Institut EURECOM
2229 route des Crêtes
B.P. 193
06904 Sophia Antipolis Cedex
FRANCE

Signal Modeling and Coding (TdS1)

Dirk T.M. Slock

Spring 2003

Telephone: +33 (0)4 93 00 26 26
Fax: +33 (0)4 93 00 26 27
Dirk Slock: +33 (0)4 93 00 26 06

E-mail: slock@eurecom.fr
www: http://www.eurecom.fr
http://wwwlab.eurecom.fr:8080/CM/Courses/TDS1/

Contents

Preface

The title of this course, Signal Modeling and Coding, is in fact a shortcut, covering quite a variety of subjects. A glance at the description of the different chapters will make this point clear.

- Chapter 0: **Background Material**

  See also the "Random Processes" course and a Linear Algebra review at http://wwwlab.eurecom.fr:8080/CM/Courses/TDS1/ .

  - probability
  - linear algebra
  - multivariate Gaussian distribution, Gauss-Markov theorem
  - complex Gaussian variables, circularity
  - vector spaces, inner products, norms
  - optimization

- Chapter 1: **Least-Squares Parameter and Signal Estimation**

  - problem formulation and solution
  - linear model, parameters, "signal" and "noise" components
  - applications of the linear model
  - interpretations of the LS solution, orthogonality principle, a priori and a posterior models, signal and noise subspaces
  - performance analysis: bias, MSE, consistency
  - model order reduction, optimal order

- Chapter 2: **Spectral Estimation**

  - non-parametric techniques, periodogram, windowing, spectral leakage, averaged periodogram, smoothed periodogram
  - parametric random process models, autoregressive (AR) processes, moving average (MA) processes, autoregressive moving average (ARMA) processes
  - parametric techniques, autoregressive processes, linear prediction, maximum entropy, Levinson algorithm, autoregressive modeling, lattice filters

  - time-frequency representations, short-time Fourier transform, QMF, subbands, perfect reconstruction filter banks, wavelet transform, hierarchical signal representation/approximation

- Chapter 3: **Image Processing**

  - two-dimensional signal processing, separable transforms, convolution and filtering in 2D
  - image enhancement, histogram modifications, edge detection, directional smoothing
  - image filtering and restoration, blur removal, reduction of additive noise

- Chapter 4: **Source Coding**

  - elements of information theory, entropy, mutual information
  - lossless source coding, Huffman codes, Lempel-Ziv coding, run-length coding
  - lossy source coding, scalar and vector quantization, rate-distortion theory
  - waveform coding: predictive coding, ADPCM, sigma-delta modulation, transform coding, examples in speech, audio and image coding

The leading thread running through and connecting the different chapters is the concept of *statistical signal processing*. Signals are regarded as realizations of stochastic processes. The statistical aspect allows us to introduce some structure in the description of the signals we are treating, that is otherwise more difficult to capture.

Statistical signal processing concepts occur in the domains of the three Eurecom departments. One key building block underlying all modern telecommunications is a modem of which the key building blocks are illustrated in Fig. 0.1. In this figure, signals appear in three representation formats: analog form (A), digital form (D) or binary/coded (B) form. A modem (short for modulator/demodulator) allows digital transmission over an analog medium. The upper part of Fig. 0.1 represents the transmitter while the lower part represents the receiver. Transmitter and receiver are linked by the transmission medium, which is commonly called the channel. The information to be transmitted (source) is presented in analog or digital form. The source is then encoded which means that the redundancy in its (binary) representation is removed as much as possible. The information is then channel encoded, which means that some redundancy is added in a controlled fashion so that errors introduced by imperfections (noise, nonlinearities) in the channel can be detected and possibly corrected by the receiver. The encoded information is then modulated which includes both spectral shaping (limiting of the bandwidth occupied) and possibly frequency translation. Demodulation performs the opposite frequency translation (which requires knowledge of the carrier frequency and possibly its phase). The demodulated signal then gets sampled, which may require synchronization of the sampling timing instants (depending on the sampling frequency among other issues). The filtering introduced by the channel gets undone by the equalizer. The binary information then gets extracted from the signal by a process called detection. The resulting bits then get treated for error detection and correction. And finally the original source gets (possibly

approximately) reconstructed from its representation bits. In practice, several of these building blocks may be combined into one single operation. The subjects treated in the TCOM course are channel coding and decoding, modulation and demodulation, sampling, detection and equalization to some extent. The subjects treated in this course are source coding and decoding, equalization/channel estimation, and parameter estimation techniques that are useful for timing and carrier synchronization. Some other issues that appear in a modem and that are not indicated in Fig. 0.1 are transmitter and receiver amplifiers, and encryption.

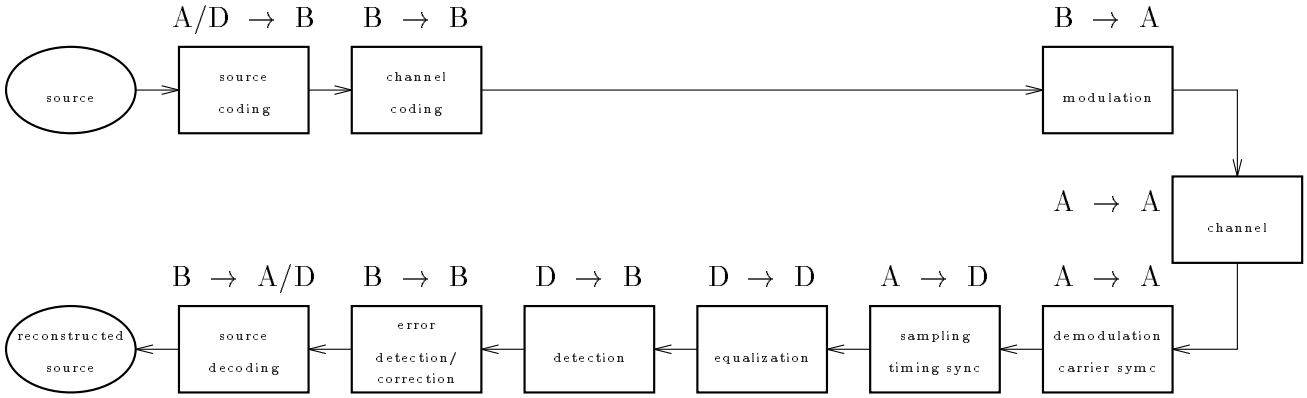

Figure 0.1: Building blocks of a modem.

The modem is a building block in all three of Eurecom's departments. The following are more specific statistical signal processing topics for each of the departments.

- **Corporate communications:**
  Even though with the advent of optical fiber communication networks have become very broadband, the appetite for information has been steadily increasing also. For example, recent developments in multimedia require the transmission of video signals, graphics, and possibly animated graphics, which are quite bandwidth consuming applications. Before transmission, it is often mandatory to eliminate redundancies in the signals to be transmitted. This operation is called source compression. The compression of information is perhaps even more important when it comes to storage (e.g. Unix commands compress and gzip).

  Recent developments in image compression allow the information to be structured in a hierarchical way. At the highest level, we have a representation of the image with a fairly low definition. At subsequent levels, we find the complementary information that allows one to recreate, using also the information at the lower definition level, the image at a higher definition level. An important property of this hierarchical information structure is the nesting: the information at a higher definition level consists of the information at a lower level plus some complementary information. This hierarchical structure is perfectly matched for transmission over a link that is of time-varying quality. When the link quality degrades, fewer bits can be transmitted correctly and it would be possible to transmit an image at a reduced level of definition, but with few errors. For applications

like this one, the corporate communications engineer needs to have some notions about (image) coding techniques.

In queuing theory, the performance of communication networks is evaluated using statistical concepts and models. For instance, the arrival of communication packets may be modeled as a Poisson process. Such statistical models introduce a certain structure in the phenomena we observe, but they often include one or more parameters of which the values depend on the specific application and need to be estimated.

- **Multimedia communications:**
  In many applications the source coding needs to be lossless or reversible, i.e. the original information can be exactly retrieved from the reduced number of bits. In many multimedia applications on the other hand, such as videoconferencing or videotelephony, lossy image and video coding are acceptable. In lossy source coding, irreversible quantization errors are introduced, which lead to a further reduction in the number of bits required for representing the source.

  Marketing studies have shown that the audio part in teleconferencing or telepresence systems is of extreme importance. For ergonomical reasons, the audioconferencing system consists of microphone and loudspeakers (and not headphones). This leads to acoustic echos and the existence of a possibly unstable closed loop for the audio signals. The solution consists of putting an adaptive filter between loudspeaker and microphone that will produce the negative of the received echo and hence will suppress it when subtracted from the microphone signal.

  More advanced multimedia systems use speech recognition for voice-controlled operation. The background noise (other speakers etc.) complicates the task of the speech recognizer. Speech enhancement and directive acoustic microphone arrays constitute further examples of statistical signal processing.

  Multimedia communications lead to a synchronization problem of multiple information streams, which are sent over possibly different routes. For instance, at the destination, audio and video information needs to be presented in a synchronized fashion.

  Transmission of continuous sources such as audio (speech) and video over packet-mode networks (internet) requires the development of source coders that are robust to packet losses and delay variations.

  Multimedia applications require lots of bandwidth and encourage the development of high-speed connectivity over the last mile to the telephone subscriber. Current voiceband modems are limited to 56Kb/s. However, telephone operators have fibre arriving up to a last switching center. To realize high-speed connectivity over the twisted pair of the subscriber loop, high-speed modem standards are being developed that are referred to as xDSL (DSL=Digital Subscriber Loop). xDSL stands for a collection of standards such as ADSL (Asymmetric), HDSL (High-speed), VDSL, which offer various bitrate and duplexing configurations. At these high rates, the twisted pair telephone line introduces quite a bit of distortion (filtering) and sophisticated techniques are required to overcome this distortion (equalization).

- **Mobile communications:**
  Digital mobile communications can be made more reliable and more flexible than analog systems. Digital transmission requires voice coding. Because of the very limited bandwidth in mobile communications, fairly sophisticated low bit rate coding techniques need to be used. The changing environment, leading to time-varying quality of service, requires the development of adaptive multi-rate coding, in which the split of a total bit rate between source and channel coding can be done adaptively.

  National regulations are evolving in the sense of prohibiting simultaneous driving and making a phone call with the classical handheld mobile phone for safety reasons. This leads to the development of handsfree telephone systems and hence again audio conferencing systems with their acoustic echo problem. Not only the talking should be handsfree but also the dialing. This requires speech recognition techniques, which have to work in the particularly noisy environment of the car interior.

  In a wireless environment, multipath propagation leads to time dispersion and mobility leads to variation in time. Hence fast channel equalization techniques are needed, possibly of limited complexity (due to limited processing power and battery life). Other signal processing operations that are required are carrier and symbol rate synchronization (timing), Doppler and fading tracking.

  In contrast to communication over telephone lines, coax cables or optical fibers, radio communication occurs in an open environment in which the presence of other users is felt much more intensely. In a cellular arrangement, the frequency range may be partitioned between different users in one particular cell, but frequency reuse occurs between different cells. A possible way to proceed may be a joint treatment of different users. Disturbance may also come from different communication systems having some overlapping frequency range. In this case, interference suppression techniques may be helpful. Another way to distinguish between different users occupying the same frequency band is to use multiple antennas. Antenna array processing allows to sort out a number of different users. Sophisticated multiple access (MA) techniques such as Spatial Division MA (SDMA) and Code Division MA (CDMA) allow for interference cancellation, which may involve so-called spatio-temporal processing.

  An emerging issue in mobile communications is user localization. National authorities are requiring cellular system operators to be able to provide the localization of any mobile user, for emergency purposes. For instance, if someone calls the emergency number 911 in the US, the police should not only get the phone call as usual, but the simple fact of dialing the number 911 should inform the police about the localization of a person in need. The localization of mobile users can be based on information provided by satellites (GPS - Global Positioning System). But it can also be provided on the basis of information available at mobile and base stations. The estimation of the location may involve a variety of signal processing methods involving attenuation trajectory profiles, timing information (TOA - time of arrival) and direction-of-arrival (DOA) information provided by base station antenna arrays.

Sections indicated with ♠ constitute advanced reading material.

Chapter 0

Background Material

0.1 Probability

For clarity, we distinguish here explicitly between the random variables $\mathbf{x}_i$ (functions on a probablity space) which are denoted in boldface, and the values x_i that these random variables may take on. Later on we shall omit this distinction.

Consider a random vector $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$. A random vector can either be considered as a vector valued mapping from a probability space or a collection of random variables. The random vector has a (joint) probability density function $f_{\mathbf{X}}(X)$. The meaning of the joint pdf (if continuous) is:

$$\Pr[\mathbf{X} \in (X, X + dX)] \;=\; f_{\mathbf{X}}(X) dX$$

where with some abuse of notation first $dX = [dx_1 \cdots dx_m]^T$ and then $dX = dx_1\, dx_2 \cdots dx_m$. The mean of $\mathbf{X}$ is

$$\begin{aligned} m_X \;&=\; E\mathbf{X} \;=\; \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\, X \\ &=\; \begin{bmatrix} \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\, x_1 \\ \vdots \\ \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\, x_m \end{bmatrix} = \begin{bmatrix} \int dx_1\, x_1 \overbrace{\int dx_2 \cdots \int dx_m f_{\mathbf{X}}(X)}^{f_{\mathbf{x}_1}(x_1)} \\ \vdots \\ \int dx_m\, x_m \underbrace{\int dx_1 \cdots \int dx_{m-1} f_{\mathbf{X}}(X)}_{f_{\mathbf{x}_m}(x_m)} \end{bmatrix} = \begin{bmatrix} m_{x_1} \\ \vdots \\ m_{x_m} \end{bmatrix}. \end{aligned} \tag{0.1}$$

Its covariance matrix is

$$C_{XX} \;=\; E\,(\mathbf{X} \Leftrightarrow m_X)\,(\mathbf{X} \Leftrightarrow m_X)^T \;=\; \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X) \underbrace{(X \Leftrightarrow m_X)\,(X \Leftrightarrow m_X)^T}_{\geq 0\,,\ \text{rank } 1} \tag{0.2}$$

where $C_{x_i x_j} \;=\; E\,(\mathbf{x}_i \Leftrightarrow m_{x_i})\left(\mathbf{x}_j \Leftrightarrow m_{x_j}\right)$. The superscript ^{T} denotes transposition. Observe $C_{XX} = C_{XX}^T \geq 0$ symmetric and positive semidefinite, as weighted average of symmetric positive semidefinite matrices. The covariance matrix (centered second-order moments) is related to the correlation matrix $R_{XX} = E\mathbf{X}\mathbf{X}^T$ (non-centered second-order moments) in the

following way

$$\begin{aligned} C_{XX} &= E\left(\mathbf{X} \Leftrightarrow m_X\right)\left(\mathbf{X} \Leftrightarrow m_X\right)^T = \left[E\left(\mathbf{x}_i \Leftrightarrow m_{x_i}\right)\left(\mathbf{x}_j \Leftrightarrow m_{x_j}\right)\right]_{i,j=1}^{m} \\ &= E\left(\mathbf{X}\mathbf{X}^T \Leftrightarrow \mathbf{X} m_X^T \Leftrightarrow m_X \mathbf{X}^T + m_X m_X^T\right) \\ &= \left(E\mathbf{X}\mathbf{X}^T\right) \Leftrightarrow (E\mathbf{X})\, m_X^T \Leftrightarrow m_X\, (E\mathbf{X})^T + m_X m_X^T \\ &= R_{XX} \Leftrightarrow m_X m_X^T \Leftrightarrow m_X m_X^T + m_X m_X^T = R_{XX} \Leftrightarrow m_X m_X^T \,. \end{aligned} \tag{0.3}$$

In this derivation we exploited the linearity of the expectation operator E, which allows us to interchange it with other linear operations.

0.2 Linear Algebra

Lemma 0.1 (Matrix Inversion Lemma) *If A and C are respectively $n \times n$ and $m \times m$ invertible matrices and B and D are respectively $n \times m$ and $m \times n$ matrices, then*

$$[A + BCD]^{-1} = A^{-1} \Leftrightarrow A^{-1} B \left[DA^{-1}B + C^{-1}\right]^{-1} DA^{-1} \tag{0.4}$$

if the inverses exist.

This lemma can be shown by straightforward verification: the product of $A+BCD$ with the RHS of the equation gives the identity matrix. This lemma is very useful when we have to compute $[A + BCD]^{-1}$, we know A^{-1}, and m is (much) smaller than n: in that case we only have to invert the $m \times m$ matrices C and $DA^{-1}B + C^{-1}$ instead of the $n \times n$ matrix $A + BCD$.

The **trace** of a $n \times n$ square matrix A is defined as the sum of its diagonal elements: $\mathrm{tr}\{A\} = \sum_{i=1}^{n} A_{ii}$. The trace of a scalar is the scalar itself since a scalar is a 1×1 matrix. An interesting property of the trace operator is that

$$\mathrm{tr}\{A\,B\} = \mathrm{tr}\{B\,A\} \tag{0.5}$$

whenever A and B are matrices such that A and B^T have the same dimensions. So AB and BA can be square matrices of different dimensions. In particular, if A is a row vector and B is a column vector, then AB is a scalar while BA is a matrix.

0.3 The Multivariate Gaussian Distribution

We shall introduce the multivariate Gaussian distribution in a constructive fashion, using the Gaussian distribution of a single variable and two postulates. Let $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ be a real random vector. For the construction of the multivariate Gaussian distribution, the only moments that are required are the moments of first and second order. So we consider available the mean m_X and the covariance matrix C_{XX}.

Eigendecomposition of a Covariance Matrix

Equation (0.2) shows that $C_{XX} = C_{XX}^T$ is symmetric. The eigenvalues λ_i and corresponding eigenvectors V_i of C_{XX} satisfy $C_{XX}V_i = \lambda_i V_i$. Symmetry of the matrix implies that its eigenvalues are real and that there exists a complete set of orthonormal eigenvectors: $V_i^T V_j = \delta_{ij}$ where δ_{ij} is the Kronecker delta (equal to 1 if $i = j$ and zero otherwise). The matrix $V = [V_1 \cdots V_m]$, the columns of which are the eigenvectors of C_{XX}, is orthogonal. Indeed, $\left[V^T V\right]_{ij} = V_i^T V_j = \delta_{ij}$, hence $V^T V = I_m$. This last identity implies that $V^{-1} = V^T$. Hence, also $V\,V^T = I_m$. Since both the matrix C_{XX} itself and its eigenvalues are real, this also implies that the eigenvectors are real. Now $I_m = V^T V$ implies that $1 = \det\left(V^T V\right) = (\det V)^2 \;\Rightarrow$ $\det V = \pm 1$. We can choose the signs of the V_i such that $\det V = 1$.

From (0.2) it follows that C_{XX} is an infinite sum (integral) of positive semidefinite (of rank one) matrices $f_{\mathbf{X}}(X)\,(X \Leftrightarrow m_X)\,(X \Leftrightarrow m_X)^T$. This implies that C_{XX} itself is positive semidefinite, which we denote as $C_{XX} \geq 0$. This means that $\forall U \in \mathcal{R}^m : U^T C_{XX} U = E\left(U^T(\mathbf{X} \Leftrightarrow m_X)\right)^2 \geq 0$ (positive definite would be $\forall U \in \mathcal{R}^m \setminus \{0\} : U^T C_{XX} U > 0$). By choosing $U = V_i$ we find $U^T C_{XX} U = V_i^T C_{XX} V_i = \lambda_i V_i^T V_i = \lambda_i \geq 0$. We can order the eigenvalues in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. If $\lambda_m = 0$, C_{XX} is singular. In this case $V_m^T C_{XX} V_m = E\left(V_m^T(\mathbf{X} \Leftrightarrow m_X)\right)^2 = 0$ or hence the auto-correlation of the random variable $V_m^T(\mathbf{X} \Leftrightarrow m_X)$ is zero. This implies that its mean and variance are zero, which in turn implies that the random variable $V_m^T(\mathbf{X} \Leftrightarrow m_X)$ is zero with probability one (w.p. 1). This means that at least one variable $\mathbf{x}_i$ is a linear combination of the other variables and 1. We shall in general exclude this possibility and hence exclude the case of singular covariance matrices. In other words, we shall assume in general that covariance matrices are positive definite: $C_{XX} > 0$, $\lambda_i > 0$, $i = 1, \ldots, m$. Note that the vectors $C_{XX}V_i \Leftrightarrow V_i\lambda_i = 0$ are the columns of the matrix $C_{XX}V \Leftrightarrow V\Lambda = 0$ where $\Lambda = \mathrm{diag}\,\{\lambda_1, \ldots, \lambda_m\}$. Using $V^{-1} = V^T$, we find

$$C_{XX} \;=\; V\Lambda V^T \;=\; [V_1 \cdots V_M]\,\mathrm{diag}\,\{\lambda_1, \ldots, \lambda_m\}\,[V_1 \cdots V_M]^T \;=\; \sum_{i=1}^{m} \lambda_i V_i V_i^T \;. \tag{0.6}$$

Derivation of the Multivariate Gaussian Distribution

In most textbooks, the expression for the multivariate distribution is first given and then some well-known properties are derived from it. These properties are that uncorrelated Gaussian variables are independent and that Gaussian variables remain Gaussian under linear transformations. Here, we shall take the opposite approach. We shall show that using these properties, it is possible to derive the expression for the multivariate Gaussian distribution from the distribution of a scalar Gaussian variable.

To this end, consider the following linear transformation of variables: $\mathbf{Z} = V^T(\mathbf{X} \Leftrightarrow m_X)$. Then we find for the first two moments

$$\begin{aligned} m_Z &= E\,V^T(\mathbf{X} \Leftrightarrow m_X) = V^T(E\mathbf{X} \Leftrightarrow m_X) = V^T(m_X \Leftrightarrow m_X) = 0 \\ C_{ZZ} &= E\,(\mathbf{Z} \Leftrightarrow m_Z)\,(\mathbf{Z} \Leftrightarrow m_Z)^T = E\mathbf{Z}\mathbf{Z}^T = EV^T\,(\mathbf{X} \Leftrightarrow m_X)\,(\mathbf{X} \Leftrightarrow m_X)^T\,V \\ &= V^T\left(E\,(\mathbf{X} \Leftrightarrow m_X)\,(\mathbf{X} \Leftrightarrow m_X)^T\right)V = V^T C_{XX} V = V^T V\,\Lambda\,V^T V = \Lambda \end{aligned} \tag{0.7}$$

Or hence $E\mathbf{z}_i\mathbf{z}_j = \lambda_i\delta_{ij}$: the variables $\mathbf{z}_i$ are zero mean and uncorrelated. At this point we have only specified the first two moments of the variables $\mathbf{z}_i$. We will now specify the rest of their distribution by stating that the $\mathbf{z}_i$ are jointly Gaussian. We furthermore postulate that zero mean uncorrelated Gaussian random variables are independent (note that this postulate is backwards compatible since in general independent zero mean random variables are uncorrelated; on the other hand, uncorrelated zero mean random variables are not independent in general). Hence the joint distribution of the independent Gaussian random variables $\mathbf{z}_i$ is

$$f_{\mathbf{Z}}(Z) = \prod_{i=1}^{m} f_{\mathbf{z}_i}(z_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\,\lambda_i}} \exp\left(\Leftrightarrow\frac{z_i^2}{2\lambda_i}\right) = (2\pi)^{-\frac{m}{2}} (\det\Lambda)^{-\frac{1}{2}} \exp\left(\Leftrightarrow\frac{1}{2} Z^T\Lambda^{-1}Z\right) \quad (0.8)$$

At this point we furthermore postulate that a linear transformation of jointly Gaussian random variables produces again jointly Gaussian random variables. Since the Jacobian ($\det V^T$) of the linear transformation between $\mathbf{X}$ and $\mathbf{Z}$ equals one, we get for the joint distribution of the variables $\mathbf{x}_i$

$$\begin{aligned} f_{\mathbf{X}}(X) &= f_{\mathbf{Z}}\left(V^T(X \Leftrightarrow m_X)\right) \\ &= (2\pi)^{-\frac{m}{2}} \left(\det(V^T C_{XX} V)\right)^{-\frac{1}{2}} \exp\left(\Leftrightarrow\tfrac{1}{2}\left[V^T\left(X\Leftrightarrow m_X\right)\right]^T \Lambda^{-1}\left[V^T\left(X\Leftrightarrow m_X\right)\right]\right) \\ &= (2\pi)^{-\frac{m}{2}} \left((\det V^T)(\det C_{XX})(\det V)\right)^{-\frac{1}{2}} \exp\left(\Leftrightarrow\tfrac{1}{2}\left[X\Leftrightarrow m_X\right]^T V\Lambda^{-1}V^T\left[X\Leftrightarrow m_X\right]\right) \\ &= (2\pi)^{-\frac{m}{2}} (\det C_{XX})^{-\frac{1}{2}} \exp\left(\Leftrightarrow\tfrac{1}{2}\left[X\Leftrightarrow m_X\right]^T C_{XX}^{-1}\left[X\Leftrightarrow m_X\right]\right) \end{aligned} \quad (0.9)$$

This is the general expression for a multivariate Gaussian probability density function (pdf). The fact that the variables $\mathbf{x}_i$ are jointly Gaussianly distributed will be denoted as $\mathbf{X} \sim \mathcal{N}(m_X, C_{XX})$, which emphasizes the fact that the Gaussian distribution is completely specified in terms of the first and second-order moments. Because of the two postulates we introduced, it should not be surprising that one can show that zero mean uncorrelated Gaussian random variables are independent and that a linear transformation (and in particular linear filtering) preserves the Gaussian distribution.

Derivation of the Conditional Multivariate Gaussian Distribution

Now let $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ and $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_n]^T$ be $m+n$ random variables that are jointly Gaussianly distributed. So we have $m+n$ variables which are arbitrarily split into two groups, $\mathbf{X}$ and $\mathbf{Y}$. The joint Gaussian (or normal) distribution of these $m+n$ variables can be stated as

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_X \\ m_Y \end{bmatrix}, \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}\right) \quad (0.10)$$

where the mean is defined as

$$\begin{bmatrix} m_X \\ m_Y \end{bmatrix} = E\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \quad (0.11)$$

and the covariance matrix is defined as

$$C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} = E\begin{bmatrix} \mathbf{X}\Leftrightarrow m_X \\ \mathbf{Y}\Leftrightarrow m_Y \end{bmatrix}\begin{bmatrix} \mathbf{X}\Leftrightarrow m_X \\ \mathbf{Y}\Leftrightarrow m_Y \end{bmatrix}^T . \quad (0.12)$$

The joint Gaussian pdf can be written as

$$f_{\mathbf{X},\mathbf{Y}}(X,Y) \;=\; (2\pi)^{-\frac{m+n}{2}} (\det C)^{-\frac{1}{2}} \exp\left(\Leftrightarrow \frac{1}{2} \begin{bmatrix} X \Leftrightarrow m_X \\ Y \Leftrightarrow m_Y \end{bmatrix}^T \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} X \Leftrightarrow m_X \\ Y \Leftrightarrow m_Y \end{bmatrix} \right). \tag{0.13}$$

We recall the meaning of the joint pdf, which is

$$\Pr[\mathbf{X} \in (X, X+dX), \mathbf{Y} \in (Y, Y+dY)] \;=\; f_{\mathbf{X},\mathbf{Y}}(X,Y) dX dY \tag{0.14}$$

where with some abuse of notation dX once means $dX = [dx_1 \cdots dx_m]^T$ and the second time it stands for $dX = dx_1\, dx_2 \cdots dx_m$ (and similarly for dY). In words, for continuously valued random variables (as opposed to discretely valued random variables), the probability that $\mathbf{x}_1,\ldots,\mathbf{y}_n$ fall in an infinitesimally small $(m+n)$-dimensional cube with corner point $x_1,\ldots,y_n$ and sides $dx_1,\ldots,dy_n$ is equal to the probability density at the corner point times the volume $dx_1 \cdots dy_n$ of the cube.

We are interested in finding the conditional pdf $f_{\mathbf{X}|\mathbf{Y}}(X|Y)$ of which the meaning is

$$\Pr[\mathbf{X} \in (X, X+dX) | \mathbf{Y} = Y] \;=\; f_{\mathbf{X}|\mathbf{Y}}(X|Y) dX \quad . \tag{0.15}$$

The conditional distribution is defined by Bayes' rule:

$$f_{\mathbf{X}|\mathbf{Y}}(X|Y) \;=\; \frac{f_{\mathbf{X},\mathbf{Y}}(X,Y)}{f_{\mathbf{Y}}(Y)} \tag{0.16}$$

where the marginal distribution $f_{\mathbf{Y}}(Y)$ is obtained from the joint distribution by integrating out X:

$$\begin{aligned} f_{\mathbf{Y}}(Y) &= \int f_{\mathbf{X},\mathbf{Y}}(X,Y) dX \\ &= \int \cdots \int f_{\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{y}_1,\ldots,\mathbf{y}_n}(x_1,\ldots,x_m,y_1,\ldots,y_n) dx_1 \cdots dx_m \\ &= (2\pi)^{-\frac{n}{2}} (\det C_{YY})^{-\frac{1}{2}} \exp\left(\Leftrightarrow \frac{1}{2} [Y \Leftrightarrow m_Y]^T C_{YY}^{-1} [Y \Leftrightarrow m_Y] \right). \end{aligned} \tag{0.17}$$

This last expression can be written down immediately since we know the mean and the covariance matrix of the Gaussian variables $\mathbf{Y}$. In order to carry out the division in (0.16), we shall try to factor out $f_{\mathbf{Y}}(Y)$ from $f_{\mathbf{X},\mathbf{Y}}(X,Y)$. For this we need to consider the block Upper Diagonal Lower (UDL) triangular factorization of the covariance matrix C:

$$C \;=\; \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} \;=\; \begin{bmatrix} I & K \\ 0 & I \end{bmatrix} \begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix} \begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix} \tag{0.18}$$

with

$$K \;=\; C_{XY} C_{YY}^{-1} \;,\quad P \;=\; C_{XX} \Leftrightarrow C_{XY} C_{YY}^{-1} C_{YX} \;. \tag{0.19}$$

To verify this factorization, it suffices to check that the product of the three factors on the RHS (right hand side) of (0.18) gives C. From the UDL factorization of C, we can obtain the

LDU factorization of C^{-1} by taking the inverse of both sides of (0.18), viz.

$$\begin{aligned} C^{-1} &= \begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix}^{-1} \begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} I & K \\ 0 & I \end{bmatrix}^{-1} \\ &= \begin{bmatrix} I & 0 \\ \Leftarrow K^T & I \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} I & \Leftarrow K \\ 0 & I \end{bmatrix} \end{aligned} \tag{0.20}$$

where we used the property that $(AB)^{-1} = B^{-1}A^{-1}$ and the expression given for the inverse of a 2×2 block triangular matrix can be easily verified. By taking determinants of both sides of (0.18), we also obtain

$$\begin{aligned} \det C &= \det \begin{bmatrix} I & K \\ 0 & I \end{bmatrix} \det \begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix} \det \begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix} \\ &= 1 \ . \ \det \begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix} \ . \ 1 \ = \ \det P \ \ \det C_{YY} \end{aligned} \tag{0.21}$$

where we used the properties $\det(AB) = \det A \ \det B$, the determinant of a triangular matrix is the product of the diagonal elements and more generally, the determinant of a block diagonal (or triangular) matrix is the product of the determinants of the diagonal blocks. Using (0.20), we can rewrite the exponent of the joint distribution as

$$\begin{aligned} &\begin{bmatrix} X \Leftarrow m_X \\ Y \Leftarrow m_Y \end{bmatrix}^T \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} X \Leftarrow m_X \\ Y \Leftarrow m_Y \end{bmatrix} \\ &= \begin{bmatrix} X \Leftarrow m_X \\ Y \Leftarrow m_Y \end{bmatrix}^T \begin{bmatrix} I & 0 \\ \Leftarrow K^T & I \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} I & \Leftarrow K \\ 0 & I \end{bmatrix} \begin{bmatrix} X \Leftarrow m_X \\ Y \Leftarrow m_Y \end{bmatrix} \\ &= \begin{bmatrix} X \Leftarrow KY \Leftarrow (m_X \Leftarrow K m_Y) \\ Y \Leftarrow m_Y \end{bmatrix}^T \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} X \Leftarrow KY \Leftarrow (m_X \Leftarrow K m_Y) \\ Y \Leftarrow m_Y \end{bmatrix} \\ &= [X \Leftarrow KY \Leftarrow (m_X \Leftarrow K m_Y)]^T P^{-1} [X \Leftarrow KY \Leftarrow (m_X \Leftarrow K m_Y)] + [Y \Leftarrow m_Y]^T C_{YY}^{-1} [Y \Leftarrow m_Y] . \end{aligned} \tag{0.22}$$

Using (0.13),(0.17),(0.21) and (0.22), we can carry out the division in (0.16) and prove the following theorem:

Theorem 0.1 (Gauss-Markov) *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *have the joint Gaussian distribution indicated in (0.10),(0.13), then the conditional distribution is*

$$\begin{aligned} &f_{\mathbf{X}|\mathbf{Y}}(X|Y) \ = \\ &(2\pi)^{-\frac{m}{2}} (\det P)^{-\frac{1}{2}} \exp\left(\Leftarrow \tfrac{1}{2} [X \Leftarrow KY \Leftarrow (m_X \Leftarrow K m_Y)]^T P^{-1} [X \Leftarrow KY \Leftarrow (m_X \Leftarrow K m_Y)] \right) \ . \end{aligned} \tag{0.23}$$

The conditional distribution is again Gaussian with conditional mean

$$E_{\mathbf{X}|\mathbf{Y}} \mathbf{X} \ = \ E[\mathbf{X}|\mathbf{Y} = Y] \ = \ m_X + C_{XY} C_{YY}^{-1} (Y \Leftarrow m_Y) \tag{0.24}$$

and conditional covariance matrix

$$E_{\mathbf{X}|\mathbf{Y}} \left[\mathbf{X} \Leftarrow E_{\mathbf{X}|\mathbf{Y}} \mathbf{X} \right] \left[\mathbf{X} \Leftarrow E_{\mathbf{X}|\mathbf{Y}} \mathbf{X} \right]^T \ = \ P \ = \ C_{XX} \Leftarrow C_{XY} C_{YY}^{-1} C_{YX} \ . \tag{0.25}$$

0.4 Complex Gaussian Variables

In general, complex random variables are best treated by considering their real and imaginary parts separately. Hence problems involving complex random variables and parameters can always be reduced to problems involving only real quantities. Nevertheless, when the complex random variables are **circular**, then it might be much more compact/convenient to work with the complex variables directly. An example of circular complex random variables is the case in which the real and imaginary parts are i.i.d. When working with circular random vectors, it often suffices to replace the transpose operator by a Hermitian (complex conjugate) transpose operator to extend results from the real case to the complex case.

0.5 Vector Spaces

The most familiar vector space is the Euclidean vector space. In this case, the vectors are in fact column vectors filled with (real) scalar constants. The Euclidean inner product between two vectors X and Y is defined as $X^T Y = \sum_{i=1}^{n} x_i y_i$. The Euclidean norm is the square-root of $\|X\|_2^2 = X^T X$. Using the trace operator, we can also write $X^T Y = \mathrm{tr}\{Y X^T\}$ which might be useful for certain operations.

Another possible vector space is a space of (real) random variables. In this case the inner product between two random variables x and y is their correlation $r_{xy} = E\,xy$. The norm in this case is the square-root of the power $r_{xx} = E\,x^2 = \sigma_x^2 + m_x^2$. Two random variables are called orthogonal if they are uncorrelated.

0.6 Optimization

For the minimization of cost functions involving a vector of coefficients, we need to consider the gradient w.r.t. the coefficient vector. If $g(\phi)$ is a $1 \times l$ row vector function then the gradient w.r.t. ϕ is defined as

$$\frac{\partial\, g(\phi)}{\partial \phi} \;=\; \begin{bmatrix} \frac{\partial\, g(\phi)}{\partial \phi_1} \\ \vdots \\ \frac{\partial\, g(\phi)}{\partial \phi_m} \end{bmatrix} \tag{0.26}$$

which is a $m \times l$ matrix. If $g(\phi)$ is a scalar ($l = 1$), then $\frac{\partial\, g(\phi)}{\partial \phi}$ is a column vector of the same dimensions as ϕ. The gradient operator commutes with linear operations. Some often-used formulas are the following.

$$\frac{\partial\, \phi^T}{\partial \phi} \;=\; \left[\frac{\partial\, \phi_j}{\partial \phi_i}\right] \;=\; I_m\ . \tag{0.27}$$

Let X be a $m \times 1$ column vector.

$$\frac{\partial}{\partial \phi}\left(\phi^T X\right) \;=\; \left(\frac{\partial\, \phi^T}{\partial \phi}\right) X \;=\; I_m\, X \;=\; X\ . \tag{0.28}$$

Since a scalar equals its transpose, we get

$$\frac{\partial}{\partial \phi}\left(X^T \phi\right) \;=\; \frac{\partial}{\partial \phi}\left(\phi^T X\right) \;=\; X \;. \tag{0.29}$$

If A is a $m \times l$ matrix, then we get

$$\frac{\partial}{\partial \phi}\left(\phi^T A\right) \;=\; \left(\frac{\partial\, \phi^T}{\partial \phi}\right) A \;=\; I_m\, A \;=\; A \;. \tag{0.30}$$

In the following, we use an extension of the well-known rule for the scalar case, $(uv)' = u'\,v + u\,v'$, to the vector case. Let $g(\phi)$ and $h(\phi)$ both be $l \times 1$ column vector functions of ϕ. Since $g^T(\phi)h(\phi) = \left(g^T(\phi)h(\phi)\right)^T = h^T(\phi)g(\phi)$, we get

$$\frac{\partial}{\partial \phi}\left(g^T(\phi)h(\phi)\right) \;=\; \left(\frac{\partial g^T(\phi)}{\partial \phi}\right) h(\phi) + \left(\frac{\partial h^T(\phi)}{\partial \phi}\right) g(\phi) \;. \tag{0.31}$$

A particular application of this rule with $g(\phi) = \phi$ and $h(\phi) = A\phi$ (with A a square matrix) leads to

$$\frac{\partial}{\partial \phi}\left(\phi^T A\phi\right) \;=\; \left(\frac{\partial\, \phi^T}{\partial \phi}\right) A\phi + \left(\frac{\partial\, \phi^T}{\partial \phi}\right) A^T\phi \;=\; \left(A + A^T\right)\phi \;. \tag{0.32}$$

When A is symmetric, this gradient reduces to $2\,A\phi$.

To know the nature of an extremum, it is important to investigate the Hessian (matrix of second-order derivatives):

$$\frac{\partial}{\partial \phi}\left[\frac{\partial}{\partial \phi}\left(\phi^T A\phi\right)\right]^T \;=\; \frac{\partial}{\partial \phi}\left[\phi^T\left(A + A^T\right)\right] \;=\; A + A^T \tag{0.33}$$

which again reduces to $2\,A$ when A is symmetric. In general, if the Hessian evaluated at an extremum is positive definite, negative definite, or indefinite, the extremum is a local minimum, a local maximum, or a saddle point respectively.

Chapter 2

Spectral Estimation

In practice, stochastic processes are usually characterized in terms of their first and second order moments only. Since most processes have zero mean, this means that the crucial quantity to be considered is the power spectral density function (psdf). This motivates us to investigate techniques for estimating the psdf. Furthermore, many of the techniques we shall see below will be useful also for optimal and adaptive filtering on one hand and signal coding on the other (in later chapters). Indeed, we shall see that in order to perform optimal filtering, we need to have a fairly good idea about the psdf of the stochastic processes involved. Furthermore, optimal filtering requires an operation called spectral factorization. This operation, as we shall see in this chapter, is closely related to linear prediction, a key ingredient in parametric spectral estimation. Linear prediction is also a key building block in source coding.

The psdf is defined for a stationary process only and in fact, spectral estimation can only be carried out in a meaningful way for stationary processes (there exist extensions though to processes that are close to stationary, cyclostationary, asymptotically stationary, etc.). For a stationary process we can, with the help of the additional assumption of ergodicity, replace averaging over an ensemble of realizations by averaging over different time segments of a single realization. This time averaging will allow us to obtain meaningful estimates from a single realization of the process under consideration.

For general nonstationary processes, we have no ergodicity properties, and we cannot average in time. In this case, the "spectrum" changes in time and we can only get a very noisy idea about it. In fact, for a general nonstationary process, spectral estimation is about different ways of representing a single realization of the stochastic process (which can be appropriately considered as a deterministic signal) as a function of time and frequency jointly. This leads to so-called time-frequency representations. These time-frequency representation will again be important for source coding.

2.1 Spectral Descriptions of Stationary Processes

A discrete-time random process can be regarded as a sequence of random variables that take on real values. The random process $\{y_k\}$ is *wide-sense stationary* (WSS) if its mean

$$E\, y_n \;=\; m_y \tag{2.1}$$

does not depend on n and its *autocorrelation function* (acf)

$$r_{yy}(k) \;=\; E\, y_{n+k} y_n \tag{2.2}$$

only depends on the time lag between the two samples and not on their absolute positions. The corresponding central moment is the *autocovariance function*

$$c_{yy}(k) \;=\; E\,(y_{n+k} \Leftrightarrow m_y)(y_n \Leftrightarrow m_y) \;=\; r_{yy}(k) \Leftrightarrow m_y^2 \;\; . \tag{2.3}$$

More generally, two jointly WSS random processes $\{x_k\}$ and $\{y_k\}$ have a *cross-correlation function* (ccf)

$$r_{xy}(k) \;=\; E\, x_{n+k} y_n \tag{2.4}$$

and a *cross-covariance function*

$$c_{xy}(k) \;=\; E\,(x_{n+k} \Leftrightarrow m_x)(y_n \Leftrightarrow m_y) \;=\; r_{xy}(k) \Leftrightarrow m_x m_y \;\; . \tag{2.5}$$

Unless stated otherwise, we shall consider below processes with zero mean so that the autocorrelation function and the autocovariance function are equal. Nevertheless, the rest of this section holds for the general case. The following symmetry properties follow immediately from the definition:

$$\begin{aligned} r_{xy}(k) &= r_{yx}(\Leftrightarrow k) \\ r_{yy}(k) &= r_{yy}(\Leftrightarrow k) \;\; . \end{aligned} \tag{2.6}$$

The following boundedness properties follow from the Cauchy-Schwarz inequality which says that for vectors x and y in a vector space with inner product $< x, y >$ (and norm $||x|| = \sqrt{< x, x >}$) we have $| < x, y > |^2 \leq < x, x >< y, y >$. We apply this to a vector space in which the vectors are random variables and the inner product is the correlation. Then $|E\, x_{k+n} y_n|^2 \leq (E\, x_{k+n} x_{k+n})(E\, y_n y_n)$ or hence

$$\begin{aligned} r_{xx}(0)\, r_{yy}(0) &\geq |r_{xy}(k)|^2 \\ r_{yy}(0) &\geq |r_{yy}(k)| \;\geq\; 0 \end{aligned} \tag{2.7}$$

The acf is furthermore a *positive semidefinite function*, meaning the following. Take any positive integer M. Let $\{a_j,\, j = 1, \ldots, M\}$ be any set of M real numbers, not all zero, and $\{k_j \in \mathcal{Z},\, j = 1, \ldots, M\}$ any set of M distinct time instants. Then

$$0 \;\leq\; E\,|\sum_{j=1}^{M} a_j\, y_{k_j}|^2 \;=\; \sum_{i=1}^{M}\sum_{j=1}^{M} a_i a_j r_{yy}(k_i \Leftrightarrow k_j) \;. \tag{2.8}$$

This means also that the symmetric matrix T with elements $T_{ij} = r_{yy}(k_i \Leftrightarrow k_j)$ is positive semidefinite. T is Toeplitz (elements along a diagonal are constant) if the k_j are equidistant. For instance, if $k_i = i$, then $T_{ij} = r_{yy}(i \Leftrightarrow j)$.

We can consider the z-transforms of the acf and the ccf

$$\begin{aligned} \mathrm{S}_{yy}(z) &= \sum_{k=-\infty}^{\infty} r_{yy}(k)\, z^{-k} \\ \mathrm{S}_{xy}(z) &= \sum_{k=-\infty}^{\infty} r_{xy}(k)\, z^{-k} \;\; . \end{aligned} \tag{2.9}$$

When evaluated on the unit circle ($z = e^{j2\pi f}$), $\mathrm{S}_{yy}(z)$ and $\mathrm{S}_{xy}(z)$ become the *(auto-)psdf*, $S_{yy}(f) = \mathrm{S}_{yy}(e^{j2\pi f})$, and the *cross-psdf*, $S_{xy}(f) = \mathrm{S}_{xy}(e^{j2\pi f})$, or more explicitly,

$$\begin{aligned} S_{yy}(f) &= \mathrm{S}_{yy}(e^{j2\pi f}) = \sum_{k=-\infty}^{\infty} r_{yy}(k)\, e^{-j2\pi fk} \quad \text{Wiener-Khinchin relation} \\ S_{xy}(f) &= \mathrm{S}_{xy}(e^{j2\pi f}) = \sum_{k=-\infty}^{\infty} r_{xy}(k)\, e^{-j2\pi fk} \ . \end{aligned} \tag{2.10}$$

Because we normalized the sampling period to 1, the sampling frequency is also normalized to 1 and hence $S_{xy}(f)$ and $S_{yy}(f)$ are periodic with period 1. We see that estimating $S_{yy}(f)$ and estimating $r_{yy}(k)$ are equivalent since both are related by the Fourier transform, a one-to-one transformation. Before giving an interpretation to the psdf, we shall consider an example and some properties.

A basic discrete-time random process is white noise, which is defined as having the acf

$$r_{yy}(k) = \sigma_y^2\, \delta_{k0} \tag{2.11}$$

where δ_{kn} is the Kronecker delta. This means that the white process is zero mean and all samples are uncorrelated. The psdf becomes

$$S_{yy}(f) = \sigma_y^2 \tag{2.12}$$

which is constant.

The filtering of a WSS process by a linear time-invariant (LTI) filter produces another WSS process. Indeed, let $y_n = \sum_m h_{n-m} x_m$ where $\{x_k\}$ is WSS. Then we get for the mean

$$E\, y_n = \sum_m h_m\, E\, x_{n-m} = m_x \sum_m h_m = H(0)\, m_x = m_y \tag{2.13}$$

which does indeed not depend on n. Furthermore,

$$\begin{aligned} E\, y_{k+n} x_n &= \sum_m h_{k+n-m}\, E\, x_m x_n = \sum_m h_{k+n-m}\, r_{xx}(m \Leftrightarrow n) = \sum_m h_{k-m}\, r_{xx}(m) \\ &= h_k * r_{xx}(k) = r_{yx}(k) \end{aligned} \tag{2.14}$$

and

$$\begin{aligned} E\, x_{k+n} y_n &= \sum_m h_{n-m}\, E\, x_{k+n} x_m = \sum_m h_{n-m}\, r_{xx}(k+n \Leftrightarrow m) = \sum_m h_{-m}\, r_{xx}(k \Leftrightarrow m) \\ &= h_{-k} * r_{xx}(k) = r_{xy}(k) \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} E\, y_{k+n} y_n &= \sum_m h_{k+n-m} \sum_l h_{n-l}\, E\, x_m x_l = \sum_m h_{k+n-m} \sum_l h_{n-l}\, r_{xx}(m \Leftrightarrow l) \\ &= \sum_m h_{k-m} \sum_l h_{-l}\, r_{xx}(m \Leftrightarrow l) = \sum_m h_{k-m} r_{xy}(m) = h_k * r_{xy}(k) \\ &= h_k * h_{-k} * r_{xx}(k) = r_{yy}(k) \end{aligned} \tag{2.16}$$

in which the correlation considered is each time independent of n. This shows that not only $\{y_k\}$ is WSS but furthermore $\{x_k\}$ and $\{y_k\}$ are jointly WSS. Taking z-transforms of these convolution relations yields

$$\begin{array}{rclcrcl} \mathrm{S}_{yx}(z) &=& \mathrm{H}(z)\,\mathrm{S}_{xx}(z) & & S_{yx}(f) &=& H(f)\,S_{xx}(f) \\ \mathrm{S}_{xy}(z) &=& \mathrm{S}_{xx}(z)\,\mathrm{H}(1/z) & & S_{xy}(f) &=& S_{xx}(f)\,H(\Leftrightarrow f) \\ \mathrm{S}_{yy}(z) &=& \mathrm{H}(z)\,\mathrm{S}_{xx}(z)\,\mathrm{H}(1/z) & & S_{yy}(f) &=& H(f)\,S_{xx}(f)\,H(\Leftrightarrow f) \;=\; |H(f)|^2 S_{xx}(f) \end{array} \tag{2.17}$$

where $\mathrm{H}(z) = \sum_k h_k z^{-k}$ and $H(\Leftrightarrow f) = H^*(f)$ for real h_k. The last relation in (2.17) leads to the interpretation of $S_{xx}(f)$ as power spectral density function. Indeed, on the one hand we find from the Fourier relationship

$$r_{xx}(0) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(f)df \tag{2.18}$$

which means that the total power $r_{xx}(0) = E\,x^2(k)$ is made up of the sum of the contributions at all frequencies. Now let $\{x_k\}$ and $\{y_k\}$ be the input-output pair associated with a LTI filter with impulse response

$$H(f) \;=\; \begin{cases} 1 & ,\; f \in [f_0, f_0 + \Delta f] \\ 0 & ,\; \text{elsewhere in } [\Leftrightarrow\frac{1}{2}, \frac{1}{2}] \end{cases} \tag{2.19}$$

where Δf is arbitrarily small, see Fig. 2.1. Assume that $S_{xx}(f)$ is continuous at f_0. From (2.17), we get

$$r_{yy}(0) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(f)df \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} |H(f)|^2\, S_{xx}(f)df \;=\; \int_{f_0}^{f_0+\Delta f} S_{xx}(f)df \;=\; S_{xx}(f_0)\,\Delta f \tag{2.20}$$

or $S_{xx}(f_0) = r_{yy}(0)/\Delta f$ which first of all implies $S_{xx}(f) \geq 0$ and secondly justifies the name power spectral density of $S_{xx}(f)$.

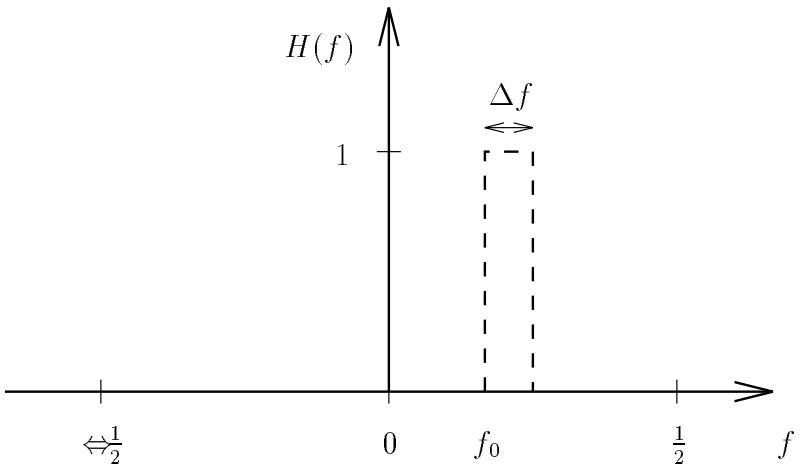

Figure 2.1: Filtering out the frequency component around f_0.

For the special case of $\{x_k\}$ being a white noise input, the psdf of the output process $\{y_k\}$ of a LTI filter with transfer function $H(f) = \mathrm{H}(e^{j2\pi f})$ becomes

$$S_{yy}(f) \;=\; \sigma_x^2 \; |H(f)|^2 \tag{2.21}$$

where $|.|$ denotes the magnitude of a complex number.

We have shown independently that the acf is a positive semidefinite function and that the psdf is nonnegative. It is possible to show that these two properties imply each other. Indeed, if $r_{xx}(k)$ and $S_{xx}(f)$ are two functions that form a Fourier transform pair, then the following theorem holds.

Theorem 2.1 (Bochner's theorem) *$r_{xx}(k)$ is real, even and positive (semi)definite iff $S_{xx}(f)$ is real, even and positive (nonnegative).*

Proof: If. Assume $S_{xx}(f)$ to be real, even and positive. Then $r_{xx}(k)$ is real and even since the inverse Fourier transform of a real and even function is real and even. Now take any positive integer M. Let $\{a_j,\, j = 1, \ldots, M\}$ be any set of M real numbers, not all zero, and $\{k_j \in \mathcal{Z},\, j = 1, \ldots, M\}$ any set of M distinct time instants. Then

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{n=1}^{M} a_m a_n r_{xx}(k_m \Leftrightarrow k_n) \\
&\quad= \sum_{m=1}^{M}\sum_{n=1}^{M} a_m a_n \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(f)\, e^{j2\pi f(k_m-k_n)} df \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} df\; S_{xx}(f) \sum_{m=1}^{M}\sum_{n=1}^{M} a_m a_n e^{j2\pi f(k_m-k_n)} df \\
&\quad= \int_{-\frac{1}{2}}^{\frac{1}{2}} df\; S_{xx}(f) \left|\sum_{m=1}^{M} a_m e^{j2\pi f k_m}\right|^2 \;\geq\; \underbrace{\min_{f\in[-\frac{1}{2},\frac{1}{2}]} S_{xx}(f)}_{>\,0} \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\sum_{m=1}^{M} a_m e^{j2\pi f k_m}\right|^2 df\;.
\end{aligned} \tag{2.22}$$

Now consider the special case $S_{xx}(f) = 1$, $r_{xx}(k) = \delta_{k0}$ in which case (2.22) leads to

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\sum_{m=1}^{M} a_m e^{j2\pi f k_m}\right|^2 df \;=\; \sum_{m=1}^{M} a_m^2 \;>\; 0 \tag{2.23}$$

where the strict inequality follows from the fact that the a_m are not all zero. From (2.22) and (2.23) follows that

$$\sum_{m=1}^{M}\sum_{n=1}^{M} a_m a_n r_{xx}(k_m \Leftrightarrow k_n) \;>\; 0\;. \tag{2.24}$$

The proof in the case of positive semidefiniteness is now immediate.

Only if. Now assume that $r_{xx}(k)$ is real, even and positive definite. Again, $S_{xx}(f)$ is real and even since the Fourier transform of a real and even function is real and even. We shall prove the positivity of $S_{xx}(f)$ by contradiction. Assume that $S_{xx}(f) \leq 0$ in some frequency band. Then let $A(f) = \sum_m a_m e^{-j2\pi f m}$ pick out that frequency band as illustrated in Fig. 2.2.

Then we get

$$\sum_m \sum_n a_m a_n r_{xx}(m \Leftrightarrow n) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(f)\; |A(f)|^2\, df \;<\; 0 \tag{2.25}$$

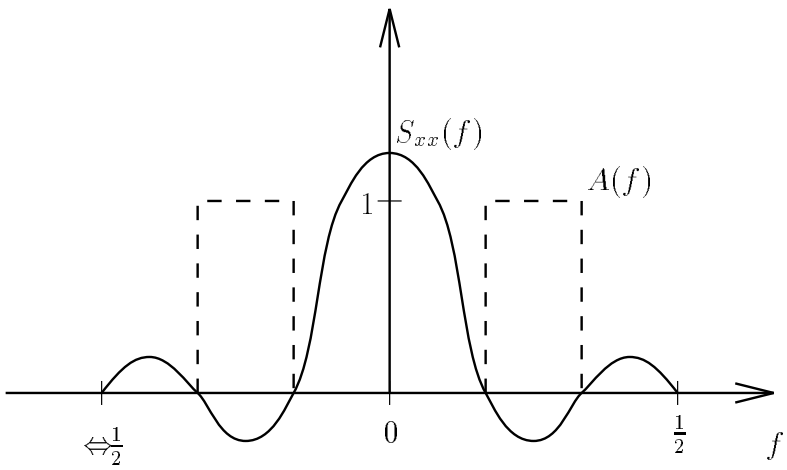

Figure 2.2: Filtering out the frequency components where $S_{xx}(f) \leq 0$.

which contradicts the positive definiteness of $r_{xx}(k)$. Hence $S_{xx}(f) > 0$, $\forall f$. Again, the proof in the case of positive semidefiniteness follows in a similar vein. □

We shall now consider the Fourier transforms $X(f) = \mathrm{X}(e^{j2\pi f}) = \sum_k x_k e^{-j2\pi fk}$ and $Y(f) = \mathrm{Y}(e^{j2\pi f}) = \sum_k y_k e^{-j2\pi fk}$ of the jointly stationary processes $\{x_k\}$ and $\{y_k\}$. Since the processes are random, their Fourier transforms are also random and the following theorem says something about their correlation structure.

Theorem 2.2 (Fourier transform correlation)
Let $X(f) = \sum_{k=-\infty}^{\infty} x_k e^{-j2\pi fk}$ *and* $Y(f) = \sum_{k=-\infty}^{\infty} y_k e^{-j2\pi fk}$. *Then*

$$E\, X(f)Y^*(f_1) \;=\; S_{xy}(f)\,\delta_1(f \Leftrightarrow f_1) \tag{2.26}$$

where we have introduced the impulse train

$$\delta_{f_0}(f) \;=\; \sum_{n=-\infty}^{\infty} \delta(f \Leftrightarrow n f_0) \tag{2.27}$$

and the superscript * denotes complex conjugate. In particular, $E\, Y(f)Y^*(f_1) \;=\; S_{yy}(f)\,\delta_1(f \Leftrightarrow f_1)$ which means that a random process $y_k \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} Y(f)e^{j2\pi fk}$ can be regarded as a superposition of exponentials $e^{j2\pi fk}$ with random complex amplitudes $Y(f)$ that are uncorrelated at different frequencies and that have a power proportional to $S_{yy}(f)$.

Proof:

$$\begin{aligned} E\,X(f)Y^*(f_1) &= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} e^{-j2\pi fn}e^{j2\pi f_1 m}\,E\,x_n y_m \;=\; \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} e^{-j2\pi fn}e^{j2\pi f_1 m}\,r_{xy}(n\Leftrightarrow m) \\ &= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} e^{-j2\pi fn}e^{j2\pi f_1 m}\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xy}(f_2)e^{j2\pi f_2(n-m)}df_2 \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} df_2 S_{xy}(f_2)\left(\sum_{n=-\infty}^{\infty} e^{j2\pi n(f_2-f)}\right)\left(\sum_{m=-\infty}^{\infty} e^{j2\pi m(f_1-f_2)}\right) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} df_2 S_{xy}(f_2)\delta_1(f\Leftrightarrow f_2)\,\delta_1(f_2\Leftrightarrow f_1) \;=\; S_{xy}(f)\,\delta_1(f\Leftrightarrow f_1) \end{aligned} \tag{2.28}$$

where we used the facts that the Fourier transform of $e^{j2\pi f_2 k}$ is $\sum_k e^{j2\pi k(f_2-f)} = \delta_1(f\Leftrightarrow f_2)$ and that $\delta_1(f)$ contains exactly one peak in the interval $(\Leftrightarrow\frac{1}{2}, \frac{1}{2}]$. □

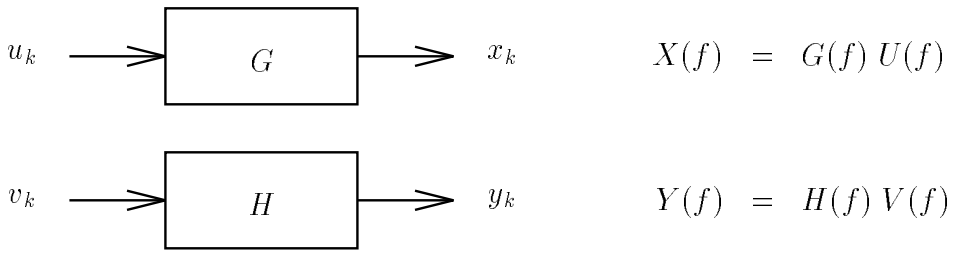

Figure 2.3: Transformation of the cross-psdf by linear filtering.

Using the result of theorem 2.2, it is straightforward to show the effect on the auto- or cross-psdf of linear filtering. Indeed consider the processes u_k, v_k, x_k and y_k which are related by linear filtering as shown in Fig. 2.3. Then we get

$$\begin{aligned} S_{xy}(f)\delta_1(f\Leftrightarrow f_1) &= E\,X(f)Y^*(f_1) \;=\; G(f)\;(E\,U(f)V^*(f_1))\,H^*(f_1) \\ &= G(f)S_{uv}(f)H^*(f_1)\delta_1(f\Leftrightarrow f_1)\;. \end{aligned} \tag{2.29}$$

Integrating both sides, $\int_{-\frac{1}{2}}^{\frac{1}{2}} df_1$, yields

$$S_{xy}(f) \;=\; G(f)S_{uv}(f)H^*(f)\;. \tag{2.30}$$

Using the relations in (2.17), we can extend this property to the z-transforms:

$$\mathrm{S}_{xy}(z) \;=\; \mathrm{G}(z)\mathrm{S}_{uv}(z)\mathrm{H}(1/z) \tag{2.31}$$

from which (2.30) would also follow by taking $z = e^{j2\pi f}$.

In this course, we shall almost exclusively deal with real scalar valued random processes. However, for completeness, we give here some extensions to complex vector valued processes. If x_k and y_k are two jointly WSS complex (column-)vector valued processes (not necessarily of

the same dimension), then their cross-correlation and cross-covariance functions are defined as

$$r_{xy}(k) = E\, x_{n+k} y_n^H \ , \ \ c_{xy}(k) = E\,(x_{n+k} \Leftrightarrow m_x)(y_n \Leftrightarrow m_y)^H \ = \ r_{xy}(k) \Leftrightarrow m_x\, m_y^H \tag{2.32}$$

where superscript ^{H} denotes Hermitian transpose (complex conjugate transpose): $y_k^H = y_k^{*T}$. Even though $r_{xy}(k)$ is a matrix in this case, we continue to call it the correlation *function* since it is a function of the correlation lag k. We reserve the term correlation *matrix* for the case where we consider $R_{XY} = E\, X\, Y^T$ (or $E\, X\, Y^H$ in the complex case) where X and Y are in some sense unique random vectors, and not samples of vector valued random processes, or in the case where X and Y are samples of vector valued random processes x_n and y_n but we only consider $R_{XY} = r_{xy}(0)$, i.e. we only consider the correlation function at lag zero. For the z-transforms, we get the relationship

$$\mathrm{S}_{xy}(z) \ = \ \mathrm{G}(z)\mathrm{S}_{uv}(z)\mathrm{H}^\dagger(z) \ , \quad \text{where } \ \mathrm{H}^\dagger(z) = \mathrm{H}^H(1/z^*) \tag{2.33}$$

denotes the *paraconjugate* version of $\mathrm{H}(z)$.

We now turn to the spectral estimation problem. Some applications of spectral estimation are:

- The spectrum transmitted by modems needs to satisfy certain restrictions, especially if the modem needs to comply with a certain standard. For instance voiceband modems over the telephone line need to be restricted to the [300Hz,3400Hz] band. Wireless modems or any radio transmitter needs to occupy a rectricted bandwidth since the radio medium is scarce. In these considerations, not only the (effective) bandwidth of the signal is of importance but also the spectral roll-off. Typically, the emitted spectrum needs to fall under a certain spectral mask (plot of the maximum allowed power spectral density as a function of frequency). To verify this, the spectrum of the emitted signal needs to be estimated.

- We shall see that the optimization of the parameters of most source coding techniques depends on the spectrum or equivalently the correlation structure of the source to be coded. Hence, spectral estimation is again required. Such sources can be speech, audio, images or video.

- Spectrum estimation also allows to evaluate certain parameters about transmitted signals such as the frequency offset and the frequency-selective distortion introduced by the channel, or the spectrum of the interfering signals or noise.

There exist two big methodologies for estimating the psdf of a stationary process: *non-parametric* techniques and *parametric* techniques. In non-parametric or *classical* spectral estimation techniques, no constraints are imposed on the possible form of $S_{yy}(f)$. This implies that an infinite number of degrees of freedom need to be estimated: the $r_{yy}(k)$. This will lead to a very high estimation variance if we have to work with a finite amount of data. In the parametric or *high-resolution* techniques, a parametric model is assumed for $S_{yy}(f)$ and the spectral estimatiom problem reduces to the estimation of the parameters describing the parametric model. Since the parametric form can only perfectly model a limited class of functions, there will be some bias in the estimate of $S_{yy}(f)$. However, to compensate for this bias, a (large) reduction in variance is achievable compared to the non-parametric techniques.

2.2 Non-parametric Spectral Estimation

The first non-parametric technique is based on the Fourier transform. Consider the Fourier transform $Y(f)$ of the WSS process y_k with zero mean. $Y(f)$ is a random function with mean

$$E\,Y(f) \;=\; \sum_{k=-\infty}^{\infty} e^{-j2\pi fk}\,\underbrace{E\,y_k}_{=0} \;=\; 0\;. \tag{2.34}$$

The correlation between the Fourier transform at frequencies f and f_1 is

$$E\,Y(f)Y^*(f_1) \;=\; S_{yy}(f)\delta_1(f \Leftrightarrow f_1)\;. \tag{2.35}$$

So the Fourier transform at different frequencies is uncorrelated, while its magnitude squared is proportional to the power spectral density function.

2.2.1 The Periodogram

In practice, we are given only a finite number of N samples $\{y_0, y_1, \ldots, y_{N-1}\}$. The periodogram was introduced by Schuster in 1898 and is defined as the scaled magnitude squared of the Fourier transform of this finite number of samples:

$$\widehat{S}_{yy}(f) \;=\; \widehat{S}_{PER}(f) \;=\; \frac{1}{N}\left|\sum_{n=0}^{N-1} y_n\, e^{-j2\pi fn}\right|^2 \tag{2.36}$$

where the normalization factor $\frac{1}{N}$ has been introduced to avoid that things go to infinity as $N \to \infty$ (compare to (2.35)). Since we want to investigate what happens for a particular frequency f, let us denote f as f_0. We can interpret the periodogram by rewriting it as follows

$$\widehat{S}_{PER}(f_0) \;=\; N\left|\sum_{n=0}^{N-1} h^o_{k-n} y_n\right|^2_{k=0} \tag{2.37}$$

where $h^o_k = h_k e^{j2\pi f_0 k}$ with

$$h_k \;=\; \begin{cases} \frac{1}{N} & ,\; k = \Leftrightarrow(N \Leftrightarrow 1), \ldots, \Leftrightarrow 1, 0 \\ 0 & ,\; \text{otherwise.} \end{cases} \tag{2.38}$$

h^o_k is an anticausal FIR filter consisting of a modulated rectangular window, see Fig. 2.4.

The frequency responses of the filter h_k and h^o_k are related by $H^o(f) = H(f \Leftrightarrow f_0)$ and

$$\begin{aligned} H(f) \;&=\; \sum_{k=-\infty}^{\infty} h_k\, e^{-j2\pi fk} \;=\; \frac{1}{N}\sum_{k=-(N-1)}^{0} e^{-j2\pi fk} \;=\; \frac{1}{N}\sum_{k=0}^{N-1} e^{j2\pi fk} \;=\; \frac{1}{N}\frac{1 \Leftrightarrow e^{j2\pi fN}}{1 \Leftrightarrow e^{j2\pi f}} \\ &=\; \frac{1}{N}\frac{e^{-j\pi fN} \Leftrightarrow e^{j\pi fN}}{e^{-j\pi f} \Leftrightarrow e^{j\pi f}}\,\frac{e^{j\pi fN}}{e^{j\pi f}} \;=\; \frac{\sin N\pi f}{N \sin \pi f}\, e^{j(N-1)\pi f} \end{aligned} \tag{2.39}$$

$H(f)$ is of course periodic with period 1. It satisfies furthermore

$$\begin{aligned} H(k) &= 1 \quad ,\; k \in \mathcal{Z} \\ H(\tfrac{k}{N}) &= 0 \quad ,\; k \in \mathcal{Z} \setminus N\mathcal{Z}\;. \end{aligned} \tag{2.40}$$

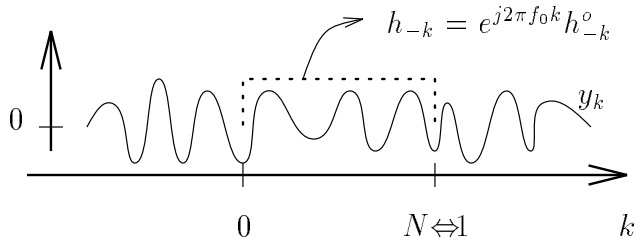

Figure 2.4: The modulated rectangular window.

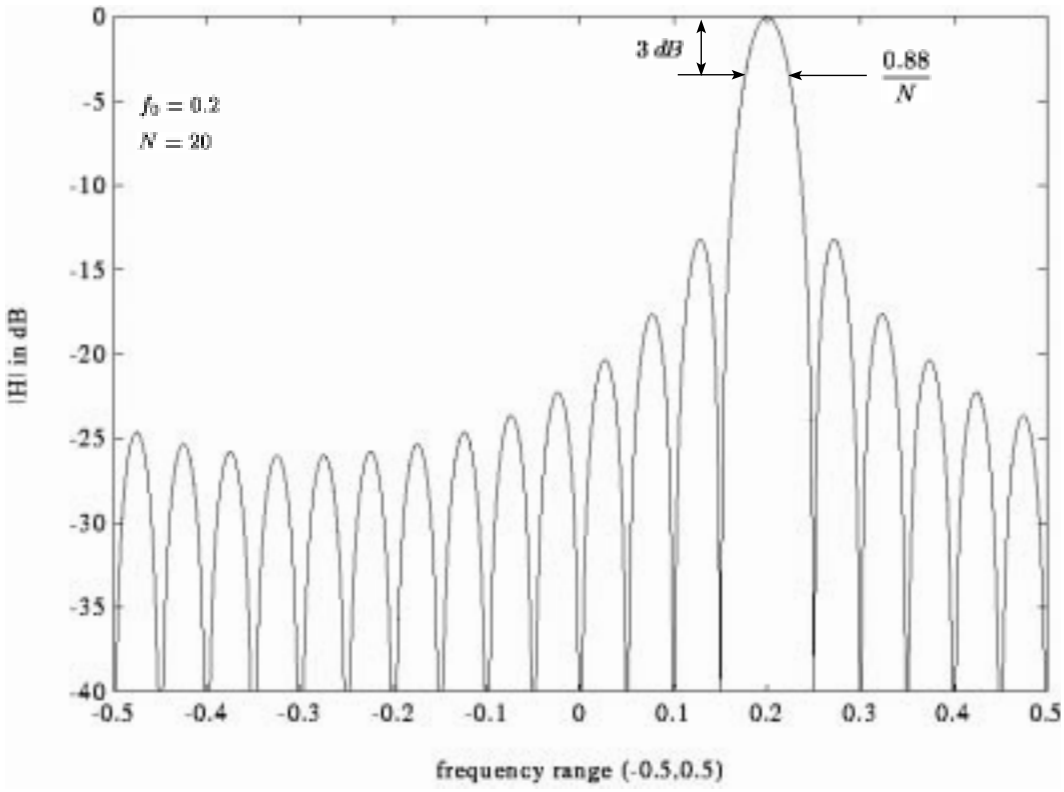

Figure 2.5: The periodogram bandpass filter $|H^o(f)| = |H(f \Leftrightarrow f_0)|$ for $f_0 = 0.2$ and $N = 20$.

$H^o(f) = H(f \Leftrightarrow f_0)$ is the frequency response of a bandpass filter with center frequency f_0, see Fig. 2.5. The 3dB bandwidth is $\frac{0.88}{N}$ which allows us to say that loosely speaking, the bandwidth is $\frac{1}{N}$.

Hence the periodogram estimates the power in $\{y_k\}$ at the frequency f_0 by filtering the data with a bandpass filter, sampling the output at time $k = 0$, and computing the magnitude squared. When multiplied by N $(= \frac{1}{\Delta f} = \frac{1}{1/N})$ (to account for the bandwidth of the bandpass filter), this yields the power spectral density estimate $\widehat{S}_{PER}(f_0)$ (compare to the psdf interpretation).

It might be supposed that if enough data are available, say $N \to \infty$, then for any f_0, $\widehat{S}_{PER}(f_0) \to S_{yy}(f_0)$ where one should keep in mind that $\widehat{S}_{PER}(f_0)$ is a random variable for every f_0 while $S_{yy}(f_0)$ is a deterministic function of f_0. Such a convergence would mean that the Periodogram would be a consistent estimator of the psdf. To test this conjecture, we consider in Fig. 2.6 the periodogram of real zero-mean white Gaussian noise for several increasing record lengths N. It appears that the random fluctuations or variance of the periodogram does not decrease with N, prompting us to conclude that the periodogram is *not* a consistent estimator of the psdf. Nevertheless, for this white noise example, the periodogram appears to fluctuate around a constant value (the true constant value of the psdf), which would imply that the periodogram is unbiased for white noise. But the variance does not tend to zero as $N \to \infty$.

Intuitively, in order to have a consistent estimator of a set of parameters, we need to have lots more data than the number of parameters to be estimated. The number of parameters is infinite here. Indeed, the Wiener-Khinchin relation (2.10) can also be written as

$$S_{yy}(f) \;=\; r_{yy}(0) + 2\sum_{n=1}^{\infty} r_{yy}(n)\,\cos(2\pi f n) \tag{2.41}$$

which shows that the $\{r_{yy}(k),\, k \geq 0\}$ can be considered as the infinite set of parameters parameterizing $S_{yy}(f)$. The number of data points N, even if it goes to infinity, cannot largely exceed the number of parameters. Hence, the variance in estimating those parameters cannot go to zero.

We now investigate in some detail the mean and variance of the periodogram. First consider the mean. We get for the convolution of h_k^o and y_k

$$\sum_n h_{k-n}^o y_n \;=\; h_k^o * y_k \;=\; \mathcal{F}^{-1}\left\{H^o(f)Y(f)\right\} \tag{2.42}$$

where h_k^o and $H^o(f) = H(f \Leftrightarrow f_0)$ depend on the particular frequency f_0 we are investigating and $\mathcal{F}^{-1}\{.\}$ denotes inverse Fourier transform. From (2.42), we get

$$\left[\sum_{n=0}^{N-1} h_{k-n}^o y_n\right]_{k=0} \;=\; \left[\int_{-\frac{1}{2}}^{\frac{1}{2}} H^o(f)Y(f)e^{j2\pi fk}df\right]_{k=0} \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} H^o(f)Y(f)df\ . \tag{2.43}$$

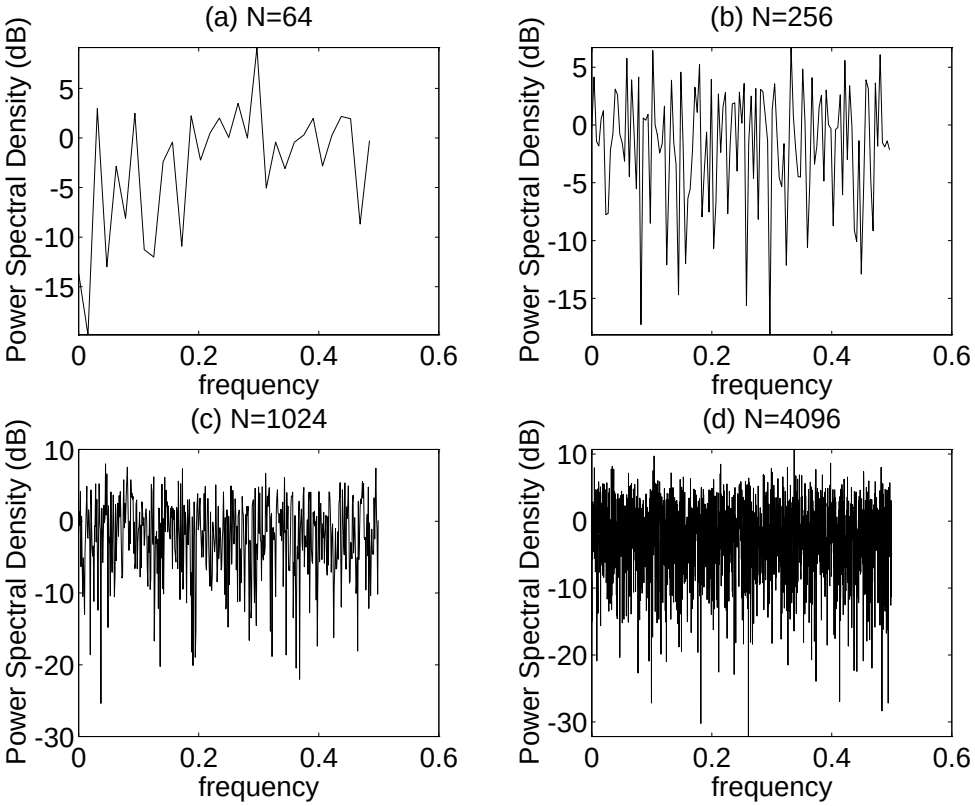

Figure 2.6: Illustrating the inconsistency of the periodogram for white Gaussian noise ($\sigma_y^2 = 1$): $10 \log_{10} \widehat{S}_{PER}(f)$ for (a) $N = 64$, (b) $N = 256$, (c) $N = 1024$, (d) $N = 4096$.

So we get

$$\begin{aligned} E\,\widehat{S}_{PER}(f_0) &= N\,E\left|\left[\sum_{n=0}^{N-1} h^o_{k-n} y_n\right]_{k=0}\right|^2 = N\,E\left|\int_{-\frac{1}{2}}^{\frac{1}{2}} H^o(f)Y(f)df\right|^2 \\ &= N\,E\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} H^o(f)Y(f)df\right)\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} H^{o*}(f_1)Y^*(f_1)df_1\right) \\ &= N\int_{-\frac{1}{2}}^{\frac{1}{2}} df H^o(f)\int_{-\frac{1}{2}}^{\frac{1}{2}} df_1 H^{o*}(f_1)\underbrace{E\,Y(f)Y^*(f_1)}_{=S_{yy}(f)\delta_1(f-f_1)} \\ &= N\int_{-\frac{1}{2}}^{\frac{1}{2}} df H^o(f)S_{yy}(f)\underbrace{\int_{-\frac{1}{2}}^{\frac{1}{2}} df_1 H^{o*}(f_1)\delta_1(f \Leftrightarrow f_1)}_{=H^{o*}(f)} \\ &= N\int_{-\frac{1}{2}}^{\frac{1}{2}} |H^o(f)|^2 S_{yy}(f)df = N\int_{-\frac{1}{2}}^{\frac{1}{2}} |H(f \Leftrightarrow f_0)|^2 S_{yy}(f)df \\ &= N\int_{-\frac{1}{2}}^{\frac{1}{2}} |H(f_0 \Leftrightarrow f)|^2 S_{yy}(f)df \end{aligned} \tag{2.44}$$

Or hence,

$$E\,\widehat{S}_{PER}(f_0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} W_B(f_0 \Leftrightarrow f)S_{yy}(f)df = W_B(f_0) * S_{yy}(f_0) \tag{2.45}$$

where

$$W_B(f) = N\,|H(f)|^2 = \frac{1}{N}\left(\frac{\sin \pi f N}{\sin \pi f}\right)^2 . \tag{2.46}$$

Since $W_B(f)$ is (apart from the scaling factor N) the square of the Fourier transform of a rectangular window, it is also the Fourier transform of the convolution of a rectangular window with itself which is a triangular window. Hence

$$w_B(k) = \mathcal{F}^{-1}\{W_B(f)\} = \begin{cases} 1 \Leftrightarrow \frac{|k|}{N} & , |k| \le N \Leftrightarrow 1 \\ 0 & , |k| \ge N\ . \end{cases} \tag{2.47}$$

Such a window is called a *Bartlett window*, see Fig. 2.7. From (2.46) we conclude that the average periodogram is the convolution of the true psdf with the Fourier transform of the Bartlett window, yielding on the average a smoothed version of the psdf.

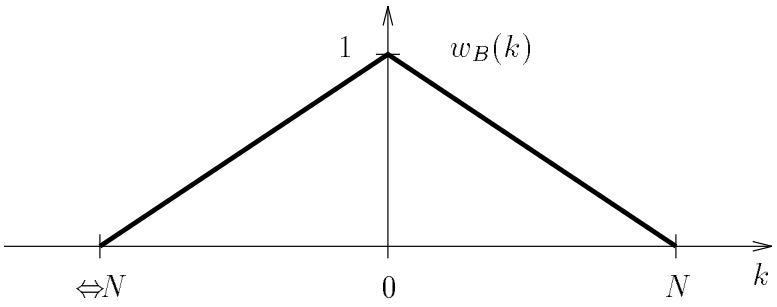

Figure 2.7: The Bartlett window.

Now since in general $W_B(f) * S_{yy}(f) \neq S_{yy}(f)$, the periodogram is biased in general for finite data records. However, the periodogram is asymptotically unbiased. Indeed

$$\begin{aligned} \lim_{N\to\infty} E\,\widehat{S}_{PER}(f) &= \lim_{N\to\infty} W_B(f) * S_{yy}(f) = \lim_{N\to\infty} \mathcal{F}\{w_B(k) r_{yy}(k)\} \\ &= \mathcal{F}\{\underbrace{\lim_{N\to\infty} w_B(k)}_{=1}\, r_{yy}(k)\} = \mathcal{F}\{r_{yy}(k)\} = S_{yy}(f)\,. \end{aligned} \tag{2.48}$$

Note that

$$w_B(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} W_B(f)df = 1\,,\ \ W_B(0) = N\,,\ \ W_B(f) \geq 0\,,\ \ 6dB \text{ bandwidth of } W_B(f) = \frac{0.88}{N}\,. \tag{2.49}$$

As $N \to \infty$, $W_B(f)$ becomes a narrow and high peak of constant surface. So $W_B(f) \to \delta(f)$. For the special case of white noise, the periodogram is unbiased even for finite data records.

As far as the covariance of the periodogram is concerned, the following can be shown exactly for white Gaussian noise. The same result holds true approximately for more general processes when the data record is large.

$$Cov\left[\widehat{S}_{PER}(f_1), \widehat{S}_{PER}(f_2)\right] \approx S_{yy}(f_1)S_{yy}(f_2)\left[W_B(f_1+f_2) + W_B(f_1 \Leftrightarrow f_2)\right]\frac{1}{N}\,. \tag{2.50}$$

The variance at the frequency f then follows as

$$Var\left[\widehat{S}_{PER}(f)\right] = Cov\left[\widehat{S}_{PER}(f), \widehat{S}_{PER}(f)\right] \approx S_{yy}^2(f)\left[1 + \frac{W_B(2f)}{N}\right] \geq S_{yy}^2(f)\,. \tag{2.51}$$

For frequencies not near 0 or $\pm\frac{1}{2}$, we can further approximate as

$$Var\left[\widehat{S}_{PER}(f)\right] \approx S_{yy}^2(f) \tag{2.52}$$

which is independent of the record length N! This means that the standard deviation of the periodogram is about equal to the psdf, the quantity to be estimated by the periodogram. Furthermore, modulo certain conditions, the following result holds exactly

$$\lim_{N\to\infty} Var\left[\widehat{S}_{PER}(f)\right] = \begin{cases} 2\,S_{yy}^2(f) & ,\ f = 0, \frac{1}{2} \\ S_{yy}^2(f) & ,\ f \in (0, \frac{1}{2})\,. \end{cases} \tag{2.53}$$

Hence the periodogram is an unreliable estimator since its standard deviation is as large as the (non-negative) quantity to be estimated.

From (2.50), we get also

$$Cov\left[\widehat{S}_{PER}(f_1), \widehat{S}_{PER}(f_2)\right] \approx 0 \text{ if } f_1 \neq f_2 \text{ are integer multiples of } \tfrac{1}{N}. \tag{2.54}$$

The values of the periodogram at integer multiples of $\frac{1}{N}$ are uncorrelated. Coupled with the constant variance (as a function of N), this implies that as N increases the periodogram fluctuates rapidly as illustrated in Fig. 2.6. In fact, it can also be shown that

$$\lim_{N\to\infty} Cov\left[\widehat{S}_{PER}(f_1), \widehat{S}_{PER}(f_2)\right] = 0\,,\ \ f_1 \neq f_2 \tag{2.55}$$

so that asymptotically, the periodogram is uncorrelated at different frequencies, just like the Fourier transform of the whole realization y_k.

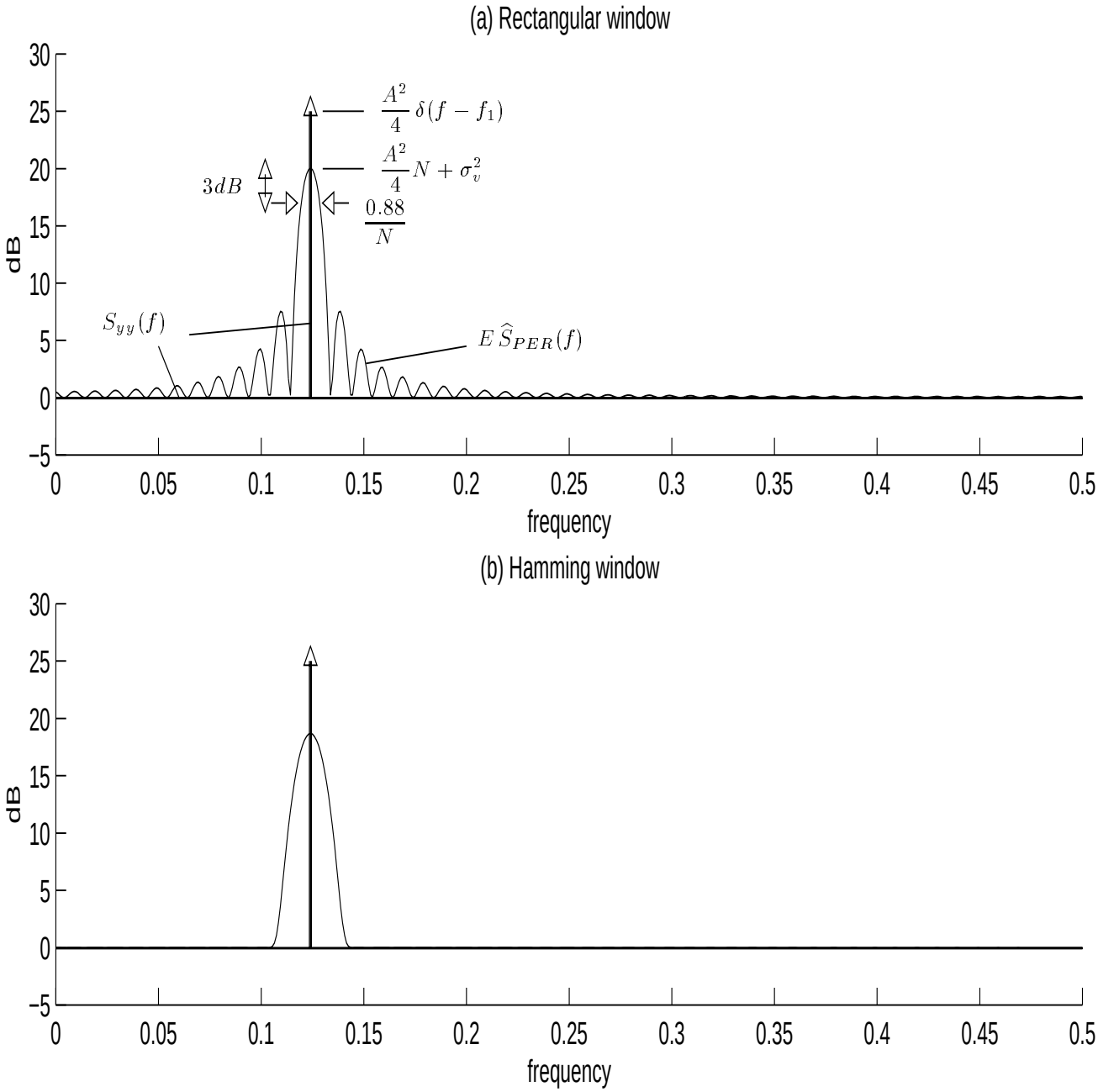

Figure 2.8: Illustrating the spectral leakage of the periodogram: the mean of the periodogram of a sinusoid with amplitude 2 and frequency 0.125 in white noise of unit variance, and the true psdf. $N = 100$. In (a) the usual periodogram with a rectangular window is shown, while in (b) the windowed periodogram with a Hamming window appears.

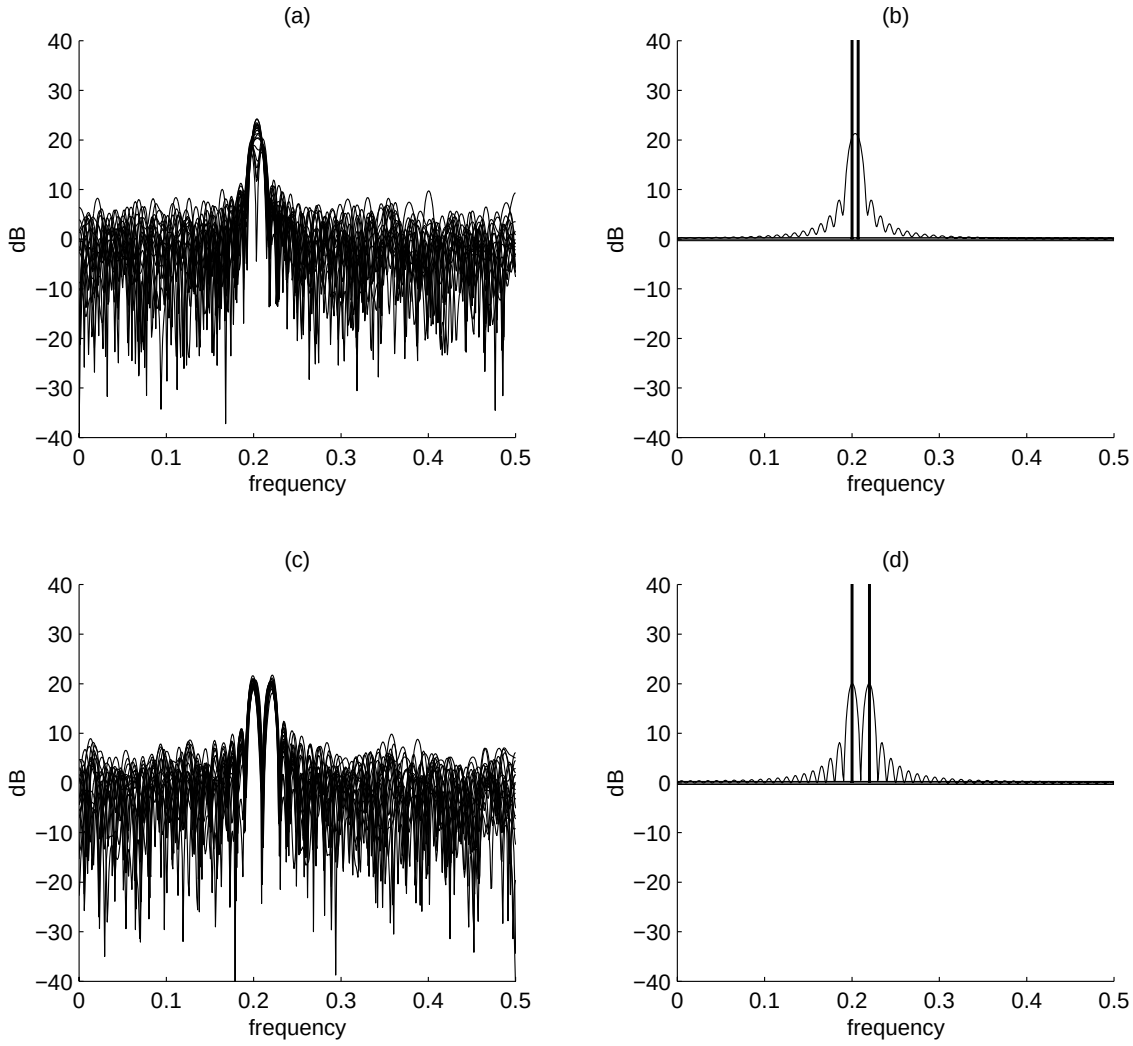

Figure 2.9: Illustrating the resolution of the periodogram: the periodogram of two sinusoids of amplitude 2 in white noise of unit variance. $N = 100$. The frequences are 0.2 and 0.207 in (a) and (b) and 0.2 and 0.22 in (c) and (d). (a) and (c) show the overlay of the periodogram for 20 realizations, while (b) and (d) show the mean of the periodogram and the true psdf in bold lines.

Spectral Leakage and Data Weighting

Consider a random process y_k consising of a (complex) sinusoid with frequency f_0, $y_k = e^{j2\pi f_0 k+\phi}$. This signal is random due to the random phase ϕ which we take to be uniformly distributed over $[0, 2\pi)$. This process can be shown to be WSS since its zero mean does not depend on time and its autocorrelation function $r_{yy}(n) = E\, y_{k+n}y_k^* = e^{j2\pi f_0 n}$ only depends on the time difference. The psdf of this process is hence $S_{yy}(f) = \delta(f \Leftrightarrow f_0)$ which is only non-zero for $f = f_0$. The mean of the periodogram for this process would be $W_B(f \Leftrightarrow f_0) = N|H^o(f)|^2$ where $|H^o(f)|$ was sketched in Fig. 2.5. We remark the presence of strong sidelobes. The effect of these sidelobes is that the average periodogram has non-zero contributions at frequencies where the psdf has no contributions. This phenomenon is called *spectral leakage*.

Consider similarly a process consisting of a real sinusoid with frequency f_1 and uniformly distributed random phase ϕ plus white noise with unit variance, $y_k = A\cos(2\pi f_1 k + \phi) + v_k$, $\sigma_v^2 = 1$. This process is again WSS with zero mean, autocorrelation function $r_{yy}(n) = \frac{A^2}{2}\cos(2\pi f_1 n) + \sigma_v^2\,\delta_{n0}$ and psdf $S_{yy}(f) = \sigma_v^2 + \frac{A^2}{4}(\delta(f \Leftrightarrow f_1) + \delta(f + f_1))$. The spectrum and the periodogram mean for such a process are sketched in Fig. 2.8(a). Again, the psdf contribution at $f = f_1$ can be seen to be leaking to neighboring frequencies due to the sidelobes of $|H(f)|^2$.

In order to reduce this spectral leakage, it is advantageous to replace the rectangular window with another window that shows less discontinuities near the edges and hence that has weaker sidelobes. The price to pay is that the bandwidth of the main lobe increases, reducing the *resolution*. Resolution is the ability to observe distinct spectral peaks for two comparable level sinusoids. For maximum resolution, no data windowing should be used. Considering the 3 dB bandwidth of $|H(f)|$, a common rule of thumb is that two equiamplitude sinusoids are resolvable if their normalized frequencies are spaced more than $1/N$ apart. In Fig. 2.9, the periodogram of several realizations of two real sinusoids in white noise and the periodogram mean are plotted for two cases of frequency separation. In the first case, $f_2 \Leftrightarrow f_1 = \frac{0.7}{N} < \frac{1}{N}$ and we note that often only a single peak can be observed in the periodogram (see the periodogram mean also). In the second case, $f_2 \Leftrightarrow f_1 = \frac{2}{N}\frac{1}{N}$ and two separate peaks can be observed for all realizations. If the data consists of only one sinusoid (or several sinusoids spaced much more than $1/N$ apart) embedded in white Gaussian noise, then the optimal (ML) frequency estimator is the periodogram without data windowing.

Since the windows used in practice are symmetric w.r.t. the center of the data record, we shall assume for the rest of this windowing discussion that $N = 2M+1$ and that the available data are $y_{-M}, \ldots, y_M$. Because of the weighting function of the window, h_k will now be denoted by w_k. So we apply a symmetric window $w_k = w_{-k}$ before computing the periodogram. The weighted periodogram becomes:

$$\widehat{S}_{PER,w}(f) \;=\; \frac{1}{2M+1}\left|\sum_{k=-M}^{M} w_k y_k e^{-j2\pi fk}\right|^2 \quad . \tag{2.56}$$

Some popular windows are given in Table 2.1 and their shape and Fourier transform magnitude are sketched in Figures 2.10 and 2.11 respectively. We note in the table that as the sidelobe level decreases, the bandwidth of the main lobe tends to increase. We note in the figures that as the window becomes more continuous in the time domain, it decreases faster in the frequency domain. Applying the windowed periodogram to a sinusoid in white noise, we find

the periodogram mean in Fig. 2.8(b) for the Hamming window. We that in this example that the leakage has been eliminated in the sense that no spurious components appear at frequencies different from f_1. The price to pay is an increased *smearing* of the periodogram due to the increased bandwidth of the main lobe: the signal contribution at f_1 gets smeared out over a wider bandwidth.

Name	Definition: w_k =	Fourier transform: $W(f)$ =	$\mathcal{B}_{3dB}$	A_{SL}
Rectangular	1	$W_R(f) = \frac{\sin \pi f(2M+1)}{\sin \pi f}$	$\frac{0.88}{N}$	-13
Bartlett	$1-\frac{\|k\|}{M}$	$W_B(f) = \frac{1}{M}\left(\frac{\sin \pi fM}{\sin \pi f}\right)^2$	$\frac{1.28}{N}$	-27
Hanning	$\frac{1}{2}+\frac{1}{2}\cos\frac{\pi k}{M}$	$\frac{1}{2}W_R(f)+$ $\frac{1}{4}(W_R(f-\frac{1}{2M})+W_R(f+\frac{1}{2M}))$	$\frac{1.44}{N}$	-32
Hamming	$0.54+0.46\cos\frac{\pi k}{M}$	$0.54W_R(f)+$ $0.23(W_R(f-\frac{1}{2M})+W_R(f+\frac{1}{2M}))$	$\frac{1.30}{N}$	-43
Parzen (M even)	$\begin{cases} 2(1-\frac{\|k\|}{M})^3-(1-2\frac{\|k\|}{M})^3, & \|k\| \le \frac{M}{2} \\ 2(1-\frac{\|k\|}{M})^3, & \frac{M}{2} < \|k\| \le M \end{cases}$	$\frac{8}{M^3}\left(\frac{3}{2}\frac{\sin^4 \pi fM/2}{\sin^4 \pi f}-\frac{\sin^4 \pi fM/2}{\sin^2 \pi f}\right)$	$\frac{1.84}{N}$	-53
Blackman	$0.42+0.5\cos\frac{\pi k}{M}+0.08\cos\frac{2\pi k}{M}$	$0.42W_R(f)+$ $0.25(W_R(f-\frac{1}{2M})+W_R(f+\frac{1}{2M}))$ $+0.04(W_R(f-\frac{1}{M})+W_R(f+\frac{1}{M}))$	$\frac{1.68}{N}$	-58

Table 2.1: Common data windows. Every window satisfies $w_k = 0$, $|k| > M$. $\mathcal{B}_{3dB}$ is the 3dB bandwidth and A_{SL} is the sidelobe attenuation in dB.

Bias Reduction by Data Prewhitening

If the psdf is constant, then the periodogram is unbiased, even for finite data record length N. This leads to the following idea for bias reduction. Suppose we have some a priori idea about the psdf (or perhaps, after a first application of the periodogram, we have such information). Then we may design a filter $A(f)$ such that the result x_k obtained by filtering y_k with $A(f)$ is approximately white (the technique of linear prediction that we shall see further provides one convenient way of doing this). Then we determine the periodogram $\widehat{S}_{xx,PER}(f)$ of the filtered data. The bias in $\widehat{S}_{xx,PER}(f)$ will be small since x_k is approximately white. Then as $S_{yy}(f) = \frac{S_{xx}(f)}{|A(f)|^2}$, the psdf estimate for the original data y_k can be taken to be

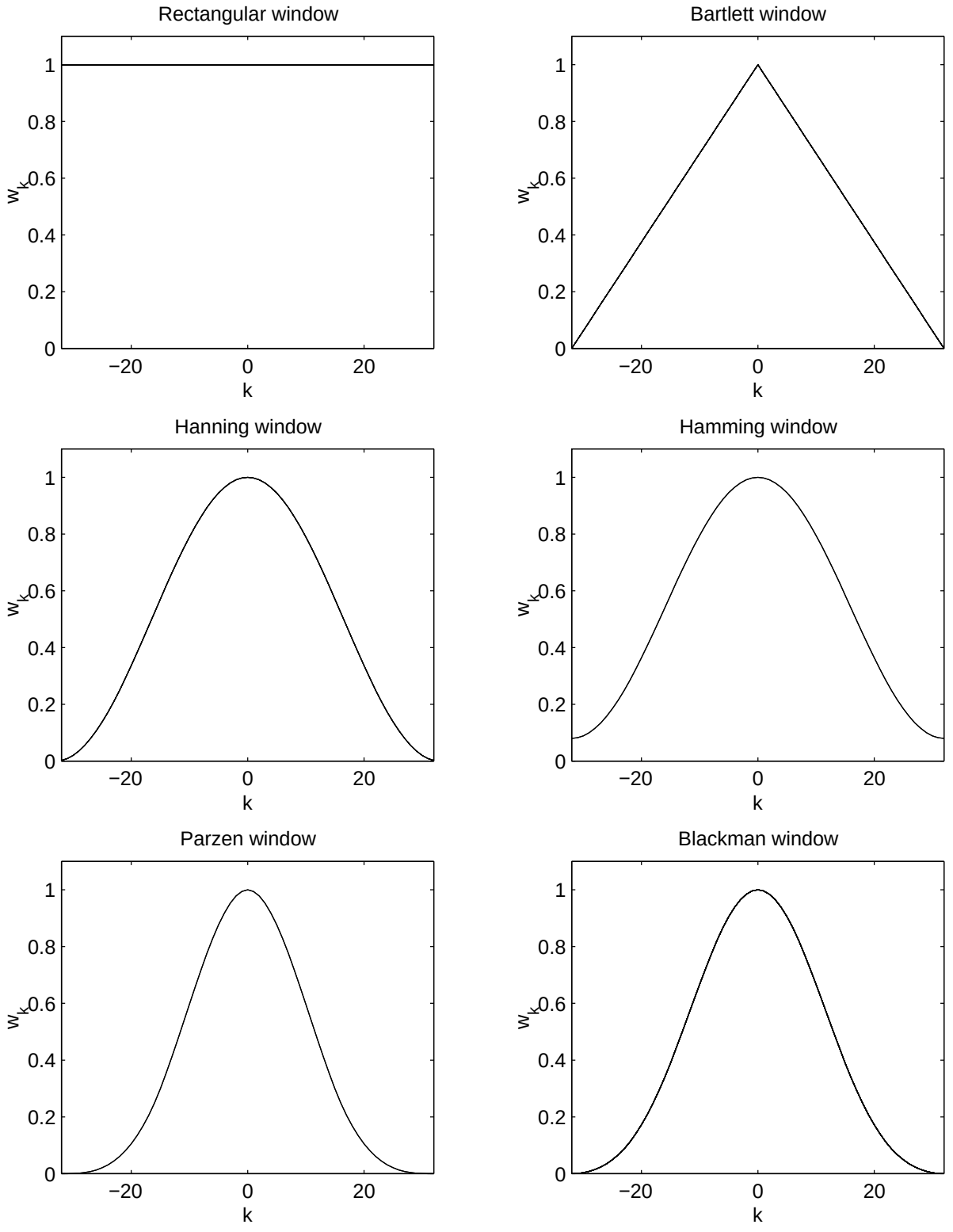

Figure 2.10: Illustrating various window functions ($N = 65$, $M = 32$).

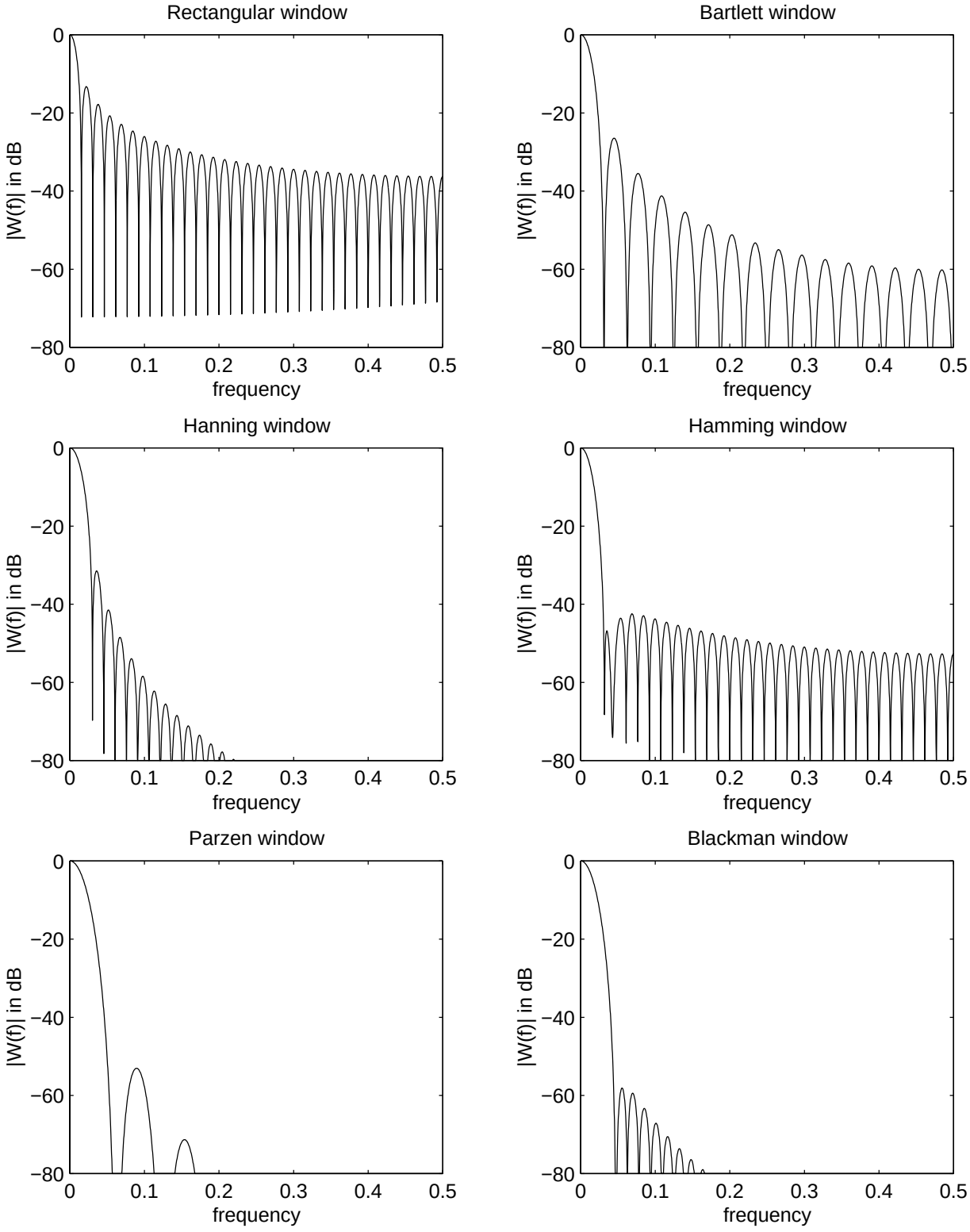

Figure 2.11: Fourier transform magnitude of various window functions ($N = 65$, $M = 32$).

$\widehat{S}_{yy}(f) = \frac{\widehat{S}_{xx,PER}(f)}{|A(f)|^2}$. Then the bias in $\widehat{S}_{yy}(f)$

$$E\,\widehat{S}_{yy}(f) \Leftrightarrow S_{yy}(f) \;=\; \frac{\widehat{S}_{xx,PER}(f) \Leftrightarrow S_{xx}(f)}{|A(f)|^2} \tag{2.57}$$

is expected to be roughly a constant small fraction of $S_{yy}(f)$.

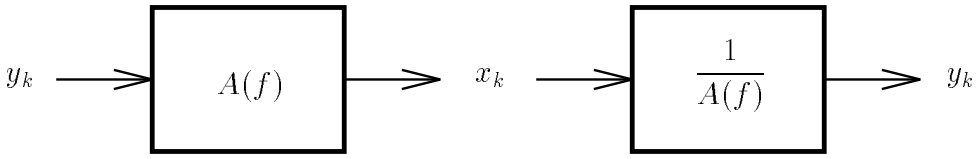

Figure 2.12: Data prewhitening.

Use of the DFT and Zero Padding

The computation of the periodogram strictly speaking requires the evaluation of the Fourier transform of the windowed data over a continuum of frequencies. The computer can only handle numbers and hence a discrete set of frequencies. The existence of an efficient algorithm, the Fast Fourier Transform (FFT), incites us to use the Discrete Fourier Transform (DFT). For consideration of the DFT/FFT, it is customary to consider the frequency interval $[0, 1]$ rather than $[\Leftrightarrow\frac{1}{2}, \frac{1}{2}]$. The DFT evaluates the Fourier transform of a signal of length N at N equispaced frequencies $f_k = k/N$, $k = 0, 1, \ldots, N\Leftrightarrow 1$. So we get

$$\widehat{S}_{PER}(f_k) \;=\; \frac{1}{N}\left|\sum_{n=0}^{N-1} y_n\, e^{-j2\pi f_k n}\right|^2 \;=\; \frac{1}{N}\left|\sum_{n=0}^{N-1} y_n\, e^{-j2\pi \frac{k}{N} n}\right|^2 \quad , \quad k = 0, 1, \ldots, N\Leftrightarrow 1\ . \tag{2.58}$$

We can obtain a finer frequency spacing by padding the data with $N'\Leftrightarrow N$ zeros and then applying and N'-point DFT. The effective data set becomes

$$y'_n \;=\; \begin{cases} y_n & ,\ n = 0, 1, \ldots, N\Leftrightarrow 1 \\ 0 & ,\ n = N, N{+}1, \ldots, N'\Leftrightarrow 1 \end{cases} \tag{2.59}$$

which has the same Fourier transform as the original data set. The frequency spacing of the DFT on the data set in (2.59) will be $\frac{1}{N'} < \frac{1}{N}$. No extra resolution is offered by zero padding, only a finer evaluation of the periodogram.

2.2.2 The Averaged Periodogram

The main problem with the periodogram is its large variance. Expression (2.35) may inspire us to introduce an averaging operation in order to reduce the variance. Assume we have K independent data records of length L available (K realizations of the same process). So the

data in record i are $y_n^{(i)}$, $n = 0, 1, \ldots, L\Leftrightarrow 1$ for $i = 0, 1, \ldots, K\Leftrightarrow 1$. The *averaged periodogram* is defined as

$$\widehat{S}_{AVPER}(f) \;=\; \frac{1}{K}\sum_{i=0}^{K-1} \widehat{S}_{PER}^{(i)}(f) \tag{2.60}$$

where $\widehat{S}_{PER}^{(i)}(f)$ is the periodogram for data record i:

$$\widehat{S}_{PER}^{(i)}(f) \;=\; \frac{1}{L}\left|\sum_{n=0}^{L-1} y_n^{(i)} e^{-j2\pi f n}\right|^2 \quad . \tag{2.61}$$

Since the data records are i.i.d., the mean of the averaged periodogram is also the mean of the periodogram of any record. Hence

$$E\,\widehat{S}_{AVPER}(f) \;=\; E\,\widehat{S}_{PER}^{(0)}(f) \;=\; W_{B,L}(f) * S_{yy}(f)\;, \quad W_{B,L}(f) = \frac{1}{L}\left(\frac{\sin \pi f L}{\sin \pi f}\right)^2 \quad . \tag{2.62}$$

The variance on the other hand will be decreased by a factor K. Indeed

$$Var\left[\widehat{S}_{AVPER}(f)\right] \;=\; \frac{1}{K}\,Var\left[\widehat{S}_{PER}^{(0)}(f)\right] \;. \tag{2.63}$$

In practice, we normally don't have K independent records but only one record of length N on which to base the spectral estimator. A common approach is to segment the data into K nonoverlapping contiguous blocks of length L so that $N = KL$. In this way, the data records become

$$y_n^{(i)} \;=\; y_{n+iL}\;, \quad n = 0, 1, \ldots, L\Leftrightarrow 1; \quad i = 0, 1, \ldots, K\Leftrightarrow 1\;. \tag{2.64}$$

Contiguous data records cannot be independent unless the y_k are i.i.d. This dependence does not influence the mean of the averaged periodogram. However, the variance reduction obtained by the averaged periodogram is generally less than a factor K. For processes with rapidly decaying acf, the correlation between different data blocks will be weak. If the data are furthermore Gaussian, the data blocks will also be roughly independent.

With the sectioning of a long data record into K shorter data records, the main design issue becomes one of finding a good compromise. Indeed, as the number of data blocks K increases, the variance of the averaged periodogram decreases. However, also the length $L = N/K$ of the data blocks decreases which means that the bandwidth $\approx \frac{1}{L}$ of the Bartlett window increases. This means that the smearing of the spectral estimate (bias) increases. One approach is to start with a large value for K (and hence small L) and to observe the averaged periodogram as K decreases. The smearing will decrease and hence more spectral details become apparent. One can stop reducing K when no more details are transpiring. This procedure is called *window closing*. This procedure is not without risk since the variance increases as K decreases and hence spurious details that appear may simply be due to estimation variance. We have already discussed before the use of data prewhitening as a technique for reducing the bias problem. This technique may help to alleviate the bias problem in the averaged periodogram.

The averaged periodogram can be a consistent spectrum estimator. For that, it suffices that $L \to \infty$ and $K \to \infty$ as $N \to \infty$. Indeed, as $L \to \infty$, the bias in the averagd periodogram disappears. And as $K \to \infty$, the variance shrinks to zero. A possible choice is $L = K = \sqrt{N}$.

A Variant due to Welch

Welch proposed to introduce a window in the computation of the periodogram corresponding to each data block, and furthermore to let the data blocks be overlapping (since the data are windowed, the interaction between consecutive data blocks is less severe than their amount of overlap may suggest). The suggested overlap is as much as 50 or even 75 %. Due to the windowing, the spectral leakage gets decreased. Furthermore, due to the overlap, a larger number of data records becomes available, leading to some extra variance reduction.

2.2.3 The Blackman-Tukey Spectral Estimator

The poor performance of the periodogram may also be illuminated by considering the following equivalent form for the periodogram

$$\begin{aligned} \widehat{S}_{PER}(f) &= \tfrac{1}{N}\left|\sum_{n=0}^{N-1} y_n\, e^{-j2\pi fn}\right|^2 = \tfrac{1}{N}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} y_n y_m\, e^{-j2\pi f(n-m)} \\ &= \sum_{k=-(N-1)}^{N-1} e^{-j2\pi fk}\,\frac{1}{N}\sum_{n=0}^{N-1-|k|} y_{n+|k|}y_n = \sum_{k=-(N-1)}^{N-1} \widehat{r}_{yy}(k)e^{-j2\pi fk} \end{aligned} \tag{2.65}$$

where

$$\widehat{r}_{yy}(k) = \begin{cases} \frac{1}{N}\sum_{n=0}^{N-1-k} y_{n+k}y_n & , \quad k = 0, 1, \ldots, N\Leftrightarrow 1 \\ \widehat{r}_{yy}(\Leftrightarrow k) & , \quad k = \Leftrightarrow(N\Leftrightarrow 1), \ldots, \Leftrightarrow 1. \end{cases} \tag{2.66}$$

Hence the periodogram may be seen to be an estimator of the psdf by equivalently estimating the acf and using the Wiener-Khinchin relation. The poor performance of the periodogram may be understood by considering the quality of the acf estimate. At lag $N\Leftrightarrow 1$ for instance, we get $\widehat{r}_{yy}(N\Leftrightarrow 1) = \frac{1}{N}y_{N-1}y_0$ which apart from being strongly biased is also highly variable due to the lack of averaging. Furthermore, $\widehat{r}_{yy}(k) = 0, \quad |k| > N\Leftrightarrow 1$. For general lags, the mean of the acf estimate is

$$E\,\widehat{r}_{yy}(k) = \left(1 \Leftrightarrow \frac{|k|}{N}\right) r_{yy}(k) = w_{B,N}(k) r_{yy}(k)\,, \quad |k| \le N\Leftrightarrow 1\,. \tag{2.67}$$

So the mean is equal to the true value weighted by the Bartlett window and hence the estimator is biased (except for $k = 0$). We could use an unbiased acf estimator in (2.65) by replacing the $\frac{1}{N}$ factor in (2.66) by $\frac{1}{N-|k|}$. However, this choice leads to a spectral estimate with higher variance. Furthermore, this $\widehat{S}(f)$ may be negative at certain frequencies since the unbiased acf estimate does not necessarily correspond to a positive semidefinite sequence.

So the estimated acf in (2.66) has a bias that increases linearly with lag and a variance that increases also with lag (number of terms in the averaging decreases). Hence, the trustworthiness of the estimated acf decreases with lag. This motivated Blackman and Tukey to introduce a weighting sequence w_k that reflects the quality of the acf estimates: w_k decreases with $|k|$. This leads to *Blackman-Tukey (BT) spectral estimator*

$$\widehat{S}_{BT}(f) = \sum_{k=-(N-1)}^{N-1} w_k\, \widehat{r}_{yy}(k)\, e^{-j2\pi fk} \tag{2.68}$$

where w_k is a real sequence termed *lag window* that satisfies the following properties:

1. $0 \leq w_k \leq w_0 = 1$
2. $w_{-k} = w_k$
3. $w_k = 0$ for $|k| > M$

where $M \leq N \Leftrightarrow 1$. Due to this last property, we may rewrite (2.68) as

$$\widehat{S}_{BT}(f) \;=\; \sum_{k=-M}^{M} w_k \, \widehat{r}_{yy}(k) \, e^{-j2\pi fk} \tag{2.69}$$

The BT spectral estimator is equivalent to the periodogram if $w_k = 1$ for $|k| \leq M = N \Leftrightarrow 1$. The BT estimator is also sometimes called a *weighted covariance* estimator. The weighting will reduce the variance of the spectral estimate, but again by increasing the bias (smearing). Some popular windows can again be found in Table 2.1. We must be careful however that the window chosen will always lead to a nonnegative spectral estimate. Remark that the BT estimator can be rewritten as

$$\widehat{S}_{BT}(f) \;=\; \mathcal{F}\left\{w_k \, \widehat{r}_{yy}(k)\right\} \;=\; W(f) * \widehat{S}_{PER}(f) \tag{2.70}$$

since $\mathcal{F}\left\{\widehat{r}_{yy}(k)\right\} \;=\; \widehat{S}_{PER}(f)$. Although $\widehat{S}_{PER}(f) \geq 0$, if $W(f)$ is negative at certain frequencies, then the convolution in (2.70) may produce negative values. To avoid this, it is preferable that $W(f) \geq 0$, $\forall f$ or equivalently that w_k is a nonnegative sequence. Only the Bartlett and Parzen windows in Table 2.1 have nonnegative Fourier transforms.

We now investigate the bias and variance of the BT estimator. From (2.70), we get for the mean

$$\begin{aligned} E\,\widehat{S}_{BT}(f) \;&=\; W(f) * E\widehat{S}_{PER}(f) \;=\; W(f) * (W_{B,N}(f) * S_{yy}(f)) \\ &=\; \underbrace{(W(f) * W_{B,N}(f))}_{=\mathcal{F}\left\{\underbrace{w_k w_B(k)}_{\approx w_k}\right\}} * S_{yy}(f) \;\approx\; W(f) * S_{yy}(f) \end{aligned} \tag{2.71}$$

where we assumed that $N \gg M$. So the true psdf gets smeared again, this time leaving a spectral resolution of about $\frac{1}{M}$. Again, prewhitening the data will help reduce the bias. For the variance, we get under the assumption that the psdf is smooth on a frequency scale of $\frac{1}{M}$, for frequencies not near 0 or $\frac{1}{2}$

$$Var\left[\widehat{S}_{BT}(f)\right] \;\approx\; S_{yy}^2(f) \, \frac{1}{N} \sum_{k=-M}^{M} w_k^2 \; . \tag{2.72}$$

Again the bias-variance trade-off becomes apparent by considering jointly (2.71) and (2.72) For a small bias, M should be chosen large since that will cause the spectral window $W(f)$ to behave as a Dirac delta function. On the other hand, a small variance imposes a small M. A maximum value of $M = N/5$ is usually recommended. As an example, for the Bartlett window

$$Var\left[\widehat{S}_{BT}(f)\right] \;\approx\; \frac{2M}{3N} \, S_{yy}^2(f) \tag{2.73}$$

so that $M = N/5$ results in a variance reduction by a factor 7.5 compared to the periodogram. Much of the art in classical spectral estimation is in choosing an appropriate window, both the type and the length (M).

Also the BT spectral estimator can yield a consistent estimate. For that we require $M \to \infty$ so that the bias disappears and $\frac{N}{M} \to \infty$ as $N \to \infty$ so that the variance shrinks to zero. Again, a possible choice is $M = \sqrt{N}$.

Smoothed Periodogram

From (2.70) we can interpret the BT estimator as the convolution of the periodogram with the spectral window $W(f)$. The role of the spectral window is to smoothen the periodogram, thus possibly increasing the bias but also reducing the variance. This suggest computing the periodogram and smoothing it in frequency. Assuming a N' point DFT is used to compute the periodogram, a discrete version of (2.70) with uniform spectral weighting is

$$\widehat{S}_{DAN}(f_k) \;=\; \frac{1}{2L+1} \sum_{i=-L}^{L} \widehat{S}_{PER}(f_k + \frac{i}{N'}) \;, \quad f_k = \frac{k}{N'} \tag{2.74}$$

where we have approximately the correspondence: bandwidth $= \frac{2L+1}{N'} = \frac{1}{M}$. This smoothed periodogram is called *Daniell's spectral estimator*. Many other smoothed periodograms are possible by choosing different spectral weightings. Remark that it is quite easy to control the nonnegativity of the spectral weighting so that a positive psdf estimate can be guaranteed.

2.3 Parametric Random Process Models

The generic parametric model we shall consider is depicted in Fig. 2.13. White noise drives a linear time-invariant causal and rational system

$$\mathrm{H}(z) \;=\; \sum_{k=0}^{\infty} h_k \, z^{-k} \;=\; \frac{\prod_i (1 \Leftrightarrow z_i z^{-1})}{\prod_k (1 \Leftrightarrow p_k z^{-1})} \; . \tag{2.75}$$

The zero mean of the input leads to a zero mean output. The variance of the output

$$\sigma_y^2 \;=\; \sigma_e^2 \sum_{k=0}^{\infty} h_k^2 \tag{2.76}$$

will be finite if the system is stable ($|p_k| < 1$: poles inside the unit circle). In that case, due to the wide-sense stationarity of the input, the output will be wide-sense stationary with psdf

$$S_{yy}(f) \;=\; \sigma_e^2 \; |H(f)|^2 \; . \tag{2.77}$$

For the modeling of $S_{yy}(f)$, only the magnitude response $|H(f)|$ is important. However, we desire to be able to determine $H(f)$ completely from $S_{yy}(f)$ and hence from $|H(f)|$. This is desirable because $H(f)$ is parameterized by a finite set of parameters and we wish that with

a certain $H(f)$ (and hence $|H(f)|$) corresponds a unique set of parameter values. This can be achieved if we assume $H(f)$ to be minimum-phase (which requires stability plus $|z_i| < 1$: zeros inside the unite circle). The minimum-phase property of the system implies that it is causally invertible ($\mathrm{H}^{-1}(z)$ is causal). Due to the uncorrelatedness of the input, we get

$$\mathrm{S}_{ye}(z) \;=\; \mathrm{H}(z)\,\sigma_e^2 \;\;\Rightarrow\;\; r_{ye}(k) \;=\; E\,y_{n+k}e_n \;=\; h_k\,\sigma_e^2 \;\;(=0\,,\;k<0) \tag{2.78}$$

which shows in particular that the output is uncorrelated with future inputs. We shall distinguish between specific parametric models on the basis of various rational models for the transfer function $\mathrm{H}(z)$. For each of these specific models, we shall be interested in how we can get the impulse response h_k from the parameters and vice versa, and how we can get the autocorrelation sequence $r_{yy}(k)$ or equivalently the psdf $\mathrm{S}_{yy}(z)$ from the parameters and vice versa.

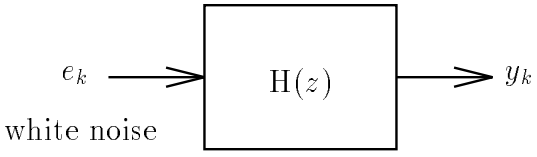

Figure 2.13: The generic parametric model for a purely random process.

2.3.1 Autoregressive (AR) Processes

An Autoregressive process of order n (AR(n)) is obtained by taking an n-th order *all-pole* transfer function

$$\mathrm{H}(z) \;=\; \frac{1}{\mathrm{A}(z)}\,,\;\; \mathrm{A}(z) \;=\; \sum_{i=0}^{n} A_i\,z^{-i}\,,\;\; A_0 = 1\,. \tag{2.79}$$

$\mathrm{A}(z)$ with $A_0 = 1$ is called a *monic* polynomial in z^{-1}. $\mathrm{A}(z)$ as a function of z needs to have all its roots inside the unit circle for $\mathrm{H}(z)$ to be minimum-phase. Input-output relations for linear time-invariant systems are easy to write down in the z-domain. To transform these relations in the time domain, we shall introduce the delay operator q^{-1} : $q^{-1}\,y_k = y_{k-1}$ (and q is the advance operator: $q\,y_k = y_{k+1}$). In fact, q^{-1} appears in the same way as z^{-1} but z^{-1} in front of a z-transform means multiplication of this z-transform with the complex number z^{-1}. On the other hand, the notation q^{-1} in front of the signal corresponding to the z-transform just mentioned means that that signal gets delayed by one sample. So, writing out the input-output relation, we get the following difference equation

$$\begin{aligned} \mathrm{Y}(z) \;&=\; \mathrm{H}(z)\,\mathrm{E}(z) \;=\; \frac{1}{\mathrm{A}(z)}\,\mathrm{E}(z) \;\Rightarrow\; \mathrm{A}(z)\,\mathrm{Y}(z) \;=\; \mathrm{E}(z) \\ \text{or } \mathrm{A}(q)\,y_k \;&=\; \sum_{i=0}^{n} A_i\,q^{-i}\,y_k \;=\; \sum_{i=0}^{n} A_i\,y_{k-i} \;=\; y_k + A_1 y_{k-1} + \cdots + A_n y_{k-n} \;=\; e_k\,. \end{aligned} \tag{2.80}$$

The name autoregression becomes more obvious when we rewrite (2.80) as

$$y_k \;=\; \Leftrightarrow A_1 y_{k-1} \Leftrightarrow \cdots \Leftrightarrow A_n y_{k-n} + e_k \tag{2.81}$$

which expresses y_k as a linear regression on its own past plus independent noise. The impulse response of the system satisfies the following recursion:

$$\mathrm{A}(z)\,\mathrm{H}(z) \;=\; 1 \quad \Rightarrow \quad \mathrm{A}(q)\,h_k \;=\; \delta_{k0}\;. \tag{2.82}$$

This allows one to find h_k recursively from $\mathrm{A}(z)$. In particular $h_0 = 1$. From the expression for the psdf, we can find

$$\mathrm{S}_{yy}(z) \;=\; \frac{\sigma_e^2}{\mathrm{A}(z)\mathrm{A}(1/z)} \quad \Rightarrow \quad \mathrm{A}(z)\,\mathrm{S}_{yy}(z) \;=\; \sigma_e^2\,\mathrm{H}(1/z) \quad \text{or} \quad \mathrm{A}(q)\,r_{yy}(k) \;=\; \sigma_e^2\,h_{-k} \tag{2.83}$$

which are the so-called *Yule-Walker equations*. The Yule-Walker equations for $k = 0, 1, \ldots, n$ constitute $n+1$ linear equations that allow one to obtain $r_{yy}(0), \ldots, r_{yy}(n)$ from $\sigma_e^2, A_1, \ldots, A_n$ or vice versa. If for example the $r_{yy}(0), \ldots, r_{yy}(n)$ have been obtained from $\sigma_e^2, A_1, \ldots, A_n$, then further Yule-Walker equations can be used to obtain the rest of the covariance sequence recursively:

$$\mathrm{A}(q)\,r_{yy}(k) \;=\; \sum_{i=0}^{n} A_i\,r_{yy}(k \Leftrightarrow i) \;=\; 0\,,\; k > n\;. \tag{2.84}$$

2.3.2 Moving Average (MA) Processes

A Moving Average process of order m (MA(m)) is obtained by taking a m-th order *all-zero* transfer function

$$\mathrm{H}(z) \;=\; \mathrm{B}(z) \;=\; \sum_{i=0}^{m} B_i\,z^{-i}\;,\quad B_0 = 1\;. \tag{2.85}$$

Again, $\mathrm{B}(z)$ is a monic polynomial in z^{-1}. $\mathrm{B}(z)$ as a function of z needs again to have all its roots inside the unit circle for $\mathrm{H}(z)$ to be minimum-phase. Writing out the input-output relation, we get the following difference equation

$$y_k \;=\; \mathrm{B}(q)\,e_k \;=\; e_k + B_1 e_{k-1} + \cdots + B_m e_{k-m}\;. \tag{2.86}$$

The name moving average stems from the fact that y_k is computed as a sliding (moving) weighted linear combination (average) of the $m+1$ last inputs. We get for the psdf

$$\mathrm{S}_{yy}(z) \;=\; \sigma_e^2\,\mathrm{B}(z)\mathrm{B}(1/z) \tag{2.87}$$

or in the time domain

$$r_{yy}(k) \;=\; \sigma_e^2\,B_k * B_{-k} \tag{2.88}$$

which implies in particular that

$$r_{yy}(k) \;=\; 0\;,\quad |k| > m\;. \tag{2.89}$$

Due to the particular form of $\mathrm{S}_{yy}(z)$ shown in (2.87), $\mathrm{S}_{yy}(z)$ has precisely $2m$ zeros which are such that if z_i is a zero, then so is $1/z_i$. Hence, σ_e^2 and $\mathrm{B}(z)$ can be identified from the $r_{yy}(k)$ by finding the zeros of $\mathrm{S}_{yy}(z)$ and assigning the minimum-phase zeros ($|z_i| \leq 1$) to $\mathrm{B}(z)$, and $\sigma_e^2 = \mathrm{S}_{yy}(1)/\mathrm{B}^2(1)$. This process is called *spectral factorization*.

Remark that the Blackman-Tukey spectral estimator implictly assumes a MA(M) process since the weighted acf is put equal to zero for lags bigger than M.

2.3.3 Autoregressive Moving Average (ARMA) Processes

An autoregressive moving average (ARMA(n,m)) process is obtained by taking a rational transfer function

$$\mathrm{H}(z) \;=\; \mathrm{B}(z)/\mathrm{A}(z) \tag{2.90}$$

where $\mathrm{A}(z)$ and $\mathrm{B}(z)$ are monic minimum-phase polynomials as before. Writing out the input-output relation, we get the following difference equation

$$\mathrm{A}(q)\,y_k \;=\; \mathrm{B}(q)\,e_k \;\text{ or }\; y_k + A_1 y_{k-1} + \cdots + A_n y_{k-n} \;=\; e_k + B_1 e_{k-1} + \cdots + B_m e_{k-m}\;. \tag{2.91}$$

This process is clearly a combination of autoregression and moving average. The impulse response of the system satisfies the following recursion:

$$\mathrm{A}(z)\,\mathrm{H}(z) \;=\; \mathrm{B}(z) \;\;\Rightarrow\;\; \mathrm{A}(q)\,h_k \;=\; B_k\;. \tag{2.92}$$

This allows one to find h_k recursively from $\mathrm{A}(z)$ and $\mathrm{B}(z)$. In particular $h_0 = 1$. From the expression for the psdf, we can find

$$\mathrm{S}_{yy}(z) \;=\; \sigma_e^2\,\frac{\mathrm{B}(z)\mathrm{B}(1/z)}{\mathrm{A}(z)\mathrm{A}(1/z)} \;\;\Rightarrow\;\; \mathrm{A}(z)\,\mathrm{S}_{yy}(z) \;=\; \sigma_e^2\,\mathrm{B}(z)\mathrm{H}(1/z) \;\;\text{or}\;\; \mathrm{A}(q)\,r_{yy}(k) \;=\; \sigma_e^2\,B_k * h_{-k} \tag{2.93}$$

which are again the *Yule-Walker equations*. Given σ_e^2, $\mathrm{A}(z)$ and $\mathrm{B}(z)$, one can obtain the acf from the Yule-Walker equations in pretty much the same way as for an AR process. Given the acf, one can determine $\mathrm{A}(z)$ from n equations of the form

$$\mathrm{A}(q)\,r_{yy}(k) \;=\; 0\,,\; k > m\;. \tag{2.94}$$

σ_e^2 and $\mathrm{B}(z)$ can then be obtained by spectral factorization of $\mathrm{A}(z)\,\mathrm{A}(1/z)\,\mathrm{S}_{yy}(z)$.

2.4 Linear Prediction

In what follows, we shall assume that y_k is a zero mean WSS random process. Later on, we shall restrict y_k to be an AR process.

2.4.1 Forward Linear Prediction

Consider predicting the sample y_k linearly from the n previous samples. So we form the prediction as a linear combination of the form

$$\widehat{y}_k \;=\; \Leftrightarrow\!\sum_{i=1}^{n} A_{n,i}\,y_{k-i} \tag{2.95}$$

where we have given the combination coefficients $A_{n,i}$ a double index since they will depend on the total number of previous samples n involved in the prediction. We shall adjust the coefficients $A_{n,i}$ to minimize the prediction error $y_k \Leftrightarrow \widehat{y}_k$. Of course, for a given sample y_k, it is always possible to find coefficients $A_{n,i}$ such that the prediction error is zero! However, we don't want to choose totally different coefficients for each sample, because then our coefficients

would simply be a nonunique nonlinear transformation of our signal and they would not extract any important characteristic of our signal. We want to minimize the prediction error, not instantaneously, but on the average. In fact, linear prediction is just a special case of LMMSE estimation of $\theta = y_k$ from $Y = [y_{k-1} \cdots y_{k-n}]^T$.

One may wonder whether this attempt to predict a random signal can be successful at all. We shall see that white noise is perfectly unpredictible. On the other end of the spectrum, since we have seen that a sinusoidal signal satisfies a homogeneous linear difference equation with constant coefficients of order two, it is perfectly possible to linearly predict a sinusoidal signal from its previous two samples without any error at all! For general colored random processes, the predictibility will be somewhere in between. As we shall see, signals with peaks in their spectrum are fairly predictible (sinusoids are an extreme example of this). The speech signal, as it is generated by air passing through the resonating oral cavity, show peaks in its spectrum. On the average, the variance of its prediction error is about ten times smaller than the variance of the speech signal. This will be useful for coding the signal. A historical motivation for prediction arose in the second world war, when Norbert Wiener at MIT was studying the prediction of the trajectories of planes to automatically adjust the positioning of anti-aircraft guns. Modern applications of linear prediction can also be found in econometrics, where for instance a clear interest exists in predicting the value of stocks on the stock market. Such prediction problems are more complex though since they involve multiple signals on the basis of which the prediction is generated.

In the spirit of LMMSE estimation, we shall minimize the prediction error variance (MSE):

$$\min_{A_{n,i}} \|y_k \Leftrightarrow \widehat{y}_k\|^2 \;=\; \min_{A_{n,i}} E\,(y_k \Leftrightarrow \widehat{y}_k)^2 \;=\; \min_{A_{n,i}} E\left(y_k + \sum_{i=1}^{n} A_{n,i} y_{k-i}\right)^2 . \tag{2.96}$$

Let us introduce the following notation for this forward prediction error of order n at time instant k

$$f_{n,k} \;=\; y_k \Leftrightarrow \widehat{y}_k \;=\; y_k + \sum_{i=1}^{n} A_{n,i} y_{k-i} \;=\; \sum_{i=0}^{n} A_{n,i} y_{k-i} \tag{2.97}$$

where we introduced $A_{n,0} = 1$. Then the solution to the least-squares problem (2.96) is characterized by the orthogonality condition: the point $\widehat{y}_k$ in the subspace spanned by $y_{k-1}, \ldots, y_{k-n}$ that is closest to y_k is the one that is the orthogonal projection of y_k onto that subspace, see Fig. 2.14.

The orthogonality conditions can be written as

$$\langle f_{n,k}, y_{k-i} \rangle \;=\; E f_{n,k} y_{k-i} \;=\; \sum_{j=0}^{n} A_{n,j} E y_{k-i} y_{k-j} \;=\; \sum_{j=0}^{n} r_{|i-j|} A_{n,j} \;=\; 0 \;,\; i = 1, \ldots, n \tag{2.98}$$

where $r_{|i-j|} = E y_{k-i} y_{k-j}$ and we introduced the shorthand notation $r_k = r_{yy}(k)$. We see that due to the stationarity of the signal, all quantities in this system of equations do not depend on the absolute time. So the optimal prediction coefficients are time invariant. Before looking in detail at this system of equations, we shall investigate the minimal value of the criterion:

$$\begin{aligned} \sigma^2_{f,n} \;&=\; E\, f^2_{n,k} \;=\; \min_{A_{n,i}} E(y_k + \sum_{i=1}^{n} A_{n,i} y_{k-i})^2 \\ &=\; E f_{n,k} y_k \;+\; \sum_{j=1}^{n} A_{n,j} \underbrace{E f_{n,k} y_{k-j}}_{=0} \;=\; E f_{n,k} y_k \end{aligned} \tag{2.99}$$

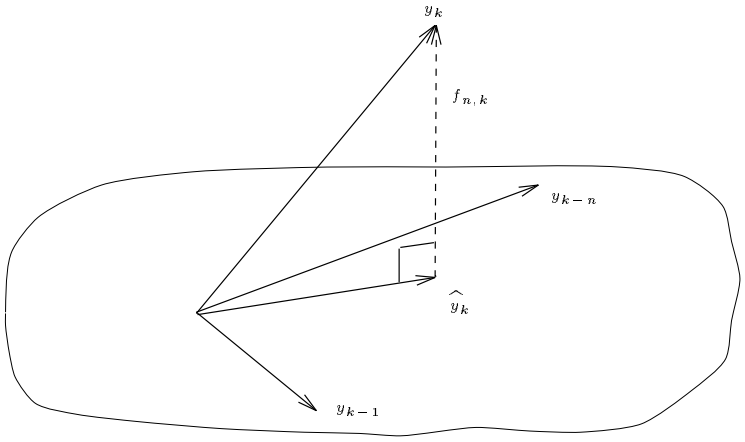

Figure 2.14: Geometric interpretation of the forward prediction problem.

where we assumed that when the expression $\min_{A_{n,i}}$ gets dropped, the optimal coefficients $A_{n,i}$ are being used. Similarly, $f_{n,k}$ normally means $\mathrm{A}_n(q)y_k$ in which the optimal coefficients $A_{n,i}$ are used and $\sigma^2_{f,n}$ normally corresponds to the variance of this optimal $f_{n,k}$. Let us introduce the following vectors

$$Y_{n+1}(k) \;=\; \begin{bmatrix} y_k \\ y_{k-1} \\ \vdots \\ y_{k-n} \end{bmatrix} \;, \quad A_n \;=\; \begin{bmatrix} 1 \\ A_{n,1} \\ \vdots \\ A_{n,n} \end{bmatrix} \;, \quad \Rightarrow \quad f_{n,k} \;=\; Y^T_{n+1}(k)A_n \;. \tag{2.100}$$

Now assembling the orthogonality conditions with the expression for the minimal variance, we get

$$E\,Y_{n+1}(k)f_{n,k} \;=\; \begin{bmatrix} Ey_k f_{n,k} \\ Ey_{k-1} f_{n,k} \\ \vdots \\ Ey_{k-n} f_{n,k} \end{bmatrix} \;=\; \begin{bmatrix} \sigma^2_{f,n} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \;. \tag{2.101}$$

On the other hand

$$E\,Y_{n+1}(k)f_{n,k} \;=\; \Big(E\,Y_{n+1}(k)Y^T_{n+1}(k)\Big)\,A_n \;=\; R_{n+1}A_n \;=\; \begin{bmatrix} \sigma^2_{f,n} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.102}$$

where we introduced the covariance matrix

$$R_{n+1} \;=\; E\,Y_{n+1}(k)Y^T_{n+1}(k) \;=\; \begin{bmatrix} r_0 & r_1 & \cdots & r_n \\ r_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_n & \cdots & r_1 & r_0 \end{bmatrix} \;. \tag{2.103}$$

Due to the stationarity of y_k, R_{n+1} is Toeplitz. The equations (2.102) are called the *normal equations* or *Yule-Walker equations*. They are indeed of the same form as the equations (2.83) taken for $k = 0, 1, \ldots, n$. They are a bit unusual in that they are $n+1$ equations in $n+1$ unknowns, but n unknowns are on the LHS , while 1 unknown ($\sigma^2_{f,n}$) is on the RHS . However, we can rewrite (2.102) as follows

$$\begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_n \end{bmatrix} + \begin{bmatrix} r_1 & \cdots & r_n \\ r_0 & \ddots & \vdots \\ \vdots & \ddots & r_1 \\ r_{n-1} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} A_{n,1} \\ \vdots \\ A_{n,n} \end{bmatrix} = \begin{bmatrix} \sigma^2_{f,n} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \Rightarrow \quad \left[\begin{array}{c|ccc} \Leftrightarrow 1 & r_1 & \cdots & r_n \\ \hline 0 & & & \\ \vdots & & R_n & \\ 0 & & & \end{array}\right] \begin{bmatrix} \sigma^2_{f,n} \\ A_{n,1} \\ \vdots \\ A_{n,n} \end{bmatrix} = \Leftrightarrow \begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_n \end{bmatrix} \tag{2.104}$$

We see that the matrix of coefficients of this last set of equations is block triangular and hence is nonsingular since R_n (a covariance matrix) is assumed to be nonsingular. One solves the last n equations for the n unknowns $A_{n,1}, \ldots, A_{n,n}$, which then get substituted in the first equation to find $\sigma^2_{f,n}$. However, to solve a system of n equations in n unknowns takes on the order of n^3 operations (multplications, additions). We shall see a fast algorithm for solving the Yule-Walker equations, which takes about n^2 operations. The details of that algorithm will lead to the lattice filter, a numerically robust filter structure, an alternative parameterization of the prediction coefficients, the reflection coefficients, and we shall also see a connection with transmission lines.

The fast algorithm for solving the normal equations for the forward prediction problem makes use of the backward prediction problem.

2.4.2 Backward Prediction Problem

Consider now the sense of the time axis as going backward in time, but we shall still work with the $n+1$ most recent samples of y_k. So consider the problem of linearly predicting y_{k-n} backward, i.e. from the n samples that come immediately afterward:

$$\widehat{y}_{k-n} \;=\; \Leftrightarrow \sum_{i=1}^{n} B_{n,i} y_{k-n+i} \;. \tag{2.105}$$

We want again to adjust the backward prediction coefficients $B_{n,i}$ to minimize the prediction error variance:

$$\min_{B_{n,i}} \|y_{k-n} \Leftrightarrow \widehat{y}_{k-n}\|^2 \;=\; \min_{B_{n,i}} E\left(y_{k-n} \Leftrightarrow \widehat{y}_{k-n}\right)^2 \;=\; \min_{B_{n,i}} E\left(y_{k-n} + \sum_{i=1}^{n} B_{n,i} y_{k-n+i}\right)^2 \;. \tag{2.106}$$

Let us introduce the following notation for the backward prediction error of order n at time instant k

$$b_{n,k} \;=\; y_{k-n} \Leftrightarrow \widehat{y}_{k-n} \;=\; y_{k-n} + \sum_{i=1}^{n} B_{n,i} y_{k-n+i} \;=\; \sum_{i=0}^{n} B_{n,i} y_{k-n+i} \tag{2.107}$$

where we introduced $B_{n,0} = 1$. Then the solution to the least-squares problem (2.106) is again characterized by the orthogonality condition: the point $\widehat{y}_{k-n}$ in the subspace spanned by $y_k, \ldots, y_{k-n+1}$ that is closest to y_{k-n} is the one that is the orthogonal projection of y_{k-n} onto that subspace, see Fig. 2.15.

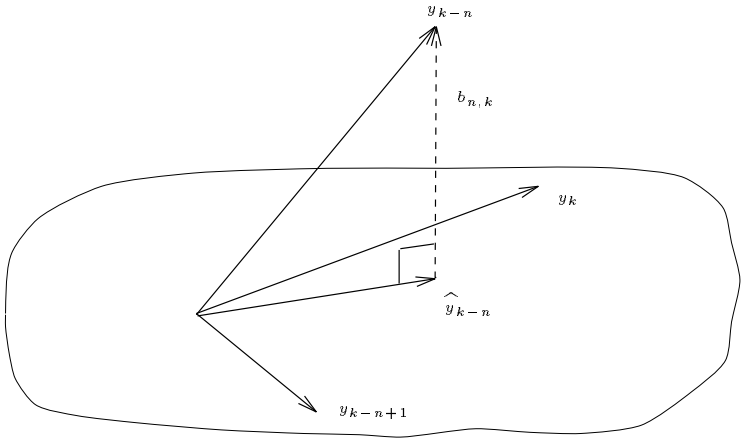

Figure 2.15: Geometric interpretation of the backward prediction problem.

The orthogonality conditions can be written as

$$\langle b_{n,k}, y_{k-n+i} \rangle \;=\; E b_{n,k} y_{k-n+i} \;=\; \sum_{j=0}^{n} B_{n,j} E y_{k-n+i} y_{k-n+j} \;=\; \sum_{j=0}^{n} r_{|i-j|} B_{n,j} \;=\; 0 \;,\; i = 1, \ldots, n \;. \tag{2.108}$$

Again the optimal prediction coefficients are constant (as a function of time). We investigate again the minimal value of the criterion:

$$\begin{aligned} \sigma_{b,n}^2 \;&=\; E\, b_{n,k}^2 \;=\; \min_{B_{n,i}} E(y_{k-n} + \sum_{i=1}^{n} B_{n,i} y_{k-n+i})^2 \\ &=\; E b_{n,k} y_{k-n} \;+\; \sum_{j=1}^{n} B_{n,j} \underbrace{E b_{n,k} y_{k-n+j}}_{=0} \;=\; E b_{n,k} y_{k-n} \end{aligned} \tag{2.109}$$

where we assumed again that when the expression $\min_{B_{n,i}}$ gets dropped, the optimal coefficients $B_{n,i}$ are being used. Now assembling the orthogonality conditions with the expression for the minimal variance, we get

$$E Y_{n+1}(k) b_{n,k} \;=\; \begin{bmatrix} E y_k b_{n,k} \\ \vdots \\ E y_{k-n+1} b_{n,k} \\ E y_{k-n} b_{n,k} \end{bmatrix} \;=\; \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{b,n}^2 \end{bmatrix} \;. \tag{2.110}$$

On the other hand, $b_{n,k} = Y_{n+1}^T(k) B_n$ where

$$B_n \;=\; \begin{bmatrix} B_{n,n} \\ \vdots \\ B_{n,1} \\ 1 \end{bmatrix} \;. \tag{2.111}$$

So we can write

$$E\,Y_{n+1}(k)b_{n,k} \;=\; \left(E\,Y_{n+1}(k)Y_{n+1}^T(k)\right)B_n \;=\; R_{n+1}B_n \;=\; \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{b,n}^2 \end{bmatrix} . \tag{2.112}$$

These equations are the normal equations for the backward prediction problem.

2.4.3 Relationship between the Forward and Backward Prediction Quantities

Let J be the reverse identity matrix:

$$J \;=\; \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix} . \tag{2.113}$$

J reverses the order of rows or columns:

$$J \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \;=\; \begin{bmatrix} y_n \\ \vdots \\ y_2 \\ y_1 \end{bmatrix} , \quad [y_1 \;\; y_2 \cdots y_n]\; J \;=\; [y_n \cdots y_2 \;\; y_1] . \tag{2.114}$$

Then it should now be clear that a symmetric Toeplitz matrix is *persymmetric* (symmetric w.r.t. the antidiagonal), i.e.

$$J\,R_{n+1}\,J \;=\; R_{n+1}^T \;=\; R_{n+1} \tag{2.115}$$

where the second identity implies that R_{n+1} is also *centrosymmetric* (symmetric and persymmetric). From the backward normal equations (2.112), we have

$$R_{n+1}\,JB_n \;=\; J\,R_{n+1}\,J\,JB_n \;=\; J\,R_{n+1}B_n \;=\; J \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{b,n}^2 \end{bmatrix} \;=\; \begin{bmatrix} \sigma_{b,n}^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} . \tag{2.116}$$

When we compare to the forward normal equations (2.102), we see that the $B_{n,i}$ and $\sigma_{b,n}^2$ satisfy exactly the same equations as the $A_{n,i}$ and $\sigma_{f,n}^2$ and hence

$$B_n \;=\; J\,A_n \text{ or } B_{n,i} = A_{n,i}\,,\; i = 1, \ldots, n\,, \quad \sigma_{b,n}^2 = \sigma_{f,n}^2\,. \tag{2.117}$$

These results are a consequence of the more general property that if we have a stationary signal and we flip the time axis around, we get a signal with exactly the same statistical properties back.

2.4.4 Levinson Algorithm

In 1948, Levinson (an assistant to Prof. Wiener at MIT) published the following fast algorithm for solving the normal equations of the prediction problem (as an appendix in a book by Wiener). The algorithm is recursive in nature. So suppose after recursion n we have A_n and $\sigma^2_{f,n}$. We now try to find the same quantities for order $n+1$. We first look at a trial solution for A_{n+1} which we obtain by appending one zero to A_n, viz.

$$\underbrace{\begin{bmatrix} r_0 & r_1 & \cdots & r_n & r_{n+1} \\ r_1 & \ddots & \ddots & & r_n \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r_n & & \ddots & \ddots & r_1 \\ r_{n+1} & r_n & \cdots & r_1 & r_0 \end{bmatrix}}_{R_{n+2}} \underbrace{\left\{ \begin{bmatrix} 1 \\ A_{n,1} \\ \vdots \\ A_{n,n} \\ 0 \end{bmatrix} + K_{n+1} \begin{bmatrix} 0 \\ A_{n,n} \\ \vdots \\ A_{n,1} \\ 1 \end{bmatrix} \right\}}_{= A_{n+1} = \begin{bmatrix} 1 \\ A_{n+1,1} \\ \vdots \\ A_{n+1,n} \\ A_{n+1,n+1} \end{bmatrix}} = \underbrace{\left\{ \begin{bmatrix} \sigma^2_{f,n} \\ 0 \\ \vdots \\ 0 \\ \Delta_{n+1} \end{bmatrix} + K_{n+1} \begin{bmatrix} \Delta_{n+1} \\ 0 \\ \vdots \\ 0 \\ \sigma^2_{f,n} \end{bmatrix} \right\}}_{= \begin{bmatrix} \sigma^2_{f,n+1} \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}} . \tag{2.118}$$

Because R_{n+2} contains R_{n+1} as its upper-left submatrix of one dimension less, and because of the form of our trial solution, the corresponding right hand side has the desired form except for the last entry, which we shall call Δ_{n+1}. If we flip our trial solution upside down (to take the corresponding trial solution for the backward prediction problem), then the right hand side also simply gets flipped upside down, because of the centrosymmetry of R_{n+2}. Because of linearity, when we take a linear combination of the two trial solutions, we get the same linear combination of the two right hand sides. Now it is pretty straightforward to see that we can choose the combination parameter K_{n+1} to get zeros everywhere in the RHS except for the first element. Namely, we should choose

$$\Delta_{n+1} + K_{n+1}\,\sigma^2_{f,n} = 0 \quad \Rightarrow \quad K_{n+1} \;=\; \Leftrightarrow \frac{\Delta_{n+1}}{\sigma^2_{f,n}} \,. \tag{2.119}$$

So the steps taken so far are of the following form:

$$\begin{aligned} R\,A &= C \\ \Rightarrow \quad R\,J\,A &= J\,C \\ \Rightarrow \quad R\,(A + J\,A) &= C + J\,C \end{aligned} \tag{2.120}$$

where the second step is similar to the reasoning in (2.116). Now it becomes clear that the combination of the two trial solutions in (2.118) has itself the right structure (1 as first element) and when multiplied by R_{n+2} gives a right hand side that has the right structure if K_{n+1} is chosen as in (2.119). Hence

$$A_{n+1} \;=\; \begin{bmatrix} A_n \\ 0 \end{bmatrix} + K_{n+1} \begin{bmatrix} 0 \\ J_{n+1} A_n \end{bmatrix} \;=\; (I_{n+2} + K_{n+1}\,J_{n+2}) \begin{bmatrix} A_n \\ 0 \end{bmatrix} \tag{2.121}$$

and in particular, $A_{n+1,n+1} = K_{n+1}$. Since we have found A_{n+1}, the top element of the RHS must be $\sigma^2_{f,n+1}$. Hence,

$$\underbrace{\sigma^2_{f,n+1}}_{\geq 0} = \sigma^2_{f,n} + K_{n+1}\Delta_{n+1} = \underbrace{\sigma^2_{f,n}}_{>0}\underbrace{\left(1 \Leftrightarrow K^2_{n+1}\right)}_{\geq 0} \tag{2.122}$$

from which it follows that

$$|K_{n+1}| \leq 1 \ . \tag{2.123}$$

We have assumed that $\sigma^2_{f,n} > 0$ (strictly positive). In fact, if it turns out that $\sigma^2_{f,n+1} = 0$, then the recursion stops. In that case, $|K_{n+1}| = 1 \Rightarrow K_{n+1} = \pm 1$. In fact, one finds in that case that A_{n+1} is either symmetric ($A_{n+1} = J\,A_{n+1}$ if $K_{n+1} = 1$) or antisymmetric ($A_{n+1} = \Leftrightarrow J\,A_{n+1}$ if $K_{n+1} = \Leftrightarrow 1$). In either case, one finds that $\Delta_{n+2} = 0$ and hence

$$A_{n+2} = \begin{bmatrix} A_{n+1} \\ 0 \end{bmatrix} \ , \ \sigma^2_{f,n+2} = 0 \tag{2.124}$$

is the solution. Solutions of higher order can be found by adding more zeros at the bottom of the filter coefficient vector. The following conditions are equivalent to describe this singular condition:

$$|K_{n+1}| = 1 \ \Leftrightarrow \ \sigma^2_{f,n+1} = 0 \ \Leftrightarrow \ \det\,(R_{n+2}) = 0 \ . \tag{2.125}$$

In this case the covariance matrix R_{n+2} is singular. A covariance matrix can only be singular when the signal consists of a sum of sinusoids. The rank of the covariance matrix augments by two per sinusoid that is present in the signal. Since $\sigma^2_{f,n+1} = 0$, a sum of m sinusoids is perfectly predictible if the prediction order is at least $2m$.

We can summarize the developments above into the Levinson algorithm as follows.

Levinson algorithm:

$$\begin{Bmatrix} A_n \\ \sigma^2_{f,n} \end{Bmatrix} \Rightarrow \begin{Bmatrix} A_{n+1} \\ \sigma^2_{f,n+1} \end{Bmatrix}$$

$$\begin{aligned} \Delta_{n+1} &= [r_{n+1} \cdots r_1]\, A_n \\ K_{n+1} &= \Leftrightarrow \tfrac{\Delta_{n+1}}{\sigma^2_{f,n}} \\ A_{n+1} &= \begin{bmatrix} A_n \\ 0 \end{bmatrix} + K_{n+1} \begin{bmatrix} 0 \\ J\,A_n \end{bmatrix} \\ \sigma^2_{f,n+1} &= \sigma^2_{f,n}\left(1 \Leftrightarrow K^2_{n+1}\right) \end{aligned} \tag{2.126}$$

Initialization: $A_0 = [1]$, $\sigma^2_{f,0} = r_0$.

Per recursion, the Levinson algorithm needs about $2n$ multiplications and a similar amount of additions. So when the algorithm is run up to some full order N, the total computational complexity is

$$\sum_{n=1}^{N} 2n = \frac{2N(N+1)}{2} \approx N^2 \ . \tag{2.127}$$

So we really have a fast algorithm for finding A_N. Furthermore, due to the *order-recursiveness* of the algorithm, not only the full order filter coefficients A_N are found bu also all lower order solutions A_n, $1 \leq n \leq N$.

The coefficients $\Leftrightarrow K_{n+1}$ have an interpretation as correlation coefficients. Indeed, from (2.118), one can establish

$$\begin{aligned} \Leftrightarrow K_{n+1} &= \tfrac{\Delta_{n+1}}{\sigma^2_{f,n}} = \sigma^{-2}_{f,n}\left[\sigma^2_{f,n}\ 0\cdots 0\ \Delta_{n+1}\right][0\ A_{n,n}\cdots A_{n,1}\ 1]^T = \sigma^{-2}_{f,n}\left[A_n^T\ 0\right] R_{n+2}\begin{bmatrix} 0 \\ J\,A_n \end{bmatrix} \\ &= \tfrac{EA_n^T Y_{n+1}(k) Y^T_{n+1}(k-1) B_n}{\sigma_{f,n}\ \sigma_{b,n}} = \tfrac{Ef_{n,k} b_{n,k-1}}{\sqrt{Ef^2_{n,k}}\ \sqrt{Eb^2_{n,k-1}}} = \tfrac{\langle f_{n,k}, b_{n,k-1}\rangle}{\|f_{n,k}\|\ \|b_{n,k-1}\|} \end{aligned} \tag{2.128}$$

This coefficient is in fact called *Partial Correlation* (PARCOR) coefficient because it describes the partial correlation between y_k and y_{k-n-1}, partial because the influence of $Y_n(k\Leftrightarrow 1)$ in between those two is removed in (2.128), see Fig. 2.16. When we apply the Cauchy-Schwarz formula to (2.128), then we find immediately $|K_{n+1}| \leq 1$ back.

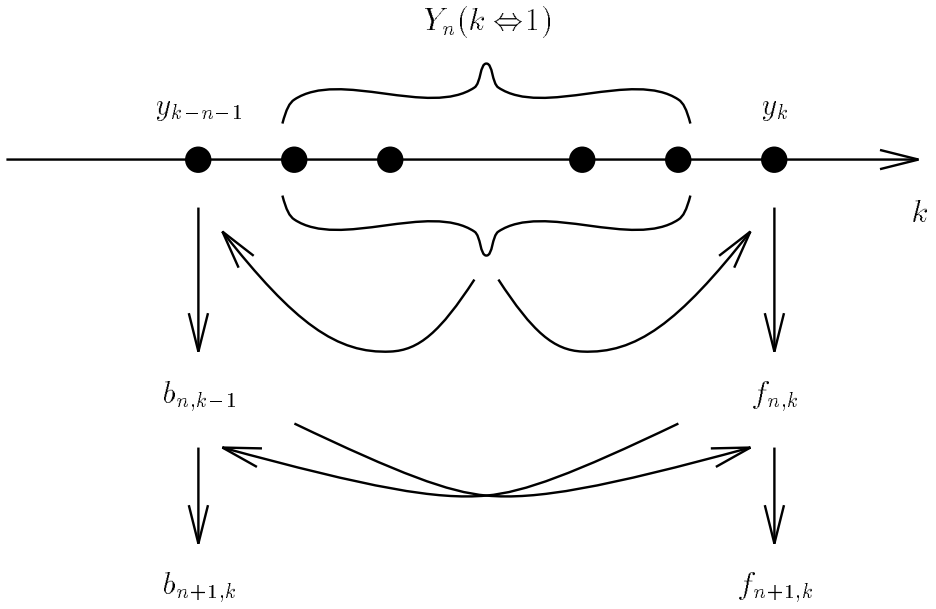

Figure 2.16: Forward and backward prediction errors.

Let us consider the z-transforms of the prediction error filter impulse responses:

$$[\mathrm{A}_n(z)\ \ \mathrm{B}_n(z)] = \left[1\ \ z^{-1}\cdots z^{-n}\right][A_n\ \ B_n]\ . \tag{2.129}$$

The property $B_n = J\,A_n$ translates to

$$\mathrm{B}_n(z) = z^{-n}\,\mathrm{A}_n(z^{-1})\ . \tag{2.130}$$

The prediction error filters have the following property.

> The positive definiteness of R_{n+1} has as a consequence that the filter $\mathrm{A}_n(z)$ is minimum-phase, i.e. has all its zeros inside the unit circle (on the unit circle when R_{n+1} is singular). This implies in particular that $1/\mathrm{A}_n(z)$ is guaranteed to be an exponentially stable filter.

Proof: Let z_i be any zero of $\mathrm{A}_n(z)$: $\mathrm{A}_n(z) = (1 \Leftrightarrow z_i\, z^{-1})\mathrm{P}_{n-1}^{(i)}(z)$, where $\mathrm{P}_{n-1}^{(i)}(z)$ is a monic polynomial in z^{-1} of order $n \Leftrightarrow 1$. Let $x_k^{(i)} = \mathrm{P}_{n-1}^{(i)}(q)\, y_k = P_{n-1}^{(i)T}\, Y_n(k)$, see Fig. 2.17. Note that z_i and hence $P_{n-1}^{(i)}$, $x_k^{(i)}$ can be complex. Using the orthogonality property of MMSE linear prediction, we find the following orthogonality

$$E\, x_{k-1}^{(i)}\, f_{n,k} \;=\; P_{n-1}^{(i)T}\, \underbrace{E\, Y_n(k \Leftrightarrow 1) f_{n,k}}_{=\,0} \;=\; 0\;. \tag{2.131}$$

From $f_{n,k} = x_k^{(i)} \Leftrightarrow z_i\, x_{k-1}^{(i)}$ follows $x_k^{(i)} = f_{n,k} + z_i\, x_{k-1}^{(i)}$. By taking the variance of both sides of this expression we find (all processes involved are zero mean)

$$\begin{aligned} \sigma^2_{x^{(i)}} \;&=\; E\,|x_k^{(i)}|^2 \;=\; E\,\left(f_{n,k}+z_i\, x_{k-1}^{(i)}\right)^*\left(f_{n,k}+z_i\, x_{k-1}^{(i)}\right) \\ &=\; \sigma^2_{f,n} + |z_i|^2\sigma^2_{x^{(i)}} + 2\Re\left\{ z_i\, \underbrace{E\, x_{k-1}^{(i)} f_{n,k}}_{=\,0} \right\} \\ \Rightarrow\quad (1 \Leftrightarrow |z_i|^2)\;\sigma^2_{x^{(i)}} \;&=\; \sigma^2_{f,n}\;. \end{aligned} \tag{2.132}$$

If we assume $R_m > 0,\ \forall m$ then

$$\left\{ \begin{array}{rcl} \sigma^2_{f,n} &=& A_n^T R_{n+1}\, A_n \;>\; 0 \\ \sigma^2_{x^{(i)}} &=& P_{n-1}^{(i)*T} R_n\, P_{n-1}^{(i)} \;>\; 0 \end{array} \right. \tag{2.133}$$

since $A_n \neq 0,\ P_{n-1}^{(i)} \neq 0$. Combining (2.132) and (2.133), we find

$$1 \Leftrightarrow |z_i|^2 > 0 \quad\Rightarrow\quad |z_i| < 1\,,\ i = 1,\ldots,n \quad\Rightarrow\quad \mathcal{A}_n(z) \text{ minimum-phase.} \tag{2.134}$$

If on the other hand $R_n > 0,\ n = 1,\ldots,N$ and R_{N+1} is singular, then $\sigma^2_{x^{(i)}} > 0$ as before. The singularity of R_{n+1} implies $\exists X = [x_0\ x_1 \cdots x_N]^T \in \mathcal{R}^{N+1}, X \neq 0 :\ R_{N+1}X = 0 \ \Rightarrow$ $X^T R_{N+1} X = 0$. Since $R_N > 0$, then necessarily $x_0 \neq 0$. Then $\sigma^2_{f,N} = \min\limits_{A_{N,1}\cdots A_{N,N}} A_N^T R_{N+1} A_N =$ 0 with $A_N = \frac{1}{x_0} X$. Hence $|z_i| = 1,\ i = 1,\ldots,N$. This means that $z_i = e^{\pm j\omega_i}$ where we indicated explicitly that the zeros of a polynomial with real coefficients are either real or occur in complex conjugate pairs. Since $E\, f_{N,k} = E\, f_{N,k}^2 = 0$, it follows that $f_{N,k} \equiv 0$ w.p. 1. The solution to the homogeneous difference equation $\mathrm{A}_N(q) y_k = \prod\limits_{i=1}^{N} (1 \Leftrightarrow z_i q^{-1})\ y_k = f_{N,k} = 0$ can be written as

$$y_k \;=\; \sum_i \left(\frac{A_i}{2} e^{j\phi_i} z_i^k + \frac{A_i}{2} e^{-j\phi_i} z_i^{*k}\right) \;=\; \sum_i \left(\frac{A_i}{2} e^{j\phi_i} e^{j\omega_i k} + \frac{A_i}{2} e^{-j\phi_i} e^{-j\omega_i k}\right) \;=\; \sum_i A_i \cos(\omega_i k + \phi_i) \tag{2.135}$$

where we assumed that all the roots z_i have multiplicity one and the number of sinusoids is such that the number of exponentials (= the number of roots z_i) is equal to N. We also exploited the fact that in order for the signal y_k to be real, the coefficients $\frac{A_i}{2} e^{j\phi_i}, \frac{A_i}{2} e^{-j\phi_i}$ corresponding to a complex conjugate pair of roots z_i, z_i^* have indeed to occur as a complex conjugate pair. The quantities A_i, ϕ_i may be random. In particular, for the process y_k to be random, the angles ϕ_i need to be distributed uniformly (for $\omega_i \in (0, \pi)$). We conclude that

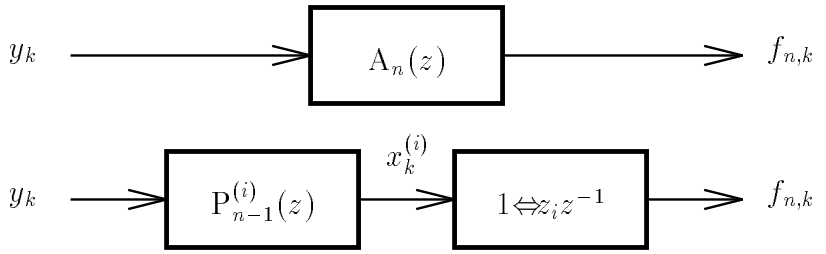

Figure 2.17: Splitting off one zero of the prediction error filter.

the only way for a stationary covariance matrix to be singular is for the signal to consist of a sum of sinusoids. In that case, the rank of the matrix corresponds to the number of the exponentials in the signal. □

The minimum phase property of $A_N(z)$ is also related to the boundedness of the PARCORs and in fact leads to the following property.

> Schur-Cohn Test: A polynomial $A_N(z)$ is minimum-phase if and only if the sequence of PARCORs is bounded by one: $|K_n| < 1,\ n = N, N\Leftrightarrow 1, \ldots, 1$.

In order to be able to apply the Schur-Cohn test, we have to know how to find the PARCORs from the filter $A_n(z)$. Whereas the Levinson algorithm is essentially the following *step-up* procedure:

$$A_{n+1} \;=\; \begin{bmatrix} A_n \\ 0 \end{bmatrix} \;+\; K_{n+1} \begin{bmatrix} 0 \\ J\,A_n \end{bmatrix} \;=\; (I + K_{n+1}\,J) \begin{bmatrix} A_n \\ 0 \end{bmatrix} \tag{2.136}$$

this procedure can be inverted to yield the following *step-down* procedure:

$$\begin{bmatrix} A_n \\ 0 \end{bmatrix} \;=\; (I + K_{n+1}\,J)^{-1}\,A_{n+1} \;=\; \frac{1}{1 \Leftrightarrow K_{n+1}^2}\,(I \Leftrightarrow K_{n+1}\,J)\,A_{n+1}\;,\quad K_{n+1} = A_{n+1,n+1} \tag{2.137}$$

which can be reiterated to yield all PARCORs, starting from the highest order polynomial.

In fact, we have the following set of equivalences

$$\begin{array}{ccc} \boxed{R_{N+1} > 0} & \overset{(a)}{\Leftrightarrow} & \boxed{\begin{array}{c}\sigma_{f,n}^2 > 0 \\ n = 0, 1, \ldots, N\end{array}} \\ (c)\ \downarrow\uparrow & & \downarrow\uparrow\ (b) \\ \boxed{\begin{array}{c}A_N(z) \text{ minimum phase} \\ \sigma_{f,N}^2 > 0\end{array}} & \overset{(d)}{\Leftrightarrow} & \boxed{\begin{array}{c}r_0 = \sigma_{f,0}^2 > 0\,,\quad |K_n| < 1 \\ n = N, N\Leftrightarrow 1, \ldots, 1\end{array}} \end{array} \tag{2.138}$$

The Schur-Cohn test, which is a subset of equivalence (d), can perhaps most easily be shown by using Rouché's theorem of complex variable theory. We mentioned equivalence (c) before. Equivalence (b) follows straightforwardly from (2.122). The equivalences in (2.138) are not

only equivalences of properties but also reveal equivalent parameterizations. Indeed the following three sets of parameters are equivalent in the sense that one can be found from the other one: $\{r_0, r_1, \ldots, r_N\}$, $\left\{\sigma^2_{f,N}, A_{N,1}, \ldots, A_{N,N}\right\}$, $\{r_0, K_1, \ldots, K_N\}$. The set of parameters $\left\{\sigma^2_{f,0}, \sigma^2_{f,1}, \ldots, \sigma^2_{f,N}\right\}$ is not equivalent since it contains a bit less information that the other three parameterizations (the missing information is the sign of the K_n).

2.4.5 Lattice Filter Realizations

We can write out the step-up (Levinson) procedure jointly for forward and backward prediction error filters (using $B_{n+1} = J\,A_{n+1}$):

$$\begin{aligned} A_{n+1} &= \begin{bmatrix} A_n \\ 0 \end{bmatrix} + K_{n+1} \begin{bmatrix} 0 \\ B_n \end{bmatrix} \\ B_{n+1} &= \begin{bmatrix} 0 \\ B_n \end{bmatrix} + K_{n+1} \begin{bmatrix} A_n \\ 0 \end{bmatrix} \end{aligned} \tag{2.139}$$

If we multiply all sides with $[1 \;\; z^{-1} \;\; \cdots \;\; z^{-n-1}]$, we get

$$\begin{aligned} \mathrm{A}_{n+1}(z) &= \mathrm{A}_n(z) + K_{n+1}\, z^{-1}\, \mathrm{B}_n(z) \\ \mathrm{B}_{n+1}(z) &= K_{n+1} \mathrm{A}_n(z) + z^{-1}\, \mathrm{B}_n(z) \end{aligned} \tag{2.140}$$

which can be rewritten as

$$\begin{bmatrix} \mathrm{A}_{n+1}(z) \\ \mathrm{B}_{n+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & K_{n+1} \\ K_{n+1} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \begin{bmatrix} \mathrm{A}_n(z) \\ \mathrm{B}_n(z) \end{bmatrix}, \quad \begin{bmatrix} \mathrm{A}_0(z) \\ \mathrm{B}_0(z) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{2.141}$$

This formula describes one lattice section. From this formula, it is straightforward to draw the realization of the complete lattice filter, by cascading lattice sections and taking the proper initialization into account:

$$\begin{bmatrix} \mathrm{A}_N(z) \\ \mathrm{B}_N(z) \end{bmatrix} = \begin{bmatrix} 1 & K_N \\ K_N & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & K_1 \\ K_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{2.142}$$

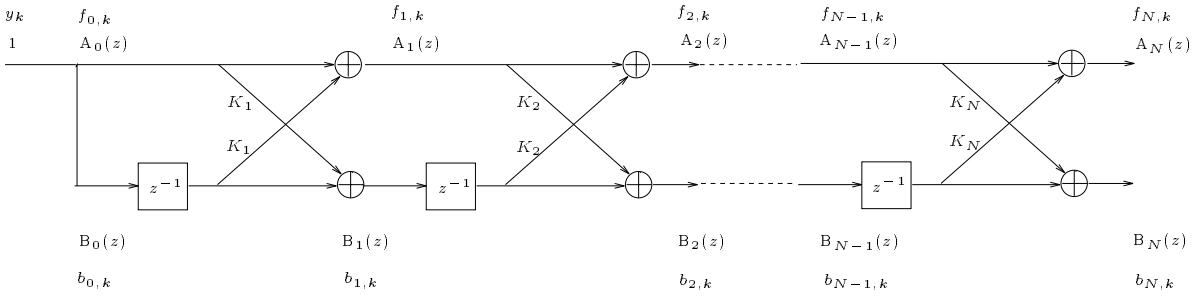

Figure 2.18: Realization of forward and backward prediction via the analysis lattice filter.

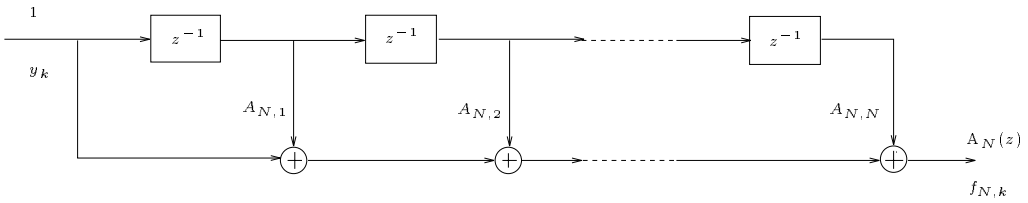

Figure 2.19: The equivalent "tapped delay line" or transversal filter realization of the FIR prediction error filter.

By multiplying both sides of the equations (2.139) to the left with $Y_{n+1}^T(k)$ or equivalently, by replacing z by q in equation (2.141) and letting both sides of the equation operate on the signal y_k we find

$$\begin{bmatrix} \mathrm{A}_{n+1}(q)y_k \\ \mathrm{B}_{n+1}(q)y_k \end{bmatrix} = \begin{bmatrix} 1 & K_{n+1} \\ K_{n+1} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & q^{-1} \end{bmatrix} \begin{bmatrix} \mathrm{A}_n(q)y_k \\ \mathrm{B}_n(q)y_k \end{bmatrix}, \quad \begin{bmatrix} \mathrm{A}_0(q)y_k \\ \mathrm{B}_0(q)y_k \end{bmatrix} = \begin{bmatrix} y_k \\ y_k \end{bmatrix}. \tag{2.143}$$

Since by definition $f_{n,k} = \mathrm{A}_n(q)y_k$, $b_{n,k} = \mathrm{B}_n(q)y_k$, we can write (2.143) as

$$\begin{bmatrix} f_{n+1,k} \\ b_{n+1,k} \end{bmatrix} = \begin{bmatrix} 1 & K_{n+1} \\ K_{n+1} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & q^{-1} \end{bmatrix} \begin{bmatrix} f_{n,k} \\ b_{n,k} \end{bmatrix} = \begin{bmatrix} f_{n,k} + K_{n+1}b_{n,k-1} \\ b_{n,k-1} + K_{n+1}f_{n,k} \end{bmatrix} \tag{2.144}$$

and $f_{0,k} = b_{0,k} = y_k$. The prediction errors of all orders up to the maximum order show up in the lattice filter if we excite it with the signal y_k as indicated in Fig. 2.18. This property is the *order-recursiveness* of the lattice filter structure.

Geometric Derivation of the Levinson Algorithm

The order recursions for the prediction errors found in (2.144), namely

$$\begin{cases} f_{n+1,k} &= f_{n,k} + K_{n+1}b_{n,k-1} \\ b_{n+1,k} &= b_{n,k-1} + K_{n+1}f_{n,k} \end{cases} \tag{2.145}$$

could in fact have been derived directly. Indeed, consider Fig. 2.16. $f_{n,k}$ is the part of y_k that is uncorrelated with (orthogonal to) $Y_n(k\Leftrightarrow 1) = [y_{k-1} \cdots y_{k-n}]^T$. $f_{n+1,k}$ is furthermore uncorrelated with y_{k-n-1}. To obtain $f_{n+1,k}$ from $f_{n,k}$, we need to orthogonalize $f_{n,k}$ w.r.t. y_{k-n-1}. However, we should orthogonalize w.r.t. y_{k-n-1} without destroying the orthogonality to $Y_n(k\Leftrightarrow 1)$! So we should orthogonalize $f_{n,k}$ w.r.t. the part of y_{k-n-1} that is orthogonal to $Y_n(k\Leftrightarrow 1)$ also, which is $b_{n,k-1}$. Orthogonalizing $f_{n,k}$ w.r.t. $b_{n,k-1}$ is done by LMMSE estimation, so $f_{n+1,k}$ is obtained as the estimation error in LMMSE estimating $f_{n,k}$ from $b_{n,k-1}$, so we get

$$f_{n+1,k} = \widetilde{f_{n,k}} = f_{n,k} \Leftrightarrow \widehat{f_{n,k}} = f_{n,k} \Leftrightarrow \frac{E\, f_{n,k}b_{n,k-1}}{E\, b_{n,k-1}^2}\, b_{n,k-1} = f_{n,k} \Leftrightarrow K_{n+1}\, b_{n,k-1} \tag{2.146}$$

and similarly for the update from $b_{n,k-1}$ to $b_{n+1,k}$. By writing now the prediction errors in (2.145) as outputs of prediction error filters and realizing that the relations thus obtained hold

for any realization of the inputs y_k, we obtain the corresponding recursions for the prediction filters, which constitute the Levinson recursions. This alternative derivation for the Levinson algorithm constitutes a geometric derivation since it uses the geometric (orthogonality in a vector space of random variables) arguments of LMMSE estimation.

Advantages of Lattice Filters

Now let's return to the filter realizations. The lattice filter has some numerical advantages (especially in fixed-point implementation). First of all, the multipliers that appear in the lattice filter are the PARCORs which are bounded by 1 in magnitude: $|K_n| < 1$. Secondly, the various prediction errors have lower variance than the input signal: $\sigma^2_{f,n} \leq \sigma^2_{f,0} = \sigma^2_y$. Therefore, if the input signal is scaled to be (approximately) in the interval $[\Leftrightarrow 1, +1]$ (by taking e.g. $\sigma_y = 0.25$; for a Gaussian random variable y_k, $\Pr(|y_k| > 4\sigma_y) \approx 0$), then so will be (approximately) all the signals appearing at all the internal nodes in the filter (those signals are all forward or backward prediction errors). The transfer function has also fairly low sensitivity to perturbations of the filter coefficients K_n. This is in sharp contrast with the transversal filter realization. First of all, the coefficients $A_{N,n}$ of a minimum-phase filter are bounded, but the bound is a function of n and increases with N. Furthermore, the signal appearing at the nodes in the transversal filter circuit is the partial sum $\sum_{i=0}^{n} A_{N,i} y_{k-i}$ for $n = 0, 1, \ldots, N$. For $n = 1, \ldots, N \Leftrightarrow 1$, this signal is not related to $f_{n,k}$ and can have a variance that gets larger than σ^2_y.

Synthesis Lattice Filters

Whereas for linear prediction we are interested in the analysis lattice filter that realizes the FIR filter $\mathrm{A}(z)$, for modeling we are interested in the synthesis lattice filter that realizes the all-pole filter $1/\mathrm{A}(z)$. We are interested in the synthesis lattice filter because it allows us to realize the transfer function $1/\mathrm{A}(z)$ with the same coefficients K_n as for the analysis lattice filter. And we are intersted in realizing the transfer function $1/\mathrm{A}(z)$ in order to generate an AR process starting from white noise. The synthesis lattice can be obtained from the analysis lattice by straightforward flowgraph manipulations. The roles of input and output get interchanged. So we change the direction of the flow as indicated in Fig. 2.20.

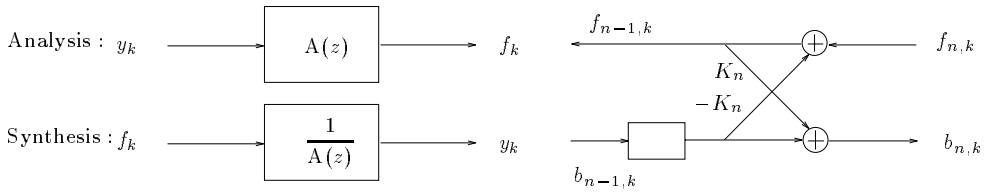

Figure 2.20: Analysis and synthesis configurations and a synthesis lattice section.

Since it is usual to have the input on the left and the output on the right, we shall flip the above synthesis lattice section around. This yields the result in Fig. 2.21, which can

also be transformed into the so-called 3-multiplier lattice section (whereas the lattice sections considered so far are 2-multiplier sections).

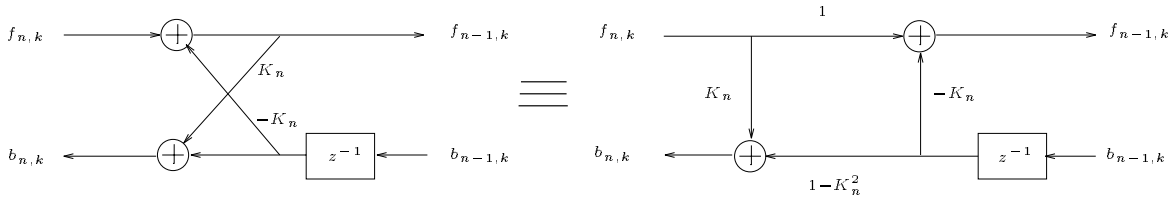

Figure 2.21: Alternative synthesis lattice section and its equivalent three-multiplier form.

The general presentation of the all-pole IIR synthesis lattice filter is given in Fig. 2.22 and may help to understand the following manipulations. The advantage of the lattice filter realization of $1/\mathrm{A}(z)$ is that it suffices to contol the magnitudes of the coefficients K_n ($|K_n| < 1$) for $1/\mathrm{A}(z)$ to be stable.

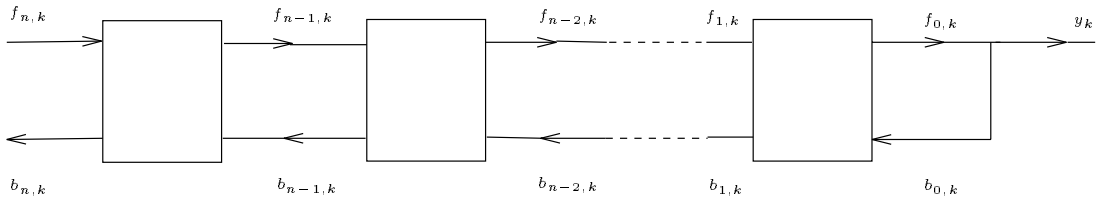

Figure 2.22: The synthesis lattice filter.

Suppose we factor the gain $1 \Leftrightarrow K_n^2$ of the lower branch into two factors, one of which (c_k) is moved to the upper branch as indicated in Fig. 2.23, then this will not change the loop gain of the feedback loop and so the dynamics remain unaltered. However, the net effect of such an operation is that all signals to the right are amplified by the factor c_k appearing in the upper branch. As a result, the overall transfer function becomes

$$\frac{\prod_{n=1}^{N} c_n}{\mathrm{A}_N(z)} \,. \tag{2.147}$$

Two interesting choices for the factor c_n emerge and are depicted in Figures 2.24 and 2.25. The first choice corresponds to $c_n = \sqrt{1 \Leftrightarrow K_n^2}$ and leads to the so-called 4-multiplier lattice or *normalized* lattice. This lattice has very good numerical properties since the input-output behavior of the static part of the lattice section is a 2×2 orthogonal rotation, which conserves energy. In a normalized synthesis lattice, all signals have the same variance if they are the prediction error signals for which the K_n are the PARCORs. Furthermore, in a normalized synthesis lattice all signals still have the same variance if the filter is driven by white noise ($= f_{N,k}$ for an AR(N) process). So the choice $c_n = \sqrt{1 \Leftrightarrow K_n^2}$ leads to a synthesis lattice structure with attractive properties for fixed-point implementation.

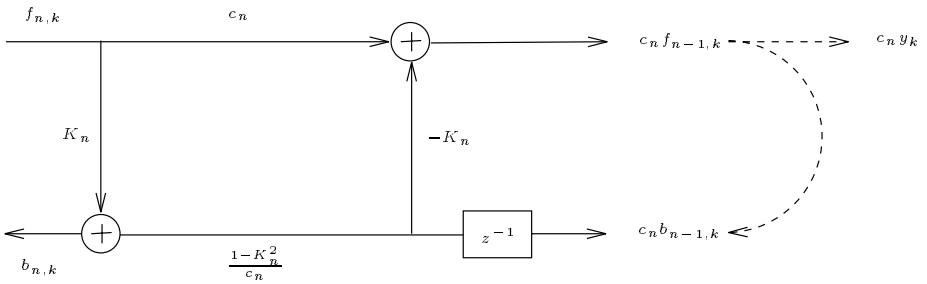

Figure 2.23: Transformations of the synthesis lattice sections.

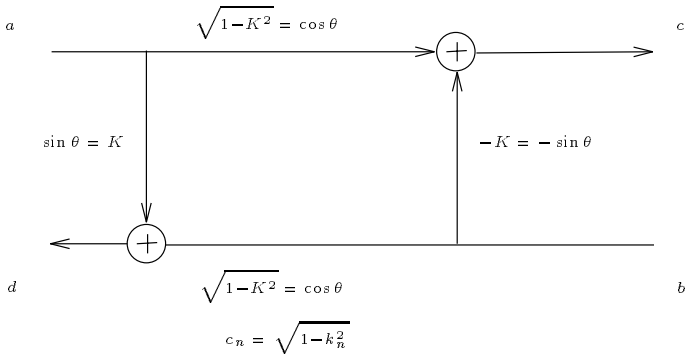

Figure 2.24: The normalized lattice section.

The second choice we shall consider is irrelevant for implementation considerations but shall allow us to give a physical interpretation for the coefficients K_n. The second choice correponds to $c_n = 1 \Leftrightarrow K_n$ and leads to the Kelly-Lochbaum lattice which corresponds exactly to a section of a transmission line. For this reason, the PARCORs are also called *reflection coefficients*. In Figure 2.25, we consider an acoustic transmission line consisting of a cascade of coaxial pieces of tube with cross-sectional surface A_n and all of the same length. Acoustic waves are travelling forward and backward. They get partially reflected when they meet a discontinuity in cross-section. If the portion K gets reflected, then the portion $1 \Leftrightarrow K$ continues ($K + 1 \Leftrightarrow K = 1$ for continuity reasons). Coming from the other side, we see exactly the opposite discontinuity and so the reflection coefficient there is $\Leftrightarrow K$. Similarly to the introduction of c_k, we can factor the delay z^{-1} in the bottom line of the lattice into $z^{-1} = z^{-1/2} z^{-1/2}$ and move one factor $z^{-1/2}$ to the top line without altering the amplitude response of the lattice filter. In this way we obtain an equal delay in bottom and top lines of the lattice section, which corresponds to the delay for the acoustic waves to travel over the length of one tube section. This delay is indeed independent of the direction (forward or backward) of the wave. This acoustic tube model arises in speech production analysis. So there exists a close relationship between the all-pole synthesis lattice and a speech production model.

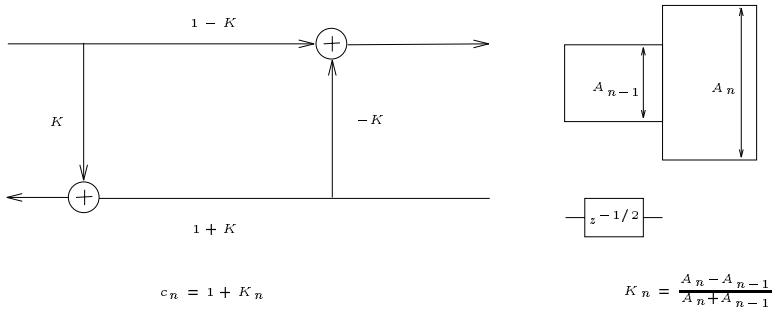

Figure 2.25: The Kelly-Lochbaum lattice section.

2.5 Spectral Estimation by AR Modeling: Motivations and Interpretations

2.5.1 Linear Prediction of an AR(N) process

It will be most interesting to see what happens when applying linear prediction to an autoregressive process. By comparing the normal equations (2.102) of linear prediction with the Yule-Walker equations (2.83) for the coefficients of an autoregressive process, we see that they are identical. Since a nonsingular set of linear equations has a unique solution, we find that for an AR(N) process, the prediction filter $\mathrm{A}_N(z)$ coincides with the numerator of the all-pole filter that generates the autoregressive process from white noise.

There is alternative way to come to this conclusion, which explains why for an autoregressive process minimizing the prediction error variance at the same time whitens the prediction

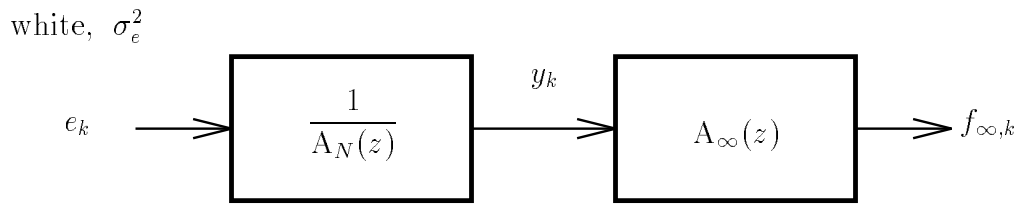

Figure 2.26: Linear prediction of an autoregressive process.

error. Consider linear prediction with an infinite prediction order applied to an AR(N) process, as illustrated in Figure 2.26. Let us denote the overall transfer function from the white input e_k to the prediction error signal $f_{\infty,k}$ as $\mathrm{H}(z) = \mathrm{A}_\infty(z)/\mathrm{A}_N(z) = \sum_{i=0}^{\infty} h_i z^{-i}$ where $h_0 = 1$ since $\mathrm{A}_N(z)$ and $\mathrm{A}_\infty(z)$ are monic. Note that $1/\mathrm{A}_N(z)$ is causal since $\mathrm{A}_N(z)$ is minimum-phase. Then we can write the variance of the prediction error as

$$\begin{aligned} \sigma_{f,\infty}^2 &= E\, f_{\infty,k}^2 = E\left(\sum_{i=0}^{\infty} h_i e_{k-i}\right)^2 = E\sum_{i=0}^{\infty}\sum_{l=0}^{\infty} h_i h_l e_{k-i} e_{k-l} \\ &= \sum_{i=0}^{\infty}\sum_{l=0}^{\infty} h_i h_l r_{ee}(i \Leftrightarrow l) = \sigma_e^2 \sum_{i=0}^{\infty}\sum_{l=0}^{\infty} h_i h_l \delta_{il} = \sigma_e^2 \sum_{i=0}^{\infty} h_i^2 = \sigma_e^2\,(1 + \sum_{i=1}^{\infty} h_i^2)\,. \end{aligned} \tag{2.148}$$

Minimization leads to

$$\min_{A_{\infty,i}, i>0} E\, f_{\infty,k}^2 = \min_{h_i, i>0} E\, f_{\infty,k}^2 = \sigma_e^2 \quad \text{for} \;\; h_i = 0\,,\; i > 0\,. \tag{2.149}$$

Hence $\mathrm{H}(z) = 1$ and $\mathrm{A}_\infty(z) = \mathrm{A}_N(z)$. We see that it was not really necessary to take an infinite prediction order to obtain this result. Indeed, for any prediction order $n \geq N$ we get

$$\mathrm{A}_n(z) \;=\; \mathrm{A}_N(z)\,, \quad n \geq N \tag{2.150}$$

where $\mathrm{A}_N(z)$ is the numerator of the all-pole filter. So for an AR(N) process, we get

$$K_n \;=\; A_{n,n} \;=\; 0\,, \quad n > N \tag{2.151}$$

which can be shown either by verifying that the Yule-Walker equations (2.83) lead to $\Delta_n = 0\,,\; n > N$, or by observing that $K_n = A_{n,n}$ from which (2.151) follows using (2.150). For an AR(N) process we also get

$$f_{n,k} \;=\; e_k \;=\; \text{white !}\,, \quad n \geq N\,. \tag{2.152}$$

A particular autoregressive process is white noise, which is indeed AR(0). Hence $\mathrm{A}_n(z) = 1,\; K_n = 0,\; \sigma_{f,n}^2 = r_0,\; f_{n,k} = y_k,\; n > 0$. In particular also $\widehat{y}_{n,k} = y_k \Leftrightarrow f_{n,k} = 0$: white noise is perfectly unpredictible.

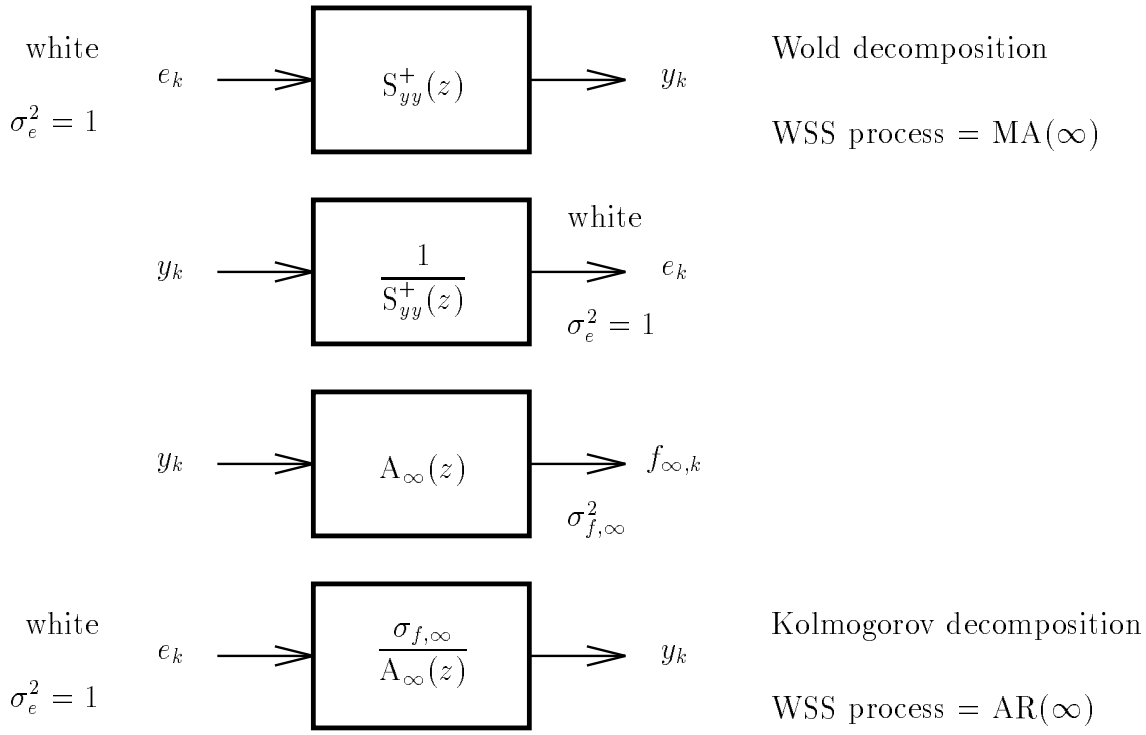

Figure 2.27: The Wold and Kolmogorov decompositions.

2.5.2 Asymptotics

Under some regularity conditions, the psdf $S_{yy}(z)$ of a wide-sense stationary (WSS) process y_k can be factored as

$$S_{yy}(z) \;=\; S_{yy}^+(z)\, S_{yy}^+(1/z) \tag{2.153}$$

where $S_{yy}^+(z)$ is a causal minimum-phase transfer function. Then automatically $S_{yy}^-(z) = S_{yy}^+(1/z)$ is anti-causal and maximum-phase. $S_{yy}^+(z)$ is unique and is called the *spectral factor* of $S_{yy}(z)$. The operation indicated in (2.153) is called *spectral factorization.*

It now is possible to interpret y_k as the output of the causal filter $S_{yy}^+(z)$ driven by white noise e_k with unit variance

$$y_k \;=\; S_{yy}^+(q)\, e_k\ , \quad \sigma_e^2 = 1 \tag{2.154}$$

as indicated in Figure 2.27. Indeed, if e_k with psdf $S_{ee}(f) = \sigma_e^2 = 1$ is driving a filter $S_{yy}^+(z)$ then its output y_k has as psdf

$$S_{yy}^+(z)S_{ee}(z)S_{yy}^+(1/z) \;=\; S_{yy}^+(z)\sigma_e^2 S_{yy}^+(1/z) \;=\; S_{yy}^+(z)S_{yy}^-(z) \;=\; S_{yy}(z) \tag{2.155}$$

which is indeed the psdf of y_k. This is the so-called *Wold decomposition* which represents any WSS process as a moving-average process of infinite order.

Alternatively, consider applying linear prediction of infinite order to y_k, yielding $f_{\infty,k}$. Let span $\{y_i,\ i \le k\}$ and span $\{f_{\infty,i},\ i \le k\}$ indicated the vector spaces generated (spanned) by the indicated random variables. Then since $f_{\infty,k}$ is obtained by causal linear filtering of y_k, the $f_{\infty,i},\ i \le k$, are linear combinations of the $y_i,\ i \le k$, and hence span $\{f_{\infty,i},\ i \le k\} \subset$ span $\{y_i,\ i \le k\}$. But $A_\infty(z)$ is minimum-phase and hence its inverse is causal. So the y_k can

be obtained by causal linear filtering of the $f_{\infty,k}$. Thus the $y_i,\ i \le k$, are linear combinations of the $f_{\infty,i},\ i \le k$, and hence span $\{f_{\infty,i},\ i \le k\} \supset$ span $\{y_i,\ i \le k\}$. By combining the above results, we find

$$\text{span}\,\{f_{\infty,i},\ i \le k\} \ = \ \text{span}\,\{y_i,\ i \le k\}\ . \tag{2.156}$$

Now, by the orthogonality condition of least-squares, we have

$$f_{\infty,k} \perp \text{span}\,\{y_i,\ i < k\} = \text{span}\,\{f_{\infty,i},\ i < k\} \ \Rightarrow \ f_{\infty,k} \perp \text{span}\,\{f_{\infty,i},\ i < k\} \tag{2.157}$$

which implies that $f_{\infty,k}$ is a white process! Hence

$$\sigma^2_{f,\infty} \ = \ \mathrm{S}_{f_\infty f_\infty}(z) \ = \ \mathrm{A}_\infty(z)\mathrm{S}_{yy}(z)\mathrm{A}_\infty(1/z) \tag{2.158}$$

$$\Rightarrow\ \mathrm{S}_{yy}(z) \ = \ \frac{\sigma^2_{f,\infty}}{\mathrm{A}_\infty(z)\mathrm{A}_\infty(1/z)} \tag{2.159}$$

$$\Rightarrow\ \mathrm{S}^+_{yy}(z) \ = \ \frac{\sigma_{f,\infty}}{\mathrm{A}_\infty(z)}\ , \quad \mathrm{A}_\infty(\infty) \ = \ 1 \tag{2.160}$$

$$\Rightarrow\ \sigma_{f,\infty} \ = \ \mathrm{S}^+_{yy}(\infty)\ , \quad \mathrm{A}_\infty(z) \ = \ \frac{\mathrm{S}^+_{yy}(\infty)}{\mathrm{S}^+_{yy}(z)} \tag{2.161}$$

All this means that any WSS process can be represented as an autoregressive process of infinite order as indicated in Figure 2.27. This is the so-called *Kolmogorov decomposition*. It can furthermore be shown that

$$\sigma^2_{f,\infty} \ = \ e^{\int_{-0.5}^{0.5} \ln S_{yy}(f)df}\ . \tag{2.162}$$

For purely random processes, $\sigma^2_{f,\infty} > 0$. Examples of not purely random processes are sinusoids or bandlimited processes ($\sigma^2_{f,\infty} = 0$). For white noise, we get again $\sigma^2_{f,\infty} \ = \ \sigma^2_y$.

As an example, consider an ARMA process $y_k = \dfrac{\mathrm{B}(q)}{\mathrm{A}(q)}e_k$ where $\mathrm{A}(z)$ and $\mathrm{B}(z)$ are monic and minimum phase and e_k is white. Let $\dfrac{\mathrm{B}(z)}{\mathrm{A}(z)} = \mathrm{H}(z)$ and $\dfrac{\mathrm{A}(z)}{\mathrm{B}(z)} = \mathrm{G}(z)$. $\mathrm{H}(z)$ and $\mathrm{G}(z)$ are again monic but of infinite length. We have

$$\mathrm{S}^+_{yy}(z) = \mathrm{H}(z)\sigma_e\ , \quad y_k = \mathrm{H}(q)\sigma_e\,\frac{e_k}{\sigma_e}\ , \quad y_k = \frac{\sigma_e}{\mathrm{G}(q)}\,\frac{e_k}{\sigma_e} \tag{2.163}$$

where the last two identities are the Wold and Kolmogorov representations respectively.

2.5.3 Autoregressive Modeling via Linear Prediction

Since $0 \le \sigma^2_{f,n} = \sigma^2_{f,n-1}(1 \Leftrightarrow K^2_n) \le \sigma^2_{f,n-1}$, $\sigma^2_{f,n}$ is a non-increasing function of n that reaches some non-negative asymptotic value $\sigma^2_{f,\infty}$, see Figure 2.28. When applied to a stationary segment of speech signal for example, the PARCORs are in general significantly different from zero for low prediction orders, but become small for n around 8 to 10. This means that the curve of $\sigma^2_{f,n}$ decreases rapidly at low orders, but starts to flatten out at n around 8 to 10: $\sigma^2_{f,10} \gtrapprox \sigma^2_{f,\infty}$. In practice, either 8 or 10 are normally used as prediction order.

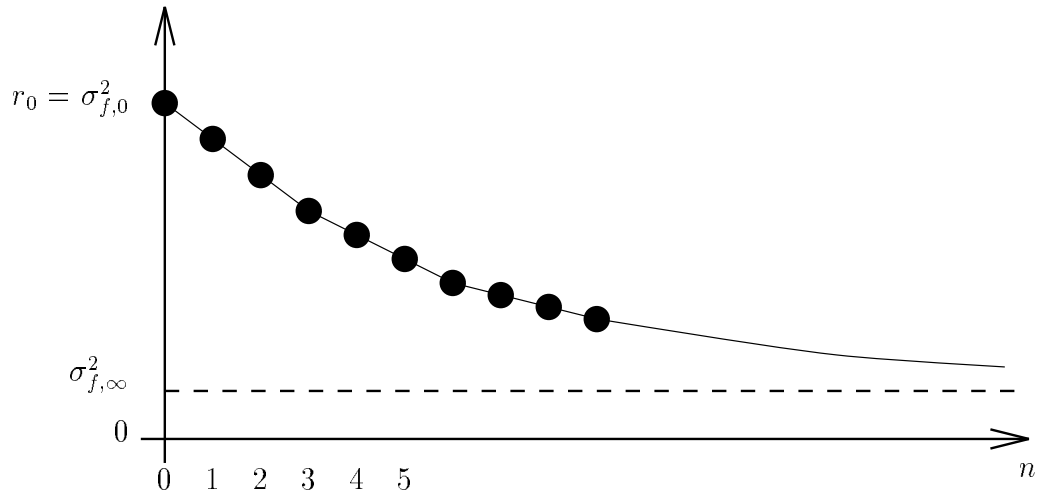

Figure 2.28: The evolution of the prediction error variance as a function of the prediction order.

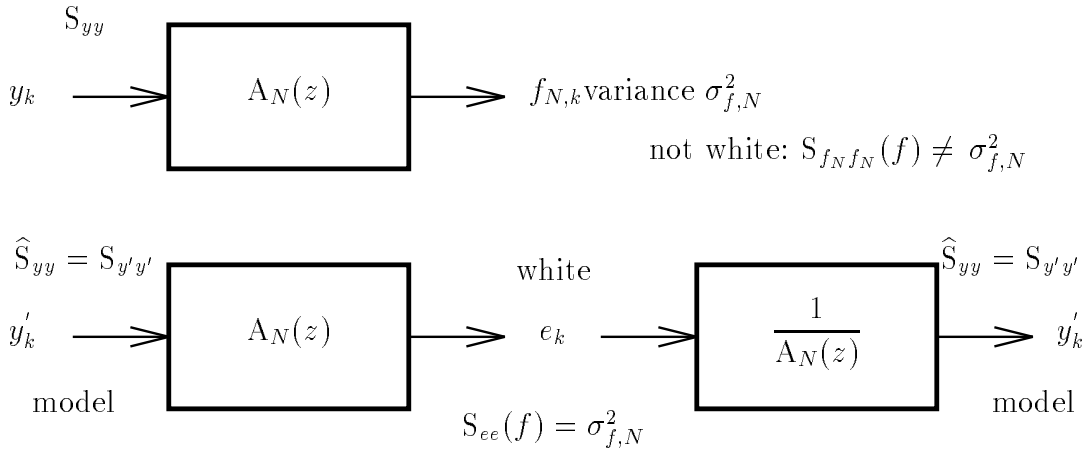

Figure 2.29: Autoregressive modeling.

The autoregressive modeling approach is depicted in Figure 2.29. The motivation is the following. The infinite-order prediction error $f_{\infty,k}$ is a white noise process and hence is unpredictible. The original process y_k is considered very predictible if $\sigma^2_{f,\infty} \ll r_0$. Assume now that for a certain finite order N we have $\sigma^2_{f,N} \gtrapprox \sigma^2_{f,\infty}$. Then $f_{N,k}$ is no longer very predictible and can therefore be considered to be approximately white. The autoregressive modeling operation consists in considering $f_{N,k}$ to be exactly white with variance $\sigma^2_{f,N}$. This implies that the power spectral density of the signal y_k can be approximated by the inverse of the magnitude squared of the Fourier transform of the prediction error filter. This means that $\mathrm{S}_{yy}(z)$ is modeled (approximated) by

$$\widehat{\mathrm{S}}_{yy}(z) \;=\; \frac{\widehat{\mathrm{S}}_{f_N f_N}(z)}{\mathrm{A}_N(z)\mathrm{A}_N(1/z)} \;=\; \frac{\sigma^2_{f,N}}{\mathrm{A}_N(z)\mathrm{A}_N(1/z)} \;\approx\; \frac{\mathrm{S}_{f_N f_N}(z)}{\mathrm{A}_N(z)\mathrm{A}_N(1/z)} \;=\; \mathrm{S}_{yy}(z)\;. \qquad (2.164)$$

The hat should here not be interpreted in the sense of estimation but as denoting an approximation. In fact, when y_k is an AR(n) process of order $n \leq N$, then $f_{N,k}$ will be exactly white and the approximation just mentioned will hold exactly.

We shall now consider five more motivations and interpretations of this autoregressive modeling procedure.

2.5.4 AR Modeling: Spectral Interpretation

Some effects of the AR modeling operation can be understood by considering the prediction error variance minimization in the frequency domain:

$$\sigma^2_{f,N} = E\, f^2_{N,k} = \int_{-0.5}^{0.5} S_{f_N f_N}(f)df = \min_{A_{N,i}\, i=1,\ldots,N} \int_{-0.5}^{0.5} |A_N(f)|^2\, S_{yy}(f)df\;. \qquad (2.165)$$

The degrees of freedom $A_{N,i}$ are used to minimize the total contribution of $|A_N(f)|^2\, S_{yy}(f) \geq 0$. $\sigma^2_{f,N}$ is made small by making $|A_N(f)|$ small (this helps to explain intuitively the minimum-phase property of the optimal $\mathrm{A}_N(z)$). However, with a limited number of parameters $A_{N,i}$, $|A_N(f)|$ cannot be made small at all frequencies. Hence, $|A_N(f)|$ should preferably be small at those frequencies where $S_{yy}(f)$ is big in order to minimize $\sigma^2_{f,N}$ well (this explains why $\mathrm{A}_N(z)$ has its zeros on the unit circle if y_k consists of at most N complex sinusoids, in which case $\sigma^2_{f,N} = 0$). So $|A_N(f)|$ will be such that $|A_N(f)|^2\, S_{yy}(f)$ will not exceed its average value by much when $\sigma^2_{f,N}$ when $S_{yy}(f)$ is big, but $|A_N(f)|$ will not do much to avoid that $|A_N(f)|^2\, S_{yy}(f)$ is small when $S_{yy}(f)$ is small. Hence $\widehat{S}_{yy}(f) = \widehat{\mathrm{S}}_{yy}(e^{j2\pi f})$ will match $S_{yy}(f)$ closely at the peaks of $S_{yy}(f)$. We say that the AR model follows the (most significant) peaks of $S_{yy}(f)$.

2.5.5 AR Modeling: Covariance Matching

Suppose we want to model y_k by an AR(N) process y'_k. This means that we want to approximate $\mathrm{S}_{yy}(z)$ by

$$\widehat{\mathrm{S}}_{yy}(z) \;=\; \mathrm{S}_{y'y'} \;=\; \frac{\sigma^2_{f,N}}{\mathrm{A}_N(z)\mathrm{A}_N(1/z)}\;. \qquad (2.166)$$

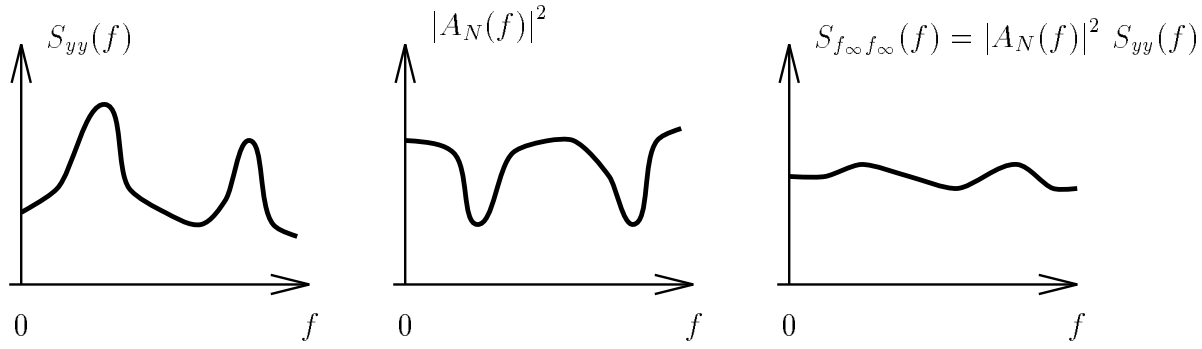

Figure 2.30: The spectral interpretation of autoregressive modeling.

This means that we choose a particular parametric form for $\widehat{\mathrm{S}}_{yy}(z)$ in which the parameters $\sigma^2_{f,N},\ A_{N,1},\ldots,A_{N,N}$ are to be determined. Consider now fixing these parameters by introducing $N+1$ covariance matching constraints:

$$\int_{-0.5}^{0.5} \widehat{S}_{yy}(f)e^{j2\pi fk}df = \widehat{r}_k \ = \ r_k = \int_{-0.5}^{0.5} S_{yy}(f)e^{j2\pi fk}df\ ,\quad k=0,1,\ldots,N\ . \tag{2.167}$$

Then the parameters $\sigma^2_{f,N},\ A_{N,1},\ldots,A_{N,N}$ that make the first $N+1$ covariance lags match are the ones that are found from linear prediction! Indeed, the AR(N) process y'_k satisfies the Yule-Walker equations in (2.83) with covariance sequence $\widehat{r}_k$. But since $\widehat{r}_k = r_k,\ k=0,1,\ldots,N$, the Yule-Walker equations (2.83) for lags $0,1,\ldots,N$ become the normal equations (2.102) of linear prediction. Hence, the AR(N) model determined by linear prediction matches the first $N+1$ covariance lags.

2.5.6 AR Modeling: Itakura-Saito Distance Minimization

Itakura and Saito introduced the following distance measure between two power spectral densities:

$$d(S,\widehat{S}) = \int_{-0.5}^{0.5} \left\{ \frac{S(f)}{\widehat{S}(f)} \Leftrightarrow \ln\frac{S(f)}{\widehat{S}(f)} \Leftrightarrow 1 \right\} df\ . \tag{2.168}$$

In order to investigate this distance measure, note that the function $x \Leftrightarrow 1 \Leftrightarrow \ln x \geq 0$ for $x \geq 0$ and is only zero for $x=1$. Based on this, we can investigate the three properties that a valid distance function should have:

(i) $d(S,S)=0$.

(ii) $d(S,\widehat{S}) \geq 0,\ d(S,\widehat{S}) > 0$ if $S \neq \widehat{S}$.

(iii) $d(S,\widehat{S}) \neq d(\widehat{S},S)$. Because the symmetry property is not satisfied, the Itakura-Saito distance is not a true distance function. Nevertheless it is a useful measure.

Assume again an AR(N) model:

$$\widehat{\mathrm{S}}_{yy}(z) \ = \ \frac{\sigma^2_{f,N}}{\mathrm{A}_N(z)\mathrm{A}_N(1/z)}\ . \tag{2.169}$$

This time, we shall determine the parameters $\sigma^2_{f,N}$, $A_{N,1}, \ldots, A_{N,N}$ by

$$\min_{\sigma^2_{f,N}, A_{N,i}} \; d(S_{yy}, \widehat{S}_{yy}) \; . \tag{2.170}$$

It can be shown that the parameters $\sigma^2_{f,N}$, $A_{N,1}, \ldots, A_{N,N}$ that minimize the Itakura-Saito distance measure are the ones that satisfy the normal equations (2.102) of linear prediction.

2.5.7 AR Modeling: Maximum Entropy

Suppose that the only statistical information we have about the process y_k is r_n, $n = 0, 1, \ldots, N$ (in practice, we have estimated them from data). In classical spectral estimation theory (see the Blackman-Tukey spectral estimator), we take $\widehat{r}_n = r_n$, $n = 0, 1, \ldots, N$, $\widehat{r}_n = 0$, $n > N$. This leads to a spectral estimate with limited resolution capability.

Suppose we want to model y_k with another process y'_k with covariance lags $\widehat{r}_n$ such that $\widehat{r}_n = r_n$, $n = 0, 1, \ldots, N$, and $\widehat{r}_n$, $n > N$ is taken such that the modeled y_k is as random as possible: we don't want to introduce any further assumptions about the process (the $\widehat{r}_n$) other than $\widehat{r}_n = r_n$, $n = 0, 1, \ldots, N$. The measure of randomness we shall use here is the *entropy*. We shall develop the notion of entropy in more detail in a later chapter. Here, we shall simply use some key results. One is that when we maximize the entropy w.r.t. a distribution function subject to constraints on some second-order moments of the distibution, the resulting distribution is Gaussian. For a stationary Gaussian process, the entropy rate per sample is given by $\log \sigma^2_{f,\infty}$. Hence, to maximize the entropy, we have to maximize $\sigma^2_{f,\infty}$ for a Gaussian process.

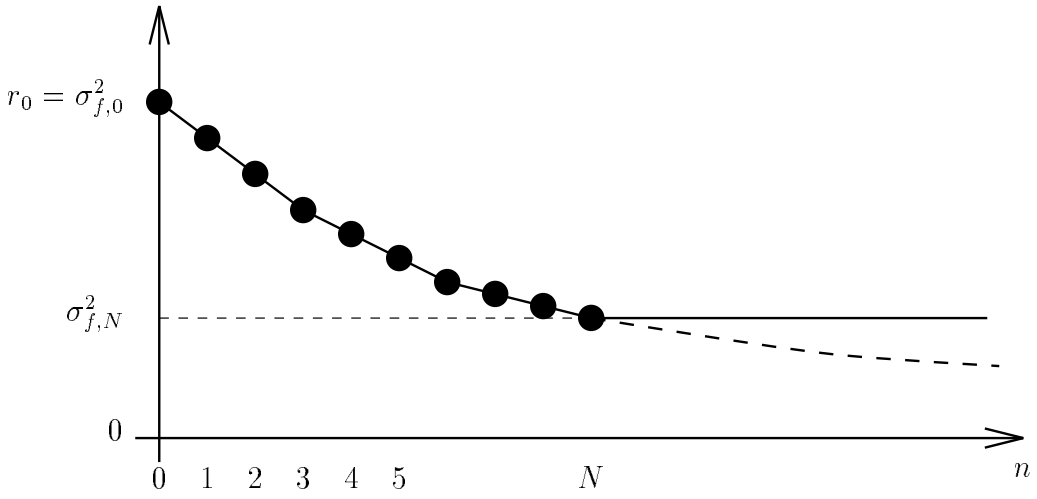

Figure 2.31: Maximizing $\sigma^2_{f,\infty}$ when r_n, $n = 0, 1, \ldots, N$ are given.

Since $\widehat{r}_n = r_n$, $n = 0, 1, \ldots, N$ are given, $\sigma^2_{f,n}$, $n = 0, 1, \ldots, N$ are determined. We then have

$$\sigma^2_{f,\infty} \;=\; \sigma^2_{f,N} \prod_{n=N+1}^{\infty} (1 \Leftrightarrow K_n^2) \;\leq\; \sigma^2_{f,N} \; . \tag{2.171}$$

We have $\widehat{r}_n,\ n > N$ to play with which alternatively means that we can choose $K_n,\ n > N$. From (2.171), we see that $\sigma^2_{f,\infty}$ is maximized by taking $K_n = 0,\ n > N$ yielding $\sigma^2_{f,\infty} = \sigma^2_{f,N}$, see Figure 2.31 also. Hence we find a Gaussian AR(N) process that satisfies the covariance matching property and hence its coefficients satisfy the normal equations of linear prediction. Note that we obtain the maximum entropy covariance extension $\widehat{r}_n,\ n > N$:

$$\begin{aligned} \widehat{r}_n &= \Leftrightarrow \sum_{i=1}^{N} A_{N,i}\widehat{r}_{n-i}\ ,\quad n > N \\ \widehat{r}_n &= r_n\ ,\quad n = 0, 1, \ldots, N \end{aligned} \tag{2.172}$$

This contrasts with MA(N) model matching for which we would have $\widehat{r}_n = 0,\ n > N$.

2.5.8 AR Modeling: Spectral Flatness Measure

We shall first introduce the spectral flatness measure for a covariance matrix. Consider the covariance matrix R_N of the stationary y_k, with positive real eigenvalues $\lambda_1 \ldots \lambda_N$. For white noise, we have $R_N = \sigma_y^2 I_N$ and hence $\lambda_1 = \cdots = \lambda_N = \sigma_y^2$. For a general covariance matrix, we may wonder how close the process is to white noise. The closeness to white noise can be measured by the flatness measure FM of the distribution of the λ_i:

$$\begin{aligned} FM &= \frac{\text{geometric average}}{\text{arithmetic average}} = \frac{(\prod_{i=1}^{N}\lambda_i)^{1/N}}{\frac{1}{N}\sum_{i=1}^{N}\lambda_i} \le 1 \quad (= 1 \text{ iff } \lambda_i = \sigma_y^2) \\ &= \frac{e^{\ln(\prod_{i=1}^{N}\lambda_i)^{1/N}}}{\frac{1}{N}\sum_{i=1}^{N}\lambda_i} = \frac{e^{\frac{1}{N}\sum_{i=1}^{N}\ln\lambda_i}}{\frac{1}{N}\sum_{i=1}^{N}\lambda_i}\ . \end{aligned} \tag{2.173}$$

The boundedness by one can be shown by using the following inequality.

Jensen's Inequality. Let $f(.)$ be a convex function and X a random variable. Then

$$f(E\,X) \le E\,f(X)\ . \tag{2.174}$$

If $f(.)$ is strictly convex, then strict inequality holds unless the distribution of X is concentrated in one point. If X has a discrete distribution, taking on the M values x_i with probabilities $\alpha_i > 0$, $\sum_{i=1}^{M}\alpha_i = 1$ then (2.174) can be written as

$$f(\sum_{i=1}^{M}\alpha_i\,x_i) \le \sum_{i=1}^{M}\alpha_i\,f(x_i)\ . \tag{2.175}$$

In words: the function value of a convex combination is bounded above by the convex combination of the function values. This property can be shown recursively. For $M = 2$,

(2.175) corresponds to the definition of a convex function. Then

$$\begin{aligned} f(\sum_{i=1}^{M}\alpha_i\,x_i) &= f(\alpha_1 x_1 + \underbrace{\frac{(1\Leftrightarrow\alpha_1)}{\sum_{k=2}^{M}\alpha_k}}_{=1}) = f(\alpha_1 x_1 + (1\Leftrightarrow\alpha_1)\underbrace{\sum_{i=2}^{M}\frac{\alpha_i}{\sum_{k=2}^{M}\alpha_k}\,x_i}_{=1}) \\ &\leq \alpha_1 f(x_1) + (1\Leftrightarrow\alpha_1)\, f(\sum_{i=2}^{M}\frac{\alpha_i}{\sum_{k=2}^{M}\alpha_k}\,x_i) \\ &\leq \cdots \leq \sum_{i=1}^{M}\alpha_i\, f(x_i)\ . \end{aligned} \tag{2.176}$$

The property (2.174) for a continuous distribution can be shown by taking the limit of (2.175) as $M \to \infty$.

To apply Jensen's inequality, take $f(x) = \Leftrightarrow\ln x$ (strictly convex), $x_i = \lambda_i$, $\alpha_i = 1/N$, $M = N$. Then (2.175) yields

$$\begin{aligned} \Leftrightarrow\ln(\tfrac{1}{N}\sum_{i=1}^{N}\lambda_i) &\leq \Leftrightarrow\tfrac{1}{N}\sum_{i=1}^{N}\ln\lambda_i \\ e^{\frac{1}{N}\sum_{i=1}^{N}\ln\lambda_i} &\leq \tfrac{1}{N}\sum_{i=1}^{N}\lambda_i \end{aligned} \tag{2.177}$$

from which we get the inequality in (2.173).

As $N \to \infty$, the distribution of the λ_i behaves similarly as $S_{yy}(f)$. For instance

$$\lim_{N\to\infty}\frac{\max\limits_{i=1,\ldots,N}\lambda_i}{\min\limits_{i=1,\ldots,N}\lambda_i} = \frac{\max\limits_{f\in[0,0.5]}S_{yy}(f)}{\min\limits_{f\in[0,0.5]}S_{yy}(f)}\ . \tag{2.178}$$

Similarly, the flatness measure in (2.173) becomes in the limit as $N \to \infty$ the spectral flatness measure

$$\frac{e^{\frac{1}{N}\sum_{i=1}^{N}\ln\lambda_i}}{\frac{1}{N}\sum_{i=1}^{N}\lambda_i} \overset{N\to\infty}{\Leftrightarrow} \xi_y = \frac{e^{\int_{-0.5}^{0.5}\ln S_{yy}(f)df}}{\int_{-0.5}^{0.5}S_{yy}(f)df} = \frac{\sigma^2_{f,\infty}}{\sigma^2_y} \in [0,1]\ . \tag{2.179}$$

Note that if e_k is white noise, then $\xi_e = 1$. We shall consider the spectral flatness measure applied to $f_{N,k}$ for which we have $S_{f_N f_N}(f) = S_{yy}(f)\,|A_N(f)|^2$. We get

$$\xi_{f_N} = \frac{\sigma^2_{f,\infty}}{\sigma^2_{f,N}}\ . \tag{2.180}$$

Hence, we try to choose the coefficients $A_{N,i}$, $i = 1,\ldots,N$ so as to make $S_{f_N f_N}(f)$ as flat as possible, or

$$\max_{A_{N,i},\,i=1,\ldots,N}\xi_{f_N} \quad\leftrightarrow\quad \min_{A_{N,i},\,i=1,\ldots,N}\sigma^2_{f,N}\ . \tag{2.181}$$

We conclude that the linear prediction in which we minimize the prediction error variance also makes the prediction error as white as possible.

2.5.9 Spectral Estimation Qualities of Autoregressive Modeling

So far: AR(N) modeled from $r_{yy}(k)$, $k = 0, 1, \ldots, N$. In practice we estimate $\sigma^2_{f,N}, A_{N,1}, \ldots, A_{N,N}$ from M samples $y_0, y_1, \ldots, y_{M-1}$. Given all the previous observations, we can conclude that the AR(N) model obtained with linear prediction gives a good spectral estimate for N sufficiently high. More precisely we can state:

- *bias*: the AR(N) model is only unbiased for AR(n) processes with $n \leq N$. Bias smaller for spectral peaks than for spectral valleys. In general, the bias disappears as $N \to \infty$.

- *variance*: so far we have assumed $r_0, \ldots, r_N$ known. In practice, they will have to be estimated from data. However, since they represent $N+1$ parameters to be estimated, we can state that the variance will be roughly proportional to N/M. So the variance will be low if $N \ll M$.

- *resolution*: due to the all-pole filter, we can in principle model arbitrarily closely spaced spectral peaks (sinusoids). There is no spectral smearing. By putting the poles of the all-pole model arbitrarily close to the unit circle, we can get arbitrarily sharp peaks. By putting furthermore two poles of the all-pole model arbitrarily close to each other, we can get two arbitrarily closely spaced peaks. For this reason, the AR modeling is also called a *high resolution* technique.

2.6 Spectral Estimation by AR Modeling: Techniques

As just mentioned before, the correlation sequence $r_0, \ldots, r_N$, which we have assumed available so far, needs to be estimated in practice. What we typically have at our disposal in practice are M samples $y_0, y_1, \ldots, y_{M-1}$ of the process of which we need to estimate the power spectral density. One way to proceed would be to estimate the correlation lags $r_0, \ldots, r_N$ and to determine the AR model parameters $\sigma^2_{f,N}, A_{N,1}, \ldots, A_{N,N}$ from the correlation estimates using the normal equations. This would be an application of the method of moments. More general techniques proceed however by estimating the parameters $\sigma^2_{f,N}, A_{N,1}, \ldots, A_{N,N}$ directly. The most popular of these techniques are based on the method of least-squares. In this method, we consider the prediction error as an equation error and we minimize the sum of squared prediction errors. Or, in a slightly different point of view, we replace the statistical average in the minimum prediction error variance criterion by a temporal average (or just the temporal sum). Using the forward prediction errors for the data samples $y_0, y_1, \ldots, y_{M-1}$, we arrive at

the following overdetermined set of equations:

$$\begin{bmatrix} f_{N,0} \\ \vdots \\ \\ f_{N,N} \\ \vdots \\ f_{N,M-1} \\ \vdots \\ \\ f_{N,M+N-1} \end{bmatrix} = \underbrace{\begin{bmatrix} y_0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ & & \ddots & 0 \\ y_N & & \cdots & y_0 \\ \vdots & & & \vdots \\ y_{M-1} & \cdots & & y_{M-N+1} \\ 0 & \ddots & & \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & y_{M-1} \end{bmatrix}}_{= \mathcal{Y} \text{ (Toeplitz data matrix)}} \begin{bmatrix} 1 \\ A_{N,1} \\ \vdots \\ A_{N,N} \end{bmatrix} \begin{matrix} \} \mathcal{Y}^{cov} \\ \} \mathcal{Y}^{pre} \\ \} \mathcal{Y}^{po} \\ \} \mathcal{Y}^{corr} \end{matrix} \tag{2.182}$$

As indicated, different least-squares methods are possible, depending on exactly which equations are taken into account. These different least-squares methods are all consistent and asymptotically equivalent. However, differences appear for a small sample size M. More generally, the following AR modeling techniques exist.

1. *correlation method* (pre- and postwindowed least-squares)

   The only explanation for the name of this method is that it was called this way in the 1970s. It corresponds to a pre- and postwindowed least-squares method. Prewindowing means that the missing samples y_k, $k < 0$, are assumed to be zero. Postwindowing refers to the fact that the missing samples y_k, $k \geq M$, are assumed to be zero. The correlation method criterion is the sum of the squares of the prediction errors at all times of the pre- and postwindowed signal. In fact, only a finite number of the prediction errors are non-zero and actually contribute to the sum of squares:

$$\min_{A_{N,1} \cdots A_{N,N}} \sum_{k=0}^{M+N-1} f_{N,k}^2 \;=\; \min_{A_{N,1} \cdots A_{N,N}} A_N^T \mathcal{Y}^T \mathcal{Y} \, A_N \;=\; (M+N)\, \sigma_{f,N}^2 \tag{2.183}$$

   where $\mathcal{Y} = \mathcal{Y}^{corr}$ from (2.182). The minimization leads to the normal equations

$$\mathcal{Y}^T \mathcal{Y} \, A_N \;=\; \begin{bmatrix} (M+N)\sigma_{f,N}^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T . \tag{2.184}$$

   In (2.184),(2.183), the factor $M+N$ appears since this least-squares method considers the sum of $M+N$ squares of prediction errors. One property of a pre- and postwindowed Toeplitz data matrix $\mathcal{Y}$ is that the resulting (multiple of a) sample covariance matrix $\mathcal{Y}^T\mathcal{Y}$ is Toeplitz also. $\mathcal{Y}^T\mathcal{Y}$ is furthermore symmetric and positive semidefinite by construction. Also, $\mathcal{Y}$ has full column rank by construction, so $\mathcal{Y}^T\mathcal{Y}$ is non-singular and hence positive definite. So the normal equations (2.184) are exactly of the same form as

the theoretical normal equations of linear prediction with known correlation sequence. Because of the correspondence of (2.184) with the theoretical normal equations, an alternative interpretation for its derivation is the substitution of the correlation sequence r_n in the theoretical normal equations with estimates $\hat{r}_n = \frac{1}{M}\sum_{k=0}^{M-n-1} y_{k+n}y_k$. In that case, the factor $M+N$ appearing in (2.184),(2.183) gets replaced by M however.

Also, because of the correspondence of (2.184) with the theoretical normal equations, we can apply the Levinson algorithm and the resulting filter estimate $\mathrm{A}_N(z)$ is guaranteed to be minimum-phase. However, as only a finite amount of data is used, the resulting parameter estimates do not equal their values that can be obtained from the true correlation sequence. We saw in the analysis of the periodogram that the use of a rectangular window on the data y_k lead to spectral leakage. Related estimation errors occur in the AR model obtained with the correlation method. One can intuitively understand that the replacement of data samples with zeros at the beginning and the end of the window must lead to artefacts that appear especially for short window lengths M. For instance, in the extreme case of only $M = 1$ sample, we get a white noise model, which can be obtained from any power spectral density by an extreme application of smearing. In order to reduce the spectral leakage, the data is often multiplied with a window with smoother transitions at its edges (see the discussion of the periodogram) before being used for the estimation of the correlations.

2. *covariance method* (unwindowed least-squares)

In this approach, we only take equations for the least-squares criterion that contain actual data samples (we don't take equations where artificial zeros appear):

$$\min_{A_{N,1}\cdots A_{N,N}} \sum_{k=N}^{M-1} f_{N,k}^2 \;=\; \min_{A_{N,1}\cdots A_{N,N}} A_N^T \mathcal{Y}^T\mathcal{Y}\, A_N \;=\; (M \Leftrightarrow N)\, \sigma_{f,N}^2 \tag{2.185}$$

where $\mathcal{Y} = \mathcal{Y}^{cov}$ from (2.182). The minimization leads to the normal equations

$$\mathcal{Y}^T\mathcal{Y}\, A_N \;=\; \begin{bmatrix} (M \Leftrightarrow N)\sigma_{f,N}^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T . \tag{2.186}$$

In this case $\mathcal{Y}^T\mathcal{Y} \;\neq$ Toeplitz, hence we cannot apply the Levinson algorithm and $\mathrm{A}_N(z)$ not guaranteed to be minimum-phase. Nevertheless, in general better estimation quality is obtained than with the correlation method, due to the absence of windowing, especially for short data lengths M ($\frac{1}{M-N}\mathcal{Y}^T\mathcal{Y}$ is an unbiased sample covariance matrix). We may remark that even though $\mathcal{Y}^T\mathcal{Y}$ is not Toeplitz, it is close to Toeplitz in a certain sense due to the fact that $\mathcal{Y}$ is Toeplitz. This allows for the derivation of a certain generalized Levinson algorithm. This derivation was one one of the starting points for the development of the rich area of fast algorithms for least-squares problems.

3. *modified covariance method*

Using a stationary covariance sequence, the forward and backward prediction problems

lead to identical prediction coefficients and variances. Due to the deviation of the sample covariance matrix in the covariance method from a Toeplitz matrix, it is helpful to consider the forward and the backward prediction problems jointly. Using the (forward) prediction coefficients, backward prediction errors can be obtained simply by performing the prediction in a backward sense:

$$\begin{bmatrix} \vdots \\ b_{N,k} \\ \vdots \end{bmatrix} = \mathcal{Y} \begin{bmatrix} A_{N,N} \\ \vdots \\ A_{N,1} \\ 1 \end{bmatrix} = \mathcal{Y}\, J \begin{bmatrix} 1 \\ A_{N,1} \\ \vdots \\ A_{N,N} \end{bmatrix} \tag{2.187}$$

(compare to (2.182). By considering the equations for the forward and the backward prediction errors jointly, we can double the number of equations and hence improve the estimation accuracy. So consider

$$\min_{A_{N,1}\cdots A_{N,N}} \sum_{k=N}^{M-1} (f_{N,k}^2 + b_{N,k}^2) \;=\; \min_{A_{N,1}\cdots A_{N,N}} A_N^T \left(\mathcal{Y}^T\mathcal{Y} + J\,\mathcal{Y}^T\mathcal{Y}\,J\right) A_N \;=\; 2(M \Leftrightarrow N)\,\sigma_{f,N}^2 \tag{2.188}$$

where $\mathcal{Y} = \mathcal{Y}^{cov}$ from (2.182). The minimization leads to the normal equations

$$\left(\mathcal{Y}^T\mathcal{Y} + J\,\mathcal{Y}^T\mathcal{Y}\,J\right) A_N \;=\; \begin{bmatrix} 2(M \Leftrightarrow N)\sigma_N^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T . \tag{2.189}$$

We can observe that $\mathcal{Y}^T\mathcal{Y} + J\mathcal{Y}^T\mathcal{Y}J$ is centro-symmetric (just like the theoretical covariance matrix), which leads to a further improved estimate (esp. for small M). Due to the doubling of the number of equations, the estimation variance of the $A_{N,n}$ gets improved, especially if only few data are available. The improvement disappears asymptotically (same information in forward and backward equations except at the borders of the data window).

4. *Itakura-Saito method*

The starting point of the Itakura-Saito method is to keep (impose) the Levinson recursions for the forward and backward prediction errors:

$$\begin{cases} f_{n+1,k} &= f_{n,k} + K_{n+1} b_{n,k-1} \\ b_{n+1,k} &= b_{n,k-1} + K_{n+1} f_{n,k} \end{cases} \tag{2.190}$$

where the range for the time index k depends on whether one takes the correlation or the covariance formulation (in the correlation case, the range for k increases as the prediction order increases, whereas the range decreases for the covariance case). The value to be used for the PARCOR in (2.190) is the one given by its usual expression in (2.128) but with statistical averages replaced by temporal averages:

$$K_{n+1}^{I-S} \;=\; \Leftrightarrow \frac{\sum\limits_k f_{n,k} b_{n,k-1}}{\sqrt{\sum\limits_k f_{n,k}^2}\sqrt{\sum\limits_k b_{n,k-1}^2}} \tag{2.191}$$

5. *Burg method* take the Levinson recursions for the prediction errors:

$$\begin{cases} f_{n+1,k} &= f_{n,k} + K_{n+1} b_{n,k-1} \\ b_{n+1,k} &= b_{n,k-1} + K_{n+1} f_{n,k} \end{cases}$$

and take the modified covariance criterion

$$\min_{K_{n+1}} \sum_k \left(f_{n+1,k}^2 + b_{n+1,k}^2 \right) \;\Rightarrow\; K_{n+1} \;=\; \Leftrightarrow \frac{\sum_k f_{n,k} b_{n,k-1}}{\frac{1}{2}\left(\sum_k f_{n,k}^2 + \sum_k b_{n,k-1}^2 \right)}$$

One can show that

$$\left| K_n^{Burg} \right| \;\leq\; \left| K_n^{Ita-S} \right| \;\leq\; 1 \;\;\Rightarrow\; \mathrm{A}_N(z) \;\text{ minimum-phase}$$

first inequality: arithmetic average $\geq$ geometric average
second inequality: Cauchy-Schwarz

2.7 Linear Time-Frequency Representations for Non-Stationary Processes

2.8 Problems

2.1. *Periodogram spectral leakage in the case of a sinusoid.*
Assume we have $y_k \;=\; A\cos(2\pi f_0 k + \phi)\,,\; k = 0, \ldots, N\Leftrightarrow 1$. We compute the periodogram (with a rectangular window) and we use the DFT to evaluate the periodogram at the frequencies $f = \frac{n}{N},\; n = 0, 1, \ldots, N\Leftrightarrow 1$ (no zero padding). Explain why the (discrete) periodogram shows exactly one non-zero component (in the range $f \in [0, \frac{1}{2}]$) when f_0 is a multiple of $\frac{1}{N}$, and multiple components otherwise.

2.2. *Use of windows in the Periodogram:*
Assume that the data available are $y_{-M}, \ldots, y_M$ and that we apply a symmetric window $w_k = w_{-k}$ before computing the periodogram. The periodogram with weighting becomes:

$$\widehat{S}_{PER,w}(f) \;=\; \frac{1}{2M+1} \left| \sum_{k=-M}^{M} w_k y_k e^{-j2\pi f k} \right|^2 .$$

Show that we get for the expected value

$$E\,\widehat{S}_{PER,w}(f) \;=\; \frac{1}{2M+1} \left| W(f) \right|^2 * S_{yy}(f)$$

where $W(f)$ is the Fourier transform of w_k and $*$ denotes convolution.

Bibliography

[1] H.L. Van Trees. *Detection, Estimation and Modulation Theory*, volume 1. Wiley, New York, 1968.

[2] L. Scharf. *Statistical Signal Processing.* Addison-Wesley, Reading, MA, 1991.

[3] S.M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prenctice Hall, 1993.

[4] B. Porat. *Digital Processing of Random Signals: Theory and Methods.* Prentice Hall, 1994.

[5] J.M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications and Control.* Prentice Hall, 1995.

[6] C.W. Therrien. *Discrete Random Signals and Statistical Signal Processing.* Prentice Hall, 1992.

[7] T. Kailath. *Lectures on Wiener and Kalman Filtering.* Springer-Verlag, Wien – New York, 1981.

[8] Lennart Ljung. *System Identification: Theory for the User.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[9] T. Söderström and P.G. Stoica. *System Identification.* Prentice Hall, 1989.