Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications. Alfred Mertins

Print ISBN 0-471-98626-7 Electronic ISBN 0-470-84183-4

Chapter 5

Transforms and Filters for Stochastic Processes

In this chapter, we consider the optimal processing of random signals. We start with transforms that have optimal approximation properties, in the least-squares sense, for continuous and discrete-time signals, respectively. Then we discuss the relationships between discrete transforms, optimal linear estimators, and optimal linear filters.

5.1 The Continuous-Time Karhunen–Lo`eve Transform

Among all linear transforms, the *Karhunen–Lo ève transform* (KLT) is the one which best approximates a stochastic process in the least squares sense. Furthermore, the KLT is a signal expansion with uncorrelated coefficients. These properties make it interesting for many signal processing applications such as coding and pattern recognition. The transform can be formulated for continuous-time and discrete-time processes. In this section, we sketch the continuous-time case [81], [149]. The discrete-time case will be discussed in the next section in greater detail.

Consider a real-valued continuous-time random process $x(t)$, $a \leq t \leq b$.

We may not assume that every sample function of the random process lies in $L_2(a,b)$ and can be represented exactly via a series expansion. Therefore, a weaker condition is formulated, which states that we are looking for a series expansion that represents the stochastic process in the mean:[1]

$$x(t) = \underset{N\to\infty}{\text{l.i.m}} \sum_{i=1}^{N} x_i \, \varphi_i(t) \tag{5.1}$$

The "unknown" orthonormal basis $\{\varphi_i(t);\ i = 1,2,\ldots\}$ has to be derived from the properties of the stochastic process. For this, we require that the coefficients

$$x_i = \langle \boldsymbol{x}, \boldsymbol{\varphi}_i \rangle = \int_a^b x(t) \, \varphi_i(t) \, dt \tag{5.2}$$

of the series expansion are uncorrelated. This can be expressed as

$$\begin{aligned} E\{x_i x_j\} &= E\left\{\langle \boldsymbol{x}, \boldsymbol{\varphi}_i \rangle \langle \boldsymbol{x}, \boldsymbol{\varphi}_j \rangle\right\} \\ &= E\left\{\left(\int_a^b x(t)\, \varphi_i(t)\, dt\right) \cdot \left(\int_a^b x(u)\, \varphi_j(u)\, du\right)\right\} \\ &= E\left\{\int_a^b \varphi_i(t) \int_a^b x(t)\, x(u)\, \varphi_j(u)\, du\, dt\right\} \\ &= \int_a^b \varphi_i(t) \left(\int_a^b E\{x(t)\, x(u)\}\, \varphi_j(u)\, du\right) dt \\ &\overset{!}{=} \lambda_j \, \delta_{ij}. \end{aligned} \tag{5.3}$$

The kernel of the integral representation in (5.3) is the autocorrelation function

$$r_{xx}(t,u) = E\{x(t)\, x(u)\}. \tag{5.4}$$

We see that (5.3) is satisfied if

$$\lambda_j \delta_{ij} = \int_a^b \varphi_i(t) \left(\int_a^b r_{xx}(t,u)\, \varphi_j(u)\, du\right) dt. \tag{5.5}$$

Comparing (5.5) with the orthonormality relation $\delta_{ij} = \int_a^b \varphi_i(t)\, \varphi_j(t)\, dt$, we realize that

$$\int_a^b r_{xx}(t,u)\, \varphi_j(u)\, du = \lambda_j \, \varphi_j(t) \tag{5.6}$$

[1] l.i.m=limit in the mean[38].

must hold in order to satisfy (5.5). Thus, the solutions $\varphi_j(t)$, $j = 1, 2, \ldots$ of the integral equation (5.6) form the desired orthonormal basis. These functions are also called eigenfunctions of the integral operator in (5.6). The values λ_j, $j = 1, 2, \ldots$ are the eigenvalues. If the kernel $r_{xx}(t, u)$ is positive definite, that is, if $\int\int r_{xx}(t, u)x(t)x(u)\ dt\ du > 0$ for all $x(t) \in L_2(a, b)$, then the eigenfunctions form a complete orthonormal basis for $L_2(a, b)$. Further properties and particular solutions of the integral equation are for instance discussed in [149].

Signals can be approximated by carrying out the summation in (5.1) only for $i = 1, 2, \ldots, M$ with finite M. The mean approximation error produced thereby is the sum of those eigenvalues λ_j whose corresponding eigenfunctions are not used for the representation. Thus, we obtain an approximation with minimal mean square error if those eigenfunctions are used which correspond to the largest eigenvalues.

In practice, solving an integral equation represents a major problem. Therefore the continuous-time KLT is of minor interest with regard to practical applications. However, theoretically, that is, without solving the integral equation, this transform is an enormous help. We can describe stochastic processes by means of uncorrelated coefficients, solve estimation or recognition problems for vectors with uncorrelated components and then interpret the results for the continuous-time case.

5.2 The Discrete Karhunen–Loève Transform

We consider a real-valued zero-mean random process

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{5.7}$$

The restriction to zero-mean processes means no loss of generality, since any process $\boldsymbol{z}$ with mean $\boldsymbol{m}_z$ can be translated into a zero-mean process $\boldsymbol{x}$ by

$$\boldsymbol{x} = \boldsymbol{z} - \boldsymbol{m}_z. \tag{5.8}$$

With an orthonormal basis $\boldsymbol{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$, the process can be written as

$$\boldsymbol{x} = \boldsymbol{U}\ \boldsymbol{\alpha}, \tag{5.9}$$

where the representation

$$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T \tag{5.10}$$

is given by

$$\boldsymbol{\alpha} = \boldsymbol{U}^T \, \boldsymbol{x}. \tag{5.11}$$

As for the continuous-time case, we derive the KLT by demanding uncorrelated coefficients:

$$E\{\alpha_i \alpha_j\} = \lambda_j \, \delta_{ij}, \qquad i,j = 1,\ldots,n. \tag{5.12}$$

The scalars λ_j, $j = 1,\ldots,n$ are unknown real numbers with $\lambda_j \geq 0$. From (5.9) and (5.12) we obtain

$$E\left\{\boldsymbol{u}_i^T \boldsymbol{x} \, \boldsymbol{x}^T \boldsymbol{u}_j\right\} = \lambda_j \, \delta_{ij}, \qquad i,j = 1,\ldots,n. \tag{5.13}$$

With

$$\boldsymbol{R}_{xx} = E\left\{\boldsymbol{x}\boldsymbol{x}^T\right\} \tag{5.14}$$

this can be written as

$$\boldsymbol{u}_i^T \, \boldsymbol{R}_{xx} \, \boldsymbol{u}_j = \lambda_j \, \delta_{ij}, \quad i,j = 1,\ldots,n. \tag{5.15}$$

We observe that because of $\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$, equation (5.15) is satisfied if the vectors $\boldsymbol{u}_j$, $j = 1,\ldots,n$ are solutions to the eigenvalue problem

$$\boldsymbol{R}_{xx} \, \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \quad j = 1,\ldots,n. \tag{5.16}$$

Since $\boldsymbol{R}_{xx}$ is a covariance matrix, the eigenvalue problem has the following properties:

1. Only real eigenvalues λ_i exist.
2. A covariance matrix is positive definite or positive semidefinite, that is, for all eigenvalues we have $\lambda_i \geq 0$.
3. Eigenvectors that belong to different eigenvalues are orthogonal to one another.
4. If multiple eigenvalues occur, their eigenvectors are linearly independent and can be chosen to be orthogonal to one another.

Thus, we see that n orthogonal eigenvectors always exist. By normalizing the eigenvectors, we obtain the orthonormal basis of the Karhunen–Loève transform.

Complex-Valued Processes. For complex-valued processes $\boldsymbol{x} \in \mathbb{C}^n$, condition (5.12) becomes

$$E\left\{\alpha_i \alpha_j^*\right\} = \lambda_j \, \delta_{ij}, \; i,j = 1,\ldots,n.$$

This yields the eigenvalue problem

$$\boldsymbol{R}_{xx}\ \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j,\ j = 1, \ldots, n$$

with the covariance matrix

$$\boldsymbol{R}_{xx} = E\left\{\boldsymbol{x}\boldsymbol{x}^H\right\}.$$

Again, the eigenvalues are real and non-negative. The eigenvectors are orthogonal to one another such that $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ is unitary.

From the uncorrelatedness of the complex coefficients we cannot conclude that their real and imaginary parts are also uncorrelated; that is, $E\{\Re\{\alpha_i\}\ \Im\{\alpha_j\}\} = 0,\ i, j = 1, \ldots, n$ is not implied.

Best Approximation Property of the KLT. We henceforth assume that the eigenvalues are sorted such that $\lambda_1 \geq \ldots \geq \lambda_n$. From (5.12) we get for the variances of the coefficients:

$$E\left\{|\alpha_i|^2\right\} = \lambda_i, \qquad i = 1, \ldots, n. \tag{5.17}$$

For the mean-square error of an approximation

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{m} \alpha_i\ \boldsymbol{u}_i, \quad m < n, \tag{5.18}$$

we obtain

$$\begin{aligned} E\left\{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2\right\} &= E\left\{\left\|\sum_{i=m+1}^{n} \alpha_i \boldsymbol{u}_i\right\|^2\right\} \\ &= \sum_{i=m+1}^{n} E\left\{|\alpha_i|^2\right\} \\ &= \sum_{i=m+1}^{n} \lambda_i. \end{aligned} \tag{5.19}$$

It becomes obvious that an approximation with those eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, which belong to the largest eigenvectors leads to a minimal error.

In order to show that the KLT indeed yields the smallest possible error among all orthonormal linear transforms, we look at the maximization of $\sum_{i=1}^{m} E\{|\alpha_i|\}$ under the condition $\|\boldsymbol{u}_i\| = 1$. With $\alpha_i = \boldsymbol{u}_i^H \boldsymbol{x}$ this means

$$\sum_{i=1}^{m} E\left\{\boldsymbol{u}_i^H \boldsymbol{x}\boldsymbol{x}^H \boldsymbol{u}_i\right\} - \gamma_i \boldsymbol{u}_i^H \boldsymbol{u}_i = \sum_{i=1}^{m} \boldsymbol{u}_i^H \boldsymbol{R}_{xx} \boldsymbol{u}_i - \gamma_i \boldsymbol{u}_i^H \boldsymbol{u}_i \stackrel{!}{=} \max, \tag{5.20}$$

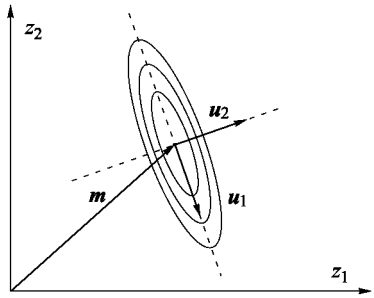

Figure 5.1. Contour lines of the pdf of a process $\boldsymbol{z} = [z_1, z_2]^T$.

where γ_i are Lagrange multipliers. Setting the gradient to zero yields

$$\boldsymbol{R}_{xx}\boldsymbol{u}_i = \gamma_i \boldsymbol{u}_i, \tag{5.21}$$

which is nothing but the eigenvalue problem (5.16) with $\gamma_i = \lambda_i$.

Figure 5.1 gives a geometric interpretation of the properties of the KLT. We see that $\boldsymbol{u}_1$ points towards the largest deviation from the center of gravity $\boldsymbol{m}$.

Minimal Geometric Mean Property of the KLT. For any positive definite matrix $\boldsymbol{X} = X_{ij},\ i,j = 1,\ldots,n$ the following inequality holds [7]:

$$\det\{\boldsymbol{X}\} \leq \prod_{k=1}^{n} X_{kk}. \tag{5.22}$$

Equality is given if $\boldsymbol{X}$ is diagonal. Since the KLT leads to a diagonal covariance matrix of the representation, this means that the KLT leads to random variables with a minimal geometric mean of the variances. From this, again, optimal properties in signal coding can be concluded [76].

The KLT of White Noise Processes. For the special case that $\boldsymbol{R}_{xx}$ is the covariance matrix of a white noise process with

$$\boldsymbol{R}_{xx} = \sigma^2\, \boldsymbol{I}$$

we have

$$\lambda_1 = \lambda_2 = \ldots = \lambda_n = \sigma^2.$$

Thus, the KLT is not unique in this case. Equation (5.19) shows that a white noise process can be optimally approximated with any orthonormal basis.

Relationships between Covariance Matrices. In the following we will briefly list some relationships between covariance matrices. With

$$\boldsymbol{\Lambda} = E\left\{\boldsymbol{\alpha}\,\boldsymbol{\alpha}^H\right\} = \begin{bmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{bmatrix}, \tag{5.23}$$

we can write (5.15) as

$$\boldsymbol{\Lambda} = \boldsymbol{U}^H \boldsymbol{R}_{xx} \boldsymbol{U}. \tag{5.24}$$

Observing $\boldsymbol{U}^H = \boldsymbol{U}^{-1}$, we obtain

$$\boldsymbol{R}_{xx} = \boldsymbol{U}\ \boldsymbol{\Lambda}\ \boldsymbol{U}^H. \tag{5.25}$$

Assuming that all eigenvalues are larger than zero, $\boldsymbol{\Lambda}^{-1}$ is given by

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{\lambda_n} \end{bmatrix} = \boldsymbol{U}^H \boldsymbol{R}_{xx}^{-1} \boldsymbol{U}. \tag{5.26}$$

Finally, for $\boldsymbol{R}_{xx}^{-1}$ we obtain

$$\boldsymbol{R}_{xx}^{-1} = \boldsymbol{U} \boldsymbol{\Lambda}^{-1} \boldsymbol{U}^H. \tag{5.27}$$

Application Example. In pattern recognition it is important to classify signals by means of a few concise features. The signals considered in this example are taken from inductive loops embedded in the pavement of a highway in order to measure the change of inductivity while vehicles pass over them. The goal is to discriminate different types of vehicle (car, truck, bus, etc.). In the following, we will consider the two groups car and truck. After appropriate pre-processing (normalization of speed, length, and amplitude) we obtain the measured signals shown in Figure 5.2, which are typical examples of the two classes. The stochastic processes considered are $\boldsymbol{z}_1$ (car) and $\boldsymbol{z}_2$ (truck). The realizations are denoted as ${}^i\mathbf{z}_1$, ${}^i\mathbf{z}_2$, $i = 1 \ldots N$.

In a first step, zero-mean processes are generated:

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{z}_1 - \boldsymbol{m}_1, \\ \boldsymbol{x}_2 &= \boldsymbol{z}_2 - \boldsymbol{m}_2. \end{aligned} \tag{5.28}$$

The mean values can be estimated by

$$\boldsymbol{m}_1 = E\left\{\boldsymbol{z}_1\right\} \approx \frac{1}{N} \sum_{i=1}^{N} {}^i\mathbf{z}_1 \tag{5.29}$$

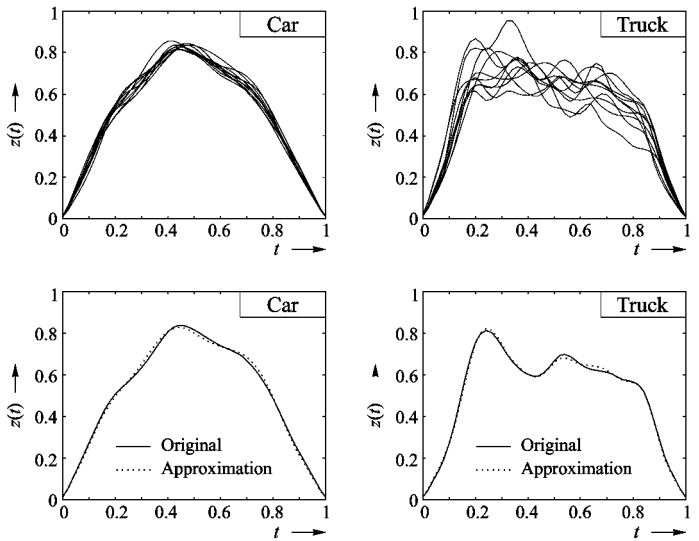

Figure 5.2. Examples of sample functions; (a) typical signal contours; (b) two sample functions and their approximations.

and

$$\boldsymbol{m}_2 = E\{\boldsymbol{z}_2\} \approx \frac{1}{N}\sum_{i=1}^{N} {}^i\mathbf{z}_2. \tag{5.30}$$

Observing the *a priori* probabilities of the two classes, p_1 and p_2, a process

$$\boldsymbol{x} = p_1\boldsymbol{x}_1 + p_2\boldsymbol{x}_2 \tag{5.31}$$

can be defined. The covariance matrix $\boldsymbol{R}_{xx}$ can be estimated as

$$\boldsymbol{R}_{xx} = E\{\boldsymbol{x}\boldsymbol{x}^T\} \approx \frac{p_1}{N+1}\sum_{i=1}^{N} {}^i\mathbf{x}_1\, {}^i\mathbf{x}_1^T + \frac{p_2}{N+1}\sum_{i=1}^{N} {}^i\mathbf{x}_2\, {}^i\mathbf{x}_2^T, \tag{5.32}$$

where ${}^i\mathbf{x}_1$ and ${}^i\mathbf{x}_2$ are realizations of the zero-mean processes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

The first ten eigenvalues computed from a training set are:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
212923	55460	20559	15790	10230	5262	5036	3139	2551	968

We see that by using only a few eigenvectors a good approximation can be expected. To give an example, Figure 5.2 shows two signals and their

approximations

$$\begin{aligned} {}^1\hat{\mathbf{z}}_1 &= \boldsymbol{U}\boldsymbol{U}^T({}^1\mathbf{z}_1 - \boldsymbol{m}_1) + \boldsymbol{m}_1, \\ {}^1\hat{\mathbf{z}}_2 &= \boldsymbol{U}\boldsymbol{U}^T({}^1\mathbf{z}_2 - \boldsymbol{m}_2) + \boldsymbol{m}_2 \end{aligned} \tag{5.33}$$

with the basis $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4\}$.

In general, the optimality and usefulness of extracted features for discrimination is highly dependent on the algorithm that is used to carry out the discrimination. Thus, the feature extraction method described in this example is not meant to be optimal for all applications. However, it shows how a high proportion of information about a process can be stored within a few features. For more details on classification algorithms and further transforms for feature extraction, see [59, 44, 167, 58].

5.3 The KLT of Real-Valued AR(1) Processes

An *autoregressive process* of order p (AR(p) process) is generated by exciting a recursive filter of order p with a zero-mean, stationary white noise process. The filter has the system function

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} \rho(i)\ z^{-i}}, \qquad \rho(p) \neq 0. \tag{5.34}$$

Thus, an AR(p) process $x(n)$ is described by the difference equation

$$x(n) = w(n) + \sum_{i=1}^{p} \rho(i)\ x(n-i), \tag{5.35}$$

where $w(n)$ is white noise. The AR(1) process with difference equation

$$x(n) = w(n) + \rho\ x(n-1) \tag{5.36}$$

is often used as a simple model. It is also known as a *first-order Markov process*. From (5.36) we obtain by recursion:

$$x(n) = \sum_{i=0}^{\infty} \rho^i\ w(n-i). \tag{5.37}$$

For determining the variance of the process $x(n)$, we use the properties

$$m_w = E\{w(n)\} = 0 \quad \rightarrow \quad m_x = E\{x(n)\} = 0 \tag{5.38}$$

and

$$r_{ww}(m) = E\{w(n)w(n+m)\} = \sigma^2 \delta_{m0}, \tag{5.39}$$

where δ_{m0} is the Kronecker delta. Supposing $|\rho| < 1$, we get

$$\begin{aligned} \sigma_x^2 &= E\{|x(n)|^2\} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \rho^i \, \rho^j \, E\{w(n-i)w(n-j)\} \\ &= \sigma^2 \sum_{i=0}^{\infty} \rho^{2i} \\ &= \frac{\sigma^2}{1-\rho^2}. \end{aligned} \tag{5.40}$$

For the autocorrelation sequence we obtain

$$\begin{aligned} r_{xx}(m) &= E\{x(n)x(n+m)\} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \rho^i \, \rho^j \, E\{w(n-i)w(n-j+m)\} \\ &= \sigma^2 \rho^{|m|} \sum_{i=0}^{\infty} \rho^{2i} \\ &= \frac{\sigma^2}{1-\rho^2} \; \rho^{|m|}. \end{aligned} \tag{5.41}$$

We see that the autocorrelation sequence is infinitely long. However, henceforth only the values $r_{xx}(-N+1), \ldots, r_{xx}(N-1)$ shall be considered. Because of the stationarity of the input process, the covariance matrix of the AR(1) process is a Toeplitz matrix. It is given by

$$\boldsymbol{R}_{xx} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{N-1} \\ \rho & 1 & \rho & & \vdots \\ \rho^2 & \rho & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \rho \\ \rho^{N-1} & \ldots & \ldots & \rho & 1 \end{bmatrix}. \tag{5.42}$$

The eigenvectors of $\boldsymbol{R}_{xx}$ form the basis of the KLT. For real signals and even N, the eigenvalues λ_k, $k = 0, \ldots, N-1$ and the eigenvectors were analytically derived by Ray and Driver [123]. The eigenvalues are

$$\lambda_k = \frac{1}{1 - 2\,\rho\,\cos(\alpha_k) + \rho^2}, \qquad k = 0, \ldots, N-1, \tag{5.43}$$

where $\alpha_k,\ k = 0, \ldots, N-1$ denotes the real positive roots of

$$\tan(N\alpha_k) = -\frac{(1-\rho^2)\sin(\alpha_k)}{\cos(\alpha_k) - 2\rho + \rho\ \cos(\alpha_k)}. \tag{5.44}$$

The components of the eigenvectors $\boldsymbol{u}_k,\ k = 0, \ldots, N-1$ are given by

$$u_k(n) = \frac{2}{N+\lambda_k}\ \sin\Big(\alpha_k\Big(n - \frac{N-1}{2}\Big) + (k+1)\frac{\pi}{2}\Big), \qquad n, k = 0, \ldots, N-1. \tag{5.45}$$

5.4 Whitening Transforms

In this section we are concerned with the problem of transforming a colored noise process into a white noise process. That is, the coefficients of the representation should not only be uncorrelated (as for the KLT), they should also have the same variance. Such transforms, known as whitening transforms, are mainly applied in signal detection and pattern recognition, because they lead to a convenient process representation with additive white noise.

Let $\boldsymbol{n}$ be a process with covariance matrix

$$\boldsymbol{R}_{nn} = E\left\{\boldsymbol{n}\boldsymbol{n}^H\right\} \quad \neq \quad \sigma^2\boldsymbol{I}. \tag{5.46}$$

We wish to find a linear transform $\boldsymbol{T}$ which yields an equivalent process

$$\tilde{\boldsymbol{n}} = \boldsymbol{T}\boldsymbol{n} \tag{5.47}$$

with

$$E\left\{\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^H\right\} = E\left\{\boldsymbol{T}\boldsymbol{n}\boldsymbol{n}^H\boldsymbol{T}^H\right\} = \boldsymbol{T}\boldsymbol{R}_{nn}\boldsymbol{T}^H = \boldsymbol{I}. \tag{5.48}$$

We already see that the transform cannot be unique since by multiplying an already computed matrix $\boldsymbol{T}$ with an arbitrary unitary matrix, property (5.48) is preserved.

The covariance matrix can be decomposed as follows (KLT):

$$\boldsymbol{R}_{nn} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^H\boldsymbol{U}^H. \tag{5.49}$$

For $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ we have

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_N} \end{bmatrix}.$$

Possible transforms are

$$\boldsymbol{T} = \boldsymbol{\Sigma}^{-1}\boldsymbol{U}^H \tag{5.50}$$

or

$$\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^H. \tag{5.51}$$

This can easily be verified by substituting (5.50) into (5.48):

$$E\left\{\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^H\right\} = \boldsymbol{T}\boldsymbol{R}_{nn}\boldsymbol{T}^H = \boldsymbol{\Sigma}^{-1}\underbrace{\boldsymbol{U}^H\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^H\underbrace{\boldsymbol{U}^H\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Sigma}^{H^{-1}} = \boldsymbol{I}. \tag{5.52}$$

Alternatively, we can apply the *Cholesky decomposition*

$$\boldsymbol{R}_{nn} = \boldsymbol{L}\boldsymbol{L}^H, \tag{5.53}$$

where $\boldsymbol{L}$ is a lower triangular matrix. The whitening transform is

$$\boldsymbol{T} = \boldsymbol{L}^{-1}. \tag{5.54}$$

For the covariance matrix we again have

$$E\left\{\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^H\right\} = \boldsymbol{T}\boldsymbol{R}_{nn}\boldsymbol{T}^H = \boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{L}^H\boldsymbol{L}^{H^{-1}} = \boldsymbol{I}. \tag{5.55}$$

In signal analysis, one often encounters signals of the form

$$\boldsymbol{r} = \boldsymbol{s} + \boldsymbol{n}, \tag{5.56}$$

where $\boldsymbol{s}$ is a known signal and $\boldsymbol{n}$ is an additive colored noise processes. The whitening transforms transfer (5.56) into an equivalent model

$$\tilde{\boldsymbol{r}} = \tilde{\boldsymbol{s}} + \tilde{\boldsymbol{n}} \tag{5.57}$$

with

$$\begin{aligned} \tilde{\boldsymbol{r}} &= \boldsymbol{T}\boldsymbol{r}, \\ \tilde{\boldsymbol{s}} &= \boldsymbol{T}\boldsymbol{s}, \\ \tilde{\boldsymbol{n}} &= \boldsymbol{T}\boldsymbol{n}, \end{aligned} \tag{5.58}$$

where $\tilde{\boldsymbol{n}}$ is a white noise process of variance $\sigma_{\tilde{n}}^2 = 1$.

5.5 Linear Estimation

In estimation the goal is to determine a set of parameters as precisely as possible from noisy observations. We will focus on the case where the estimators are linear, that is, the estimates for the parameters are computed as linear combinations of the observations. This problem is closely related to the problem of computing the coefficients of a series expansion of a signal, as described in Chapter 3.

Linear methods do not require precise knowledge of the noise statistics; only moments up to the second order are taken into account. Therefore they are optimal only under the linearity constraint, and, in general, non-linear estimators with better properties may be found. However, linear estimators constitute the globally optimal solution as far as Gaussian processes are concerned [149].

5.5.1 Least-Squares Estimation

We consider the model

$$\boldsymbol{r} = \boldsymbol{S}\,\boldsymbol{a} + \boldsymbol{n}, \tag{5.59}$$

where $\boldsymbol{r}$ is our observation, $\boldsymbol{a}$ is the parameter vector in question, and $\boldsymbol{n}$ is a noise process. Matrix $\boldsymbol{S}$ can be understood as a basis matrix that relates the parameters to the clean observation $\boldsymbol{Sa}$.

The requirement to have an unbiased estimate can be written as

$$E\{\hat{\boldsymbol{a}}(\boldsymbol{r})|\mathbf{a}\} = \mathbf{a}, \tag{5.60}$$

where $\mathbf{a}$ is understood as an arbitrary non-random parameter vector. Because of the additive noise, the estimates $\hat{\boldsymbol{a}}(\boldsymbol{r})|\mathbf{a}$ again form a random process.

The linear estimation approach is given by

$$\hat{\boldsymbol{a}}(\boldsymbol{r}) = \boldsymbol{A}\,\boldsymbol{r}. \tag{5.61}$$

If we assume zero-mean noise $\boldsymbol{n}$, matrix $\boldsymbol{A}$ must satisfy

$$\boldsymbol{A}\,\boldsymbol{S} = \boldsymbol{I} \tag{5.62}$$

in order to ensure unbiased estimates. This is seen from

$$\begin{aligned} E\{\hat{\boldsymbol{a}}(\boldsymbol{r})|\mathbf{a}\} &= E\{\boldsymbol{A}\,\boldsymbol{r}|\mathbf{a}\} \\ &= \boldsymbol{A}\,E\{\boldsymbol{r}|\mathbf{a}\} \\ &= \boldsymbol{A}\,E\{\boldsymbol{S}\mathbf{a}+\boldsymbol{n}\} \\ &= \boldsymbol{A}\,\boldsymbol{S}\,\mathbf{a} \\ &\stackrel{!}{=} \mathbf{a}. \end{aligned} \tag{5.63}$$

The *generalized least-squares estimator* is derived from the criterion

$$\|\mathbf{r}-\boldsymbol{S}\boldsymbol{\alpha}\|\Big|_{\boldsymbol{\alpha}=\hat{\mathbf{a}}(\mathbf{r})} \stackrel{!}{=} \min, \tag{5.64}$$

where an arbitrary weighting matrix $\boldsymbol{G}$ may be involved in the definition of the inner product that induces the norm in (5.64). Here the observation $\mathbf{r}$ is considered as a single realization of the stochastic process $\boldsymbol{r}$. Making use of the fact that orthogonal projections yield a minimal approximation error, we get

$$\hat{\mathbf{a}}(\mathbf{r}) = [\boldsymbol{S}^H\boldsymbol{G}\boldsymbol{S}]^{-1}\boldsymbol{S}^H\boldsymbol{G}\mathbf{r} \tag{5.65}$$

according to (3.95). Assuming that $[\boldsymbol{S}^H\boldsymbol{G}\boldsymbol{S}]^{-1}$ exists, the requirement (5.65) to have an unbiased estimator is satisfied for arbitrary weighting matrices, as can easily be verified.

If we choose $\boldsymbol{G}=\boldsymbol{I}$, we speak of a *least-squares estimator*. For weighting matrices $\boldsymbol{G}\neq\boldsymbol{I}$, we speak of a *generalized least-squares estimator*. However, the approach leaves open the question of how a suitable $\boldsymbol{G}$ is found.

5.5.2 The Best Linear Unbiased Estimator (BLUE)

As will be shown below, choosing $\boldsymbol{G}=\boldsymbol{R}_{nn}^{-1}$, where

$$\boldsymbol{R}_{nn} = E\{\boldsymbol{n}\boldsymbol{n}^H\} \tag{5.66}$$

is the correlation matrix of the noise, yields an unbiased estimator with minimal variance. The estimator, which is known as the *best linear unbiased estimator* (BLUE), then is

$$\boldsymbol{A} = [\boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}]^{-1}\boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}. \tag{5.67}$$

The estimate is given by

$$\hat{\boldsymbol{a}}(\boldsymbol{r}) = [\boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}]^{-1}\boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\,\boldsymbol{r}. \tag{5.68}$$

The variances of the individual estimates can be found on the main diagonal of the covariance matrix of the error $\boldsymbol{e} = \hat{\boldsymbol{a}}(\boldsymbol{r}) - \mathbf{a}$, given by

$$\boldsymbol{R}_{ee} = [\boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}]^{-1}. \tag{5.69}$$

Proof of (5.69) and the optimality of (5.67). First, observe that with $\boldsymbol{A}\boldsymbol{S} = \boldsymbol{I}$ we have

$$\begin{aligned} \hat{\boldsymbol{a}}(\boldsymbol{r}) - \mathbf{a}|\mathbf{a} &= \boldsymbol{A}\,\boldsymbol{S}\,\mathbf{a} + \boldsymbol{A}\,\boldsymbol{n} - \mathbf{a} \\ &= \boldsymbol{A}\,\boldsymbol{n}. \end{aligned} \tag{5.70}$$

Thus,

$$\begin{aligned} \boldsymbol{R}_{ee} &= \boldsymbol{A} E\{\boldsymbol{n}\boldsymbol{n}^H\} \boldsymbol{A}^H \\ &= \boldsymbol{A}\boldsymbol{R}_{nn}\boldsymbol{A}^H \\ &= [\boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}]^{-1} \boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{R}_{nn} \boldsymbol{R}_{nn}^{-1} \boldsymbol{S} [\boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}]^{-1} \\ &= [\boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}]^{-1}. \end{aligned} \tag{5.71}$$

In order to see whether $\boldsymbol{A}$ according to (5.67) is optimal, an estimation

$$\tilde{\boldsymbol{a}}(\boldsymbol{r}) = \tilde{\boldsymbol{A}}\,\boldsymbol{r} \tag{5.72}$$

with

$$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{D} \tag{5.73}$$

will be considered. The unbiasedness constraint requires that

$$\tilde{\boldsymbol{A}}\,\boldsymbol{S} = \boldsymbol{I}. \tag{5.74}$$

Because of $\boldsymbol{A}\,\boldsymbol{S} = \boldsymbol{I}$ this means

$$\boldsymbol{D}\,\boldsymbol{S} = \mathbf{0} \qquad \text{(null matrix)}. \tag{5.75}$$

For the covariance matrix of the error $\tilde{\boldsymbol{e}}(\boldsymbol{r}) = \tilde{\boldsymbol{a}}(\boldsymbol{r}) - \mathbf{a}$ we obtain

$$\begin{aligned} \boldsymbol{R}_{\tilde{e}\tilde{e}} &= \tilde{\boldsymbol{A}}\boldsymbol{R}_{nn}\tilde{\boldsymbol{A}}^H \\ &= [\boldsymbol{A} + \boldsymbol{D}]\boldsymbol{R}_{nn}[\boldsymbol{A} + \boldsymbol{D}]^H \\ &= \boldsymbol{A}\boldsymbol{R}_{nn}\boldsymbol{A}^H + \boldsymbol{A}\boldsymbol{R}_{nn}\boldsymbol{D}^H + \boldsymbol{D}\boldsymbol{R}_{nn}\boldsymbol{A}^H + \boldsymbol{D}\boldsymbol{R}_{nn}\boldsymbol{D}^H. \end{aligned} \tag{5.76}$$

With

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{R}_{nn}\boldsymbol{D}^H)^H = \boldsymbol{D}\boldsymbol{R}_{nn}\boldsymbol{A}^H &= \boldsymbol{D}\boldsymbol{R}_{nn}\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}[\boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}]^{-1} \\ &= \underbrace{\boldsymbol{D}\boldsymbol{S}}_{\boldsymbol{0}}[\boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}]^{-1} \\ &= \boldsymbol{0}\end{aligned} \tag{5.77}$$

(5.76) reduces to

$$\boldsymbol{R}_{\tilde{e}\tilde{e}} = \boldsymbol{A}\boldsymbol{R}_{nn}\boldsymbol{A}^H + \boldsymbol{D}\boldsymbol{R}_{nn}\boldsymbol{D}^H. \tag{5.78}$$

We see that $\boldsymbol{R}_{\tilde{e}\tilde{e}}$ is the sum of two non-negative definite expressions so that minimal main diagonal elements of $\boldsymbol{R}_{\tilde{e}\tilde{e}}$ are yielded for $\boldsymbol{D} = \boldsymbol{0}$ and thus for $\boldsymbol{A}$ according to (5.67). □

In the case of a white noise process $\boldsymbol{n}$, (5.68) reduces to

$$\hat{\boldsymbol{a}}(\boldsymbol{r}) = [\boldsymbol{S}^H\boldsymbol{S}]^{-1}\boldsymbol{S}^H\boldsymbol{r}. \tag{5.79}$$

Otherwise the weighting with $\boldsymbol{G} = \boldsymbol{R}_{nn}^{-1}$ can be interpreted as an implicit whitening of the noise. This can be seen by using the Cholesky decomposition $\boldsymbol{R}_{nn} = \boldsymbol{L}\boldsymbol{L}^H$ and and by rewriting the model as

$$\tilde{\boldsymbol{r}} = \tilde{\boldsymbol{S}}\boldsymbol{a} + \tilde{\boldsymbol{n}}, \tag{5.80}$$

where

$$\tilde{\boldsymbol{r}} = \boldsymbol{L}^{-1}\boldsymbol{r},\ \tilde{\boldsymbol{S}} = \boldsymbol{L}^{-1}\boldsymbol{S},\ \tilde{\boldsymbol{n}} = \boldsymbol{L}^{-1}\boldsymbol{n}. \tag{5.81}$$

The transformed process $\tilde{\boldsymbol{n}}$ is a white noise process. The equivalent estimator then is

$$\hat{\boldsymbol{a}}(\tilde{\boldsymbol{r}}) = [\tilde{\boldsymbol{S}}^H\tilde{\boldsymbol{S}}]^{-1}\tilde{\boldsymbol{S}}^H\ \tilde{\boldsymbol{r}}. \tag{5.82}$$

5.5.3 Minimum Mean Square Error Estimation

The advantage of the linear estimators considered in the previous section is their unbiasedness. If we dispense with this property, estimates with smaller mean square error may be found. We will start the discussion on the assumptions

$$E\{\boldsymbol{r}\} = \boldsymbol{0}, \qquad E\{\boldsymbol{a}\} = \boldsymbol{0}. \tag{5.83}$$

Again, the linear estimator is described by a matrix $\boldsymbol{A}$:

$$\hat{\boldsymbol{a}}(\boldsymbol{r}) = \boldsymbol{A}\ \boldsymbol{r}. \tag{5.84}$$

Here, $\boldsymbol{r}$ is somehow dependent on $\boldsymbol{a}$, but the inner relationship between $\boldsymbol{r}$ and $\boldsymbol{a}$ need not be known however. The matrix $\boldsymbol{A}$ which yields minimal main diagonal elements of the correlation matrix of the estimation error $\boldsymbol{e} = \boldsymbol{a} - \hat{\boldsymbol{a}}$ is called the *minimum mean square error (MMSE) estimator.*

In order to find the optimal $\boldsymbol{A}$, observe that

$$\begin{aligned} \boldsymbol{R}_{ee} &= E\left\{[\hat{\boldsymbol{a}} - \boldsymbol{a}]\,[\hat{\boldsymbol{a}} - \boldsymbol{a}]^H\right\} \\ &= E\left\{\boldsymbol{a}\boldsymbol{a}^H\right\} - E\left\{\hat{\boldsymbol{a}}\boldsymbol{a}^H\right\} - E\left\{\boldsymbol{a}\hat{\boldsymbol{a}}^H\right\} + E\left\{\hat{\boldsymbol{a}}\hat{\boldsymbol{a}}^H\right\}. \end{aligned} \tag{5.85}$$

Substituting (5.84) into (5.85) yields

$$\boldsymbol{R}_{ee} = \boldsymbol{R}_{aa} - \boldsymbol{A}\boldsymbol{R}_{ar} - \boldsymbol{R}_{ra}\boldsymbol{A}^H + \boldsymbol{A}\boldsymbol{R}_{rr}\boldsymbol{A}^H \tag{5.86}$$

with

$$\begin{aligned} \boldsymbol{R}_{aa} &= E\left\{\boldsymbol{a}\boldsymbol{a}^H\right\}, \\ \boldsymbol{R}_{ar} &= \boldsymbol{R}_{ra}^H = E\left\{\boldsymbol{r}\boldsymbol{a}^H\right\}, \\ \boldsymbol{R}_{rr} &= E\left\{\boldsymbol{r}\boldsymbol{r}^H\right\}. \end{aligned} \tag{5.87}$$

Assuming the existence of $\boldsymbol{R}_{rr}^{-1}$, (5.86) can be extended by

$$\boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{-1}\boldsymbol{R}_{ar} - \boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{-1}\boldsymbol{R}_{ar}$$

and be re-written as

$$\boldsymbol{R}_{ee} = \left[\boldsymbol{A} - \boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{-1}\right]\boldsymbol{R}_{rr}\left[\boldsymbol{A}^H - \boldsymbol{R}_{rr}^{-1}\boldsymbol{R}_{ar}\right] - \boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{-1}\boldsymbol{R}_{ar} + \boldsymbol{R}_{aa}. \tag{5.88}$$

Clearly, $\boldsymbol{R}_{ee}$ has positive diagonal elements. Since only the first term on the right-hand side of (5.88) is dependent on $\boldsymbol{A}$, we have a minimum of the diagonal elements of $\boldsymbol{R}_{ee}$ for

$$\boldsymbol{A} = \boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{-1}. \tag{5.89}$$

The correlation matrix of the estimation error is then given by

$$\boldsymbol{R}_{ee} = \boldsymbol{R}_{aa} - \boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{-1}\boldsymbol{R}_{ar}. \tag{5.90}$$

Orthogonality Principle. In Chapter 3 we saw that approximations $\hat{\boldsymbol{x}}$ of signals $\boldsymbol{x}$ are obtained with minimal error if the error $\hat{\boldsymbol{x}} - \boldsymbol{x}$ is orthogonal to $\hat{\boldsymbol{x}}$. A similar relationship holds between parameter vectors $\boldsymbol{a}$ and their MMSE estimates. With $\boldsymbol{A}$ according to (5.89) we have

$$\boldsymbol{R}_{ra} = \boldsymbol{A}\,\boldsymbol{R}_{rr}, \quad \text{i.e.} \quad E\left\{\boldsymbol{a}\boldsymbol{r}^H\right\} = \boldsymbol{A}\,E\left\{\boldsymbol{r}\boldsymbol{r}^H\right\}. \tag{5.91}$$

This means that the following orthogonality relations hold:

$$\begin{aligned} E\left\{[\hat{\boldsymbol{a}}-\boldsymbol{a}]\,\hat{\boldsymbol{a}}^H\right\} &= \boldsymbol{R}_{\hat{a}\hat{a}}-\boldsymbol{R}_{\hat{a}a} \\ &= [\,\boldsymbol{A}\boldsymbol{R}_{rr}-\boldsymbol{R}_{ra}\,]\,\boldsymbol{A}^H \\ &= \boldsymbol{0}. \end{aligned} \tag{5.92}$$

With $\boldsymbol{A}\,\boldsymbol{r}=\hat{\boldsymbol{a}}$ the right part of (5.91) can also be written as

$$E\left\{\boldsymbol{a}\boldsymbol{r}^H\right\}=E\left\{\hat{\boldsymbol{a}}\boldsymbol{r}^H\right\}, \tag{5.93}$$

which yields

$$E\left\{[\hat{\boldsymbol{a}}-\boldsymbol{a}]\,\boldsymbol{r}^H\right\}=\boldsymbol{0}. \tag{5.94}$$

The relationship expressed in (5.94) is referred to as the *orthogonality principle*. The orthogonality principle states that we get an MMSE estimate if the error $\hat{\boldsymbol{a}}(\boldsymbol{r})-\boldsymbol{a}$ is uncorrelated to all components of the input vector $\boldsymbol{r}$ used for computing $\hat{\boldsymbol{a}}(\boldsymbol{r})$.

Singular Correlation Matrix. There are cases where the correlation matrix $\boldsymbol{R}_{rr}$ becomes singular and the linear estimator cannot be written as

$$\boldsymbol{A}=\boldsymbol{R}_{ra}\,\boldsymbol{R}_{rr}^{-1}. \tag{5.95}$$

A more general solution, which involves the replacement of the inverse by the pseudoinverse, is

$$\boldsymbol{A}=\boldsymbol{R}_{ra}\,\boldsymbol{R}_{rr}^{+}. \tag{5.96}$$

In order to show the optimality of (5.96), the estimator

$$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{D} \tag{5.97}$$

with $\boldsymbol{A}$ according to (5.96) and an arbitrary matrix $\boldsymbol{D}$ is considered. Using the properties of the pseudoinverse, we derive from (5.97) and (5.86):

$$\begin{aligned} \boldsymbol{R}_{ee} &= \boldsymbol{R}_{aa}-\tilde{\boldsymbol{A}}\boldsymbol{R}_{ar}-\boldsymbol{R}_{ra}\tilde{\boldsymbol{A}}^H+\tilde{\boldsymbol{A}}\boldsymbol{R}_{rr}\tilde{\boldsymbol{A}}^H \\ &= \boldsymbol{R}_{aa}-\boldsymbol{R}_{ra}\boldsymbol{R}_{rr}^{+}\boldsymbol{R}_{ar}+\boldsymbol{D}\boldsymbol{R}_{rr}^{+}\boldsymbol{D}^H. \end{aligned} \tag{5.98}$$

Since $\boldsymbol{R}_{rr}^{+}$ is at least positive semidefinite, we get a minimum of the diagonal elements of $\boldsymbol{R}_{ee}$ for $\boldsymbol{D}=\boldsymbol{0}$, and (5.96) constitutes one of the optimal solutions.

Additive Uncorrelated Noise. So far, nothing has been said about possible dependencies between $\boldsymbol{a}$ and the noise contained in $\boldsymbol{r}$. Assuming that

$$\boldsymbol{r}=\boldsymbol{S}\,\boldsymbol{a}+\boldsymbol{n}, \tag{5.99}$$

where $\boldsymbol{n}$ is an additive, uncorrelated noise process, we have

$$\begin{aligned} \boldsymbol{R}_{ra} &= \boldsymbol{R}_{ar}^H = \boldsymbol{R}_{aa}\, \boldsymbol{S}^H, \\ \boldsymbol{R}_{rr} &= \boldsymbol{S}\, \boldsymbol{R}_{aa}\, \boldsymbol{S}^H + \boldsymbol{R}_{nn}, \end{aligned} \tag{5.100}$$

and $\boldsymbol{A}$ according to (5.89) becomes

$$\boldsymbol{A} = \boldsymbol{R}_{aa}\boldsymbol{S}^H \left[\boldsymbol{S}\boldsymbol{R}_{aa}\boldsymbol{S}^H + \boldsymbol{R}_{nn}\right]^{-1}. \tag{5.101}$$

Alternatively, $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \left[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\right]^{-1} \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}. \tag{5.102}$$

This is verified by equating (5.101) and (5.102), and by multiplying the obtained expression with $[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}]$ from the left and with $[\boldsymbol{S}\boldsymbol{R}_{aa}\boldsymbol{S}^H + \boldsymbol{R}_{nn}]$ from the right, respectively:

$$\left[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\right] \boldsymbol{R}_{aa}\boldsymbol{S}^H = \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}[\boldsymbol{S}\boldsymbol{R}_{aa}\boldsymbol{S}^H + \boldsymbol{R}_{nn}].$$

The equality of both sides is easily seen. The matrices to be inverted in (5.102), except $\boldsymbol{R}_{nn}$, typically have a much smaller dimension than those in (5.101). If the noise is white, $\boldsymbol{R}_{nn}^{-1}$ can be immediately stated, and (5.102) is advantageous in terms of computational cost.

For $\boldsymbol{R}_{ee}$ we get from (5.89), (5.90), (5.100) and (5.102):

$$\begin{aligned} \boldsymbol{R}_{ee} &= \boldsymbol{R}_{aa} - \boldsymbol{A}\boldsymbol{R}_{ar} \\ &= \boldsymbol{R}_{aa} - \left[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\right]^{-1} \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\boldsymbol{R}_{aa}. \end{aligned} \tag{5.103}$$

Multiplying (5.103) with $[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}]$ from the left yields

$$\begin{aligned} \left[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\right] \boldsymbol{R}_{ee} &= \left[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\right] \boldsymbol{R}_{aa} - \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\boldsymbol{R}_{aa} \\ &= \boldsymbol{I}, \end{aligned} \tag{5.104}$$

so that the following expression is finally obtained:

$$\boldsymbol{R}_{ee} = \left[\boldsymbol{R}_{aa}^{-1} + \boldsymbol{S}^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}\right]^{-1}. \tag{5.105}$$

Equivalent Estimation Problems. We partition $\boldsymbol{A}$ and $\boldsymbol{a}$ into

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix}, \qquad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}, \tag{5.106}$$

such that

$$\begin{bmatrix} \hat{\boldsymbol{a}}_1(\boldsymbol{r}) \\ \hat{\boldsymbol{a}}_2(\boldsymbol{r}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{r}. \tag{5.107}$$

If we assume that the processes $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{n}$ are independent of one another, the covariance matrix $\boldsymbol{R}_{aa}$ and its inverse $\boldsymbol{R}_{aa}^{-1}$ have the form

$$\boldsymbol{R}_{aa} = \begin{bmatrix} \boldsymbol{R}_{a_1a_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_{a_2a_2} \end{bmatrix}, \quad \boldsymbol{R}_{aa}^{-1} = \begin{bmatrix} \boldsymbol{R}_{a_1a_1}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_{a_2a_2}^{-1} \end{bmatrix}, \tag{5.108}$$

and $\boldsymbol{A}$ according to (5.102) can be written as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_1 + \boldsymbol{R}_{a_1a_1}^{-1} & \boldsymbol{S}_1^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_2 \\ \boldsymbol{S}_2^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_1 & \boldsymbol{S}_2^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_2 + \boldsymbol{R}_{a_2a_2}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{S}_1^H \boldsymbol{R}_{nn}^{-1} \\ \boldsymbol{S}_2^H \boldsymbol{R}_{nn}^{-1} \end{bmatrix}, \tag{5.109}$$

where $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2]$. Applying the matrix equation

$$\begin{aligned} \begin{bmatrix} \mathcal{E} & \mathcal{F} \\ \mathcal{G} & \mathcal{H} \end{bmatrix}^{-1} &= \begin{bmatrix} \mathcal{E}^{-1} + \mathcal{E}^{-1}\mathcal{F}\mathcal{D}^{-1}\mathcal{G}\mathcal{E}^{-1} & -\mathcal{E}^{-1}\mathcal{F}\mathcal{D}^{-1} \\ -\mathcal{D}^{-1}\mathcal{G}\mathcal{E}^{-1} & \mathcal{D}^{-1} \end{bmatrix} \\ \mathcal{D} &= \mathcal{H} - \mathcal{G}\mathcal{E}^{-1}\mathcal{F} \end{aligned} \tag{5.110}$$

yields

$$\boldsymbol{A}_1 = [\boldsymbol{S}_1^H \boldsymbol{R}_{n_1n_1}^{-1} \boldsymbol{S}_1 + \boldsymbol{R}_{a_1a_1}^{-1}]^{-1} \boldsymbol{S}_1^H \boldsymbol{R}_{n_1n_1}^{-1}, \tag{5.111}$$

$$\boldsymbol{A}_2 = [\boldsymbol{S}_2^H \boldsymbol{R}_{n_2n_2}^{-1} \boldsymbol{S}_2 + \boldsymbol{R}_{a_2a_2}^{-1}]^{-1} \boldsymbol{S}_2^H \boldsymbol{R}_{n_2n_2}^{-1} \tag{5.112}$$

with

$$\boldsymbol{R}_{n_1n_1} = \boldsymbol{R}_{nn} + \boldsymbol{S}_2 \boldsymbol{R}_{a_2a_2} \boldsymbol{S}_2^H, \tag{5.113}$$

$$\boldsymbol{R}_{n_2n_2} = \boldsymbol{R}_{nn} + \boldsymbol{S}_1 \boldsymbol{R}_{a_1a_1} \boldsymbol{S}_1^H. \tag{5.114}$$

The inverses $\boldsymbol{R}_{n_1n_1}^{-1}$ and $\boldsymbol{R}_{n_2n_2}^{-1}$ can be written as

$$\boldsymbol{R}_{n_1n_1}^{-1} = \left[\boldsymbol{R}_{nn}^{-1} - \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_2 \left(\boldsymbol{S}_2^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_2 + \boldsymbol{R}_{a_2a_2}^{-1}\right)^{-1} \boldsymbol{S}_2^H \boldsymbol{R}_{nn}^{-1}\right], \tag{5.115}$$

$$\boldsymbol{R}_{n_2n_2}^{-1} = \left[\boldsymbol{R}_{nn}^{-1} - \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_1 \left(\boldsymbol{S}_1^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}_1 + \boldsymbol{R}_{a_1a_1}^{-1}\right)^{-1} \boldsymbol{S}_1^H \boldsymbol{R}_{nn}^{-1}\right]. \tag{5.116}$$

Equations (5.111) and (5.112) describe estimations of $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ in the models

$$\boldsymbol{r} = \boldsymbol{S}_1 \boldsymbol{a}_1 + \boldsymbol{n}_1, \tag{5.117}$$

$$\boldsymbol{r} = \boldsymbol{S}_2\boldsymbol{a}_2 + \boldsymbol{n}_2 \tag{5.118}$$

with

$$\begin{aligned} \boldsymbol{n}_1 &= \boldsymbol{S}_2\boldsymbol{a}_2 + \boldsymbol{n}, \\ \boldsymbol{n}_2 &= \boldsymbol{S}_1\boldsymbol{a}_1 + \boldsymbol{n}. \end{aligned} \tag{5.119}$$

Thus, each parameter to be estimated can be understood as noise in the estimation of the remaining parameters. An exception is given if $\boldsymbol{S}_1^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}_2 = \boldsymbol{0}$, which means that $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are orthogonal to each other with respect to the weighting matrix $\boldsymbol{R}_{nn}^{-1}$. Then we get

$$\boldsymbol{A}_1 = [\boldsymbol{S}_1^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}_1 + \boldsymbol{R}_{a_1a_1}^{-1}]^{-1}\boldsymbol{S}_1^H\boldsymbol{R}_{nn}^{-1}$$

and

$$\boldsymbol{R}_{e_1e_1} = [\boldsymbol{S}_1^H\boldsymbol{R}_{nn}^{-1}\boldsymbol{S}_1 + \boldsymbol{R}_{a_1a_1}^{-1}]^{-1},$$

and we observe that the second signal component $\boldsymbol{S}_2\boldsymbol{a}_2$ has no influence on the estimate.

Nonzero-Mean Processes. One could imagine that the precision of linear estimations with respect to nonzero-mean processes $\boldsymbol{r}$ and $\boldsymbol{a}$ can be increased compared to the solutions above if an additional term taking care of the mean values of the processes is considered. In order to describe this more general case, let us denote the mean of the parameters as

$$\bar{\boldsymbol{a}} = E\{\boldsymbol{a}\}. \tag{5.120}$$

The estimate is now written as

$$\hat{\boldsymbol{a}} = \boldsymbol{A}\boldsymbol{r} + \bar{\boldsymbol{a}} + \boldsymbol{c}, \tag{5.121}$$

where $\boldsymbol{c}$ is yet unknown. Using the shorthand

$$\begin{aligned} \boldsymbol{b} &= \boldsymbol{a} - \bar{\boldsymbol{a}}, \\ \hat{\boldsymbol{b}} &= \hat{\boldsymbol{a}} - \bar{\boldsymbol{a}}, \\ \boldsymbol{M} &= [\boldsymbol{c}, \boldsymbol{A}] \\ \boldsymbol{x} &= \begin{bmatrix} 1 \\ \boldsymbol{r} \end{bmatrix}, \end{aligned} \tag{5.122}$$

(5.121) can be rewritten as:

$$\hat{\boldsymbol{b}} = \boldsymbol{M}\boldsymbol{x}. \tag{5.123}$$

The relationship between $\hat{\boldsymbol{b}}$ and $\boldsymbol{x}$ is linear as usual, so that the optimal $\boldsymbol{M}$ can be given according to (5.89):

$$\boldsymbol{M} = \boldsymbol{R}_{xb}\boldsymbol{R}_{xx}^{-1}. \tag{5.124}$$

Now let us express $\boldsymbol{R}_{xb}$ and $\boldsymbol{R}_{xx}^{-1}$ through correlation matrices of the processes $\boldsymbol{a}$ and $\boldsymbol{r}$. From (5.122) and $E\{\boldsymbol{b}\} = \boldsymbol{0}$ we derive

$$\boldsymbol{R}_{xb} = [\boldsymbol{0}, \boldsymbol{R}_{rb}] \tag{5.125}$$

with

$$\begin{aligned} \boldsymbol{R}_{rb} &= E\left\{[\boldsymbol{a} - \bar{\boldsymbol{a}}]\, \boldsymbol{r}^H\right\} \\ &= E\left\{[\boldsymbol{a} - \bar{\boldsymbol{a}}]\, [\boldsymbol{r} - \bar{\boldsymbol{r}}]^H\right\}, \end{aligned} \tag{5.126}$$

where

$$\bar{\boldsymbol{r}} = E\{\boldsymbol{r}\}. \tag{5.127}$$

$\boldsymbol{R}_{xx}$ writes

$$\boldsymbol{R}_{xx} = \begin{bmatrix} 1 & \bar{\boldsymbol{r}}^H \\ \bar{\boldsymbol{r}} & \boldsymbol{R}_{rr} \end{bmatrix}. \tag{5.128}$$

Using (5.110) we obtain

$$\boldsymbol{R}_{xx}^{-1} = \begin{bmatrix} 1 + \bar{\boldsymbol{r}}^H \left[\boldsymbol{R}_{rr} - \bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^H\right]^{-1} \bar{\boldsymbol{r}} & -\bar{\boldsymbol{r}}^H \left[\boldsymbol{R}_{rr} - \bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^H\right]^{-1} \\ -\left[\boldsymbol{R}_{rr} - \bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^H\right]^{-1} \bar{\boldsymbol{r}} & \left[\boldsymbol{R}_{rr} - \bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^H\right]^{-1} \end{bmatrix}. \tag{5.129}$$

From (5.122) – (5.129) and

$$\left[\boldsymbol{R}_{rr} - \bar{\boldsymbol{r}}\bar{\boldsymbol{r}}^H\right] = E\left\{[\boldsymbol{r} - \bar{\boldsymbol{r}}]\, [\boldsymbol{r} - \bar{\boldsymbol{r}}]^H\right\} \tag{5.130}$$

we finally conclude

$$\hat{\boldsymbol{a}} = E\left\{[\boldsymbol{a} - \bar{\boldsymbol{a}}]\, [\boldsymbol{r} - \bar{\boldsymbol{r}}]^H\right\}\; E\left\{[\boldsymbol{r} - \bar{\boldsymbol{r}}]\, [\boldsymbol{r} - \bar{\boldsymbol{r}}]^H\right\}^{-1} [\boldsymbol{r} - \bar{\boldsymbol{r}}] + \bar{\boldsymbol{a}}. \tag{5.131}$$

Equation (5.131) can be interpreted as follows: the nonzero-mean processes $\boldsymbol{a}$ and $\boldsymbol{r}$ are first modified so as to become zero-mean processes $\boldsymbol{a} - \bar{\boldsymbol{a}}$ and $\boldsymbol{r} - \bar{\boldsymbol{r}}$. For the zero-mean processes the estimation problem can be solved as usual. Subsequently the mean value $\bar{\boldsymbol{a}}$ is added in order to obtain the final estimate $\hat{\boldsymbol{a}}$.

Unbiasedness for Random Parameter Vectors. So far the parameter vector to be estimated was assumed to be non-random. If we consider $\boldsymbol{a}$ to be a random process, various other weaker definitions of unbiasedness are possible.

The straightforward requirement

$$E\{\hat{\boldsymbol{a}}(\boldsymbol{r})\} = \boldsymbol{A}\, E\{\boldsymbol{r}\} \stackrel{!}{=} E\{\boldsymbol{a}\} \tag{5.132}$$

is meaningless, because it is satisfied for any $\boldsymbol{A}$ as far as $\boldsymbol{r}$ and $\boldsymbol{a}$ are zero-mean.

A useful definition of unbiasedness in the case of random parameters is to consider one of the parameters contained in $\boldsymbol{a}$ (e.g. a_k) as non-random and to regard all other parameters $a_1, \ldots, a_{k-1}, a_{k+1}$ etc. as random variables:

$$E\{\ \hat{a}_k(\boldsymbol{r}) \mid \mathrm{a}_k\ \} \stackrel{!}{=} \mathrm{a}_k. \tag{5.133}$$

In order to obtain an estimator which is unbiased in the sense of (5.133), the equivalences discussed above may be applied. Starting with the model

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{S}\boldsymbol{a} + \boldsymbol{n} \\ &= \boldsymbol{s}_k a_k + \tilde{\boldsymbol{n}}, \end{aligned} \tag{5.134}$$

in which $\tilde{\boldsymbol{n}}$ contains the additive noise $\boldsymbol{n}$ and the signal component produced by all random parameters a_j, $j \neq k$, we can write the unbiased estimate as

$$\hat{a}_k = \boldsymbol{h}_k^H \boldsymbol{r} \tag{5.135}$$

with

$$\boldsymbol{h}_k^H = \left[\boldsymbol{s}_k^H \boldsymbol{R}_{\tilde{n}\tilde{n}}^{-1} \boldsymbol{s}_k\right]^{-1} \boldsymbol{s}_k^H \boldsymbol{R}_{\tilde{n}\tilde{n}}^{-1}. \tag{5.136}$$

Then,

$$\boldsymbol{A} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots]^H \tag{5.137}$$

is an estimator which is unbiased in the sense of (5.133).

The Relationship between MMSE Estimation and the BLUE. If $\boldsymbol{R}_{aa} = E\{\boldsymbol{a}\boldsymbol{a}^H\}$ is unknown, $\boldsymbol{R}_{aa}^{-1} = \boldsymbol{0}$ is substituted into (5.102), and we obtain the BLUE (cf. (5.67)):

$$\boldsymbol{A} = \left[\boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}\right]^{-1} \boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1}. \tag{5.138}$$

In the previous discussion it became obvious that it is possible to obtain unbiased estimates of some of the parameters and to estimate the others with minimum mean square error. This result is of special interest if no unbiased estimator can be stated for all parameters because of a singular matrix $\boldsymbol{S}^H \boldsymbol{R}_{nn}^{-1} \boldsymbol{S}$.

5.6 Linear Optimal Filters

5.6.1 Wiener Filters

We consider the problem depicted in Figure 5.3. By linear filtering of the noisy signal $r(n) = x(n) + w(n)$ we wish to make $y(n) = r(n) * h(n)$ as similar as possible to a desired signal $d(n)$. The quality criterion used for designing the optimal causal linear filter $h(n)$ is

$$E\left\{|e(n)|^2\right\} = E\left\{|d(n) - y(n)|^2\right\} \stackrel{!}{=} \min. \tag{5.139}$$

The solution to this optimization problem can easily be stated by applying the orthogonality principle. Assuming a causal FIR filter $h(n)$ of length p, we have

$$y(n) = \sum_{i=0}^{p-1} h(i)\, r(n-i). \tag{5.140}$$

Thus, according to (5.94), the following orthogonality condition must be satisfied by the optimal filter:

$$E\left\{\left[d(n) - \sum_{i=0}^{p-1} h(i)\; r(n-i)\right]\; r^*(n-j)\right\} = 0, \qquad j = 0, 1, \ldots, p-1. \tag{5.141}$$

For stationary processes $r(n)$ and $d(n)$ this yields the discrete form of the so-called *Wiener–Hopf equation*:

$$\sum_{i=0}^{p-1} h(i)\; r_{rr}(j-i) = r_{rd}(j), \qquad j = 0, 1, \ldots, p-1, \tag{5.142}$$

with

$$\begin{aligned} r_{rr}(m) &= E\left\{r^*(n)\; r(n+m)\right\}, \\ r_{rd}(m) &= E\left\{r^*(n)\; d(n+m)\right\}. \end{aligned} \tag{5.143}$$

The optimal filter is found by solving (5.142).

An application example is the estimation of data $d(n)$ from a noisy observation $r(n) = \sum_\ell c(\ell)\; d(n-\ell) + w(n)$, where $c(n)$ is a channel and $w(n)$ is noise. By using the optimal filter $h(n)$ designed according to (5.142) the data is recovered with minimal mean square error.

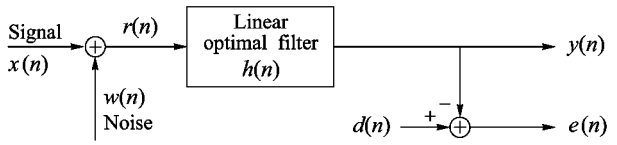

Figure 5.3. Designing linear optimal filters.

Variance. For the variance of the error we have

$$\begin{aligned}\sigma_e^2 &= E\left\{|e(n|^2\right\} \\ &= \sigma_d^2 - \sum_{i=0}^{p-1} h(i)\; r_{rd}^*(i) - \sum_{i=0}^{p-1} h^*(i)\; r_{rd}(i) \\ &+ \sum_{i=0}^{p-1}\sum_{j=0}^{p-1} h(i)\; h^*(j)\; r_{rr}(j-i)\end{aligned} \tag{5.144}$$

with $\sigma_d^2 = E\left\{|d(n|^2\right\}$. Substituting the optimal solution (5.142) into (5.144) yields

$$\sigma_{e_{\min}}^2 = \sigma_d^2 - \sum_{i=0}^{p-1} h(i)\; r_{rd}^*(i). \tag{5.145}$$

Matrix Notation. In matrix notation (5.142) is

$$\boldsymbol{R}_{rr}\; \boldsymbol{h} = \boldsymbol{r}_{rd} \tag{5.146}$$

with

$$\boldsymbol{h}^T = [h(0), h(1), \ldots, h(p-1)], \tag{5.147}$$

$$\boldsymbol{r}_{rd}^T = [r_{rd}(0), r_{rd}(1), \ldots, r_{rd}(p-1)] \tag{5.148}$$

and

$$\boldsymbol{R}_{rr} = \begin{bmatrix} r_{rr}(0) & r_{rr}(-1) & \ldots & r_{rr}(-p+1) \\ r_{rr}(1) & r_{rr}(0) & \ldots & r_{rr}(-p+2) \\ \vdots & \vdots & & \vdots \\ r_{rr}(p-1) & r_{rr}(p-1) & \ldots & r_{rr}(0) \end{bmatrix}. \tag{5.149}$$

From (5.146) and (5.145) we obtain the following alternative expressions for the minimal variance:

$$\begin{aligned} \sigma^2_{e_{\min}} &= \sigma_d^2 - \boldsymbol{r}_{rd}^H \, \boldsymbol{h} \\ &= \sigma_d^2 - \boldsymbol{r}_{rd}^H \, \boldsymbol{R}_{rr}^{-1} \, \boldsymbol{r}_{rd}. \end{aligned} \tag{5.150}$$

Special Cases. The following three cases, where the desired signal is a delayed version of the clean input signal $x(n)$, are of special interest:

(i) Filtering: $d(n) = x(n)$.

(ii) Interpolation: $d(n) = x(n + D), D < 0$.

(iii) Prediction: $d(n) = x(n + D), D > 0$. Here the goal is to predict a future value.

For the three cases mentioned above the Wiener–Hopf equation is

$$\sum_{i=0}^{p-1} h(i) \, r_{rr}(j - i) = r_{rx}(j + D), \qquad j = 0, 1, \ldots, p - 1. \tag{5.151}$$

Uncorrelated Noise. If the noise $w(n)$ is uncorrelated to $x(n)$, we have

$$r_{rr}(m) = r_{xx}(m) + r_{ww}(m) \tag{5.152}$$

and

$$r_{rd}(m) = r_{xx}(m + D), \tag{5.153}$$

and from (5.151) we derive

$$\sum_{i=0}^{p-1} h(i) \, [r_{xx}(j-i) + r_{ww}(j-i)] = r_{xx}(j+D), \qquad j = 0, 1, \ldots, p-1. \tag{5.154}$$

In matrix notation we get

$$[\boldsymbol{R}_{xx} + \boldsymbol{R}_{ww}] \, \boldsymbol{h} = \boldsymbol{r}_{xx}(D) \tag{5.155}$$

with

$$\boldsymbol{h}^T = [h(0), h(1), \ldots, h(p-1)], \tag{5.156}$$

$$\boldsymbol{r}_{xx}^T(D) = [r_{xx}(D), r_{xx}(D+1), \ldots, r_{xx}(D+p-1)] \tag{5.157}$$

and

$$\boldsymbol{R}_{xx} = \begin{bmatrix} r_{xx}(0) & r_{xx}(-1) & \dots & r_{xx}(-p+1) \\ r_{xx}(1) & r_{xx}(0) & \dots & r_{xx}(-p+2) \\ \vdots & \vdots & & \vdots \\ r_{xx}(p-1) & r_{xx}(p-1) & \dots & r_{xx}(0) \end{bmatrix}. \tag{5.158}$$

For the correlation matrix $\boldsymbol{R}_{ww}$ the corresponding definition holds.

The minimal variance is

$$\begin{aligned} \sigma^2_{e_{\min}} &= \sigma_d^2 - \boldsymbol{r}_{xx}^H(D)\, \boldsymbol{h} \\ &= \sigma_d^2 - \boldsymbol{r}_{xx}^H(D)\, [\boldsymbol{R}_{xx} + \boldsymbol{R}_{ww}]^{-1}\, \boldsymbol{r}_{xx}(D). \end{aligned} \tag{5.159}$$

5.6.2 One-Step Linear Prediction

One-step linear predictors are used in many applications such as speech and image coding (DPCM, ADPCM, LPC,...), in spectral estimation, and in feature extraction for speech recognition. Basically, they may be regarded as a special case of Wiener–Hopf filtering.

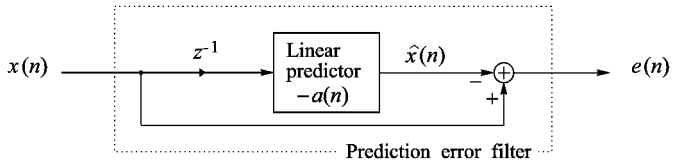

Figure 5.4. One-step linear prediction.

We consider the system in Figure 5.4. A comparison with Figure 5.3 shows that the optimal predictor can be obtained from the Wiener–Hopf equations for the special case $D = 1$ with $d(n) = x(n+1)$, while no additive noise is assumed, $w(n) = 0$. Note that the filter $a(n)$ is related to the Wiener–Hopf filter $h(n)$ as $a(n) = -h(n-1)$. With

$$\hat{x}(n) = -\sum_{i=1}^{p} a(i)\, x(n-i), \tag{5.160}$$

where p is the length of the FIR filter $a(n)$, the error becomes

$$\begin{aligned} e(n) &= x(n) - \hat{x}(n) \\ &= x(n) + \sum_{i=1}^{p} a(i)\, x(n-i). \end{aligned} \tag{5.161}$$

Minimizing the error with respect to the filter coefficients yields the equations

$$-\sum_{i=1}^{p} a(i)\; r_{xx}(j-i) = r_{xx}(j), \qquad j = 1, 2, \ldots, p, \tag{5.162}$$

which are known as the *normal equations of linear prediction.* In matrix notation they are

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(-1) & \ldots & r_{xx}(-p+1) \\ r_{xx}(1) & r_{xx}(0) & \ldots & r_{xx}(-p+2) \\ \vdots & \vdots & & \vdots \\ r_{xx}(p-1) & r_{xx}(p-2) & \ldots & r_{xx}(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{bmatrix} = - \begin{bmatrix} r_{xx}(1) \\ r_{xx}(2) \\ \vdots \\ r_{xx}(p) \end{bmatrix}, \tag{5.163}$$

that is

$$\boldsymbol{R}_{xx}\boldsymbol{a} = -\boldsymbol{r}_{xx}(1) \tag{5.164}$$

with

$$\boldsymbol{a}^T = [a(1), \ldots, a(p)]\,. \tag{5.165}$$

According to (5.159) we get for the minimal variance:

$$\begin{aligned} \sigma^2_{e_{\min}} = E\left\{|e(n)|^2\right\} &= r_{xx}(0) - \boldsymbol{r}^H_{xx}(1)\; \boldsymbol{R}^{-1}_{xx}\; \boldsymbol{r}_{xx}(1) \\ &= r_{xx}(0) + \boldsymbol{r}^H_{xx}(1)\; \boldsymbol{a}. \end{aligned} \tag{5.166}$$

Autoregressive Processes and the Yule–Walker Equations. We consider an autoregressive process of order p (AR(p) process). As outlined in Section 5.3, such a process is generated by exciting a stable recursive filter with a stationary white noise process $w(n)$. The system function of the recursive system is supposed to be[2]

$$U(z) = \frac{1}{1 + \sum\limits_{i=1}^{p} a(i)\; z^{-i}}, \qquad a(p) \neq 0. \tag{5.167}$$

The input-output relation of the recursive system may be expressed via the difference equation

$$x(n) = w(n) - \sum_{i=1}^{p} a(i)\; x(n-i). \tag{5.168}$$

[2]In order to keep in line with the notation used in the literature, the coefficients $\rho(i)$, $i = 1, \ldots, p$ introduced in (5.34) are replaced by the coefficients $-a(i)$, $i = 1, \ldots, p$.

For the autocorrelation sequence of the process $x(n)$ we thus derive

$$\begin{aligned} r_{xx}(m) &= E\{x^*(n)x(n+m)\} \\ &= r_{xw}(m) - \sum_{i=1}^{p} a(i)\, r_{xx}(m-i). \end{aligned} \tag{5.169}$$

The cross correlation sequence $r_{xw}(m)$ is

$$\begin{aligned} r_{xw}(m) &= E\{x^*(n)w(n+m)\} \\ &= \sum_{i=1}^{\infty} u^*(i)\, \underbrace{r_{ww}(i+m)}_{\sigma_w^2\,\delta(i+m)} \\ &= \sigma_w^2\, u^*(-m), \end{aligned} \tag{5.170}$$

where $u(n)$ is the impulse response of the recursive filter. Since $u(n)$ is causal ($u(n) = 0$ for $n < 0$), we derive

$$r_{xw}(m) = \begin{cases} 0, & m < 0, \\ \sigma_w^2\, u^*(-m), & m \geq 0. \end{cases} \tag{5.171}$$

By combining (5.169) and (5.171) we finally get

$$r_{xx}(m) = \begin{cases} -\sum_{i=1}^{p} a(i)\, r_{xx}(m-i), & m > 0, \\ \sigma_w^2 - \sum_{i=1}^{p} a(i)\, r_{xx}(m-i), & m = 0, \\ r_{xx}^*(-m), & m < 0. \end{cases} \tag{5.172}$$

The equations (5.172) are known as the *Yule–Walker equations*. In matrix form they are

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(-1) & r_{xx}(-2) & \dots & r_{xx}(-p) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(-1) & \dots & r_{xx}(1-p) \\ \vdots & \vdots & \vdots & & \vdots \\ r_{xx}(p) & r_{xx}(p-1) & r_{xx}(p-1) & \dots & r_{xx}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a(1) \\ \vdots \\ a(p) \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.173}$$

As can be inferred from (5.173), we obtain the coefficients $a(i)$, $i = 1, \dots, p$ by solving (5.163). By observing the power of the prediction error we can also determine the power of the input process. From (5.166) and (5.172) we have

$$\begin{aligned} \sigma_w^2 &= \sigma_{e_{\min}}^2 \\ &= r_{xx}(0) + \boldsymbol{r}_{xx}^H(1)\, \boldsymbol{a}. \end{aligned} \tag{5.174}$$

Thus, all parameters of an autoregressive process can be exactly determined from the parameters of a one-step linear predictor.

Prediction Error Filter. The output signal of the so-called *prediction error filter* is the signal $e(n)$ in Figure 5.4 with the coefficients $a(n)$ according to (5.163). Introducing the coefficient $a(0) = 1$, $e(n)$ is given by

$$e(n) = \sum_{i=0}^{p} a(i)\, x(n-i), \qquad a(0) = 1. \tag{5.175}$$

The system function of the prediction error filter is

$$A(z) = 1 + \sum_{i=1}^{p} a(i) z^{-i} = \sum_{i=0}^{p} a(i) z^{-i}, \qquad a(0) = 1. \tag{5.176}$$

In the special case that $x(n)$ is an autoregressive process, the prediction error filter $A(z)$ is the inverse system to the recursive filter $U(z) \longleftrightarrow u(n)$. This also means that the output signal of the prediction error filter is a white noise process. Hence, the prediction error filter performs a whitening transform and thus constitutes an alternative to the methods considered in Section 5.4. If $x(n)$ is not truly autoregressive, the whitening transform is carried out at least approximately.

Minimum Phase Property of the Prediction Error Filter. Our investigation of autoregressive processes showed that the prediction error filter $A(z)$ is inverse to the recursive filter $U(z)$. Since a stable filter does not have poles outside the unit circle of the z-plane, the corresponding prediction error filter cannot have zeros outside the unit circle. Even if $x(n)$ is not an autoregressive process, we obtain a minimum phase prediction error filter, because the calculation of $A(z)$ only takes into account the second-order statistics, which do not contain any phase information, cf. (1.105).

5.6.3 Filter Design on the Basis of Finite Data Ensembles

In the previous sections we assumed stationary processes and considered the correlation sequences to be known. In practice, however, linear predictors must be designed on the basis of a finite number of observations.

In order to determine the predictor filter $a(n)$ from measured data $x(1), x(2), \ldots, x(N)$, we now describe the prediction error

$$e(n) = x(n) + \sum_{i=1}^{p} a(i) x(n-i)$$

via the following matrix equation:

$$\boldsymbol{e} = \boldsymbol{X}\ \boldsymbol{a} + \boldsymbol{x}, \tag{5.177}$$

where $\boldsymbol{a}$ contains the predictor coefficients, and $\boldsymbol{X}$ and $\boldsymbol{x}$ contain the input data. The term $\boldsymbol{X}\,\boldsymbol{a}$ describes the convolution of the data with the impulse response $a(n)$.

The criterion

$$\|\boldsymbol{e}\| = \|\boldsymbol{X}\ \boldsymbol{a} + \boldsymbol{x}\| \stackrel{!}{=} \min \tag{5.178}$$

leads to the following normal equation:

$$\boldsymbol{X}^H \boldsymbol{X}\ \boldsymbol{a} = -\boldsymbol{X}^H \boldsymbol{x}. \tag{5.179}$$

Here, the properties of the predictor are dependent on the definition of $\boldsymbol{X}$ and $\boldsymbol{x}$. In the following, two relevant methods will be discussed.

Autocorrelation Method. The *autocorrelation method* is based on the following estimation of the autocorrelation sequence:

$$\hat{r}_{xx}^{(AC)}(m) = \frac{1}{N} \sum_{n=1}^{N-|m|} x^*(n)\ x(n+m). \tag{5.180}$$

As can be seen, $\hat{r}_{xx}^{(AC)}(m)$ is a biased estimate of the true autocorrelation sequence $r_{xx}(m)$, which means that $E\{\hat{r}_{xx}^{(AC)}(m)\} \neq r_{xx}(m)$. Thus, the autocorrelation method yields a biased estimate of the parameters of an autoregressive process. However, the correlation matrix $\hat{\boldsymbol{R}}_{xx}^{(AC)}$ built from $\hat{r}_{xx}^{(AC)}(m)$ has a Toeplitz structure, which enables us to efficiently solve the equation

$$\hat{\boldsymbol{R}}_{xx}^{(AC)}\ \hat{\boldsymbol{a}} = -\hat{\boldsymbol{r}}_{xx}^{(AC)}(1) \tag{5.181}$$

by means of the *Levinson–Durbin recursion* [89, 47] or the *Schur algorithm* [130]. Textbooks that cover this topic are, for instance, [84, 99, 117].

The autocorrelation method can also be viewed as the solution to the problem (5.178) with

$$\boldsymbol{x} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x(N) \\ \vdots \\ x(p+1) \\ \vdots \\ x(2) \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} & & x(N) \\ & \iddots & \vdots \\ x(N) & \dots & x(N-p+1) \\ x(N-1) & \dots & x(N-p) \\ \vdots & & \vdots \\ x(p) & \dots & x(1) \\ \vdots & \iddots & \\ x(1) & & \end{bmatrix} \tag{5.182}$$

and

$$\boldsymbol{a} = [a(1), a(2), \ldots, a(p)]^T. \tag{5.183}$$

We have

$$\hat{\boldsymbol{R}}_{xx}^{(AC)} = \boldsymbol{X}^H \boldsymbol{X} \tag{5.184}$$

and

$$\hat{\boldsymbol{r}}_{xx}^{(AC)}(1) = \boldsymbol{X}^H \boldsymbol{x}. \tag{5.185}$$

Covariance Method. The *covariance method* takes into account the prediction errors in steady state only and yields an unbiased estimate of the autocorrelation matrix. In this case $\boldsymbol{X}$ and $\boldsymbol{x}$ are defined as

$$\boldsymbol{X} = \begin{bmatrix} x(N-1) & \ldots & x(N-p) \\ \vdots & & \vdots \\ x(p+1) & \ldots & x(2) \\ x(p) & \ldots & x(1) \end{bmatrix} \tag{5.186}$$

and

$$\boldsymbol{x} = \begin{bmatrix} x(N) \\ \vdots \\ x(p+2) \\ x(p+1) \end{bmatrix}. \tag{5.187}$$

The equation to be solved is

$$\hat{\boldsymbol{R}}_{xx}^{(CV)} \, \hat{\boldsymbol{a}} = -\hat{\boldsymbol{r}}_{xx}^{(CV)}(1), \tag{5.188}$$

where

$$\hat{\boldsymbol{R}}_{xx}^{(CV)} = \boldsymbol{X}^H \boldsymbol{X}, \tag{5.189}$$

$$\hat{\boldsymbol{r}}_{xx}^{(CV)}(1) = \boldsymbol{X}^H \boldsymbol{x}. \tag{5.190}$$

Note that $\hat{\boldsymbol{R}}_{xx}^{(CV)}$ is not a Toeplitz matrix, so that solving (5.188) is much more complex than solving (5.181) via the Levinson–Durbin recursion. However, the covariance method has the advantage of being unbiased; we have

$$E\left\{\hat{\boldsymbol{R}}_{xx}^{(CV)}\right\} = \boldsymbol{R}_{xx}. \tag{5.191}$$

5.7 Estimation of Autocorrelation Sequences and Power Spectral Densities

5.7.1 Estimation of Autocorrelation Sequences

In the following, we will discuss methods for estimating the autocorrelation sequence of random processes from given sample values $x(n)$, $n = 0, \ldots, N-1$. We start the discussion with the estimate

$$\hat{r}_{xx}^{b}(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x^*(n)\, x(n+m), \tag{5.192}$$

which is the same as the estimate $\hat{r}_{xx}^{(AC)}(m)$, used in the autocorrelation method explained in the last section. As can easily be verified, the estimate $\hat{r}_{xx}^{b}(m)$ is biased with mean

$$E\left\{\hat{r}_{xx}^{b}(m)\right\} = \frac{N-|m|}{N}\, r_{xx}(m). \tag{5.193}$$

However, since

$$\lim_{N\to\infty} E\left\{\hat{r}_{xx}^{b}(m)\right\} = r_{xx}(m), \tag{5.194}$$

the estimate is asymptotically unbiased. The triangular window $\frac{N-|m|}{N}$ that occurs in (5.193) is known as the *Bartlett window.*

The variance of the estimate can be approximated as [77]

$$\mathrm{var}[\hat{r}_{xx}^{b}(m)] \approx \frac{1}{N} \sum_{n=-\infty}^{\infty} |r_{xx}(n)|^2 + r_{xx}^*(n-m)\, r_{xx}(n+m). \tag{5.195}$$

Thus, as $N \to \infty$, the variance tends to zero:

$$\lim_{N\to\infty} \mathrm{var}[\hat{r}_{xx}^{b}(m)] \to 0. \tag{5.196}$$

Such an estimate is said to be *consistent.* However, although consistency is given, we cannot expect good estimates for large m as long as N is finite, because the bias increases as $|m| \to N$.

Unbiased Estimate. An unbiased estimate of the autocorrelation sequence is given by

$$\hat{r}_{xx}^{u}(m) = \frac{1}{N-|m|} \sum_{n=0}^{N-|m|-1} x^*(n)\, x(n+m). \tag{5.197}$$

We have

$$E\left\{\hat{r}_{xx}^{u}(m)\right\} = r_{xx}(m), \quad |m| \le N-1. \tag{5.198}$$

The variance of the estimate can be approximated as [77]

$$\mathrm{var}[\hat{r}^b_{xx}(m)] \approx \frac{N}{(N-|m|)^2} \sum_{n=-\infty}^{\infty} |r_{xx}(n)|^2 + r^*_{xx}(n-m)\, r_{xx}(n+m). \quad (5.199)$$

As $N \to \infty$, this gives

$$\lim_{N\to\infty} \mathrm{var}[\hat{r}^u_{xx}(m)] \to 0, \quad (5.200)$$

which means that $\hat{r}^u_{xx}(m)$ is a consistent estimate. However, problems arise for large m as long as N is finite, because the variance increases for $|m| \to N$.

5.7.2 Non-Parametric Estimation of Power Spectral Densities

In many real-world problems one is interested in knowledge about the power spectral density of the data to be processed. Typically, only a finite set of observations $x(n)$ with $n = 0, 1, \ldots, N-1$ is available. Since the power spectral density is the Fourier transform of the autocorrelation sequence, and since we have methods for the estimation of the autocorrelation sequence, it is a logical consequence to look at the Fourier transforms of these estimates. We start with $\hat{r}^b_{xx}(m)$. The Fourier transform of $\hat{r}^b_{xx}(m)$ will be denoted as

$$P_{xx}(e^{j\omega}) = \sum_{m=-(N-1)}^{N-1} \hat{r}^b_{xx}(m)\, e^{-j\omega m}. \quad (5.201)$$

We know that $\hat{r}^b_{xx}(m)$ is a biased estimate of the true autocorrelation sequence $r_{xx}(m)$, so that we can conclude that the spectrum $P_{xx}(e^{j\omega})$ is a biased estimate of the true power spectral density $S_{xx}(e^{j\omega})$. In order to be explicit, let us recall that

$$E\left\{\hat{r}^b_{xx}(m)\right\} = w_B(m)\, r_{xx}(m), \quad (5.202)$$

with $w_B(m)$ being the Bartlett window; i.e.

$$w_B(m) = \frac{N-|m|}{N}. \quad (5.203)$$

In the spectral domain, we have

$$\begin{aligned} E\left\{P_{xx}(e^{j\omega})\right\} &= \sum_{m=-(N-1)}^{N-1} E\left\{\hat{r}_{xx}^{b}(m)\right\}\, e^{-j\omega m} \\ &= \sum_{m=-(N-1)}^{N-1} w_B(m)\, r_{xx}(m)\, e^{-j\omega m} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{xx}(e^{j\nu})\, W_B(e^{j(\omega-\nu)})\, d\nu, \end{aligned} \tag{5.204}$$

where $W_B(e^{j\omega})$ is the Fourier transform of $w_B(m)$ given by

$$W_B(e^{j(\omega)}) = \frac{1}{N}\left(\frac{\sin(\omega N/2)}{\sin(\omega/2)}\right)^2. \tag{5.205}$$

Thus, $E\left\{P_{xx}(e^{j\omega})\right\}$ is a smoothed version of the true power spectral density $S_{xx}(e^{j\omega})$, where smoothing is carried out with the Fourier transform of the Bartlett window.

A second way of computing $P_{xx}(e^{j\omega})$ is to compute the Fourier transform of $x(n)$ first and to derive $P_{xx}(e^{j\omega})$ from $X(e^{j\omega})$. By inserting (5.192) into (5.201) and rearranging the expression obtained, we get

$$P_{xx}(e^{j\omega}) = \frac{1}{N}\,|X(e^{j\omega})|^2. \tag{5.206}$$

In the form (5.206) $P_{xx}(e^{j\omega})$ is known as the *periodogram.*

Another way of deriving an estimate of the power spectral density is to consider the Fourier transform of the estimate $\hat{r}_{xx}^{u}(m)$. We use the notation $Q_{xx}(e^{j\omega})$ for this type of estimate:

$$Q_{xx}(e^{j\omega}) = \sum_{m=-(N-1)}^{N-1} \hat{r}_{xx}^{u}(m)\, e^{-j\omega m}. \tag{5.207}$$

The expected value is

$$\begin{aligned} E\left\{Q_{xx}(e^{j\omega})\right\} &= \sum_{m=-(N-1)}^{N-1} E\left\{\hat{r}_{xx}^{u}(m)\right\}\, e^{-j\omega m} \\ &= \sum_{m=-(N-1)}^{N-1} r_{xx}(m)\, e^{-j\omega m} \\ &= \sum_{m=-\infty}^{\infty} w_R(m)\, r_{xx}(m)\, e^{-j\omega m} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{xx}(e^{j\nu})\, W_R(e^{j(\omega-\nu)})\, d\nu, \end{aligned} \tag{5.208}$$

where $w_R(m)$ is the rectangular window

$$w_R(m) = \begin{cases} 1, & \text{for } |m| \le N-1 \\ 0, & \text{otherwise,} \end{cases} \tag{5.209}$$

and $W_R(e^{j\omega})$ is its Fourier transform:

$$W_R(e^{j(\omega)}) = \frac{\sin(\omega[2N-1]/2)}{\sin(\omega/2)}. \tag{5.210}$$

This means that although $\hat{r}_{xx}^{u}(m)$ is an unbiased estimate of $r_{xx}(m)$, the quantity $Q_{xx}(e^{j\omega})$ is a biased estimate of $S_{xx}(e^{j\omega})$. The reason for this is the fact that only a finite number of taps of the autocorrelation sequence is used in the computation of $Q_{xx}(e^{j\omega})$. The mean $E\left\{Q_{xx}(e^{j\omega})\right\}$ is a smoothed version of $S_{xx}(e^{j\omega})$, where smoothing is carried out with the Fourier transform of the rectangular window.

As $N \to \infty$ both estimates $\hat{r}_{xx}^{b}(m)$ and $\hat{r}_{xx}^{u}(m)$ become unbiased. The same holds for $P_{xx}(e^{j\omega})$ and $Q_{xx}(e^{j\omega})$, so that both estimates of the power spectral density are asymptotically unbiased. The behavior of the variance of the estimates is different. While the estimates of the autocorrelation sequences are consistent, those of the power spectral density are not. For example, for a Gaussian process $x(n)$ with power spectral density $S_{xx}(e^{j\omega})$, the variance of the periodogram becomes

$$\operatorname{var}\left[P_{xx}(e^{j\omega})\right] = \left[1 + \left(\frac{\sin(\omega N)}{N\sin\omega}\right)^2\right] S_{xx}^2(e^{j\omega}), \tag{5.211}$$

which yields

$$\lim_{N\to\infty} \operatorname{var}\left[P_{xx}(e^{j\omega})\right] = S_{xx}^2(e^{j\omega}). \tag{5.212}$$

Thus, the periodogram does not give a consistent estimate of $S_{xx}(e^{j\omega})$. The proof of (5.211) is straightforward and is omitted here.

Use of the DFT or FFT for Computing the Periodogram. Since the periodogram is computed from the Fourier transform of the finite data sequence, it can be efficiently evaluated at a discrete set of frequencies by using the FFT. Given a length-N sequence $x(n)$, we may consider a length-N DFT, resulting in

$$P_{xx}(e^{j\omega_k}) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi k/N} \right|^2 \tag{5.213}$$

with $\omega_k = 2\pi k/N$. In many applications, the obtained number of samples of $P_{xx}(e^{j\omega})$ may be insufficient in order to draw a clear picture of the periodogram. Moreover, the DFT length may be inconvenient for computation, because no powerful FFT algorithm is at hand for the given length. These problems can be solved by extending the sequence $x(n)$ with zeros to an arbitrary length $N' \geq N$. This procedure is known as *zero padding*. We obtain

$$P_{xx}(e^{j\omega_k}) = \frac{1}{N'} \left| \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi k/N'} \right|^2 \tag{5.214}$$

with $\omega_k = 2\pi k/N'$. The evaluation of (5.214) is typically carried out via the FFT.

Bartlett Method. Various methods have been proposed for achieving consistent estimates of the power spectral density. The *Bartlett method* does this by decomposing the sequence $x(n)$ into disjoint segments of smaller length and taking the ensemble average of the spectrum estimates derived from the smaller segments. With

$$x^{(i)}(n) = x(n + iM), \quad i = 0, 1, \ldots, K-1, \quad n = 0, 1, \ldots, M-1, \tag{5.215}$$

we get the K periodograms

$$P_{xx}^{(i)}(e^{j\omega}) = \frac{1}{M} \left| \sum_{n=0}^{M-1} x^{(i)}(n)\, e^{-j\omega n} \right|^2, \quad i = 0, 1, \ldots, K-1. \tag{5.216}$$

The Bartlett estimate then is

$$P_{xx}^{B}(e^{j\omega}) = \frac{1}{K} \sum_{i=0}^{K-1} P_{xx}^{(i)}(e^{j\omega}). \tag{5.217}$$

The expected value becomes

$$E\left\{P_{xx}^{B}(e^{j\omega})\right\} = E\left\{P_{xx}^{(i)}(e^{j\omega})\right\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{xx}(e^{j\nu})\, W_{B_M}(e^{j(\omega-\nu)})\, d\nu, \tag{5.218}$$

with $W_{B_M}(e^{j\omega})$ being the Fourier transform of the length-M Bartlett window. Assuming a Gaussian process $x(n)$, the variance becomes

$$\text{var}\,[\,P_{xx}^{B}(e^{j\omega})\,] = \frac{1}{K}\,\text{var}\,[\,P_{xx}(e^{j\omega})\,] = \frac{1}{K}\left[1 + \left(\frac{\sin(\omega M)}{M\,\sin\omega}\right)^2\right]\,S_{xx}^2(e^{j\omega}). \tag{5.219}$$

Thus, as $N, M, K \to \infty$, the variance tends to zero and the estimate is consistent. For finite N, the decomposition of $x(n)$ into K sets results in a reduced variance, but the bias increases accordingly and the spectrum resolution decreases.

Blackman-Tukey Method. Blackman and Tukey proposed windowing the estimated autocorrelation sequence prior the Fourier transform [8]. The argument is that windowing allows us to reduce the influence of the unreliable estimates of the autocorrelation sequence for large m. Denoting the window and its Fourier transform as $w(m)$ and $W(e^{j\omega})$, respectively, the estimate can be written as

$$P_{xx}^{BT}(e^{j\omega}) = \sum_{m=-(N-1)}^{N-1} w(m)\,\hat{r}_{xx}^{b}(m)\;e^{-j\omega m}. \tag{5.220}$$

In the frequency domain, this means that

$$P_{xx}^{BT}(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{xx}(e^{j\nu})\,W(e^{j(\omega-\nu)})\;d\nu. \tag{5.221}$$

The window $w(n)$ should be chosen such that

$$W(e^{j\omega}) \;>\; 0 \;\;\forall\,\omega \tag{5.222}$$

in order to ensure that $P_{xx}^{BT}(e^{j\omega})$ is positive for all frequencies.

The expected value of $P_{xx}^{BT}(e^{j\omega})$ is most easily expressed in the form

$$E\left\{P_{xx}^{BT}(e^{j\omega})\right\} = \sum_{m=-(N-1)}^{N-1} w(m)\,w_B(m)\,r_{xx}(m)\;e^{-j\omega m}. \tag{5.223}$$

Provided that $w(m)$ is wide with respect to $r_{xx}(m)$ and narrow with respect to $w_B(m)$, the expected value can be approximated as

$$E\left\{P_{xx}^{BT}(e^{j\omega})\right\} = w(0)\;S_{xx}(e^{j\omega}). \tag{5.224}$$

Thus, in order to achieve an asymptotically unbiased estimate, the window should satisfy

$$w(0) = 1. \tag{5.225}$$

For a symmetric window $w(m) = w(-m)$ the variance can be estimated as [8]

$$\operatorname{var}[P_{xx}^{BT}(e^{j\omega})] \approx \frac{1}{N}\left[\sum_{m=-(N-1)}^{N-1} w^2(n)\right] S_{xx}^2(e^{j\omega}). \tag{5.226}$$

This approximation is based on the assumption that $W(e^{j\omega})$ is wide with respect to $W_B(e^{j\omega})$ and narrow with respect to the variations of $S_{xx}(e^{j\omega})$.

Welch Method. In the *Welch method* [162] the data is divided into overlapping blocks

$$x^{(i)}(n) = x(n + iD), \quad i = 0, 1, \ldots, K-1, \quad n = 0, 1, \ldots, M-1 \tag{5.227}$$

with $D \leq M$. For $D = M$ we approach the decomposition in the Bartlett method. For $D < M$ we have more segments than in the Bartlett method.

Each block is windowed prior to computation of the periodogram, resulting in K spectral estimates

$$V_{xx}^{(i)}(e^{j\omega}) = \frac{1}{\alpha M}\left|\sum_{n=0}^{M-1} x^{(i)}(n)\, w(n)\, e^{-j\omega n}\right|^2, \quad i = 0, 1, \ldots, K-1. \tag{5.228}$$

The factor α is chosen as

$$\alpha = \frac{1}{M}\sum_{m=0}^{M-1} w^2(m) = \frac{1}{M}\,\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{ww}^E(e^{j\omega})\, d\omega, \tag{5.229}$$

which means that the analysis is carried out with a window of normalized energy. Taking the average yields the final estimate

$$P_{xx}^W(e^{j\omega}) = \frac{1}{K}\sum_{i=0}^{K-1} V_{xx}^{(i)}(e^{j\omega}). \tag{5.230}$$

The expected value becomes

$$E\left\{P_{xx}^W(e^{j\omega})\right\} = E\left\{V_{xx}^{(i)}(e^{j\omega})\right\} \tag{5.231}$$

with

$$E\left\{V_{xx}^{(i)}(e^{j\omega})\right\} = \frac{1}{\alpha M}\sum_{n=0}^{M-1}\sum_{m=0}^{M-1} w(n)\, w(m)\, r_{xx}(n-m)\, e^{-j\omega(n-m)}. \tag{5.232}$$

In the spectral domain, this can be rewritten as

$$E\left\{V_{xx}^{(i)}(e^{j\omega})\right\} = \frac{1}{\alpha M}\,\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{xx}(e^{j\nu})\, S_{ww}^E(e^{j(\omega-\nu)})\, d\nu, \tag{5.233}$$

where

$$S_{ww}^{E}(e^{j\omega}) = \left|W(e^{j\omega})\right|^2 . \tag{5.234}$$

With increasing N and M, $S_{ww}^{E}(e^{j(\omega-\nu)})$ becomes narrow with respect to $S_{xx}(e^{j\nu})$ and the expected value tends to

$$E\left\{V_{xx}^{(i)}(e^{j\omega})\right\} = \frac{1}{\alpha M}\, S_{xx}(e^{j\omega})\, \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{ww}^{E}(e^{j\omega})\, d\omega = S_{xx}(e^{j\omega}). \tag{5.235}$$

This shows that the Welch method is asymptotically unbiased.

For a Gaussian process, the variance of the estimate is

$$\operatorname{var}\left[P_{xx}^{W}(e^{j\omega})\right] = \frac{1}{K^2} \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} E\left\{V_{xx}^{(i)}(e^{j\omega})\, V_{xx}^{(j)}(e^{j\omega})\right\} - \left[E\left\{V_{xx}^{(i)}(e^{j\omega})\right\}\right]^2 . \tag{5.236}$$

If no overlap is considered ($D = M$), the expression reduces to

$$\operatorname{var}\left[P_{xx}^{W}(e^{j\omega})\right] = \frac{1}{K} \operatorname{var}\left[V_{xx}^{(i)}(e^{j\omega})\right] \approx \frac{1}{K}\, S_{xx}^{2}(e^{j\omega}). \tag{5.237}$$

For $k \to \infty$ the variance approaches zero, which shows that the Welch method is consistent.

Various windows with different properties are known for the purpose of spectral estimation. In the following, a brief overview is given.

Hanning Window.

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), & n = 0, 1, \ldots, N-1 \\ 0, & \text{otherwise.} \end{cases} \tag{5.238}$$

Hamming Window.

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & n = 0, 1, \ldots, N-1 \\ 0, & \text{otherwise.} \end{cases} \tag{5.239}$$

Blackman Window.

$$w(n) = \begin{cases} 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right), & n = 0, 1, \ldots, N-1 \\ 0, & \text{otherwise.} \end{cases} \tag{5.240}$$

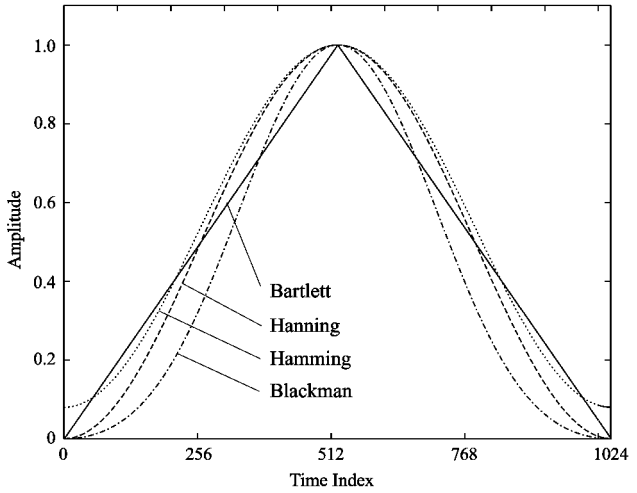

Figure 5.5. Window functions.

Figure 5.5 shows the windows, and Figure 5.6 shows their magnitude frequency responses. The spectrum of the Bartlett window is positive for all frequencies, which also means that the bias due to the Bartlett window is strictly positive. The spectra of the Hanning and Hamming window have relatively large negative side lobes, so that the estimated power spectral density may have a negative bias in the vicinity of large peaks in $S_{xx}(e^{j\omega})$. The Blackman window is a compromise between the Bartlett and the Hanning/Hamming approaches.

5.7.3 Parametric Methods in Spectral Estimation

Parametric methods in spectral estimation have been the subject of intensive research, and many different methods have been proposed. We will consider the simplest case only, which is related to the Yule–Walker equations. A comprehensive treatment of this subject would go far beyond the scope of this section.

Recall that in Section 5.6.2 we showed that the coefficients of a linear one-step predictor are identical to the parameters describing an autoregressive process. Hence the power spectral density may be estimated as

$$\hat{S}_{xx}(e^{j\omega}) = \frac{\hat{\sigma}_w^2}{\left|1 + \sum_{n=1}^{p} \hat{a}(n) e^{-j\omega n}\right|^2}. \tag{5.241}$$

The coefficients $\hat{a}(n)$ in (5.241) are the predictor coefficients determined from

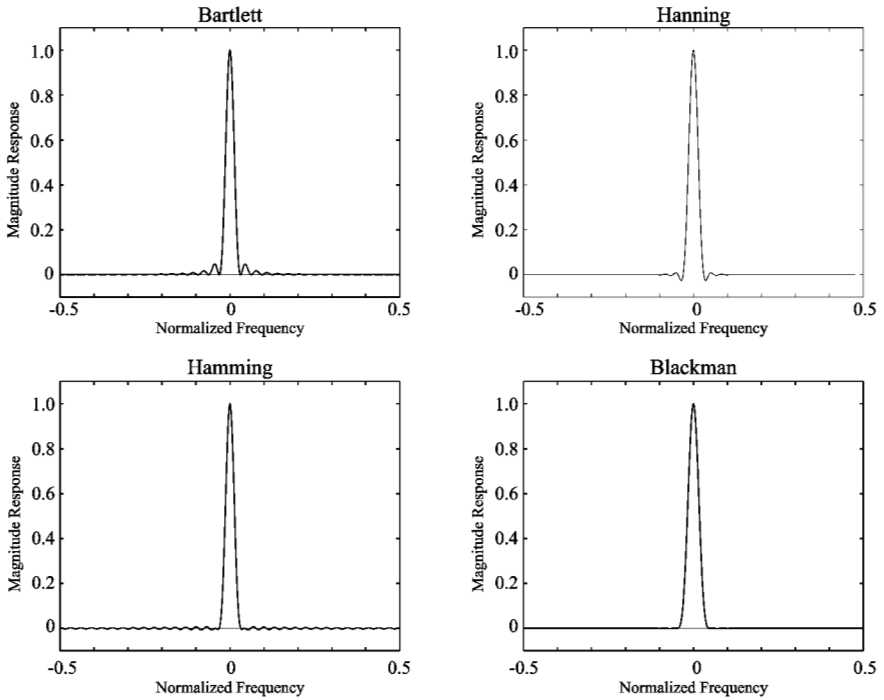

Figure 5.6. Magnitude frequency responses of common window functions.

the observed data, and $\hat{\sigma}_w^2$ is the power of the white input process estimated according to (5.174):

$$\hat{\sigma}_w^2 = \hat{r}_{xx}(0) + \hat{\boldsymbol{r}}_{xx}^H(1)\ \hat{\boldsymbol{a}}. \tag{5.242}$$

If we apply the autocorrelation method to the estimation of the predictor coefficients $\hat{a}(n)$, the estimated autocorrelation matrix has a Toeplitz structure, and the prediction filter is always minimum phase, just as when using the true correlation matrix $\boldsymbol{R}_{xx}$. For the covariance method this is not the case.

Finally, it shall be remarked that besides a forward prediction a backward prediction may also be carried out. By combining both predictors one can obtain an improved estimation of the power spectral density compared to (5.241). An example is the *Burg method* [19].