}$$

It is not immediately apparent why $V_x(Y)$ is a fundamental quantity, but as one indication of this, note that if $f(y \mid x)$ is replaced by the likelihood ratio $\Lambda(y, x)$ in (4.167), the derivative remains the same. Thus $V_x(\vec{y})$ is the partial derivative, with respect to x, of the log likelihood ratio $\mathrm{LLR}(\vec{y}, x)$.

Now suppose that some value of x is selected (x is simply a parameter, and we assume no a priori probability distribution on it), and an observation $\vec{y}$ occurs according to the density $f(\vec{y} \mid x)$. An observer, given the value $\vec{y}$, chooses an estimate $\widehat{X}(\vec{y})$ of x. Thus $\widehat{X}(\vec{y})$ is a function of $\vec{y}$, and thus, for any given parameter value x, $\widehat{X}(\vec{Y})$ can be considered as a random variable (conditional on x). The Cramer-Rao bound gives us a lower bound on $E_x[(x - \widehat{X}(\vec{Y}))^2]$, the second moment of $\widehat{X}(\vec{Y})$ around x, given x. One may think of this as a communication channel with input x and output $\vec{Y}$, with $\vec{Y}$ taking on the value $\vec{y}$ with probability density $f(\vec{y} \mid x)$. For each choice of x, one can in principle calculate $E_x[(x - \widehat{X}(\vec{Y}))^2]$, and this depends on the particular estimate $\widehat{X}(\vec{y})$ and also on the particular x. One can make this estimate better for some values of x by making it poorer for others, but no matter how one does this,$E_x[(x - \widehat{X}(\vec{Y}))^2]$ must, for each x, satisfy a soon to be derived lower bound called the Cramer-Rao bound.

The *bias*[9], $b_{\widehat{X}}(x)$, of an estimate $\widehat{X}(\vec{y})$ is defined as

$$b_{\widehat{X}}(x) = E_x[\widehat{X}(\vec{Y}) - x] \tag{4.169}$$

An estimate $\widehat{X}$ is called unbiased if $b_{\widehat{X}}(x) = 0$ for all x. Many people take it for granted that estimates should be unbiased, but there are many situations in which unbiased estimates do not exist, and others in which they exist but are not particularly desirable (see Exercise 4.16).

The Fisher information $J(x)$ is defined as the variance (conditional on x) of $V_x(\vec{Y})$.

$$J(x) = VAR_x[V_x(\vec{Y})] = E_x[(V_x(\vec{Y}))^2] \tag{4.170}$$

For the example where Y, conditional on x, is $\mathcal{N}(x, \sigma^2)$, $V_x(Y) = (Y - x)/\sigma^2$, and thus $J(x) = 1/\sigma^2$ (see Exercise 4.17). Viewing Y as a noisy measurement of x, we see that the Fisher information gets smaller as the measurement noise gets larger. The major reason for considering the Fisher information comes from the Cramer-Rao bound below, and one is advised to look at the bound and some examples rather than trying to get an abstract sense of why one might call this an information.

Theorem 4.6 (CRAMER-RAO BOUND) *For each x, and any estimator $\widehat{X}(\vec{Y})$,*

$$VAR_x\left[\widehat{X}(\vec{Y})\right] \geq \frac{\left[1 + \frac{\partial b_{\widehat{X}}(x)}{\partial x}\right]^2}{J(x)} \tag{4.171}$$

[9] When the parameter x is a sample of a known random variable X, the expected bias is $E[\widehat{X}(\vec{Y}) - X]$. The expected bias in this case is sometimes simply called bias, causing some confusion. Bias as we define it is a function of x, whereas the expected bias is a single number, formed by averaging over X.

Proof:

$$\begin{aligned}
E_x[V_x(\vec{Y})\widehat{X}(\vec{Y})] &= \int f(\vec{y} \mid x)V_x(\vec{y})\widehat{X}(\vec{y})d\vec{y} \\
&= \int \frac{\partial f(\vec{y} \mid x)}{\partial x}\widehat{X}(\vec{y})d\vec{y} \qquad \text{(from (4.167))} \\
&= \frac{\partial}{\partial x}\int f(\vec{y} \mid x)\widehat{X}(\vec{y})dy \qquad \text{(interchanging differentiation and integration)} \\
&= \frac{\partial E_x[\widehat{X}(\vec{Y})]}{\partial x} = \frac{\partial[x + b_{\widehat{X}}(x)]}{\partial x} \qquad \text{(from (4.169))} \\
&= 1 + \frac{\partial b_{\widehat{X}}(x)}{\partial x} \qquad (4.172)
\end{aligned}$$

Let $\widetilde{X_x}(\vec{Y}) = \widehat{X}(\vec{Y}) - E_x(\widehat{X}(\vec{Y}))$ be the fluctuation in $\widehat{X}(\vec{Y})$ for given x. Since $V_x(\vec{Y})$ is zero mean, we know that

$$E_x\left[V_x\vec{Y})\,\widetilde{X_x}(\vec{Y})\right] = E_x[V_x(\vec{Y})\widehat{X}(\vec{Y})] = 1 + \frac{\partial b_{\widehat{X}}(x)}{\partial x} \qquad (4.173)$$

Since the normalized covariance between V_x and $\widehat{X}(\vec{Y})$ (for given x) must be at most 1, we have

$$E_x^2[V_x(\vec{Y})\,\widetilde{X_x}(\vec{Y})] \leq VAR_x[V_x(\vec{Y})]VAR_x[\widehat{X}(\vec{Y})] = J(x)VAR_x[\widehat{X}(\vec{Y})] \qquad (4.174)$$

Combining (4.173) and (4.174) yields (4.171).

The usual quantity of interest is the mean square estimation error for any given x, i.e., $E_x[(\widehat{X}(\vec{Y}) - x)^2]$. Since this is equal to $VAR_x[\widehat{X}(\vec{Y})] + E_x^2[\widehat{X}(\vec{Y})] - x$, and since $b_{\widehat{X}} = E_x[\widehat{X}(\vec{Y})] - x$, (4.171) is equivalent to

$$E_x\left[(\widehat{X}(\vec{Y}) - x)^2\right] \geq \frac{\left[1 + \frac{\partial b_{\widehat{X}}(x)}{\partial x}\right]^2}{J(x)} + \left[b_{\widehat{X}}(x)\right]^2 \qquad (4.175)$$

Finally, if we restrict attention to the special case of unbiased estimates, (4.175) becomes

$$E_x[(\widehat{X}(\vec{Y}) - x)^2] \geq \frac{1}{J(x)} \qquad (4.176)$$

This is the usual form of the Cramer-Rao inequality. Many people mistakenly believe that unbiased estimates must be better than biased, and that therefore biased estimates also satisfy (4.176), but this is untrue, and as mentioned before, there are many situations in which no unbiased estimates exist.

4.8 EXERCISES

Exercise 4.1 Let X and Z be IID random variables and let $Y = X + Z$. For the following densities for X and Y, find the minimum mean square estimate $\widehat{X}(y)$ of the sample value of X given the observation $Y = y$.

a) $p_X(x) = 1$ for $0 \leq x \leq 1$ and $p_X(x) = 0$ elsewhere.

b) $p_X(x) = e^{-x}$ for $x \geq 0$ and $p_X(x) = 0$ for $x < 0$.

c $p_X(x)$ is arbitrary.

Exercise 4.2 **a)** Consider the joint probability density $p_{X,Z}(x, z) = e^{-z}$ for $0 \leq x \leq z$ and $p_{X,Z}(x, z) = 0$ otherwise. Find the pair x, z of values that maximize this density. Find the marginal density $p_Z(z)$ and find the value of z that maximizes this.

b) Let $p_{X,Z,Y}(x, z, y)$ be $y^2 e^{-yz}$ for $0 \leq x \leq z$, $1 \leq y \leq 2$ and be 0 otherwise. Conditional on an observation $Y = y$, find the joint MAP estimate of X, Z. Find $p_{Z|Y}(z \mid y)$, the marginal density of Z conditional on $Y = y$, and find the MAP estimate of Z conditional on $Y = y$.

c) Explain why the expected minimum cost estimate for any vector $\vec{X} = (X_1, \dots, X_n)^T$ must be the vector of the expected minimum cost estimates of the individual components $\widehat{X}_1, \dots, \widehat{X}_n$ for both the squared cost function and the absolute value cost function.

Exercise 4.3 **a)** Let $X, Z_1, Z_2, \dots$ be independent zero mean Gaussian rv's with variances $\sigma_X^2, \sigma_{Z_1}^2, \dots$ respectively. Let $Y_i = h_i X + Z_i$ for $i \geq 1$ and let $\vec{Y}_i = (Y_1, \dots Y_i)^T$. Let $\vec{a}_i$ be the row vector $\vec{a}_i = K_{X\vec{Y}_i} K_{\vec{Y}_i}^{-1}$. Multiply this on the right by $K_{\vec{Y}_i}$ to solve for $\vec{a}_i$. Then use (4.13) to show that the MMSE estimate of X from $\vec{Y}_i = \vec{y}_i = (y_1, \dots, y_i)^T$, is given by

$$\widehat{X}(\vec{y}_i) = \sum_{j=1}^{i} a_{ij} y_j \,; \quad a_{ij} = \frac{h_j/\sigma_{Z_j}^2}{(1/\sigma_X^2) + \sum_{n=1}^{i} h_n^2/\sigma_{Z_n}^2} \tag{4.177}$$

b) Let $\Xi_i = \widehat{X}(\vec{Y}_i) - X$ and use a) to show that

$$\frac{1}{\sigma_{\Xi_i}^2} = \frac{1}{\sigma_X^2} + \sum_{n=1}^{i} \frac{h_n^2}{\sigma_{Z_n}^2} \tag{4.178}$$

c) Now suppose that X has a mean, $\overline{X}$. Modify (4.177) and (4.178) to account for the mean and show that the result is (4.33).

d) Show that (4.33) is consistent with (4.31) and (4.32).

Exercise 4.4 Write out $E[(X - \vec{\alpha}^T \vec{Y})^2] = \sigma_X^2 - 2K_{X\vec{Y}}\vec{\alpha} + \vec{\alpha}^T K_{\vec{Y}} \vec{\alpha}$ as a function of $\alpha_1, \alpha_2, \dots, \alpha_n$ and take the partial derivative of the function with respect to α_i for each i, $1 \leq i \leq n$. Show that the vector of these partial derivatives is $-2K_{X\vec{Y}} + 2\vec{\alpha}^T K_{\vec{Y}}$.

Exercise 4.5 Let $\widehat{X}_A(\vec{Y})$ be an arbitrary estimate of a rv X from a r$\vec{\mathrm{v}}$ $\vec{Y}$ and let $\widehat{X}_{MMSE}(\vec{Y})$ be the MMSE estimate. Let $\Xi_A = \widehat{X}_A(\vec{Y}) - X$ and $\Xi = \widehat{X}_{MMSE}(\vec{Y}) - X$ be the corresponding estimation errors.

a Show that for any given sample value $\vec{Y} = \vec{y}$,

$$E[\Xi_A^2 \mid \vec{Y}{=}\vec{y}] = E[\Xi^2 \mid \vec{Y}{=}\vec{y}] + [\widehat{X}_A(\vec{y}) - \widehat{X}_{MMSE}(\vec{y})]^2$$

b Taking the expectation over $\vec{Y}$, show that

$$E[\Xi_A^2] = E[\Xi^2] + E\left[(\widehat{X}_A(\vec{Y}) - \widehat{X}_{MMSE}(\vec{Y}))^2\right]$$

Exercise 4.6 **a)** Show that the set of vectors $\{\theta_i = \lambda_i^{-1/2}H\phi_i\,;\ 1{\le}i{\le}m\}$ is an orthonormal set, where the vectors ϕ_i; $1{\le}i{\le}m$ satisfy (4.115) and are orthonormal and $\lambda_i > 0$ for $1 \le i \le m$.

b) Show that θ_i is an eigenvector, with eigenvalue λ_i, of HH^T.

c) Show that HH^T has $n-m$ additional orthonormal eigenvectors, each of eigenvalue 0.

d) Now assume that H^TH has only $m' < m$ orthonormal eigenvectors with positive eigenvalues and has $m - m'$ orthonormal eigenvectors with eigenvalue 0. Show that HH^T has $n - m'$ orthonormal eigenvectors with eigenvalue 0.

e) Let $\vec{Y} = H\vec{X} + \vec{Z}$. Show that for each eigenvector θ_i of HH^T with eigenvalue 0, $\theta_i^T\vec{Y} = \theta_i^T\vec{Z}$ and show that the random variable $\theta_i^T\vec{Z}$ is statistically independent of $\theta_j^T\vec{Z}$ for each eigenvector θ_j with a non-zero eigenvalue.

Exercise 4.7 For a real inner product space, show that n vectors, $Y_1, \dots, Y_n$ are linearly dependent iff the matrix of inner products, $\{\langle Y_j, Y_i\rangle;\ 1 \le i,j \le n\}$, is singular.

Exercise 4.8 Show that (4.68–4.70) agree with (4.59).

Exercise 4.9 Show that if $\vec{X}$ is a Gaussian m–r$\vec{\mathrm{v}}$ and $\vec{Z}$ is a Gaussian n–r$\vec{\mathrm{v}}$ independent of $\vec{X}$, then Y and X are jointly non-singular, where $\vec{Y} = H\vec{X} + \vec{Z}$ and H is an arbitrary m by n matrix. Hint: Show that $p_{\vec{X}}(\vec{x})$ and $p_{\vec{Y}|\vec{X}}(\vec{y} \mid \vec{x})$ are bounded and explain why this establishes the desired result.

Exercise 4.10 Assume that H is invertible and derive (4.80) and (4.81) from (4.84) and (4.85) using subsection 4.4.1

Exercise 4.11 Let $\vec{X} = (X_1, \dots, X_n)^T$ be a zero mean complex r$\vec{\mathrm{v}}$ with real and imaginary components $X_{r,j}, X_{im,j}$, $1{\le}j{\le}n$ respectively. Express $E[X_{r,j}X_{r,k}]$, $E[X_{r,j}X_{im,k}]$, $E[X_{im,j}X_{im,k}]$, $E[X_{im,j}X_{r,k}]$ as functions of the components of $K_{\vec{X}}$ and $E[\vec{X}\vec{X}^T]$.

Exercise 4.12 Prove Theorem 4.5. Hint: Modify the proof of theorem 4.4 for the complex case.

Exercise 4.13 Let $Y = Y_r + jY_{im}$ be a complex random variable. For arbitrary real numbers a, b, c, d, find complex numbers α and β such that

$$\Re\left[\alpha Y + \beta Y^*\right] = aY_r + bY_{im}$$

$$\Im\left[\alpha Y + \beta Y^*\right] = cY_r + dY_{im}$$

Exercise 4.14 (Derivation of circularly symmetric Gaussian density) Let $\vec{X} = \vec{X}_r + j\vec{X}_{im}$ be a zero mean circularly symmetric n dimensional Gaussian complex r$\vec{v}$. Let $\vec{U} = (\vec{X}_r^T, \vec{X}_{im}^T)^T$ be the corresponding $2n$ dimensional real r$\vec{v}$. Let $K_r = E[\vec{X}_r \vec{X}_r^T]$ and $K_{ri} = E[\vec{X}_r \vec{X}_{im}^T]$.

a) Show that

$$K_{\vec{U}} = \begin{bmatrix} K_r & K_{ri} \\ -K_{ri} & K_r \end{bmatrix}$$

b) Show that

$$K_{\vec{U}}^{-1} = \begin{bmatrix} B & C \\ -C & B \end{bmatrix}$$

and find the B, C for which this is true.

c) Show that $K_{\vec{X}} = 2(K_r - K_{ri})$.

d) Show that $K_{\vec{X}}^{-1} = \frac{1}{2}(B - jC)$.

e) Define $p_{\vec{X}}(\vec{x}) = p_{\vec{U}}(\vec{u})$ for $\vec{u} = (\vec{x}_r^T, \vec{x}_{im}^T)^T$ and show that

$$p_{\vec{X}}(\vec{x}) = \frac{\exp -\vec{x}^* K_{\vec{X}}^{-1} \vec{x}^T}{(2\pi)^n \sqrt{\det K_{\vec{U}}}}$$

.

f) Show that

$$\det K_{\vec{U}} = \det \begin{bmatrix} K_r + jK_{ri} & K_r i - jK_r \\ -K_{ri} & K_r \end{bmatrix}$$

Hint: Recall that elementary row operations do not change the value of a determinant.

g) Show that

$$\det K_{\vec{U}} = \begin{bmatrix} K_r + jK_{ri} & 0 \\ -K_{ri} & K_r - jK_{ri} \end{bmatrix}$$

. Hint: Recall that elementary column operations do not change the value of a determinant.

h Show that

$$\det K_{\vec{U}} = 2^{-2n} \left(\det K_{\vec{X}}\right)^2$$

and from this conclude that (4.166) is valid.

Exercise 4.15 **a)** (Alternate derivation of circularly symmetric Gaussian density).

a) Let X be a circularly symmetric zero mean complex Gaussian rv with covariance 1. Show that

$$p_X(x) = \frac{\exp -x^*x}{\pi}$$

Hint: Note that the variance of the real part is 1/2 and the variance of the imaginary part is 1/2.

b) Let $\vec{X}$ be an n dimensional circularly symmetric complex Gaussian zero mean random vector with $K_{\vec{X}} = I_n$. Show that

$$p_{\vec{X}}(\vec{x}) = \frac{\exp -\vec{x}^{*T}\vec{x}}{\pi^n}$$

c) Let $\vec{Y} = H\vec{X}$ where H is n by n and invertible. Show that

$$p_{\vec{Y}}(\vec{y}) = \frac{\exp\left[-\vec{y}^{*T}H^{-1*T}H^{-1}\vec{y}\right]}{v\pi^n}$$

where v is $|d\vec{y}|/|d\vec{x}|$, the ratio of an incremental $2n$ dimensional volume element after being transformed by H to that before being transformed.

d) Show that

$$v = \frac{|d\vec{y}|}{|d\vec{x}|} = det[K_{\vec{Y}}]$$

and thus conclude that (4.166) is valid.

Exercise 4.16 Let $Y = X^2 + Z$ where Z is a zero mean unit variance Gaussian random variable. Show that no unbiased estimate of X exists from observation of Y. Hint. Consider any $x > 0$ and compare with $-x$.

Exercise 4.17 **a)** Assume that for each parameter value x, Y is Gaussian, $\mathcal{N}(x, \sigma^2)$. Show that $V_x(y)$ as defined in (4.167) is equal to $(y-x)/\sigma^2$ and show that the Fisher information is equal to $1/\sigma^2$.

b) Show that the Cramer-Rao bound is satisfied with equality for ML estimation for this example. Show that if X is $\mathcal{N}(0, \sigma_X^2)$, the MMSE estimate satisfies the Cramer-Rao bound with equality.

Exercise 4.18 Assume that Y is $\mathcal{N}(0, x)$. Show that $V_x(y)$ as defined in(4.167) is $V_x(y) = [y^2/x - 1]/(2x)$. Verify that $V_x(Y)$ is zero mean for each x. Find the Fisher information, $J(x)$.