
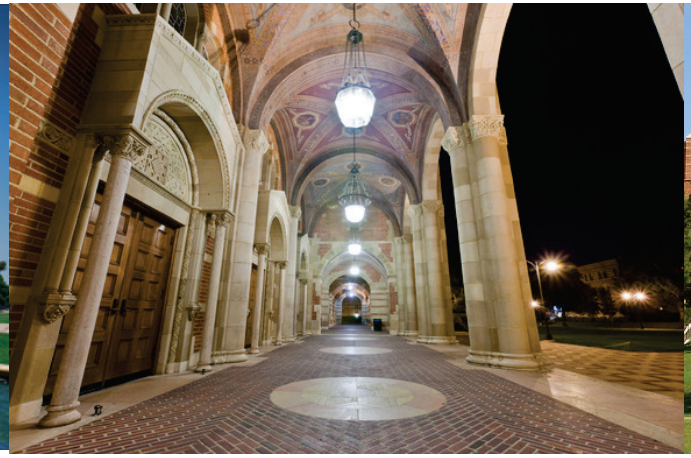


EE210A: Adaptation and Learning

Professor Ali H. Sayed

MAIN RESULT FROM LAST LECTURE

Theorem 1.1 (Optimal mean-square-error estimator) The least-mean-squares estimator (l.m.s.e.) of x given y is the conditional expectation of x given y, i.e., $\hat{x} = \mathsf{E}\,(x|y)$. The resulting estimate is

$$\hat{x} = \mathsf{E}\,(x|y = y) \;=\; \int_{\mathcal{S}_x} x f_{x|y}(x|y)\mathrm{d}x$$

where $\mathcal{S}_x$ denotes the support (or domain) of the random variable x. Moreover, the estimator is unbiased, i.e., $\mathsf{E}\,\hat{x} = \bar{x}$, and the resulting minimum cost is $\mathsf{E}\,\tilde{x}^2 \;=\; \sigma_x^2 \;-\; \sigma_{\hat{x}}^2$.

MAIN RESULT FROM LAST LECTURE

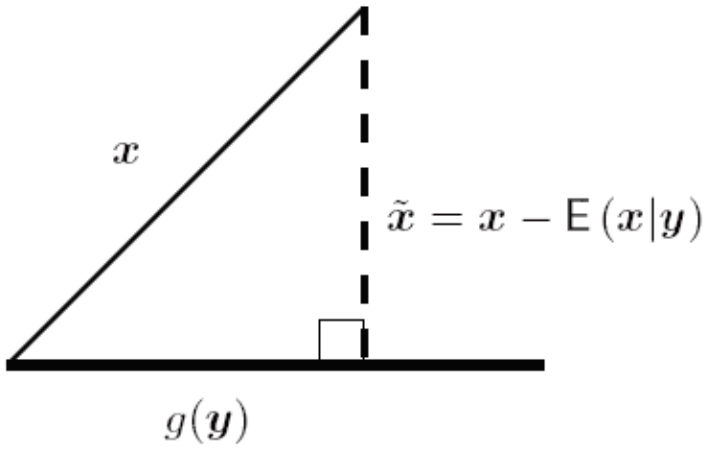

FIGURE 1.4 The orthogonality condition: $\tilde{x} \perp g(y)$.

Theorem 1.2 (Orthogonality condition) Given two random variables x and y, an estimator $\hat{x} = h(y)$ is optimal in the least-mean-squares sense (1.3) if, and only if, $\hat{x}$ is unbiased (i.e., $\mathsf{E}\,\hat{x} = \bar{x}$) and $x - \hat{x} \perp g(y)$ for any function $g(\cdot)$.

LECTURE #02

VECTOR AND COMPLEX-VALUED MEAN-SQUARE ERROR ESTIMATION

Sections in order: B.1, A.3, A.4, 2.1, 2.2, 2.3

HERMITIAN MATRICES

B.1 HERMITIAN AND POSITIVE-DEFINITE MATRICES

Hermitian matrices. The Hermitian conjugate, A^*, of a matrix A is the complex conjugate of its transpose, e.g.,

$$\text{if} \quad A = \begin{bmatrix} 1 & -j \\ 2+j & 1-j \end{bmatrix} \quad \text{then} \quad A^* = \begin{bmatrix} 1 & 2-j \\ j & 1+j \end{bmatrix}, \quad \text{where} \quad j = \sqrt{-1}$$

A Hermitian matrix is a square matrix satisfying $A^* = A$, e.g.,

$$\text{if} \quad A = \begin{bmatrix} 1 & 1+j \\ 1-j & 1 \end{bmatrix} \quad \text{then} \quad A^* = \begin{bmatrix} 1 & 1+j \\ 1-j & 1 \end{bmatrix} = A$$

so that A is Hermitian.

SPECTRAL DECOMPOSITION

Spectral decomposition. Hermitian matrices can only have *real* eigenvalues. To see this, assume u_i is an eigenvector of A corresponding to an eigenvalue λ_i, i.e., $Au_i = \lambda_i u_i$. Multiplying from the left by u_i^* we get $u_i^* A u_i = \lambda_i \|u_i\|^2$, where $\|\cdot\|$ denotes the Euclidean norm of its argument. Now the scalar quantity on the left-hand side of this equality is real since it coincides with its complex conjugate, namely $(u_i^* A u_i)^* = u_i^* A^* u_i = u_i^* A u_i$. Therefore, λ_i must be real too.

Another important property of Hermitian matrices, whose proof requires a more involved argument, is that such matrices always have a *full* set of orthonormal eigenvectors. That is, if A is $n \times n$ Hermitian, then there will exist n orthonormal eigenvectors u_i satisfying

$$Au_i = \lambda_i u_i, \quad \|u_i\|^2 = 1, \quad u_i^* u_j = 0 \ \text{ for } i \neq j$$

In compact matrix notation we can write this so-called spectral (or modal or eigen-) decomposition of A as

$$\boxed{A = U \Lambda U^*}$$

where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, $U = \begin{bmatrix} u_1 & u_2 & \ldots & u_n \end{bmatrix}$, and U satisfies

$$\boxed{UU^* = U^* U = \mathrm{I}}$$

We say that U is a unitary matrix. Here, the notation $\text{diag}\{a, b\}$ denotes a diagonal matrix with diagonal entries a and b.

POSITIVE-DEFINITE MATRICES

Positive-definite matrices. An $n \times n$ Hermitian matrix A is positive semi-definite (also called nonnegative definite) if it satisfies $x^* Ax \geq 0$ for all column vectors x. It is positive definite if $x^* Ax > 0$ except when $x = 0$. We denote a positive-definite matrix by writing $A > 0$ and a positive semi-definite matrix by writing $A \geq 0$. Among the several characterizations of positive-definite matrices, we note the following.

Lemma B.2 (Eigenvalues of positive-definite matrices) An $n \times n$ Hermitian matrix A is positive-definite if, and only if, all its eigenvalues are positive.

Proof: Let $A = U\Lambda U^*$ denote the spectral decomposition of A. Let also u_i be the $i-$th column of U with λ_i the corresponding eigenvalue, i.e., $Au_i = \lambda_i u_i$ with $\|u_i\|^2 = 1$. If we multiply this equality from the left by u_i^* we get

$$u_i^* Au_i = \lambda_i \|u_i\|^2 = \lambda_i > 0$$

where the last inequality follows from the fact that $x^* Ax > 0$ for any nonzero vector x. Therefore, $A > 0$ implies $\lambda_i > 0$. Conversely, assume all $\lambda_i > 0$ and multiply the equality $A = U\Lambda U^*$ by any nonzero vector x and its conjugate transpose, from right and left, to get $x^* Ax = x^* U\Lambda U^* x$. Now define the matrix $\Lambda^{1/2} = \text{diag}\left\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right\}$ and the vector $y = \Lambda^{1/2} U^* x$. The vector y is nonzero since U and $\Lambda^{1/2}$ are nonsingular matrices and, therefore, the product $\Lambda^{1/2} U^*$ cannot map a nonzero vector x to 0. Then the above equality becomes $x^* Ax = \|y\|^2 > 0$, which establishes that $A > 0$.

POSITIVE-DEFINITE MATRICES

In a similar vein, we can show that

$$\boxed{A \geq 0 \iff \lambda_i \geq 0}$$

Note further that since $\det A = (\det U)(\det \Lambda)(\det U^*)$ and $(\det U)(\det U^*) = 1$, we find that $\det A = \det \Lambda = \prod_{i=1}^{n} \lambda_i$. Therefore, the determinant of a positive-definite matrix is positive,

$$\boxed{A > 0 \implies \det A > 0}$$

A.3 COMPLEX-VALUED RANDOM VARIABLES

Although we have focused so far on real-valued random variables, we shall often encounter applications that deal with random variables that assume complex values. Accordingly, a complex-valued random variable is defined as one whose real and imaginary parts are *real*-valued random variables, say

$$\boxed{\boldsymbol{x} = \boldsymbol{x}_{\mathrm{r}} + j\boldsymbol{x}_{\mathrm{i}}, \quad j \triangleq \sqrt{-1}}$$

where $\boldsymbol{x}_{\mathrm{r}}$ and $\boldsymbol{x}_{\mathrm{i}}$ denote the real and imaginary parts of $\boldsymbol{x}$. Therefore, the pdf of a complex-valued random variable $\boldsymbol{x}$ can be characterized in terms of the joint pdf, $f_{\boldsymbol{x}_{\mathrm{r}},\boldsymbol{x}_{\mathrm{i}}}(\cdot,\cdot)$, of its real and imaginary parts. This means that we can regard a complex random variable as a function of two real random variables. The mean of $\boldsymbol{x}$ is

$$\begin{aligned} \mathsf{E}\,\boldsymbol{x} &\triangleq \mathsf{E}\,\boldsymbol{x}_{\mathrm{r}} + j\mathsf{E}\,\boldsymbol{x}_{\mathrm{i}} \\ &= \bar{x}_{\mathrm{r}} + j\bar{x}_{\mathrm{i}} \end{aligned}$$

in terms of the means of its real and imaginary parts. The variance of $\boldsymbol{x}$, on the other hand, is defined as

$$\boxed{\sigma_x^2 \triangleq \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^* = \mathsf{E}\,|\boldsymbol{x}-\bar{x}|^2} = \sigma_{x_{\mathrm{r}}}^2 + \sigma_{x_{\mathrm{i}}}^2 \qquad \text{(A.7)}$$

ORTHOGONAL RANDOM VARIABLES

We shall say that two complex-valued random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated if, and only if, their cross-correlation is zero, i.e.,

$$\sigma_{xy} \triangleq \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^* = 0$$

On the other hand, we shall say that they are *orthogonal* if, and only if,

$$\mathsf{E}\,\boldsymbol{x}\boldsymbol{y}^* = 0$$

It can be immediately verified that the concepts of orthogonality and uncorrelatedness coincide if at least one of the random variables has zero mean.

EXAMPLE

Example A.1 (QPSK constellation)

Consider a signal $\boldsymbol{x}$ that is chosen uniformly from a QPSK constellation, i.e., $\boldsymbol{x}$ assumes any of the values $\pm\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}$ with probability $1/4$ (see Fig. A.3). Clearly, $\boldsymbol{x}$ is a complex-valued random variable; its mean and variance are easily found to be $\bar{x} = 0$ and $\sigma_x^2 = 1$.

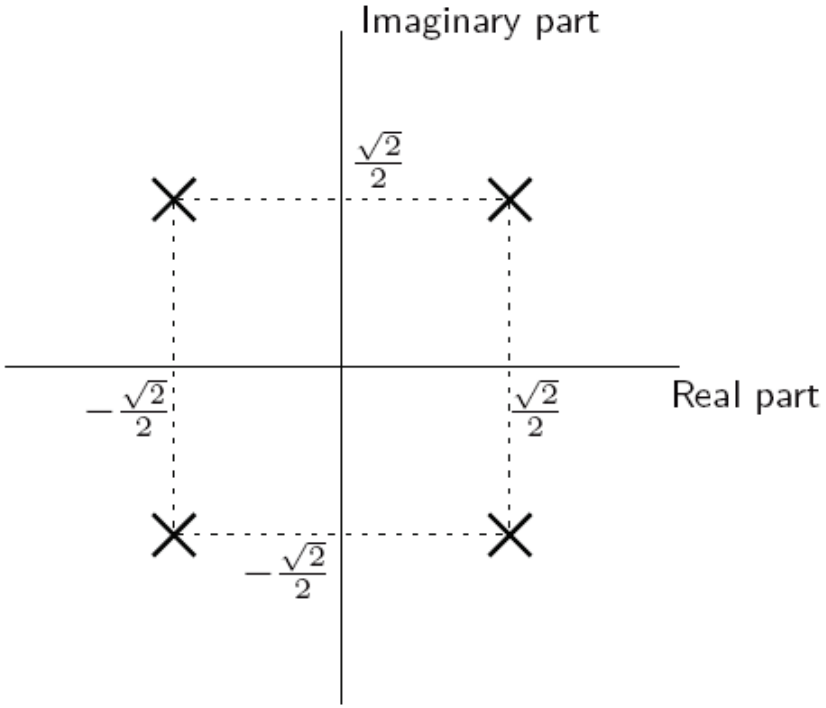

FIGURE A.3 A QPSK constellation.

A.4 VECTOR-VALUED RANDOM VARIABLES

A vector-valued random variable is a collection (in column or row vector forms) of random variables. The individual entries can be real or complex-valued themselves. For example, if $\boldsymbol{x} = \text{col}\{\boldsymbol{x}(0), \boldsymbol{x}(1)\}$ is a random vector with entries $\boldsymbol{x}(0)$ and $\boldsymbol{x}(1)$, then we shall define its mean as the vector of individual means,

$$\bar{x} = \mathsf{E}\,\boldsymbol{x} = \begin{bmatrix} \bar{x}(0) \\ \bar{x}(1) \end{bmatrix} \triangleq \begin{bmatrix} \mathsf{E}\,\boldsymbol{x}(0) \\ \mathsf{E}\,\boldsymbol{x}(1) \end{bmatrix}$$

and its covariance *matrix* as

$$\boxed{R_x \triangleq \mathsf{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^*} \tag{A.8}$$

where the symbol $*$ now denotes complex-conjugate transposition (i.e., we transpose the vector and then replace each of its entries by the corresponding conjugate value). Note that we are using parentheses to index the scalar entries of a vector, e.g., $\boldsymbol{x}(k)$ denotes the $k-$th entry of $\boldsymbol{x}$. Moreover, if $\boldsymbol{x}$ were a *row* random vector, rather than a *column* random vector, then its covariance matrix would instead be defined as

$$R_x \triangleq \mathsf{E}\,(\boldsymbol{x} - \bar{x})^*(\boldsymbol{x} - \bar{x})$$

with the conjugate term coming first. This is because in this case, it is the product $(\boldsymbol{x} - \bar{x})^*(\boldsymbol{x} - \bar{x})$ that yields a matrix, while the product $(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^*$ would be a scalar.

EXAMPLE

For the two-element column vector $x = \mathrm{col}\{x(0), x(1)\}$ we obtain

$$R_x = \begin{bmatrix} \mathsf{E}\,|x(0) - \bar{x}(0)|^2 & \mathsf{E}\,[x(0) - \bar{x}(0)][x(1) - \bar{x}(1)]^* \\ \mathsf{E}\,[x(1) - \bar{x}(1)][x(0) - \bar{x}(0)]^* & \mathsf{E}\,|x(1) - \bar{x}(1)|^2 \end{bmatrix}$$

with the individual variances of the variables $\{x(0), x(1)\}$ appearing on the diagonal and the cross-correlations between them appearing on the off-diagonal entries. In the zero-mean case, the definition of R_x, and the above expression, simplify to

$$\boxed{R_x \triangleq \mathsf{E}\,xx^*}$$

and

$$R_x = \begin{bmatrix} \mathsf{E}\,|x(0)|^2 & \mathsf{E}\,x(0)x^*(1) \\ \mathsf{E}\,x(1)x^*(0) & \mathsf{E}\,|x(1)|^2 \end{bmatrix}$$

PROPERTIES OF COVARIANCE MATRICES

It should be noted that the covariance matrix R_x is Hermitian, i.e., it satisfies

$$R_x = R_x^*$$

Moreover, R_x is a nonnegative-definite matrix, written as

$$R_x \geq 0$$

By definition, a Hermitian matrix R is said to be nonnegative definite if, and only if, $a^* R a \geq 0$ for any column vector a (real or complex-valued). In order to verify that $R_x \geq 0$, we introduce the scalar-valued random variable $y = a^*(x - \bar{x})$, where a is an arbitrary column vector. Then y has zero mean and

$$\sigma_y^2 = \mathsf{E}\,|y|^2 = a^* R_x a$$

But since the variance of any scalar-valued random variable is always nonnegative, we conclude that $a^* R_x a \geq 0$ for any a. This means that R_x is nonnegative definite, as claimed.

For real-valued data, the symbol $*$ is replaced by the transposition symbol T, and R_x is defined as

$$\boxed{R_x \triangleq \mathsf{E}\,(x - \bar{x})(x - \bar{x})^{\mathsf{T}}}$$

ESTIMATING A SCALAR FROM A VECTOR

2.1 OPTIMAL ESTIMATOR IN THE VECTOR CASE

It turns out that the optimal estimator in the general vector and complex-valued case is still given by the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$. To see this, let us start with a special case. Assume $\boldsymbol{x}$ and $\boldsymbol{y}$ are both real-valued with $\boldsymbol{x}$ a *scalar* and $\boldsymbol{y}$ a vector, say

$$\boldsymbol{y} = \mathrm{col}\{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)\}$$

As before, let $\hat{\boldsymbol{x}} = h(\boldsymbol{y})$ denote an estimator for $\boldsymbol{x}$. Since $\boldsymbol{y}$ is vector-valued, the function $h(\cdot)$ operates on the entries of $\boldsymbol{y}$ and provides a real scalar quantity as a result. More explicitly, we write

$$\hat{\boldsymbol{x}} = h\left(\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)\right)$$

The function $h(\cdot)$ is to be chosen optimally by minimizing the variance of the error $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \hat{\boldsymbol{x}}$, i.e., by solving

$$\min_{h(\cdot)} \ \mathsf{E}\ \tilde{\boldsymbol{x}}^2$$

The same argument that we used to establish Thm. 1.1 can be repeated here to verify that the optimal estimator is still given by

$$\hat{\boldsymbol{x}} \ = \ \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y}) \ = \ \mathsf{E}\,[\boldsymbol{x}|\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \tag{2.1}$$

EXAMPLE

Example 2.1 (Noisy measurements of a BPSK signal)

Let us return to Ex. 1.2, where x is a BPSK signal that is either $+1$ or -1 with probability $1/2$ each. Assume that we collect two noisy measurements $y(0)$ and $y(1)$ of x, say $y(0) = x + v(0)$ and $y(1) = x + v(1)$, where $\{v(0), v(1)\}$ are zero-mean unit-variance Gaussian random variables that are independent of each other and of x. The value of x is the same in both measurements (i.e., if it is $+1$ in the measurement $y(0)$, it is also $+1$ in the measurement $y(1)$, and similarly for -1.) We may interpret $\{y(0), y(1)\}$ as the noisy signals measured at two antennas as a result of transmitting x over two additive Gaussian-noise channels — see Fig. 2.1.

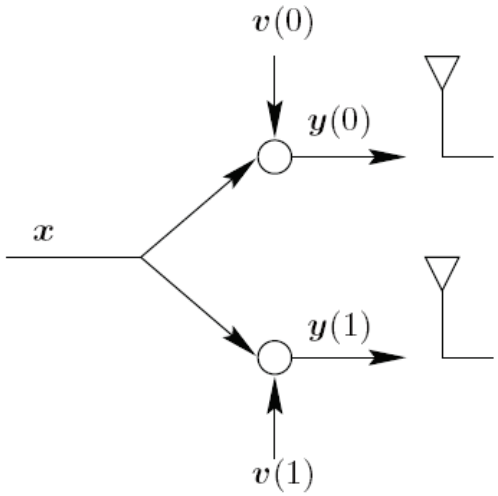

FIGURE 2.1 Reception by two antennas of a symbol x transmitted over two additive Gaussian-noise channels.

EXAMPLE

We can then pose the problem of estimating x given *both* measurements $\{y(0), y(1)\}$. According to (2.1), the solution is given by

$$\hat{x} = \mathsf{E}\,[x|y(0), y(1)]$$

The evaluation of the conditional expectation in this case is a trivial extension of the derivation given in Ex. 1.2, and it is left as an exercise to the reader — see Prob. I.13, where the more general case of multiple measurements is treated. The result of that problem shows that

$$\hat{x} = \tanh[y(0) + y(1)]$$

In the context of the two-antenna example of Fig. 2.1, this result leads to the optimal receiver structure shown in Fig. 2.2

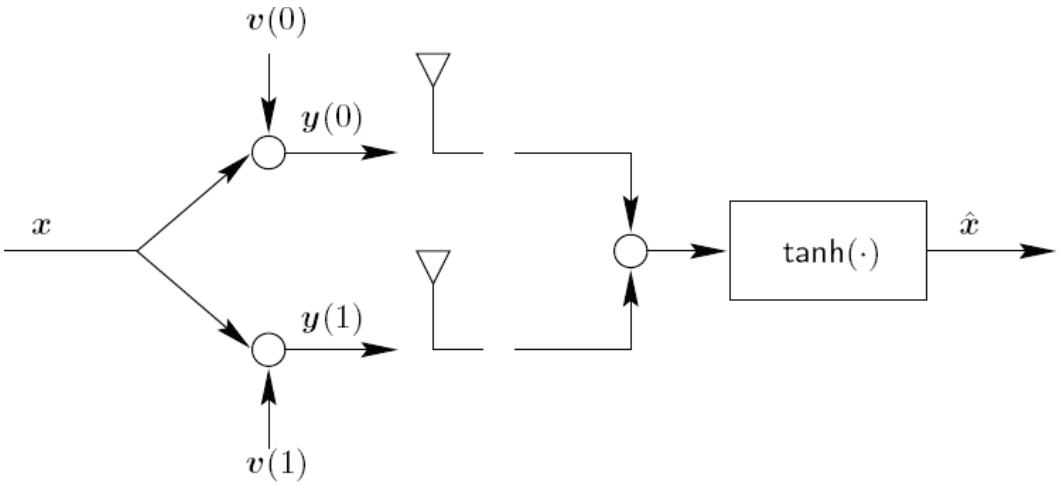

FIGURE 2.2 Optimal receiver structure for recovering a symbol x from two separate measurements over additive Gaussian-noise channels.

ESTIMATING A VECTOR FROM A VECTOR

Let us now study the general case and determine the form of the optimal estimator for a *vector-valued* random variable $\boldsymbol{x}$ given another vector-valued random variable $\boldsymbol{y}$, with both variables allowed to be complex-valued as well. Thus assume that $\boldsymbol{x}$ is $p-$dimensional while $\boldsymbol{y}$ is $q-$dimensional.

Again, let $\hat{\boldsymbol{x}} = h(\boldsymbol{y})$ denote an estimator for $\boldsymbol{x}$. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are vector-valued, the function $h(\cdot)$ operates on the entries of $\boldsymbol{y}$ and provides a vector quantity as a result. More explicitly, we can write for the individual entries of $\hat{\boldsymbol{x}}$ and $\boldsymbol{y}$,

$$\begin{bmatrix} \hat{\boldsymbol{x}}(0) \\ \hat{\boldsymbol{x}}(1) \\ \hat{\boldsymbol{x}}(2) \\ \vdots \\ \hat{\boldsymbol{x}}(p-1) \end{bmatrix} = \begin{bmatrix} h_0[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \\ h_1[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \\ h_2[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \\ \vdots \\ h_{p-1}[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \end{bmatrix}$$

where the $\{h_k(\cdot)\}$ represent the individual mappings from the observation vector $\boldsymbol{y}$ to the estimators $\{\hat{\boldsymbol{x}}(k)\}$. We can then seek optimal functions $\{h_k(\cdot)\}$ that minimize the variance of the error in each component of $\boldsymbol{x}$, namely, each $h_k(\cdot)$ is determined by solving

$$\min_{h_k(\cdot)} \quad \mathsf{E}\ |\tilde{\boldsymbol{x}}(k)|^2 \tag{2.3}$$

EQUIVALENT CRITERION

where

$$\tilde{\boldsymbol{x}}(k) \triangleq \boldsymbol{x}(k) - h_k(\boldsymbol{y})$$

Actually, this formulation is equivalent to solving over all $\{h_k(\cdot)\}$ the following problem:

$$\boxed{\min_{\{h_k(\cdot)\}} \quad \mathsf{E}\, \tilde{\boldsymbol{x}}^* \tilde{\boldsymbol{x}}} \tag{2.4}$$

This is because the quantity $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}$ in (2.4) is the sum of the individual terms $\mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2$,

$$\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}} = \mathsf{E}\,|\tilde{\boldsymbol{x}}(0)|^2 + \mathsf{E}\,|\tilde{\boldsymbol{x}}(1)|^2 + \ldots + \mathsf{E}\,|\tilde{\boldsymbol{x}}(p-1)|^2$$

with each term $\mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2$ depending only on the corresponding function $h_k(\cdot)$. In this way, minimizing the sum $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}$ over all $\{h_k(\cdot)\}$ is equivalent to minimizing each individual term, $\mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2$, over its $h_k(\cdot)$. Note further that

$$\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}} = \mathsf{Tr}\,(\mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*) = \mathsf{Tr}\,(R_{\tilde{\boldsymbol{x}}})$$

EQUIVALENT CRITERION

Then problem (2.4) is also equivalent to solving over all $\{h_k(\cdot)\}$:

$$\boxed{\min_{\{h_k(\cdot)\}} \quad \mathsf{Tr}(R_{\tilde{x}})} \tag{2.5}$$

Now the solution to the general problem (2.3) follows from the special case discussed at the beginning of this section. Indeed, if we express $\boldsymbol{x}(k)$ and $h_k(\cdot)$ in terms of their real and imaginary parts, say

$$\boldsymbol{x}(k) \triangleq \boldsymbol{x}_{\mathsf{r}}(k) + j\,\boldsymbol{x}_{\mathsf{i}}(k), \qquad h_k(\boldsymbol{y}) \triangleq h_{\mathsf{r},k}(\boldsymbol{y}) + j\,h_{\mathsf{i},k}(\boldsymbol{y})$$

then we can expand the error criterion as

$$\mathsf{E}\,|\boldsymbol{x}(k) - h_k(\boldsymbol{y})|^2 = \mathsf{E}\,[\boldsymbol{x}_{\mathsf{r}}(k) - h_{\mathsf{r},k}(\boldsymbol{y})]^2 + \mathsf{E}\,[\boldsymbol{x}_{\mathsf{i}}(k) - h_{\mathsf{i},k}(\boldsymbol{y})]^2$$

and we are reduced to minimizing the sum of two nonnegative quantities over the unknowns $\{h_{\mathsf{r},k}(\cdot), h_{\mathsf{i},k}(\cdot)\}$. This is equivalent to minimizing each term separately,

$$\min_{h_{\mathsf{r},k}(\cdot)} \mathsf{E}\,[\boldsymbol{x}_{\mathsf{r}}(k) - h_{\mathsf{r},k}(\boldsymbol{y})]^2, \qquad \min_{h_{\mathsf{i},k}(\cdot)} \mathsf{E}\,[\boldsymbol{x}_{\mathsf{i}}(k) - h_{\mathsf{i},k}(\boldsymbol{y})]^2$$

OPTIMAL ESTIMATOR

Theorem 2.1 (Optimal estimation in the vector case) The least-mean-squares estimator of a (possibly complex-valued) vector x given another (possibly complex-valued) vector y is the conditional expectation of x given y, i.e., $\hat{x} = \mathsf{E}(x|y)$. This estimator solves

$$\min_{\hat{x}} \mathsf{Tr}(R_{\tilde{x}})$$

where $R_{\tilde{x}} = \mathsf{E}\,\tilde{x}\tilde{x}^*$ and $\tilde{x} = x - \hat{x}$.

2.2 SPHERICALLY INVARIANT GAUSSIAN VARIABLES

So assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian random *vector* variables with a nonsingular covariance matrix

$$R \triangleq \begin{bmatrix} R_x & R_{xy} \\ R_{yx} & R_y \end{bmatrix}$$

where

$$\begin{aligned} R_x &= \mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^* \\ R_y &= \mathsf{E}(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^* \\ R_{xy} &= \mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^* = R_{yx}^* \end{aligned}$$

The variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ are assumed to be complex-valued with dimensions $p \times 1$ for $\boldsymbol{x}$ and $q \times 1$ for $\boldsymbol{y}$.

VECTOR GAUSSIAN VARIABLES

If $\boldsymbol{x}$ and $\boldsymbol{y}$ were *real-valued*, then their individual probability density functions, as well as their joint pdf, would be given by (see Sec. A.5):

$$f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{(2\pi)^p}} \frac{1}{\sqrt{\det R_x}} \exp\left\{-\tfrac{1}{2}(x-\bar{x})^{\mathsf{T}} R_x^{-1}(x-\bar{x})\right\}$$

$$f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{(2\pi)^q}} \frac{1}{\sqrt{\det R_y}} \exp\left\{-\tfrac{1}{2}(y-\bar{y})^{\mathsf{T}} R_y^{-1}(y-\bar{y})\right\}$$

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \frac{1}{\sqrt{(2\pi)^{p+q}}} \frac{1}{\sqrt{\det R}} \exp\left\{-\tfrac{1}{2}\begin{bmatrix} (x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}} \end{bmatrix} R^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\}$$

In particular, observe that if $\boldsymbol{x}$ and $\boldsymbol{y}$ were *uncorrelated*, i.e., if $R_{xy} = 0$, then the covariance matrix R becomes block diagonal, with entries $\{R_x, R_y\}$, and it is straightforward to verify from the above pdf expressions that in this case $f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{x}}(x) f_{\boldsymbol{y}}(y)$. In other words, uncorrelated real-valued Gaussian random variables are also independent.

CIRCULARITY ASSUMPTIONS

When, on the other hand, $\boldsymbol{x}$ and $\boldsymbol{y}$ are *complex-valued*, they need to satisfy two conditions in order for their individual and joint pdfs to have forms similar to the above in the Gaussian case. These conditions are known as *circularity* assumptions, and the need for them is explained in Sec. A.5. The conditions are as follows. Each variable is required to be *circular*, meaning that $\{\boldsymbol{x}, \boldsymbol{y}\}$ should satisfy

$$\mathsf{E}\,(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} = 0 \qquad \text{and} \qquad \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^{\mathsf{T}} = 0$$

with the transposition symbol T used instead of the conjugation symbol $*$. The variables are also required to be second-order circular, i.e.,

$$\mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} = 0$$

These circularity assumptions are not needed when the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ are real-valued.

CIRCULAR GAUSSIAN VARIABLES

These circularity assumptions are not needed when the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ are real-valued. The circularity of $\boldsymbol{x}$ in the complex case guarantees that its pdf in the Gaussian case will have the form

$$f_{\boldsymbol{x}}(x) = \frac{1}{\pi^p} \frac{1}{\det R_x} \exp\left\{-(x-\bar{x})^* R_x^{-1} (x-\bar{x})\right\}$$

Likewise, the circularity of $\boldsymbol{y}$ guarantees that its pdf will have the form

$$f_{\boldsymbol{y}}(y) = \frac{1}{\pi^q} \frac{1}{\det R_y} \exp\left\{-(y-\bar{y})^* R_y^{-1} (y-\bar{y})\right\}$$

The second-order circularity of $\boldsymbol{x}$ and $\boldsymbol{y}$ guarantees that the joint pdf of $\{\boldsymbol{x}, \boldsymbol{y}\}$ will have the form

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \frac{1}{\pi^{p+q}} \frac{1}{\det R} \exp\left\{-\begin{bmatrix} (x-\bar{x})^* & (y-\bar{y})^* \end{bmatrix} R^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\}$$

Thus observe again that if $\boldsymbol{x}$ and $\boldsymbol{y}$ were *uncorrelated*, then the above pdf expressions lead to

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{x}}(x) \cdot f_{\boldsymbol{y}}(y)$$

ESTIMATION WITH GAUSSIAN DATA

Now the least-mean-squares estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ requires that we determine the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. This can be obtained from the calculation

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)} = \frac{1}{\pi^p}\frac{\det R_y}{\det R}\frac{\exp\left\{-\begin{bmatrix}(x-\bar{x})^* & (y-\bar{y})^*\end{bmatrix}R^{-1}\begin{bmatrix}x-\bar{x}\\ y-\bar{y}\end{bmatrix}\right\}}{\exp\left\{-(y-\bar{y})^* R_y^{-1}(y-\bar{y})\right\}}$$

Following the same argument that we used earlier in Sec. 1.4, we can simplify the above expression by introducing the *block* upper-diagonal-lower triangular factorization (whose validity can again be verified, e.g., by direct calculation):

$$R \triangleq \begin{bmatrix} R_x & R_{xy} \\ R_{yx} & R_y \end{bmatrix} = \begin{bmatrix} \mathrm{I} & R_{xy}R_y^{-1} \\ 0 & \mathrm{I} \end{bmatrix}\begin{bmatrix} \Sigma & 0 \\ 0 & R_y \end{bmatrix}\begin{bmatrix} \mathrm{I} & 0 \\ R_y^{-1}R_{yx} & \mathrm{I} \end{bmatrix}$$

where Σ is the Schur complement of R_y in R, namely,

$$\Sigma = R_x - R_{xy}R_y^{-1}R_{yx}$$

TRIANGULAR FACTORIZATION

Inverting both sides of the above factorization for R we get

$$R^{-1} = \begin{bmatrix} \mathrm{I} & 0 \\ -R_y^{-1}R_{yx} & \mathrm{I} \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & R_y^{-1} \end{bmatrix} \begin{bmatrix} \mathrm{I} & -R_{xy}R_y^{-1} \\ 0 & \mathrm{I} \end{bmatrix}$$

which allows us to express the term

$$\begin{bmatrix} (x-\bar{x})^* & (y-\bar{y})^* \end{bmatrix} R^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}$$

which appears in the expression for $f_{\boldsymbol{x},\boldsymbol{y}}(x,y)$, as the following separable sum of two quadratic terms,

$$[(x-\bar{x}) - R_{xy}R_y^{-1}(y-\bar{y})]^*\Sigma^{-1}[(x-\bar{x}) - R_{xy}R_y^{-1}(y-\bar{y})] \; + \; (y-\bar{y})^*R_y^{-1}(y-\bar{y})$$

AFFINE RELATION

Substituting this equality into the expression for $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, and using

$$\det R = \det \Sigma \cdot \det R_y$$

we conclude that

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = \frac{1}{\pi^p}\frac{1}{\det \Sigma}\exp\left\{-[(x-\bar{x})-R_{xy}R_y^{-1}(y-\bar{y})]^*\Sigma^{-1}[(x-\bar{x})-R_{xy}R_y^{-1}(y-\bar{y})]\right\}$$

which can be interpreted as the pdf of a circular Gaussian random variable with covariance matrix Σ and mean value $\bar{x} + R_{xy}R_y^{-1}(y-\bar{y})$. We therefore conclude that

$$\boxed{\hat{\boldsymbol{x}} \triangleq \mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) = \bar{x} + R_{xy}R_y^{-1}(\boldsymbol{y}-\bar{y})}$$

and the resulting m.m.s.e. matrix is

$$\boxed{\text{m.m.s.e.} \triangleq R_{\tilde{x}} = R_x - R_{\hat{x}} = R_x - R_{xy}R_y^{-1}R_{yx}}$$

ZERO MEAN CASE

case. Note further that in the zero-mean case we obtain

$$\boxed{\hat{\boldsymbol{x}} = R_{xy} R_y^{-1} \boldsymbol{y}}$$

with $\{R_x, R_y, R_{xy}\}$ defined accordingly,

$$R_x = \mathsf{E}\, \boldsymbol{x}\boldsymbol{x}^*, \quad R_y = \mathsf{E}\, \boldsymbol{y}\boldsymbol{y}^*, \quad R_{xy} = \mathsf{E}\, \boldsymbol{x}\boldsymbol{y}^*$$

Lemma 2.1 (Circular Gaussian variables) If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two circular and jointly Gaussian random variables with means $\{\bar{x}, \bar{y}\}$ and covariance matrices $\{R_x, R_y, R_{xy}\}$, then the least-mean-squares estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\hat{\boldsymbol{x}} \;=\; \bar{x} + R_{xy} R_y^{-1} (\boldsymbol{y} - \bar{y})$$

and the resulting minimum cost is m.m.s.e. $= R_x - R_{xy} R_y^{-1} R_{yx}$.

EQUIVALENT CRITERION

2.3 EQUIVALENT OPTIMIZATION CRITERION

A useful fact to highlight at the end of this chapter is that the optimal estimator $\mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$ defined by (2.6), which solves problems (2.3)–(2.5), is also the optimal solution of another related *matrix-valued* error criterion (cf. (2.9) further ahead), as we now explain.

Thus consider the following alternative formulation. Assume that we pose the problem of estimating $\boldsymbol{x}$ from $\boldsymbol{y}$ by requiring that the functions $\{h_k(\cdot)\}$ be such that they minimize the variance of any *arbitrary* linear combination of the entries of the error vector, say $a^*(\boldsymbol{x} - h(\boldsymbol{y}))$ for any a. That is, assume we replace the optimization problem (2.3) by the alternative problem

$$\min_{\{h_k(\cdot)\}} \ \mathsf{E}\,|a^*\tilde{\boldsymbol{x}}|^2, \quad \text{for any column vector } a \tag{2.8}$$

The error vector $\tilde{\boldsymbol{x}}$ is dependent on the choice of h and, therefore, the covariance matrix $\mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*$ is also dependent on h. Let us indicate this fact explicitly by writing

$$R_{\tilde{x}}(h) \triangleq \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*$$

EQUIVALENT CRITERION

Now note that

$$\mathsf{E}\,|a^*\tilde{\boldsymbol{x}}|^2 = a^* R_{\tilde{x}}(h) a$$

so that problem (2.8) is in effect seeking an optimal function h^o such that, for any vector a and for any other h,

$$a^* R_{\tilde{x}}(h) a \;\geq\; a^* R_{\tilde{x}}(h^o) a$$

That is, the difference matrix $R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o)$ should be nonnegative-definite for all h. For this reason, we can equivalently interpret (2.8) as the problem of minimizing the error covariance matrix $R_{\tilde{x}}$ itself, written as

$$\boxed{\min_{h(\cdot)} \quad \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*} \qquad (2.9)$$

Comparing with (2.4) we see that we are replacing the scalar $\mathsf{E}\,\tilde{x}^*\tilde{x}$ by the matrix $\mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*$.

EQUIVALENT CRITERION

Let us now verify that the solution to (2.8), or equivalently (2.9), is again $h^o(\boldsymbol{y}) = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$. For this purpose, we recall that, for any $h(\boldsymbol{y})$,

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \hat{\boldsymbol{x}} = \boldsymbol{x} - h(\boldsymbol{y})$$

so that the covariance matrix $R_{\tilde{x}}(h)$ is given by

$$\begin{aligned} R_{\tilde{x}}(h) &= \mathsf{E}\,[\boldsymbol{x} - h(\boldsymbol{y})][\boldsymbol{x} - h(\boldsymbol{y})]^* \\ &= \mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^* - \mathsf{E}\,\boldsymbol{x}h^*(\boldsymbol{y}) - \mathsf{E}\,h(\boldsymbol{y})\boldsymbol{x}^* + \mathsf{E}\,h(\boldsymbol{y})h^*(\boldsymbol{y}) \end{aligned}$$

We now verify that

$$R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o) \geq 0$$

for any choice of h. Indeed, from the orthogonality property (2.7), we have that $\boldsymbol{x} - h^o(\boldsymbol{y})$ is uncorrelated with any function of $\boldsymbol{y}$. Hence,

$$\begin{aligned} R_{\tilde{x}}(h^o) &= \mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})][\boldsymbol{x} - h^o(\boldsymbol{y})]^* \\ &= \mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})]\boldsymbol{x}^* \\ &= \mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^* - \mathsf{E}\,h^o(\boldsymbol{y})\boldsymbol{x}^* \end{aligned}$$

EQUIVALENT CRITERION

Subtracting from $R_{\tilde{x}}(h)$ leads to

$$R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o) \quad = \quad -\mathsf{E}\, \boldsymbol{x}h^*(\boldsymbol{y}) - \mathsf{E}\, h(\boldsymbol{y})\boldsymbol{x}^* + \mathsf{E}\, h(\boldsymbol{y})h^*(\boldsymbol{y}) + \mathsf{E}\, h^o(\boldsymbol{y})\boldsymbol{x}^*$$

From the orthogonality property (2.7) we again have that

$$\mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})]h^{o*}(\boldsymbol{y}) = 0, \qquad \mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})]h^*(\boldsymbol{y}) = 0$$

so that

$$\mathsf{E}\, \boldsymbol{x}h^{o*}(\boldsymbol{y}) = \mathsf{E}\, h^o(\boldsymbol{y})h^{o*}(\boldsymbol{y}) \qquad \text{and} \qquad \mathsf{E}\, \boldsymbol{x}h^*(\boldsymbol{y}) = \mathsf{E}\, h^o(\boldsymbol{y})h^*(\boldsymbol{y})$$

These two equalities allow us to rewrite the difference $R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o)$ as a perfect square:

$$R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o) = \mathsf{E}\,[h^o(\boldsymbol{y}) - h(\boldsymbol{y})][h^o(\boldsymbol{y}) - h(\boldsymbol{y})]^*$$

The right-hand side is nonnegative-definite for all h, as desired. Finally, since the cost used in (2.4) is simply the trace of the error covariance matrix, we conclude that minimizing the error covariance matrix is equivalent to minimizing its trace.

EQUIVALENT CRITERION

Lemma 2.2 (Cost function) The conditional expectation of x given y is optimal relative to either cost

$$\min_{\hat{\boldsymbol{x}}} \mathsf{Tr}(R_{\tilde{x}}) \quad \text{or} \quad \min_{\hat{\boldsymbol{x}}} R_{\tilde{x}}$$

where $R_{\tilde{x}} = \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*$ and $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \hat{\boldsymbol{x}}$.

COMPUTER PROJECT

Project I.1 (Comparing optimal and suboptimal estimators) The purpose of this project[2] is to compare the performance of an optimal least-mean-squares estimator with three approximations for it, along the lines discussed in Ex. 1.2. Thus consider the setting of Prob. I.13.

(a) Write a MATLAB program that generates a BPSK random variable x that is equal to $+1$ with probability p and to -1 with probability $1-p$.

(b) Simulate the estimator of part (b) of Prob. I.13 for different number of observations N. Generate observations $\{y(i)\}$ and plot $\hat{x}_N$ as a function of N for $1 \leq N \leq 10$, with all observations assumed generated by the same value of x — either $+1$ or -1, and using zero-mean Gaussian noise with unit variance. Plot $\hat{x}_N$ for the cases $p = 0.1, 0.3, 0.5, 0.8$.

(c) Compare the performance of the optimal estimate $\hat{x}_N$ with the averaged estimate

$$\hat{x}_{N,\text{av}} \triangleq \frac{1}{N}\sum_{i=0}^{N-1} y(i)$$

for several values of N, say, for $1 \leq N \leq 300$, and for the same values of p in part (b). Does it take many more samples N for the averaged estimate $\hat{x}_{N,\text{av}}$ to provide a good result compared with the optimal nonlinear estimate $\hat{x}_N$?

COMPUTER PROJECT

(d) Fix $p = 1/2$, and define the nonlinear decision device:

$$\text{sign}[z] = \begin{cases} +1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases}$$

Consider also the alternative (sign-of-optimal) estimate $\hat{x}_{\text{dec}} = \text{sign}\,[\hat{x}_N]$. It is clear that $\hat{x}_{\text{dec}}$ assumes the values ± 1, whereas the optimal estimate $\hat{x}_N$ does not. Is $\hat{x}_{\text{dec}}$ a better estimate than $\hat{x}_N$? The answer in the mean-square sense is of course negative since we already know that $\hat{x}_N$ is the best estimate. To verify this fact do the following. Fix the number of observations at $N = 10$. Then perform 1000 experiments, with each experiment i resulting in an optimal estimate $\hat{x}_{10}(i)$ and an estimate $\hat{x}_{\text{dec}}(i)$. For each estimate, the value of x is fixed at either $+1$ or -1. Compute the sample variances

$$\frac{1}{1000}\sum_{i=1}^{1000} |x - \hat{x}_{10}(i)|^2, \qquad \frac{1}{1000}\sum_{i=1}^{1000} |x - \hat{x}_{\text{dec}}(i)|^2$$

Which one is smaller? Repeat for the following (sign-of-average) estimate:

$$\hat{x}_{\text{sign}} = \text{sign}\,[\hat{x}_{N,\text{av}}]$$

That is, apply the decision device to the estimate that is obtained from averaging.

COMPUTER PROJECT SOLUTION

1. psk.m This function generates a BPSK signal $\boldsymbol{x}$ that assumes the value $+1$ with probability p and the value -1 with probability $1-p$.
2. partB.m This program generates four plots of $\hat{x}_N$, as a function of N, one for each value of p — see Fig. 1.

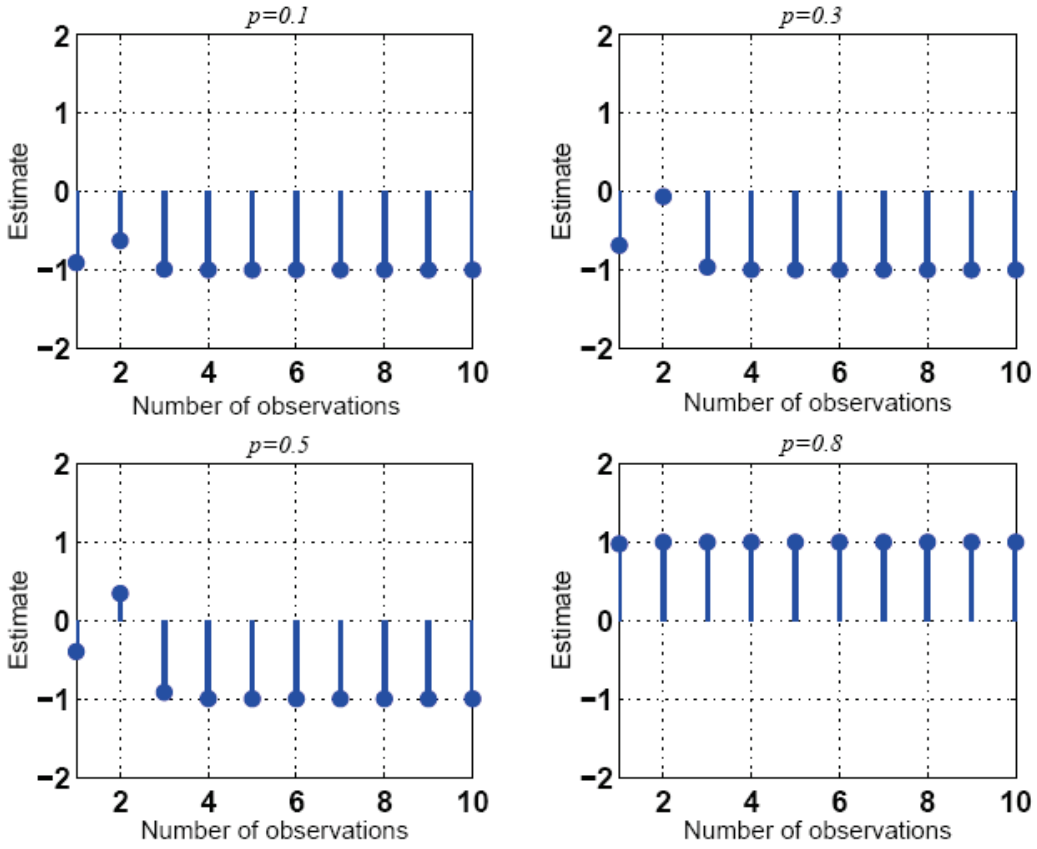

Figure I.1. The plots show the values of the optimal estimates $\hat{x}_N$ for different choices of N (the number of observations) and for different values of p (which determines the probability distribution of $\boldsymbol{x}$).

COMPUTER PROJECT SOLUTION

3. partC.m This program generates a plot that shows $\{\hat{x}_N, \hat{x}_{N,av}\}$ for four different values of p over the interval $1 \leq N \leq 300$ — see Fig. 2.

4. partD.m This program estimates the variances of $\{\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{x}}_{N,av}, \hat{\boldsymbol{x}}_{dec}, \hat{\boldsymbol{x}}_{sign}\}$. Typical values are

$$\{0.0031, 0.1027, 0.0040, 0.0040\}$$

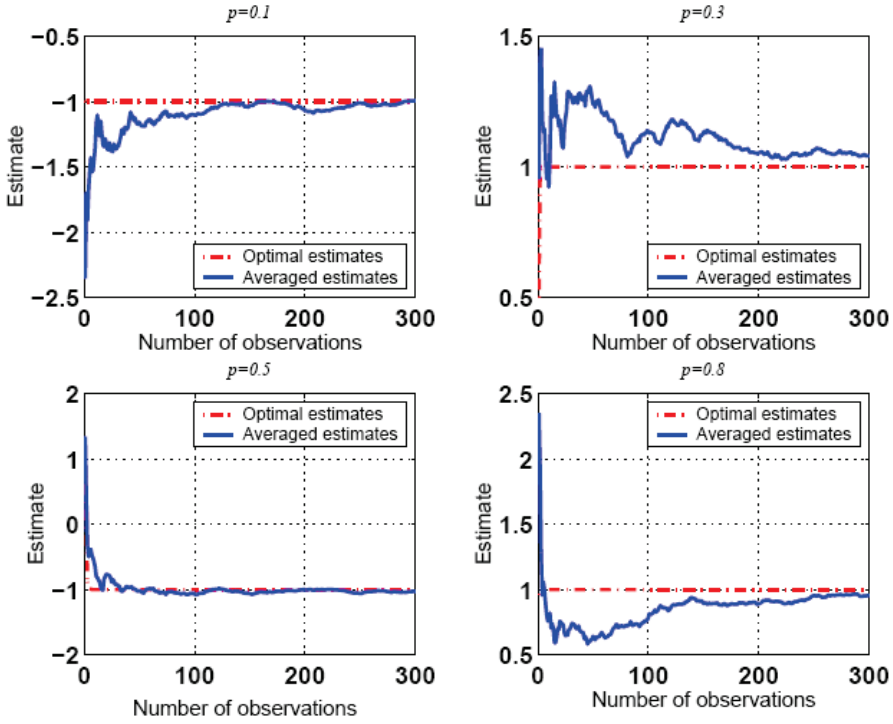

Figure I.2. The plots show the values of the optimal estimates $\hat{x}_N$ (*dotted lines*) and the averaged estimates $\hat{x}_{N,av}$ (*solid lines*) for different choices of N (the number of observations) and for different values of p (which determines the probability distribution of $\boldsymbol{x}$). Observe how the averaged estimates are significantly less reliable for a smaller number of observations.