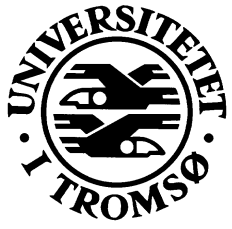
Introduction to
STATISTICAL SIGNAL PROCESSING AND DATA ANALYSIS

Alfred Hanssen
Physics Department
University of Tromsø



Lecture Notes Fys-261
Spring 2003

Contents

Chapter 1

STOCHASTIC VARIABLES

The stochastic (or random) variable is a very important concept when discussing data analysis. The theory of stochastic variables is quite extensive, but we will in this section only give a very brief review of the description of stochastic variables. We will not dwell with the fine details, but rather discuss those aspects of stochastic variables that are necessary in order to follow a discussion on stochastic processes and spectral estimation.

1.1 Univariate Case

Let X be a continuous or discrete stochastic (random) variable. A complete statistical description of X requires that its probability density function (PDF) $f_X(x)$ or its cumulative probability distribution function $F_X(x) = \int_{-\infty}^{x} f_X(x')dx'$ is known. Here, X denotes the stochastic variable itself, and x denotes a particular outcome. Given the PDF, one may readily calulate the average (or mean value, or expected value) of X according to

$$\overline{X} = E[X], \tag{1.1}$$

or its variance according to

$$\sigma_X^2 = E[(X - \overline{X})^2], \tag{1.2}$$

where $E[g(X)]$ denotes the statistical expectation of the stochastic function $g(X)$,

$$E[g(X)] \equiv \int_{-\infty}^{\infty} g(x) f_X(x) dx. \tag{1.3}$$

More generally, the moment of order n about origo of X is defined to be

$$m_X^{(n)} = E[X^n] \quad ; \quad n = 0, 1, 2, \ldots \tag{1.4}$$

and the central moments (i.e., moments about the mean value) is defined as

$$\mu_X^{(n)} = E[(X - \overline{X})^n] \quad ; \quad n = 0, 1, 2, \ldots \tag{1.5}$$

Note that all the results discussed above are valid also for discrete valued variables. Assume that the variable X can only attain N different values $x_1, x_2, .., x_N$ with probabilities $P_1, P_2, .., P_N$, respectively. If the following PDF is assigned to X

$$f_X(x) = \sum_{n=1}^{N} P_n \delta(x - x_n), \tag{1.6}$$

where $\delta(x)$ is the Dirac delta function, then all the results derived for continuous variables also apply to the discrete case.

1.2 Bivariate Case

In real systems, one usually consider several variables simultaneously. If two stochastic variables X and Y are available, then we need to know their joint probability density function $f_{X,Y}(x,y)$. Knowing the joint PDF, we may deduce several important properties of X, Y, and their interrelationship. We may compute the univariate PDF (or the "marginal" PDF) of one of the variables simply by "integrating the other variable away",

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \tag{1.7}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx. \tag{1.8}$$

If the joint PDF can be written as the product of its marginal PDF's,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \tag{1.9}$$

we say that X and Y are *statistically independent*. Note the statistical independence is a very strong assumption.

The results from the previous section are now readily generalized. E.g., the *joint moment about origo* of order $n+k$ is defined by

$$m_{XY}^{(n,k)} = E[X^n Y^k], \tag{1.10}$$

and the *joint central moment* of order $n+k$ is defined by

$$\mu_{XY}^{(n,k)} = E[(X-\overline{X})^n (Y-\overline{Y})^k]. \tag{1.11}$$

Some joint moments are particularly important, and have been given special designations. The *correlation* between X and Y is defined by

$$R_{XY} \equiv m_{XY}^{(1,1)} = E[XY]. \tag{1.12}$$

If the correlation can be written in the form

$$R_{XY} = E[X]E[Y], \tag{1.13}$$

then X and Y are said to be *uncorrelated*. If X and Y are statistically independent, they are also uncorrelated. The converse is not true in general (the important exception are Gaussian variables). If the variables are such that

$$R_{XY} = 0, \tag{1.14}$$

they are called *orthogonal*.

Another particularly important quantity is the *covariance* of X and Y defined by

$$\begin{aligned} C_{XY} \equiv \mu_{XY}^{(1,1)} &= E[(X - \overline{X})(Y - \overline{Y})] \\ &= R_{XY} - E[X]E[Y]. \end{aligned} \tag{1.15}$$

The last form of the covariance shows that if X and Y are either independent or uncorrelated, then $C_{XY} = 0$. Thus, the covariance is a measure of correlation between stochastic variables. Usually, one prefers to apply a normalized or dimensionless version of the covariance. This quantity is usually called the *correlation coefficient* between X and Y, and is defined by

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y} = E\left[\frac{(X - \overline{X})}{\sigma_X}\frac{(Y - \overline{Y})}{\sigma_Y}\right]. \tag{1.16}$$

It is easy to show that $-1 \leq \rho \leq 1$. If the variables X and Y are linearly related, $Y = aX + b$, then one can show that $\rho_{XY} = \text{sgn}\{a\}$, where $\text{sgn}\{a\}$ means the sign of the constant a. Thus, linearly related stochastic variables yield the maximum value of the correlation coefficient.

1.3 Characteristic Functions and Moments

The *characteristic function* $\Phi_X(\xi)$ of a stochastic variable X is defined as the expected value of the function $\exp(j\xi X)$, i.e.,

$$\Phi_X(\xi) = E\left[\exp(j\xi X)\right] = \int_{-\infty}^{\infty} f_X(x) \exp(j\xi x)\, dx, \tag{1.17}$$

where ξ is a transform variable.

By direct inspection, it is easy to show that the n-th order derivative of the characteristic function evaluated at $\xi = 0$ directly produces the n-th order moment of X according to

$$m_X^{(n)} = (-j)^n \frac{d\Phi_X(\xi)}{d\xi^n}|_{\xi=0}. \tag{1.18}$$

The characteristic function $\Phi_X(\xi)$ always exists for a stochastic variable, so this method ensures that the moments of variables can be found provided the derivatives of $\Phi_X(\xi)$ exist.

1.4 Cumulants

Cumulants (or *semi-invariants* as they are sometimes dubbed) are generalizations of the statistical moments, and they play a central role in describing and quantifying non-linear phenomena and variables. The *cumulant generating function* for a stochastic variable X is defined by (e.g., *Nikias and Petropulu* [1993])

$$\Psi_X(\xi) = \ln\left[\Phi_X(\xi)\right]. \tag{1.19}$$

The cumulant of order n for the variable X is then defined as

$$\kappa_X^{(n)} = (-j)^n \frac{d^n \Psi_X(\xi)}{d\xi^n}|_{\xi=0}. \tag{1.20}$$

1.4.1 Relation Between Moments and Cumulants

It is very useful to know the explicit relationship between the characteristic function and the cumulants. This can be derived as follows. Expand the exponential in eq. (1.17) in a power series around $\xi = 0$, to obtain

$$\Phi_X(\xi) = E\left[1 + \sum_{q=1}^{\infty}(j\xi)^q \frac{X^q}{q!}\right] = 1 + \sum_{q=1}^{\infty}(j\xi)^q \frac{m_X^{(q)}}{q!}. \tag{1.21}$$

The Taylor series expansion of the function $\ln(1+\alpha)$ is

$$\ln(1+\alpha) = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{\alpha^n}{n}. \tag{1.22}$$

By inserting this result into the expression for $\Phi_X(\xi)$ from eq. (1.21), we obtain

$$\Psi_X(\xi) = \left[\sum_{q=1}^{\infty}(j\xi)^q \frac{m_X^{(q)}}{q!}\right] - \frac{1}{2}\left[\sum_{q=1}^{\infty}(j\xi)^q \frac{m_X^{(q)}}{q!}\right]^2 + \frac{1}{3}\left[\sum_{q=1}^{\infty}(j\xi)^q \frac{m_X^{(q)}}{q!}\right]^3 - \cdots \tag{1.23}$$

Collecting terms of equal orders in $j\xi$, we see that the cumulant generating function can be written as

$$\Psi_X(\xi) = m_X^{(1)}\frac{j\xi}{1!} + \left[m_X^2 - \left(m_X^{(1)}\right)^2\right]\frac{(j\xi)^2}{2!} + \left[m_X^{(3)} - 3m_X^{(1)}m_X^{(2)} + 2\left(m_X^{(1)}\right)^3\right]\frac{(j\xi)^3}{3!} + \cdots \tag{1.24}$$

By examining the definition of the cumulant (1.20) and the definition of the cumulant generating function, we see that we may express the cumulant generating function as the infinite series

$$\Psi_X(\xi) = \ln\Psi_X(\xi) = \sum_{p=1}^{\infty}\kappa_X^{(p)}\frac{(j\xi)^p}{p!}, \tag{1.25}$$

where $\kappa_X^{(p)}$ is the cumulant of order p. By comparing eq. (1.24) and eq. (1.25) term by term, we find the desired relationship between moments and cumulants. The four lowest order cumulants can thus be written in terms of the moments as

$$\begin{aligned}
\kappa_X^{(1)} &= m_X^{(1)} \\
\kappa_X^{(2)} &= m_X^{(2)} - \left(m_X^{(1)}\right)^2 \\
\kappa_X^{(3)} &= m_X^{(3)} - 3m_X^{(1)}m_X^{(2)} + 2\left(m_X^{(1)}\right)^3 \\
\kappa_X^{(4)} &= m_X^{(4)} - 3\left(m_X^{(2)}\right)^2 - 4m_X^{(1)}m_X^{(3)} + 12\left(m_X^{(1)}\right)^2 m_X^{(2)} - 6\left(m_X^{(1)}\right)^4.
\end{aligned} \tag{1.26}$$

From these relationships we acknowledge the general fact that the cumulant of order p depends on all moments up to and including the same order.

For the special case of a zero mean variable, however, we see that the expressions in (1.26) simplify to

$$\begin{aligned}
\kappa_X^{(1)} &= m_X^{(1)} \\
\kappa_X^{(2)} &= m_X^{(2)} \\
\kappa_X^{(3)} &= m_X^{(3)} \\
\kappa_X^{(4)} &= m_X^{(4)} - 3\left(m_X^{(2)}\right)^2.
\end{aligned} \tag{1.27}$$

For zero mean variables, we thus see that it is first at the fourth order that cumulants differ from moments.

The single most important property of the cumulants is that for Gaussian variables, $\kappa_X^{(p)} \equiv 0$ for $p > 2$. For non-Gaussian variables, however, the cumulants take on non-zero values also for orders exceeding 2. Thus, the cumulants may be used to characterize the "non-Gaussianity" of stochastic variables.

1.5 Multivariate Case

If N real valued stochastic variables $X_1, X_2, \ldots, X_N$ are to be described statistically, we understand that the most fundamental description will be through an n-dimensional probability density function, or alternatively an n-dimensional cumulative distribution function. To simplify the notation, let $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$ be an n-dimensional stochastic vector, and let $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$ be a vector of outcomes. Then the relationship between the multidimensional PDF and the multidimensional distribution function can be written as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^N F_{\mathbf{X}}(\mathbf{x})}{\partial \mathbf{x}}, \tag{1.28}$$

where $\partial^N / \partial \mathbf{x} \equiv \partial^N / \partial x_1 \partial x_2 \cdots \partial x_N$.

Any marginal density (or distribution) may be obtained by integrating over the "unwanted" variables. E.g., if $f_{X_1}(x_1)$ is of interest, then the following $(N-1)$-dimensional integral must be performed

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_2 \cdots dx_N. \tag{1.29}$$

1.6 Density of a Sum of Stochastic Variables

Let the real valued stochastic variable W be the sum of two random variables X and Y,

$$W = X + Y \tag{1.30}$$

The cumulative distribution function for the sum variable W is therefore

$$F_W(w) = P\{X + Y \leq w\}. \tag{1.31}$$

For a chosen w, we thus see that we must evaluate the above probability for all x and y in the region where $x + y \leq w$. This can be accomplished by first fixing w, then using that $x = w - y$, and lastly let y vary over the entire real axis. Thus we may write

$$F_W(w) = \int_{-\infty}^{\infty} \int_{-\infty}^{w-y} f_{X,Y}(x, y) dx dy. \tag{1.32}$$

To obtain the PDF of W, we now differentiate eq. (1.32) with respect to w. The expression for the sought probability density is then found to be

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,Y}(w - y, y) dy. \tag{1.33}$$

If X and Y are statistically independent, then $f_{X,Y}(w-y,y) = f_X(w-y)f_Y(y)$, and the desired probability density actually simplifies to a convolution between the densities for each of the variables forming the sum,

$$f_W(w) = \int_{-\infty}^{\infty} f_X(w-y)f_Y(y)dy = f_X(w) * f_Y(w), \tag{1.34}$$

where $*$ denotes a convolution.

These results are readily generalized to a sum of N statistically independent stochastic variables

$$W = \sum_{n=1}^{N} X_n, \tag{1.35}$$

where the variable X_n has a probability density function $f_{X_n}(x_n)$. The probability density of the sum is then the $(N-1)$-fold convolution of all the variables in the sum in eq. (1.35)

$$f_W(w) = f_{X_1}(w) * f_{X_2}(w) * \cdots * f_{X_N}(w). \tag{1.36}$$

It can be shown that asymptotically ($N \to \infty$), $f_W(w)$ converges to a Gaussian probability density function, (almost) regardless of what densities the individual stochastic variables have. This is in fact a manifestation of *the central limit theorem* which has important consequences when considering a total event that is the sum of several random events.

Chapter 2

STOCHASTIC PROCESSES

In real engineering and science problems, we often have to deal with waveforms that vary as a function of time (or space). Most often, such waveforms include some random behavior. In binary communication systems, e.g., the bit stream itself can be regarded as a random message because each bit in the stram occurs randomly. In other cases, a deterministic signal may be corrupted by random noise. Thus, the ability to describe, analyse, and manipulate random waveforms is essential in real world data analysis [*Peebles*, 1993; *Shiavi*, 1999].

2.1 Basic Definition

A stochastic process is a generalization of stochastic variables to also include a time-like parameter t. Thus, instead of X, which can take on a set (of finite or infinite size) of values according to some probabilistic rule, we now talk about a *collection of waveforms* $X(t)$ chosen according to some probability law. Such a collection of waveforms is called a *stochastic process*.

In order to describe the probabilistic properties of a stochastic process, we need to augment the PDFs to also be functions of time. Now, consider the stochastic process $X(t)$ at a particular time $t = t_1$. At a particular time, the process is nothing but a stochastic variable, $X_1 = X(t_1)$. The PDF associated with the random variable X_1 will be denoted by $f_X(x_1; t_1)$. If we consider the process at two different times t_1 and t_2, we may form the two stochastic variables $X_1 = X(t_1)$ and $X_2 = X(t_2)$. We can thus define the second-order joint PDF of X_1 and X_2 as $f_X(x_1, x_2; t_1, t_2)$. In a similar manner, for N random variables $X_i = X(t_i)$, $i = 1, 2, \ldots, N$, the N-th joint PDF is $f_X(x_1, x_2, .., X_N; t_1, t_2, .., t_N)$. For a complete description of the whole process, we thus need to know f_X at number of time points, $N = 1, 2, .., \infty$.

From these definitions, it should be evident that one may interpret a stochastic process as a *sequence* of stochastic variables. The great statistician and time series expert *Tukey* used to say that a stochastic process is just "one damned thing after another".

2.2 Stationarity

It is of uttermost importance to discuss how the statistical properties of a stochastic process varies with time. Various definitions of statistical stationarity forms the basis for such a discussion. We will now discuss the most important stationarity

A stochastic process is called *stationary to first order* if its first-order probability density

function does not change with a shift in time origin. In other words,

$$f_X(x;t) = f_X(x;t+\Delta t) \tag{2.1}$$

for any t and any Δt. Thus, in this case $f_X(x;t)$ is independent of t. One important consequence of this assumption is that the mean value of the process is a constant

$$E[X(t)] = \overline{X} \tag{2.2}$$

A process is likewise called *stationary to second order* if its second-order density function satisfies

$$f_X(x_1,x_2;t_1,t_2) = f_X(x_1,x_2;t_1+\Delta t,t_2+\Delta t) \tag{2.3}$$

for any $t_1, t_2, \Delta t$. Thus, only the *time difference* $\tau = t_2 - t_1$ plays a role, and not the absloute time.

The above reasoning is readily extended to N stochastic variables, $X_1, X_2, .., X_N$. A stochastic process is said to be *stationary to N-th order* if its Nth order density function is invariant to a shift of the time origin, i.e.,

$$f_X(x_1,\ldots,x_N;t_1,\ldots,t_N) = f_X(x_1,\ldots,x_N;t_1+\Delta t,\ldots,t_N+\Delta t) \tag{2.4}$$

for all $t_1,\ldots,t_N$ and Δt. Note that stationarity of order N implies stationarity to all lower orders $k < N$. A process that is stationary to *all* orders $N = 1, 2, \ldots$, is called a *stric-sense stationary* process.

2.3 Autocorrelation Function

The correlation between X_1 and X_2 is in general a function of the two time arguments t_1 and t_2,

$$R_{XX}(t_1,t_2) = E[X(t_1)X(t_2)]. \tag{2.5}$$

The function in Eq. (2.5) is called the *autocorrelation function* (ACF) of the random process $X(t)$.

An important special case occurs if we assume that $X(t)$ is stationary to second order. Because of Eq. (2.3), it is then obvious that also the ACF is a function only of time differences. One way of formulating this, is that

$$R_{XX}(t_1,t_1+\tau) = E[X(t_1)X(t_1+\tau)] \equiv R_{XX}(\tau), \tag{2.6}$$

where $\tau = t_2 - t_1$.

2.4 Wide-Sense Stationarity

In signal analysis applications we often deal with the mean value and the ACF of a random process. Most practical problems are substantially simplified if we assume that the mean value is time invariant, and that the ACF does not depend on absolute time. Such processes are said to be *wide-sense stationary* (WSS), and they must obey the two conditions

$$E[X(t)] = \overline{X} = \text{const.} \tag{2.7}$$

$$E[X(t)X(t+\tau)] = R_{XX}(\tau). \tag{2.8}$$

It should now be obvious that a second-order stationary process will always be wide-sense stationary. However, the converse is not generally true, so second-order stationarity is a much more restrictive requirement than that of wide-sense stationarity.

2.5 Ergodicity

In most real world problems, we only have access to a single realization of a stochastic process. It may therefore be impossible to perform a statistical expectation operation on the data, since this require knowledge about the statistical properties of the full ensemble of realizations.

Under certain conditions, the expectation operation of the whole process may be replaced by the time average of a single realization of the process. Processes for which time averages may replace ensemble averages are called *ergodic*.

Let $x(t)$ be a single realization of the process $X(t)$. Then we define the *time average* of the realization to be

$$\overline{x} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt, \tag{2.9}$$

and the *time autocorrelation function* is defined by

$$\mathcal{R}_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)dt. \tag{2.10}$$

In general, $\overline{x}$ and $\mathcal{R}_{xx}(\tau)$ (for a fixed τ) are both stochastic variables, since the value of the time integral will depend upon which realization we use. If the processes $X(t)$ is ergodic, however, all the statistical information about the process is contained in one single realization. If $x(t)$ is an arbitrary realization of $X(t)$, then ergodicity implies that

$$\overline{x} = \overline{X} \tag{2.11}$$

$$\mathcal{R}_{xx}(\tau) = R_{XX}(\tau), \tag{2.12}$$

and likewise for all other time and ensemble average pairs we can think of. A process which *only* obeys Eq. (2.11) is called *time-ergodic*, and a process which only obeys Eq. (2.12) is called *autocorrelation ergodic*. A process which obeys both (2.11) and (2.12) is called *weakly ergodic*.

Note that the time-average is *never* a function of t, and the time-ACF is always a function of only τ. We thus understand that any ergodic process must also be stationary, and in particular, a weakly ergodic process is always wide-sense stationary.

2.6 Three Kinds of Autocorrelation Functions

From this discussion, we conclude that a tri-section between distinct classes of autocorrelation functions is natural. For energy signals (deterministic or stochastic pulses having finite region of support) the appropriate definition is

$$\tilde{\mathcal{R}}_{hh}(\tau) = \int_{-\infty}^{\infty} h(t)h(t+\tau)dt, \tag{2.13}$$

where $h(t)$ is the pulse. For power signals two different definitions are needed. The first is appropriate for deterministic power signals and for single realizations of stochastic processes, and is given by

$$\mathcal{R}_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)dt, \tag{2.14}$$

where $x(t)$ is the signal. The second definition is appropriate for stochastic processes, and is given by

$$R_{XX}(t,t+\tau) = E\left[X(t)X(t+\tau)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_X(x_1,x_2;t,t+\tau)dx_1 dx_2, \tag{2.15}$$

where $X(t)$ is a stochastic process, and $f_X(x_1,x_2;t,t+\tau)$ is the bivariate probability density function of the process.

2.7 Cumulant Functions

We have seen that the autocorrelation function is a generalization of correlation between stochastic variables to the quantification of correlation within a stochastic process parametrized by a time difference (for wide-sense stationary processes). Following this pattern, it is possible to define cumulant functions of a given order from the definition of statistical cumulants.

First, we define the cross-moment of order $N = k_1 + k_2 + \cdots + k_M$ between M different real valued stochastic variables $\mathbf{X} = [X_1, X_2, \ldots, X_M]^T$ as

$$m_{\mathbf{X}}^{(N)} = \mathrm{mom}\left\{X_1^{k_1}, X_2^{k_2}, \ldots, X_M^{k_M}\right\} = (-j)^N \frac{\partial^N \Phi_{\mathbf{X}}\left(\xi_1,\xi_2,\ldots,\xi_M\right)}{\partial\xi_1^{k_1}\partial\xi_2^{k_2}\cdots\partial\xi_M^{k_M}}\,|_{\xi_1=\cdots=\xi_M=0}\,. \tag{2.16}$$

The cross-cumulant of order N is then likewise defined as

$$\kappa_{\mathbf{X}}^{(N)} = \mathrm{cum}\left\{X_1^{k_1}, X_2^{k_2}, \ldots, X_M^{k_M}\right\} = (-j)^N \frac{\partial^N \Psi_{\mathbf{X}}\left(\xi_1,\xi_2,\ldots,\xi_M\right)}{\partial\xi_1^{k_1}\partial\xi_2^{k_2}\cdots\partial\xi_M^{k_M}}\,|_{\xi_1=\cdots=\xi_M=0}\,, \tag{2.17}$$

where $\Psi_{\mathbf{X}}\left(\xi_1,\xi_2,\ldots,\xi_M\right) = \ln\left\{\Phi_{\mathbf{X}}\left(\xi_1,\xi_2,\ldots,\xi_M\right)\right\}$.

It is now straightforward to define moments and cumulants for stochastic processes by using the definitions in eqs. (2.16) and (2.17) for fixed times of a sampled stochastic process. If we assume a process that is stationary at least to order N, we may define the N-th order *moment function* of the stochastic process $X(t)$ as

$$m_X^{(N)}(\tau_1,\tau_2,\ldots,\tau_{n-1}) = \mathrm{mom}\left\{X(t), X(t+\tau_1), \ldots, X(t+\tau_{N-1})\right\}, \tag{2.18}$$

and the N-th order *cumulant function* as

$$\kappa_X^{(N)}(\tau_1,\tau_2,\ldots,\tau_{n-1}) = \mathrm{cum}\left\{X(t), X(t+\tau_1), \ldots, X(t+\tau_{N-1})\right\}. \tag{2.19}$$

We see that the individual orders has been chosen according to $k_m = 1$, $m = 1,\ldots,M$, and $M = N$ in the above definitions.

A third-order ($N = 3$) cumulant function is e.g. defined by (assuming a zero-mean process $X(t)$ stationary to the third order, which simplifies matters a great deal)

$$\kappa_X^{(3)}(\tau_1,\tau_2) = E\left[X(t)X(t+\tau_1)X(t+\tau_2)\right]. \tag{2.20}$$

This function has several important symmetries, e.g.,

$$\begin{aligned}
\kappa_X^{(3)}(\tau_1,\tau_2) &= \kappa_X^{(3)}(\tau_2,\tau_1) \\
&= \kappa_X^{(3)}(-\tau_2,\tau_1-\tau_2) = \kappa_X^{(3)}(\tau_1-\tau_2,-\tau_2) \\
&= \kappa_X^{(3)}(-\tau_1,\tau_2-\tau_1) = \kappa_X^{(3)}(\tau_2-\tau_1,-\tau_1).
\end{aligned} \tag{2.21}$$

These symmetries are important since they restrict the region in (τ_1, τ_2)-space where we need to evaluate $\kappa_X^{(3)}(\tau_1, \tau_2)$. We also notice that the third order cumulant function is equal to the third order moment function for zero-mean data. For higher orders than three, this correspondence no longer holds.

2.8 Gaussian Processes

Unarguably, the Gaussian processes are the most important ones. A stochastic process $X(t)$ is said to be a Gaussian process if, for any set of fixed times $t_1, t_2, \ldots, t_N$, and any value of N, the resulting random variables X_n follow a multidimensional Gaussian distribution. Thus, a Gaussian process always has the joint PDF

$$f_X(x_1, \ldots, x_N; t_1, \ldots, t_N) = \frac{1}{\sqrt{(2\pi)^N \det\{\mathbf{C}_X\}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \overline{\mathbf{X}})^T \mathbf{C}_X^{-1} (\mathbf{x} - \overline{\mathbf{X}})\right], \qquad (2.22)$$

where the data vector is $\mathbf{x} = [x_1, \ldots, x_N]^T$, the mean value vector is $\overline{\mathbf{X}} = [E(X_1), \ldots, E(X_N)]^T$, and the covariance matrix is $\mathbf{C}_X = E[(\mathbf{x} - \overline{\mathbf{X}})(\mathbf{x} - \overline{\mathbf{X}})^T]$, and superscript T denotes a transpose. From this formulation, we thus see that a Gaussian process can be both stationary and non-stationary.

2.9 Markov Processes

Markov processes play an important role in the theory of stochastic processes, and they are important for many real-world applications. Several important physical fluctuation phenomena have a fundamental description which fits in the framework of Markov processes, and applicable models of digital signals and images often fall in the Markov class.

The formal definition is as follows. A stochastic process $X(t)$ is said to be a Markov process if its conditional probability can be written as follows

$$P\{X(t_n) \leq x_n | X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_{n-1}) = x_{n-1}\} = P\{X(t_n) \leq x_n | X(t_{n-1}) = x_{n-1}\}, \qquad (2.23)$$

where $t_1 < t_2 < \cdots < t_{n-1} < t_n$.

An immediate consequence of the definition in eq. (2.23) is that the conditional probability of a Markov process at time t_n given its values at $n-1$ previous time instants, only needs to be conditioned on its immediate past value at t_{n-1}. Thus, the process is independent of the time history prior to t_{n-1}.

If the process has discrete valued amplitudes, it is called a *Markov chain*, and if the amplitude is continuous, the process is termed a *diffusion process*.

2.10 Parametric Models

In some cases, it is possible to assume a certain model, or a certain structure of the stochastic process under study. If the model contains a set of parameters that can be applied in the description and characterization of the process, we call the estimation technique a "parametric method". In this section, we will briefly look into one particular class of parametric models.

The simplest example of a parametric model for a stochastic process, is the so-called autoregressive model of order 1 [AR(1)],

$$x[n] = -a_1 x[n-1] + \varepsilon[n] \quad ; \quad n = ..., -1, 0, 1, ... \tag{2.24}$$

Here, a_1 is a constant, and $\varepsilon[n]$ is a zero mean white and Gaussian process called the "driving noise". Thus, $E\{\varepsilon[n]\} = 0$ and $E\{\varepsilon[n]\varepsilon[n+m]\} = \sigma^2\delta_{m,0}$, where $\delta_{m,0}$ is the Kronecker delta. The interpretation of model Eq. (2.24) is that the value of the process at time step n depends linearly on the value of the process at the previous time step $n-1$, but that the linear correlation is perturbed by the driving noise. The values of a_1 and σ^2 thus determine the correlation properties (and therefore the spectral properties) of the process. In this particular case, a_1 and σ^2 are the parameters that completely specifies the process under discussion. The negative sign of the a_1 term is just a handy convention.

This way of formulating a discrete time stochastic process is called a "difference equation", since it expresses a weighted difference between subsequent samples of the process. Alternatively, we may say that this definition is a recursion of the process upon itself.

Observe the following interesting properties of the AR(1)-process above. If $a_1 < 0$, the next sample of the process will typically have the same sign as the previous sample. Thus, this case corresponds to a "slowly varying" process, or a low-pass process. If $a_1 > 0$, the next sample of the process will typically have the opposite sign as the previous sample. Thus, this case corresponds to a "rapidly varying" process, or a high-pass process. For the case $a_1 = 0$ we see that $x[n] = \varepsilon[n]$, or pure white noise. If $|a_1| > 1$ the process will diverge, since on average, $|x[n]| > |x[n-1]|$ in this case.

Autoregressive processes

A straightforward generalization of the process in Eq. (2.24) is obtained if we say that the present value of the process is a linear combination of the p previous process values, in addition to a driving noise. Such a model is called an "Autoregressive model of order p" [AR(p)], and has the form

$$x[n] = -a_1 x[n-1] - a_2 x[n-2] - \cdots - a_p x[n-p] + \varepsilon[n], \tag{2.25}$$

where $\varepsilon[n]$ is as defined in Eq. (2.24). In this case, the $p+1$ parameters $a_1, ..a_p, \sigma^2$ are needed in order to define the process completely.

In Fig. 1 we display three different AR-proceses, along with their exact power spectral densities (to be defined later in this manuscript).

Moving average processes

If the value of the process can be said to be a weighted mean ("moving average") of q previous white Gaussian noise samples, one calls the resulting process a "Moving average process of order q". Such a process can be written as

$$x[n] = \varepsilon[n] + b_1\varepsilon[n-1] + \cdots + b_q\varepsilon[n-q]. \tag{2.26}$$

Here, b_m are constants (parameters) that specify the modified driving noise. This generalization of the driving noise obviously creates correlations between different time instants, and thus the resulting driving noise in this case is a coloured Gaussian noise process. In this case, the $q+1$ parameters $b_1, ..b_q, \sigma^2$ are needed in order to define the process completely.

Autoregressive / moving average processes

A further generalization can be obtained if we combine the AR(p) and the MA(q) into one common stochastic process model. A model of this kind is called an "Autoregressive/moving average model of orders p and q" [ARMA(p,q)], and is defined by

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + \sum_{m=1}^{q} b_m \varepsilon[n-m] \tag{2.27}$$

An ARMA(p,q)-process has $p+q+1$ parameters, $a_1,..,a_p,b_1,..,b_q,\sigma^2$.

If we define $a_0 \equiv b_0 \equiv 1$, a mathematically convenient form of the ARMA(p,q) process is

$$\sum_{k=0}^{p} a_k x[n-k] = \sum_{m=0}^{q} b_m \varepsilon[n-m]. \tag{2.28}$$

We note that Eq. (2.28) is in the form of discrete convolutions on each side of the equality sign. The left hand side is a convolution of the process $x[n]$ with the parameter sequence $a_1,..,a_p$, and the right hand side is a convolution of the process $\varepsilon[n]$ with the sequence $b_1,..,b_q$. Notice also that AR(p)=ARMA$(p,0)$, and MA(q)=ARMA$(0,q)$.

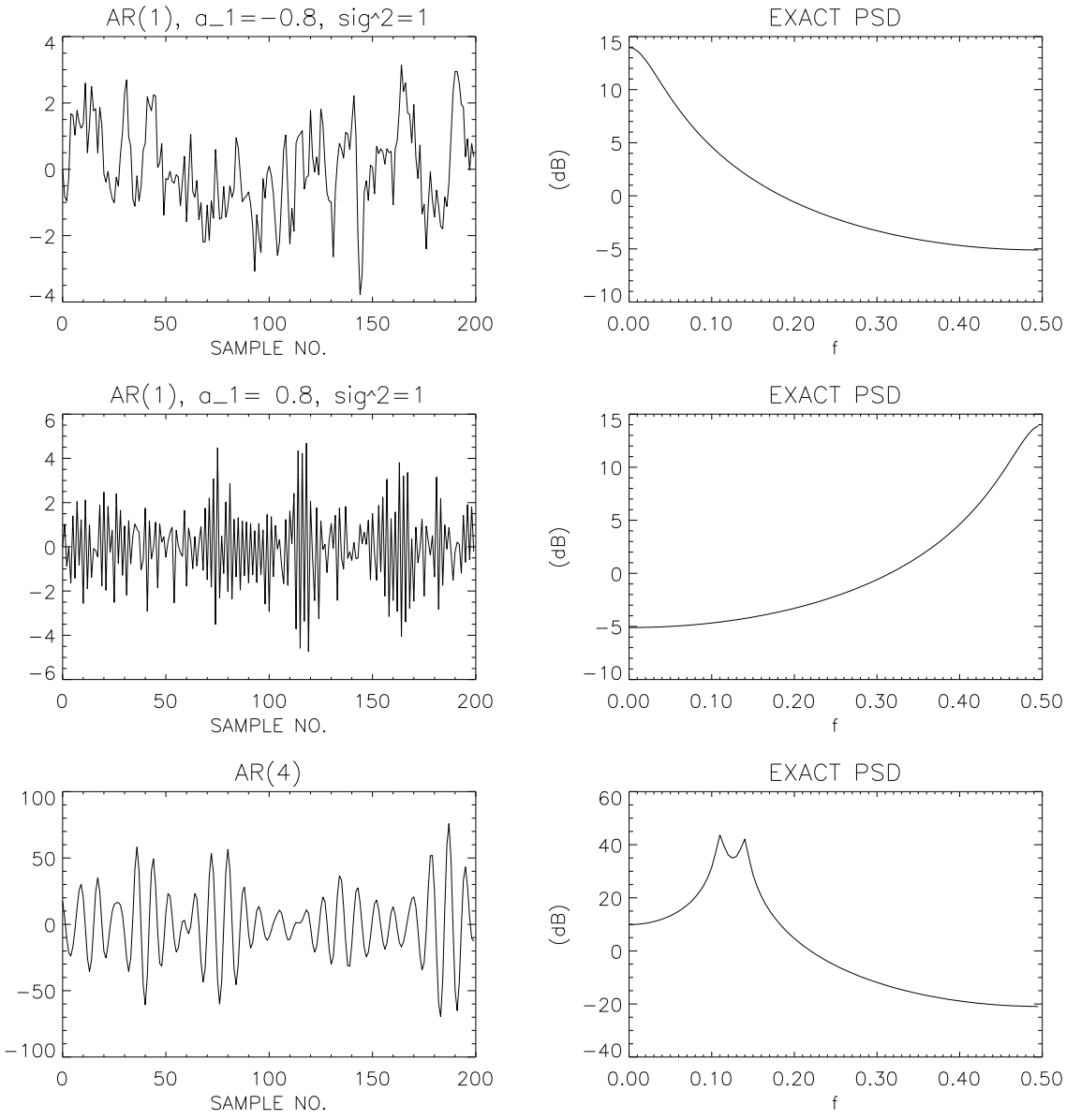


Figure 2.1: Realizations of some AR-process (left) and their exact power spectra (right). Upper panels: AR(1) with $a_1 = -0.8$ and $\sigma^2 = 1$. Central panels: AR(1) with $a_1 = 0.8$ and $\sigma^2 = 1$. Lower panel: AR(4) with $a_1 = -2.7607, a_2 = 3.8106, a_3 = -2.6535, a_4 = 0.9238$, and $\sigma^2 = 1$.

Chapter 3

FOURIER ANALYSIS AND POWER SPECTRAL DENSITY

When discussing power spectral estimation, some important results from Fourier analysis will be needed. For completeness, I will review these results in this Section. This is important also in order to introduce the notation that will be applied in this part of the course.

3.1 Continuous Time – Continuous Frequency

If $x(t)$ is absolute integrable, $\int_{-\infty}^{\infty} |x(t)|dt < \infty$, then its Fourier transform exists, and is given by

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)dt \quad ; \quad -\infty < f < \infty. \tag{3.1}$$

Here I have written the Fourier transform explicitly in terms of its frequency $f = \omega/2\pi$ instead of its angular frequency ω. This is convenient when dealing with real applications.

The inverse Fourier transform is then given by

$$x(t) = \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft)df \quad ; \quad -\infty < t < \infty. \tag{3.2}$$

Note that the factor $1/2\pi$ in front of the inverse Fourier integral has disappeared since we deal with f instead of ω, and thus $d\omega = 2\pi df$.

Signal energy and Parseval's theorem

The *energy* of a function $x(t)$ is a real valued quantity defined as

$$\mathcal{E}_x = \int_{-\infty}^{\infty} |x(t)|^2 dt. \tag{3.3}$$

By substituting the inverse Fourier transform Eq. (3.2) into Eq. (3.3), we can readily show that the energy can also be expressed in the frequency domain as

$$\mathcal{E}_x = \int_{-\infty}^{\infty} |X(f)|^2 df, \tag{3.4}$$

where $X(f)$ is the Fourier transform of $x(t)$. The equality of (3.3) and (3.4) is usually referred to as *Parseval's theorem.*

3.2 Discrete Time – Continuous Frequency

If we deal with a sampled signal, $x[n] = x(n\Delta t) \quad ; \quad n = \cdots, -1, 0, 1, \cdots$, the Fourier transform becomes periodic. We will use the following definition of the Fourier transform in this case

$$X(f) = \Delta t \sum_{n=-\infty}^{\infty} x[n] \exp(-j2\pi f n \Delta t) \quad ; \quad -\frac{1}{2\Delta t} \leq f \leq \frac{1}{2\Delta t}, \tag{3.5}$$

Note the factor Δt in front of the integral in Eq. (3.5). This is included to get the same dimension as in the continuous definition Eq. (3.1), i.e., Δt simulates the differential dt in the integral. In the book by *Oppenheim et al.*, they assumed that $\Delta t = 1$, and that is why this factor is missing in their definitions.

It is easy to show that $X(f) = X(f + f_s)$ where $f_s = 1/\Delta t$ is called the *sampling frequency* (show this yourself by insertion of $f + f_s$ in the expression for the Fourier transform).

The inverse Fourier transform in this case is

$$x[n] = \int_{-1/(2\Delta t)}^{1/(2\Delta t)} X(f) \exp(j2\pi f n \Delta t) df \quad ; \quad n = \cdots, -1, 0, 1, \cdots. \tag{3.6}$$

One way of interpreting the inverse transform in this case, is that we have taken out the transform values in a frequency window of size f_s, and reconstructed the signal values on the basis of these frequencies. The frequencies outside the band $|f| \leq f_s/2$ are not needed, since all necessary frequency information is contained in this "primary band". Note that $x[n]$ is *not* periodic.

By looking at the expressions for the transform Eq. (3.5) and the inverse transform Eq. (3.6), we can conclude that a discretization in one domain causes the result in the transform domain to be periodic.

3.3 Discrete Time – Discrete Frequency

In most real cases in digital signal processing, both the time- and frequency axes are discretized. Thus, both the signal sequence and its Fourier transform will be periodic. If the signal is known at N equidistant times separated by Δt, then one often specifies the transform at N equidistant frequency points separated by Δf. As will be shown, Δt and Δf are not independent of each other.

To find this relationship, we now look at one period of the Fourier transform, that we know from the previous Section has a length $f_s = 1/\Delta t$ in frequency space. If we have N frequency points (and time samples) available, then obviously the distance between the samples in frequency space is given by

$$\Delta f = \frac{f_s}{N} = \frac{1}{N\Delta t}. \tag{3.7}$$

Thus we have found that the frequency- and time-resolution are not independent, but always obey

$$\Delta f \Delta t = \frac{1}{N} \tag{3.8}$$

if N data values are available.

The expression for the Fourier transform in this case is

$$\begin{aligned} X[m] &= \Delta t \sum_{n=0}^{N-1} x[n] \exp(-j2\pi m \Delta f n \Delta t) \\ &= \Delta t \sum_{n=0}^{N-1} x[n] \exp\left(\frac{-j2\pi mn}{N}\right) \quad ; \quad m = 0, 1, .., N-1 \end{aligned} \tag{3.9}$$

where m is the frequency component index (i.e., $f = m\Delta f \quad ; \quad m = 0, 1, .., N-1$), and we applied Eq. (3.7) to obtain Eq. (3.9).

The inverse transform in this case is given by

$$x[n] = \Delta f \sum_{m=0}^{N-1} X[m] \exp\left(\frac{j2\pi mn}{N}\right) \quad ; \quad n = 0, 1, .., N-1. \tag{3.10}$$

The reader should now check that both $x[n]$ and $X[m]$ are periodic with a period of N by evaluation of $x[n+N]$ and $X[m+N]$.

Both Matlab and IDL contain function calls to a very important operation called a "Fast Fourier Transform" (FFT). An FFT is simply an extremely efficient way of calculating sums like those appearing in the above transform pair. Most practical solutions to digital signal analysis problems apply an FFT algorithm at some stage.

To sum up, we may organize all the four types of Fourier-transforms in the following table:

	Continuous time	**Discrete time**
Continuous frequency	Aperiodic in f and t	Periodic in f
Discrete frequency	Periodic in t	Periodic in f and t

As we can see from this table, a discretization in one domain always imply a periodization in the other domain.

3.4 The Power Spectral Density

When analyzing signals, a major concern is to know which frequencies that carry the signal power or signal energy. In the last section, we stated that only functions that are absolutely integrable have a Fourier transform. Therefore, stochastic processes are *not* Fourier transformable, since they will in general oscillate forever.

Even though the energy does not exist, it is meaningful to talk about the energy per time unit, or *power* contained in the signal. Consider now a time limited version $X_T(t)$ of the stochastic process $X(t)$,

$$X_T(t) = \begin{cases} X(t) & ; \quad -T < t < T \\ 0 & ; \quad \text{elsewhere} \end{cases} \tag{3.11}$$

This process clearly has an existing Fourier transform, $X_T(f)$. The average energy per time unit (or *average power*) for the process is now obtained by (1) dividing the energy by the time interval $2T$, (2) letting $T \to \infty$, and (3) applying the expectation operator on the resulting expression. We thus obtain the total average power of the process $X(t)$ as

$$P_{XX} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} E\left[|X(t)|^2\right] dt = \int_{-\infty}^{\infty} \lim_{T\to\infty} \frac{E\left[|X_T(f)|^2\right]}{2T} df, \tag{3.12}$$

where we used Parseval's theorem to obtain the last equality. From Eq. (3.12) we thus see that the expression

$$S_{XX}(f) = \lim_{T\to\infty} \frac{E\left[|X_T(f)|^2\right]}{2T} \tag{3.13}$$

represents the *power spectral density* (PSD) as a function of frequency. This is an extremely important quantity when discussing the properties of stochastic processes.

From Eq. (3.13) we thus see that the PSD of a stochastic process $X(t)$ can be formulated as

$$S_{XX}(f) = \lim_{T\to\infty} E\left[\frac{1}{2T}\left|\int_{-T}^{T} X(t)\exp(-j2\pi ft)dt\right|^2\right]. \tag{3.14}$$

This is the fundamental definition of a PSD to be used in these notes.

3.5 The Wiener-Khinchin Theorem

If the process under study is at least wide-sense stationary, then one can show a very important relation between the ACF and the PSD. These two important quantities are in fact a Fourier transform pair,

$$S_{XX}(f) = \mathcal{F}\{R_{XX}(\tau)\} = \int_{-\infty}^{\infty} R_{XX}(\tau)\exp(-j2\pi f\tau)d\tau \tag{3.15}$$

$$R_{XX}(\tau) = \mathcal{F}^{-1}\{S_{XX}(f)\} = \int_{-\infty}^{\infty} S_{XX}(\tau)\exp(j2\pi f\tau)df, \tag{3.16}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denotes the Fourier transform operator and its inverse, respectively. These relations are often called the *Wiener-Khinchin relations.*

3.6 Filtered Processes

Consider now a filtered stochastic process

$$Y(t) = \int_{-\infty}^{\infty} h(t-\xi)X(\xi)d\xi, \tag{3.17}$$

where $h(t)$ is the impulse response of a linear and time invariant (LTI) system.

If $X(t)$ is a wide-sense stationary process, it is straightforward to show that the expectation of the filtered process is

$$E[Y(t)] = \overline{X} \int_{-\infty}^{\infty} h(t)dt, \tag{3.18}$$

and that the ACF of the filtered process is

$$R_{YY}(\tau) = R_{XX}(\tau) * \tilde{\mathcal{R}}_{hh}(\tau), \tag{3.19}$$

where

$$\tilde{\mathcal{R}}_{hh}(\tau) = \int_{-\infty}^{\infty} h(t)h(t+\tau)dt \tag{3.20}$$

is the time autocorrelation function of the impulse response. Thus, the output process of an LTI system is also WSS if the input process is WSS.

It is now an easy to task to Fourier transform the output ACF $R_{YY}(\tau)$ in eq. (3.20) to obtain the power spectral density of the filtered process as

$$S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega), \tag{3.21}$$

where $H(\omega) = \mathcal{F}\{h(t)\}$ is the transfer function of the LTI system.

3.7 Cumulant Spectra

As a generalization of Wiener-Khinchin's theorem, we may define the *cumulant spectrum* or *polyspectrum* of order N to be the $(N-1)$-dimensional Fourier transform of the cumulant function of order N, i.e.

$$C_X^{(N)}(f_1, f_2, \ldots, f_{N-1}) = \tag{3.22}$$
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \kappa_X(\tau_1, \tau_2, \ldots, \tau_{N-1}) \exp\left[-j2\pi(\tau_1 f_1 + \cdots \tau_{N-1} f_{N-1})\right] d\tau_1 d\tau_2 \cdots d\tau_{N-1}$$

The condition for the $(N-1)$-dimensional Fourier transform to exist, is that the cumulant function is absolutely integrable.

If we assume a zero-mean process, we see that $N = 2$ yields the ordinary power spectral density. For orders $N \geq 3$ we obtain the so-called *higher-order spectra* which are important both from a fundamental point of view, and because of their impact on applications and practical data analysis.

3.7.1 The Bispectrum

The higher-order spectrum which has attracted the most attention is the *bispectrum*. This function of two frequency arguments (f_1, f_2) is the third-order cumulant spectrum $(N = 3)$, which is defined by

$$C_X^{(3)}(f_1, f_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \kappa_X^{(3)}(\tau_1, \tau_2) \exp\left[-j2\pi(\tau_1 f_1 + \tau_2 f_2)\right] d\tau_1 d\tau_2. \tag{3.23}$$

In contrast to the power spectrum (or the cumulant spectrum of order $N = 2$), we understand that the bispectrum is complex valued. This means that it retains the phase information about the underlying stochastic process, while the power spectrum is phase-blind. The phase-sensitivity of the bispectrum is also the explanation of the success and importance of bispectral analysis.

Due to the lag-symmetries in $\kappa_X^{(3)}(\tau_1, \tau_2)$, we can readily show that the bispectrum has the following symmetries

$$\begin{aligned} C_X^{(3)}(f_1, f_2) &= C_X^{(3)}(f_2, f_1) \\ &= C_X^{(3)*}(-f_2, -f_1) = C_X^{(3)}(-f_1 - f_2, f_2) \\ &= C_X^{(3)}(f_1, -f_1 - f_2) = C_X^{(3)}(-f_1 - f_2, f_1) \\ &= C_X^{(3)}(f_2, -f_1 - f_2). \end{aligned} \tag{3.24}$$

These symmetries are very important since they allow us to characterize the bispectrum in the entire (f_1, f_2)-plane only by knowledge of the bispectrum within a pizza-slice shaped sector.

The main problem concerning the use of higher-order spectra in practice, is that most of the estimators available are noisy, computationally intensive, and notoriously unreliable unless very much data are available. There are still many unsolved scientific problems in this field, and many potentially important signal processing problems that require that good polyspectral estimators are available.

Chapter 4

ESTIMATION THEORY

In general, when estimating a parameter θ from a finite sized data set, estimation errors are introduced. To quantify the quality of an estimator, it is customary to evaluate its expectation value and the statistical scatter around the expectation. In addition, if the data are Gaussian, one may discuss the significance of an estimate by means of the appropriate confidence intervals.

In general, an estimator of the parameter θ is defined as a function $g(\cdot)$ of an input data vector $\mathbf{x} = [x_0, x_1, \ldots, x_{N-1}]^T$, i.e., $\widehat{\theta} = g(\mathbf{x})$. For a given parameter, there usually exist a large number of different possible estimators $g(\cdot)$. Different estimators may e.g. be the result of different design criteria. It is therefore necessary to be able to compare different estimators and their performance.

It should be obvious that a valid estimator $\widehat{\theta} = g(\mathbf{x})$ cannot depend on the unknown value of θ. Note that some design criteria may lead to invalid estimators, so great care must be taken.

4.1 Bias and Variance of Estimators

The first and the second order central moments are convenient quantities when discussing the quality of an estimator.

It is customary to say that if the estimator $\widehat{\theta}$ is equal to θ *on average*, i.e.,

$$E[\widehat{\theta}] = \theta. \tag{4.1}$$

the estimator is centered or *unbiased*. Thus, one can define a quantity called the *bias* of the estimator by

$$b(\widehat{\theta}) = E[\widehat{\theta}] - \theta. \tag{4.2}$$

An *unbiased estimator* obviously has $b(\widehat{\theta}) = 0$.

The statistical scatter of the estimator around its mean value is conventionally quantified by the *estimation variance*

$$\mathrm{var}\{\widehat{\theta}\} = E\left[\left(\widehat{\theta} - E[\widehat{\theta}]\right)^2\right]. \tag{4.3}$$

It is desirable that the estimation variance is as small as possible, and ideally we would like $\mathrm{var}\{\widehat{\theta}\} \to 0$.

4.2 Consistent Estimators

We would like our estimator to be as good as possible in some statistical sense. A meaningful and desirable criterion of an estimator, is that is becomes better as more data are available. If $\widehat{\theta}_N$ denotes a parameter estimator based on N data samples, then $\widehat{\theta}_N$ is said to be *consistent* if

$$\lim_{N\to\infty} P\left\{\left|\widehat{\theta}_N - \theta\right| > \varepsilon\right\} = 0 \quad ; \quad \forall \varepsilon > 0. \tag{4.4}$$

This condition is often referred to as *convergence in probability*, because the probability that the estimator sequence deviates from the true value approaches zero asymptotically. It is also customary to say that the *stochastic limit* of $\widehat{\theta}_N$ is θ.

It is possible to link the asymptotic bias and variance properties to the definition of consistency by means of the well known Chebyshev inequality. The Chebyshev inequality states that

$$P\left\{\left|\widehat{\theta}_N - E\left[\widehat{\theta}_N\right]\right| \geq \varepsilon\right\} \leq \frac{\mathrm{var}\left\{\widehat{\theta}_N\right\}}{\varepsilon^2}, \tag{4.5}$$

for any $\varepsilon > 0$.

We thus understand that estimators that are asymptotically unbiased, and that has asymptotically vanishing variance

$$\lim_{N\to\infty} b(\widehat{\theta}_N) = 0 \tag{4.6}$$

$$\lim_{N\to\infty} \mathrm{var}\{\widehat{\theta}_N\} = 0. \tag{4.7}$$

are in fact consistent estimators. It is customary to disregard inconsistent estimators in practice of obvious reasons.

Chapter 5

ESTIMATION OF CORRELATION FUNCTIONS

In this section, we will look into ways of estimating an autocorrelation function (ACF) based on N samples from *a single* realization of a stochastic process $X(t)$. Recall that the basic definition of an ACF for a wide sense stationary stochastic process is [*Peebles*, 1993]

$$R_{XX}(\tau) = E[X(t)X(t+\tau)], \tag{5.1}$$

where $E[\cdot]$ denotes an ensemble average. The time shift τ is often called a *time lag* among signal analysts. Also, recall that for an ergodic process, all ensemble averages can be replaced by time averages, i.e.

$$\mathcal{R}_{XX}(\tau) = R_{XX}(\tau) \tag{5.2}$$

where the *time autocorrelation function* is defined by

$$\mathcal{R}_{XX}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)dt \tag{5.3}$$

where $x(t)$ is an arbitrary realization of the process.

It thus seems like a hopeless task to estimate a whole function $R_{XX}(\tau)$ defined on $-\infty < \tau < \infty$ on the basis of a finite sample set. We must therefore expect to make errors when we try to estimate the ACF based on real data sets.

5.1 An Unbiased ACF Estimator

Assume that a single realization $x(t)$ of the stochastic process $X(t)$ is regularly sampled (i.e., at equidistant points in time). Thus, the only data available are

$$x[n] \equiv x(n\Delta t) \quad ; \quad n = 0, 1, \ldots, N-1 \tag{5.4}$$

where Δt is the sampling interval.

How can the true ACF $R_{XX}(\tau)$ be estimated from this sampled realization? To answer this question, we shall first assume that an infinite number of samples are available from the realization, thus $x[n] = x(n\Delta t) \quad ; \quad n = \ldots, -1, 0, 1, \ldots$. The next two steps are:

1. Assume that $X(t)$ is ergodic, see Eq. (5.2).

2. Discretize the time integral in Eq. (5.3).

The discretization of the integral in Eq. (5.3) leads to

$$\begin{aligned} R_{XX}(\tau) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)dt \\ &\simeq \lim_{N\to\infty} \frac{1}{(2N+1)\Delta t} \sum_{n=-N}^{N} x[n]x[n+k]\Delta t, \end{aligned} \tag{5.5}$$

where the time shift k appearing corresponds to a time lag of $\tau = k\Delta t$ units.

An estimate for the ACF at time lag number k is thus given by

$$\widehat{R}_{XX}[k] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n]x[n+k]. \tag{5.6}$$

Note that in this equation we have assumed that an infinite number of samples were available.

In practice, however, we only have a finite number of data points available. Thus, an evaluation of the doubly infinite sum in Eq. (5.6) cannot be carried out for real world signals. Notice in the discretized integral Eq. (5.5), that we divide by the duration of the signal $(2N+1)\Delta t$, or after cancelling the common Δt factor, we divide by the number of data points $2N+1$ in Eq. (5.6).

When a finite sized sample set of length $N \not\to \infty$ is available, we must also divide by the number of products in the sum for each time lag k. Since the data set is of finite size, the number of available products will depend on which lag k that is being estimated. If the data set is $x[n] = x(n\Delta t) \quad ; \quad n = 0, 1, \ldots, N-1$, a possible estimator for the ACF is given by

$$\widehat{R}^{(1)}_{XX}[k] = \frac{1}{N-k} \sum_{n=0}^{N-1-k} x[n]x[n+k] \quad ; \quad k = 0, 1, \ldots, N-1, \tag{5.7}$$

and $\widehat{R}^{(1)}_{XX}[-k] = \widehat{R}^{(1)}_{XX}[k]$ (because of property (2) on page 174 in *Peebles* [1993]). A compact representation of the estimator that is valid for both positive and negative lags is then (check this yourself!)

$$\widehat{R}^{(1)}_{XX}[k] = \frac{1}{N-|k|} \sum_{n=0}^{N-1-|k|} x[n]x[n+|k|] \quad ; \quad k = 0, \pm 1, \ldots, \pm(N-1). \tag{5.8}$$

In a practical situation, we see that it is sufficient to calculate only the positive lags $0 \le k \le N-1$, since the negative lags have values completely given by the positive lags.

Bias and Variance

By taking the ensemble average of $\widehat{R}^{(1)}_{XX}[k]$, we now see that

$$E\left[\widehat{R}^{(1)}_{XX}[k]\right] = R_{XX}[k]. \tag{5.9}$$

This ACF estimator is therefore *unbiased*, since $b\{\widehat{R}_{XX}^{(1)}[k]\} = 0$.

Evaluating the variance properties of the ACF estimator in Eq. (5.8) is a difficult task in the general case. If we assume that $E[X(t)] = 0$, and that $X(t)$ is a Gaussian process, then it is possible to show that [e.g., *Proakis et al.*, 1992]

$$\mathrm{var}\{\widehat{R}_{XX}^{(1)}[k]\} = \frac{1}{(N-|k|)^2} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \left\{ R_{XX}^2[m-n] + R_{XX}[m-n+|k|]R_{XX}[m-n-|k|] \right\}. \tag{5.10}$$

Eq. (5.10) was obtained by using the important fact that if X_1, X_2, X_3, X_4 are four zero-mean Gaussian variables, then [see e.g., *Proakis et al.*, 1992]

$$E[X_1X_2X_3X_4] = E[X_1X_2]\,E[X_3X_4] + E[X_1X_3]\,E[X_2X_4] + E[X_1X_4]\,E[X_2X_3]. \tag{5.11}$$

One may readily derive eq. (5.11) by a four fold differentiation of the characteristic function $\Phi_{\mathbf{X}}(\xi)$ for Gaussian variables.

For a finite time lag k, we see that in the asymptotic limit $N \to \infty$ we find

$$\lim_{N\to\infty} b\{\widehat{R}_{XX}^{(1)}[k]\} = 0 \tag{5.12}$$

$$\lim_{N\to\infty} \mathrm{var}\{\widehat{R}_{XX}^{(1)}[k]\} = 0 \tag{5.13}$$

The ACF estimator $\widehat{R}_{XX}^{(1)}[k]$ is thus a consistent estimator, provided that

$$\sum_{k=-\infty}^{\infty} R_{XX}^2[k] < \infty.$$

Even though this ACF estimator has nice statistical properties asymptotically, there are severe problems when working on finite sized data sets. This can be understood by considering Eq. (5.10) for lags k approaching N (i.e., "large" lags). We see that since we divide by a factor $(N-|k|)^2$, the estimation variance increases as $|k|$ increases. Thus, the statistical errors are larger for large $|k|$ than for small $|k|$. The reason for this is of course that when estimating the larger lags, we average over considerably fewer data products. E.g., when evaluating the expression for $k = 0$, Eq. (5.8) will contain an average over N signal products. However, when $|k| = N-1$ is inserted, there is only one product available, $x[0]x[N-1]$, and no averaging at all takes place. Therefore, the larger time lags suffer from a small degree of averaging, which must be bad since the summations are approximations to time averages.

We conclude that care must be taken when applying the unbiased ACF estimator $\widehat{R}_{XX}^{(1)}[k]$, since large estimation errors are to be expected for large time lags. For small time lags, however, one should expect the estimator to work satisfactorily.

5.2 A Biased ACF Estimator

An alternative ACF estimator can be constructed by replacing the factor $(N-|k|)$ in Eq. (5.8) by N. Instead of dividing by the number of products in the summation, we now divide by the total number of samples available. All lags are thus divided by the same factor. This is the "standard" ACF estimator that is most often used in practice, and its formal definition is

$$\widehat{R}_{XX}^{(2)}[k] = \frac{1}{N} \sum_{n=0}^{N-1-|k|} x[n]x[n+|k|] \quad ; \quad k = 0, \pm 1, .., \pm(N-1). \tag{5.14}$$

This ACF estimator is in general biased, since

$$E\left[\widehat{R}_{XX}^{(2)}[k]\right] = \frac{N-|k|}{N} R_{XX}[k]. \tag{5.15}$$

The bias for this estimator is then

$$b\{\widehat{R}_{XX}^{(2)}[k]\} = -\frac{|k|}{N} R_{XX}[k]. \tag{5.16}$$

The bias increases for the larger lags, but for small lags the bias is small. Note however that for a fixed time lag, the estimator is *asymptotically* unbiased.

Again, it is difficult to evaluate the variance for a general process. If $X(t)$ is a zero-mean Gaussian process, it is however possible to show that

$$\mathrm{var}\{\widehat{R}_{XX}^{(2)}[k]\} = \frac{1}{N^2} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \left\{ R_{XX}^2[m-n] + R_{XX}[m-n+|k|] R_{XX}[m-n-|k|] \right\}. \tag{5.17}$$

By taking the appropriate limit $N \to \infty$, we see that also this ACF estimator is consistent, since is is asymptotically unbiased, and the variance approaches zero asymptotically. An important property of this estimator is seen from the expression for the variance Eq. (5.17) for finite time lags. Now, the variance does not increase as $|k|$ approaches N. We have thus removed the unwanted large statistical error in the unbiased ACF estimate $\widehat{R}_{XX}^{(1)}[k]$, but at the expense of introducing a bias in the estimate.

In Fig. 2 we show typical results obtained from two different ACF-estimators. We generated $N = 50$ samples of an autoregressive process of order one (AR(1)), with parameters $a_1 = -0.6$ and driving noise variance of $\sigma^2 = 1 - a_1^2$. The unbiased ACF-estimate is represented by the dash-dotted line, and the biased estimate is represented by the dashed line. The true ACF is the full line. We note the following facts from this figure. For small lags, the two estimators yield almost identical results. For higher lags, the unbiased ACF estimate starts to oscillate wildly, whereas the biased estimator settles with small variations from one k value to the next. The erratic behavior in the unbiased estimator is a clear manifestation of the $1/(N-|k|)^2$ dependence of the variance.

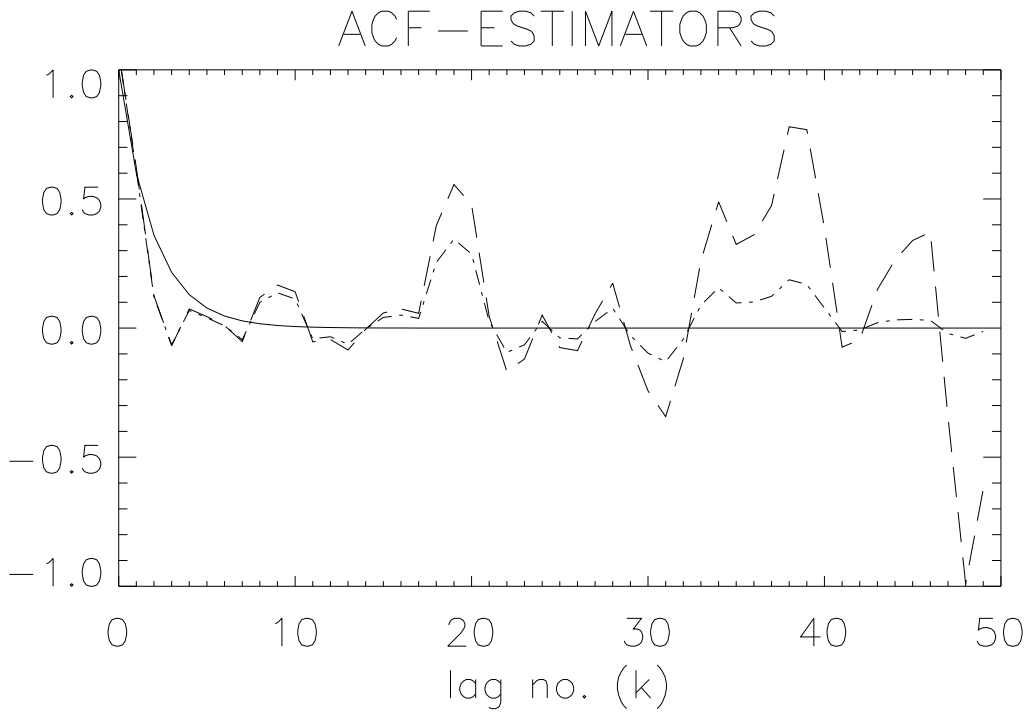


Figure 5.1: ACF-estimates of data from AR(1)-process with $a_1 = -0.6$, $\sigma^2 = 1 - a_1^2$, and $N = 50$. Full line: true ACF, dash-dotted line: biased ACF-estimate, and dashed line: unbiased ACF-estimate.

Chapter 6

ESTIMATION OF POWER SPECTRAL DENSITIES

A very important task for signal analysts is to estimate the Power Spectral Density (PSD) of a stochastic process $X(t)$ based on a finite set of samples $x[n] \quad ; \quad n = 0, 1, .., N-1$ of a realization of the process. Spectral estimation is a very active field of signal processing research, and applications exist in radar/sonar/lidar, communications, image analysis, medical signal processing, and all fields where data analysis occurs. In these Lecture Notes, we will derive some of the most basic digital spectral estimation techniques. All these techniques are simple to implement and understand, and they form the basis for other more advanced estimation techniques that you will encounter in later courses and scientific work.

These notes will basically cover three different classes of spectral estimators: (1) the periodogram, which is a discretization of the basic definition of the continuous PSD, (2) the Blackman-Tukey estimator, which is a discretization of the Wiener-Khinchin relation, and (3) autoregressive spectral estimation, which applies a specific signal model that is fitted to the data set.

6.1 The Periodogram

The so-called "periodogram" is the oldest and simplest of all power spectral estimators. It was derived already in 1898 by *Schuster*. He developed the technique as a tool to study "hidden periodicities" in astrophysical data, hence its name.

If only N samples of one single realization of the process are available, the following three things happen: (1) the expectation operator in Eq. (3.14) is irrelevant, and (2) the infinite time extent applied in Eq. (3.14) must be replaced by the actual observation time of the signal, (3) the integral must be replaced by a summation.

Let the discretized time axis be $t_n = n\Delta t$ where $n = 0, 1, .., N-1$. Then, it is obvious that the total observation interval $2T$ must be replaced by $N\Delta t$, and the integral must be replaced by the appropriate sum. The periodogram power spectral density estimator is thus given by

$$\widehat{S}_{XX}^{(per)}(f) \quad = \quad \frac{1}{N\Delta t} \left| \Delta t \sum_{n=0}^{N-1} x[n] \exp(-j2\pi f n \Delta t) \right|^2$$

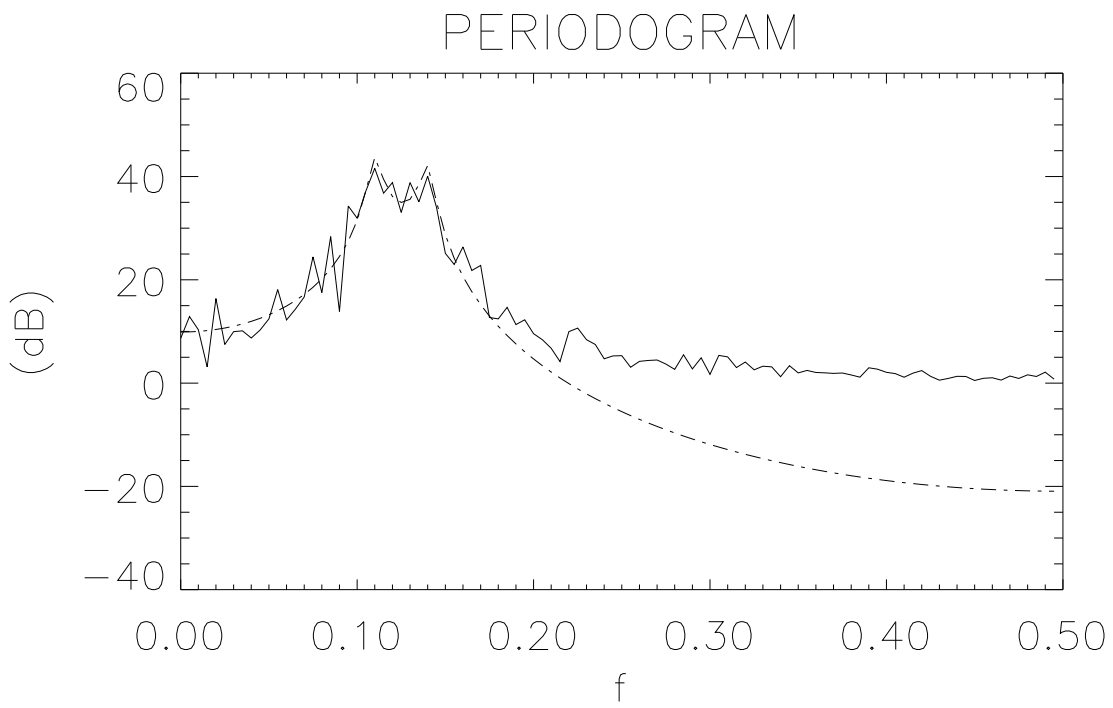


Figure 6.1: Periodogram estimate based on $N = 200$ samples of an AR(4)-process with $a_1 = -2.7607, a_2 = 3.8106, a_3 = -2.6535, a_4 = 0.9238$, and $\sigma^2 = 1$. The data are shown in Fig. 1, lower left panel. Full line: periodogram estimate, dash-dotted line: true PSD.

$$= \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} x[n]\exp(-j2\pi f n\Delta t)\right|^2 \quad ; \quad |f| \leq \frac{1}{2\Delta t} \tag{6.1}$$

The above PSD estimator is valid for a continuous frequency f. Note that this estimator is periodic in the frequency domain with a period of $f = 1/\Delta t$, due to the periodic behavior of the Discrete-Continuous Fourier transform. One should therefore plot only *one period* of the periodogram estimator, since the rest of the frequency information is redundant.

In practical situations, one often calculate the periodogram in N different frequency locations. If the frequency axis is discretized as $f = m\Delta f$, for $m = 0, 1, .., N-1$, and where $\Delta f = 1/(N\Delta t)$, then the discrete frequency version is given by

$$\widehat{S}_{XX}^{(per)}[m] = \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} x[n]\exp\left(-j\frac{2\pi mn}{N}\right)\right|^2 \quad ; \quad m = 0, 1, .., N-1. \tag{6.2}$$

This estimator forms the basis for most classical spectral estimators, despite the fact that the periodogram itself has very poor statistical properties.

Note an important fact of the periodogram estimator in Eq. (6.2). Since the frequency axis is periodic, and we only calculate "positive" frequencies ($m \geq 0$), it is obvious that the rightmost half of the periodogram estimate corresponds to the negative frequencies. In detail, $m = 0, 1, .., N/2$ corresponds to $f = 0, \Delta f, 2\Delta f, .., N\Delta f/2 = f_s/2$, and $m = N/2+1, .., N-1$ corresponds to $f = (-N/2+1)\Delta f, (-N/2+2)\Delta f, .., -\Delta f$. You should make a sketch of the

discretized frequency axis, and figure out where the correct correspondence between frequency index m and true frequency f.

Even though it was very simple to derive the periodogram estimator, its statistical properties are very undesirable as will be shown in later in this. Therefore, care must be taken when using this estimator.

6.2 The Blackman-Tukey Estimator

An alternative PSD estimator can be derived by means of the Wiener-Khinchin relation. In its continuous form, the Wiener-Khinchin representation of the PSD in terms of the ACF reads [e.g., *Peebles*, 1993]

$$S_{XX}(f) = \int_{-\infty}^{\infty} R_{XX}(\tau) \exp(-j2\pi f\tau) d\tau. \tag{6.3}$$

We now discretize the time lag as $\tau = k\Delta t$ where $k = -(N-1), .., -1, 0, 1, .., N-1$, and we assume that some ACF estimate is available at the discrete times,

$$\widehat{R}_{XX}[m] = \widehat{R}_{XX}(m\Delta t).$$

By discretizing the continuous equation (6.3), we now obtain the standard Blackman-Tukey estimator as

$$\widehat{S}_{XX}^{(BT)}(f) = \Delta t \sum_{k=-M}^{M} \widehat{R}_{XX}[m] \exp(-j2\pi f k\Delta t) \quad ; \quad |f| \leq \frac{1}{2\Delta t} \tag{6.4}$$

where $M \leq N-1$. Note that in general we can choose how many ACF-lags M we want to use in the Blackman-Tukey estimator. The significance of this will be demonstrated later through computer based problems and theoretical considerations. Note also that we have not specified *which* ACF estimator one should apply in the Blackman-Tukey estimator. This is a choice of the analyst, and different ACF estimators will give different power spectral estimates.

A very important special case occurs when we choose the (standard) biased ACF estimator from Eq. (5.14), i.e., $\widehat{R}_{XX}[k] = \widehat{R}_{XX}^{(2)}[k]$, and we use all available lags, i.e., $M = N-1$. In this special case, one can readily show that the periodogram estimator and the Blackman-Tukey estimator are identical,

$$\widehat{S}_{XX}^{(per)}(f) = \widehat{S}_{XX}^{(BT)}(f). \tag{6.5}$$

6.3 Statistical Properties

For data with general amplitude statistics, it is difficult to analyze the statistical properties of the periodogram and the Blackman-Tukey PSD estimators. The expectation value is however easy to obtain, and it is in fact independent of the probability density of the data. For the variance it is not so, and this quality measure will be discussed only for zero-mean Gaussian data.

Bias of the periodogram

If we again assume that $\widehat{R}_{XX}[k] = \widehat{R}^{(2)}_{XX}[k]$ and $m = N-1$ are applied, then it becomes easy to evaluate the bias of the Periodogram going via the expression for the BT-estimator. We now obtain

$$\begin{aligned} E\{\widehat{S}^{(BT)}_{XX}(f)\} &= \Delta t \sum_{k=-(N-1)}^{N-1} E\{\widehat{R}^{(2)}_{XX}[k]\} \exp(-j2\pi f k \Delta t) \\ &= \Delta t \sum_{k=-(N-1)}^{N-1} \left(\frac{N-|k|}{N}\right) R_{XX}[k] \exp(-j2\pi f k \Delta t), \end{aligned} \tag{6.6}$$

where we used Eq. (5.15) to obtain the last line.

If we define a "lag-window" $w_B[k]$ as

$$w_B[k] = 1 - \frac{|k|}{N} \quad ; \quad k = 0, \pm 1, .., \pm(N-1) \tag{6.7}$$

then we see that the expectation value may be formulated as

$$E\{\widehat{S}^{(BT)}_{XX}(f)\} = \Delta t \sum_{k=-(N-1)}^{N-1} w_B[k] R_{XX}[k] \exp(-j2\pi f k \Delta t). \tag{6.8}$$

The window $w_B[k]$ is often called the Bartlett-window due to its inventor, or the "triangular window" due to its shape. We will use both names interchangeably.

We now note that Eq. (6.8) is a Fourier transform of a product between $w_B[k]$ and $R_{XX}[k]$. But a Fourier transform of a product is nothing but a convolution of the Fourier transformed functions, so another useful way of expressing $E\{\widehat{S}^{(BT)}_{XX}(f)\}$ is

$$E\{\widehat{S}^{(BT)}_{XX}(f)\} = W_B(f) * S_{XX}(f) \tag{6.9}$$

where

$$\begin{aligned} W_B(f) &= \Delta t \sum_{k=-(N-1)}^{N-1} w_B[k] \exp(-j2\pi k \Delta t f) \\ &= \frac{\Delta t}{N} \frac{\sin^2(N\pi f \Delta t)}{\sin^2(\pi f \Delta t)} \quad ; \quad |f| \leq \frac{1}{2\Delta t} \end{aligned} \tag{6.10}$$

is the Fourier transform of the window sequence $w_B[k]$. The frequency function $W_B(f)$ is often called the Dirichlet kernel [*Percival and Walden*, 1993].

The above result shows that the Periodogram spectral estimator (and the Blackman-Tukey in this special case) is in general *biased*, since the expectation value is not the true spectrum $S_{XX}(f)$, but the convolution of $S_{XX}(f)$ with another function $W_B(f)$,

$$E\{\widehat{S}^{(per)}_{XX}(f)\} = W_B(f) * S_{XX}(f). \tag{6.11}$$

However, when $N \to \infty$ we see that

$$\lim_{N\to\infty} w_B[k] = 1,$$

and

$$\lim_{N\to\infty} W_B(f) = \delta(f). \tag{6.12}$$

Thus, in the asymptotic limit we find that

$$\lim_{N\to\infty} E\{\widehat{S}_{XX}^{(per)}(f)\} = \delta(f) * S_{XX}(f) = S_{XX}(f). \tag{6.13}$$

We can therefore conclude that the Periodogram is an *asymptotically unbiased* spectral estimator, but for finite N it may be substantially biased.

Variance of the periodogram

In general, it is impossible to find an intelligible expression for the variance of the periodogram estimator for an arbitrary stochastic process $X(t)$. The calculation we need to perform is obviously

$$\mathrm{var}\{\widehat{S}_{XX}^{(per)}(f)\} = E\left\{\left[\widehat{S}_{XX}^{(per)}(f)\right]^2\right\} - \left[E\left\{\widehat{S}_{XX}^{(per)}(f)\right\}\right]^2. \tag{6.14}$$

Direct calculation shows that

$$E\left\{\left[\widehat{S}_{XX}^{(per)}(f)\right]^2\right\} = \frac{\Delta t^2}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}\sum_{n_3=0}^{N-1}\sum_{n_4=0}^{N-1} E\left\{x[n_1]x[n_2]x[n_3]x[n_4]\right\} e^{-j2\pi f(n_1-n_2+n_3-n_4)\Delta t}. \tag{6.15}$$

We see that the fourth order moment of the data enters explicitly. To get any further, we now assume that the data come from a zero-mean Gaussian process. Then, the fourth moment property of the Gaussian, see Eq. (5.11) can be applied to obtain

$$\begin{aligned} E\left\{\left[\widehat{S}_{XX}^{(per)}(f)\right]^2\right\} &= \left[E\left\{\widehat{S}_{XX}^{(per)}(f)\right\}\right]^2 + \left|\frac{\Delta t}{N}\sum_n\sum_{n_1} R_{XX}[n-n_1]e^{-j2\pi f(n+n_1)\Delta t}\right|^2 \\ &+ \left|\frac{\Delta t}{N}\sum_n\sum_{n_1} R_{XX}[n-n_1]e^{-j2\pi f(n-n_1)\Delta t}\right|^2. \end{aligned} \tag{6.16}$$

The variance is therefore just given by the two last terms in the equation above. Formulating the expression in the frequency domain, we can write

$$\begin{aligned} \mathrm{var}\{\widehat{S}_{XX}^{(per)}(f)\} &= \left|\int_{-1/(2\delta t)}^{1/(2\Delta t)} S_{XX}(f)\frac{\Delta t}{N}W(f'+f)W^*(f'-f)df'\right|^2 \\ &+ \left(\int_{-1/(2\Delta t)}^{1/(2\Delta t)} S_{XX}(f')\frac{\Delta t}{N}\left|W(f'-f)\right|^2 df'\right)^2, \end{aligned} \tag{6.17}$$

where $W(f)$ is the DFT of the uniform function $w[n] = 1 \quad ; \quad n = 0, 1, \ldots, N-1$.

For $f \neq 0$ and N sufficiently large, the first term in Eq. (6.17) will be negligible, while the second term will approach $S_{XX}^2(f)$. In the asymptotic limit we thus obtain the variance as

$$\lim_{N\to\infty} \mathrm{var}\{\widehat{S}_{XX}^{(per)}(f)\} = S_{XX}^2(f). \tag{6.18}$$

The periodogram estimator is therefore *not consistent*, since $\lim_{N\to\infty} \mathrm{var}\{\widehat{S}_{XX}^{(per)}(f)\} \neq 0$.

Covariance of the periodogram

The so-called "Fourier frequencies" $f = n/(N\Delta t)$, where $n = -N/2+1, .., -1, 0, 1, .., N/2$ play a special role in spectral estimation. For instance, it is possible to show that for a zero-mean Gaussian process, the covariance between the periodogram estimate at two different frequencies f_1 and f_2 is given by

$$\mathrm{cov}\{\widehat{S}_{XX}^{(per)}(f_1), \widehat{S}_{XX}^{(per)}(f_2)\} = 0 \tag{6.19}$$

if both f_1 and f_2 are Fourier frequencies (i.e., multiples of $1/(N\Delta t)$. For all other frequency pairs, we find that

$$\mathrm{cov}\{\widehat{S}_{XX}^{(per)}(f_1), \widehat{S}_{XX}^{(per)}(f_2)\} \neq 0. \tag{6.20}$$

The conclusion is thus that the periodogram produces spectral estimates that are uncorrelated if we consider the Fourier frequencies, but that all other frequency pairs will be statistically correlated.

A word of caution

Because of the poor statistical properties in the asymptotic limit, it is correct to call the periodogram (and the Blackman-Tukey estimator for the parameters applied here) a *naive* spectral estimator. From Eq. (6.18) it is evident that the asymptotic standard deviation is as large as the quantity we want to estimate! The estimation error thus becomes larger, the larger the true spectral density is. Even though the periodogram is regretfully still used in applications by many ignorants (some people erroneously even call the periodogram for "the spectrum of the signal"!), one should simply ignore any conclusions drawn from a periodogram of a single dataset. We have shown that the periodogram is severely biased due to leakage, and that is an inconsistent estimator. The only conclusion we can reach is therefore that *one shall never apply the naive periodogram* to answer any scientific or technological questions.

Lord Rayleigh commented on the severe problems with the periodogram already around 1910, and the inconsistency has been well known for almost a century now. That many ignorants still persist in using periodograms is a confirmation that human evolution is a slow process.

Note however, that there are several practical ways to improve the periodogram substantially. Thus, even though one shall never apply the periodogram directly, one may derive meaningful spectral estimators that are (fundamental) modifications of the periodogram.

Chapter 7

IMPROVED CLASSICAL SPECTRAL ESTIMATORS

As shown in the last section, the statistical properties of the Periodogam and the Blackman-Tukey estimator for $M = N - 1$ were undesirable. We thus need to find ways of improving the spectral estimators. In this section, we will describe some standard ways of improving the classical spectral estimators.

7.1 Leakage reduction by data windows

We saw in Section 8.3 that the periodogram estimator is in general biased. The bias can be attributed to a spectral leakage from one part of the PSD to another, due to the high sidelobes of the rectangular data window. One may reduce the spectral sidelobes and hence the bias, by introducing data windows that multiply the data set prior to the Fourier transformation. See the books by *Marple* [1987] and *Kay* [1988] for details about the effect of data windows in spectral estimation.

Let the sample set be $x[0], .., x[N-1]$, and let $w[0], .., w[N-1]$ denote a real valued sequence of numbers called the *data window* or *data taper*. We now define the *windowed periodogram* or the *modified periodogram* by

$$\widehat{S}_{XX}^{(wper)}(f) = \frac{\Delta t}{N\mathcal{U}} \left| \sum_{n=0}^{N-1} w[n]x[n] \exp(-j2\pi f n \Delta t) \right|^2 . \tag{7.1}$$

Here, $\mathcal{U}$ is a normalization factor that depends on the energy contained in the data window,

$$\mathcal{U} = \frac{1}{N} \sum_{n=0}^{N-1} w^2[n]. \tag{7.2}$$

It is easy to show that the bias of this spectral estimator is controlled by the spectral properties of the window sequence $w[n]$, or

$$E\left\{\widehat{S}_{XX}^{(wper)}(f)\right\} = \int_{-1/(2\Delta t)}^{1/2(\Delta t)} Q(f - f')S_{XX}(f')df', \tag{7.3}$$

where the so-called *spectral window* or the *frequency response* is defined by [*Percival and Walden*, 1993]

$$Q(f) = \frac{\Delta t}{N\mathcal{U}} \left| \sum_{n=0}^{N-1} w[n] \exp(-j2\pi f n \Delta t) \right|^2 . \tag{7.4}$$

Thus, if we can find a data window that has a larger mainlobe-to-sidelobe ratio than does the *rectangular data window* $w[n] \equiv 1$, then we have in effect reduced the leakage through the sidelobes of the spectral widow. However, there is a price to pay: if we use a window with better leakage properties, then the *width* of mainlobe always increases. This means that the ability to separate two closely spaced components may become worse. Data windowing therefore always imply a trade-off between leakage reduction and loss of spectral resolution.

In Fig. 4 we display some standard data windows (left column), along with their energy spectra (right column). From top to bottom, we see the rectangular, triangular, Hann(ing), and Hamming windows, respectively.

The normalization factor $\mathcal{U}$ ensures that the energy of the data window is unity,

$$\int_{-1/(2\Delta t)}^{1/(2\Delta t)} Q(f) df = 1.$$

This has the effect that the energy content of the signal $x[n]$ is not altered by the windowing. Windowed periodograms are however frequently used without the normalization implied by Eq. (7.2).

In Fig. 5 we show the effect of windowing the AR(4) data from Fig. 1. Note that there is almost no difference between the results when using a triangular and a Hann window, while the Hamming case has a larger bias at frequencies approaching the Nyquist frequency.

7.2 Averaged periodogram

One of the reasons for the bad performance of the Periodogram, is the neglect of the expectation operator entering the basic PSD definition, Eq. (3.14). An intuitive way of simulating the effect of $E\{\cdot\}$, is to introduce an average over a set of spectral estimates. This technique is commonly referred to as the "Averaged Periodogram", and is constructed as follows.

Assume that N data points $x[0], x[1], .., x[N-1]$ are available. Instead of calculating the periodogram based on the N samples, we now segment the data set into K segments each consisting of M samples, such that

$$N = KM.$$

Data segment number k of length M is thus formed as

$$x_k[m] = x[m + kM] \quad ; \quad m = 0, 1, .., M-1 \tag{7.5}$$

where $k = 0, .., K-1$. The periodogram based on segment k is defined by

$$\widehat{S}_{XX}^{(per)}(f;k) = \frac{\Delta t}{M} \left| \sum_{m=0}^{M-1} x_k[m] \exp(-j2\pi f m \Delta t) \right|^2 \quad ; \quad |f| < 1/(2\Delta t), \tag{7.6}$$

and the resulting *averaged periodogram* is simply the arithmetic average of the K individual periodograms

$$\widehat{S}_{XX}^{(Mper)}(f) = \frac{1}{K}\sum_{k=0}^{K-1}\widehat{S}_{XX}^{(per)}(f;k) \quad ; \quad |f| < 1/(2\Delta t). \tag{7.7}$$

It can be shown that the variance of this estimator is given by

$$\mathrm{var}\{\widehat{S}_{XX}^{(Mper)}(f)\} \simeq \frac{1}{K}S_{XX}^2(f), \tag{7.8}$$

and the expectation value is given by

$$E\{\widehat{S}_{XX}^{(Mper)}(f)\} = S_{XX}(f) * W_B(f), \tag{7.9}$$

where $W_B(f)$ is the Fourier transform of the triangular window based on M points. From Eq. (7.8) it is evident that the variance of the average periodogram decreases drastically when segmenting the data set into blocks. The price we pay, is that the ability to resolve closely spaced spectral maxima will decrease. This can be seen from Eq. (7.9), since we know that the width of the mainlobe of the triangular window increases with decreasing size of the data set. Thus, since $W_B(f)$ is based on M points rather than N, the mainlobe becomes a factor K broader, and the convolution in Eq. (7.9) leads to a larger smearing of the true spectrum $S_{XX}(f)$. We can thus conclude that the "frequency resolution" becomes worse, but that the estimation variance decreases.

7.3 Weighted Overlapped Segment Averaging

The so-called "Weighted Overlapped Segment Averaging" (WOSA) technique allows the data blocks to overlap, and it applies a data window $w[n]$ on each data segment before Fourier transforming. This technique was developed by *Welch* (1967), and is therefore also commonly referred to as Welch's method or "Welch's Overlapped Segment Averaging". This is therefore an estimation technique that combines the bias reduction property of data windowing, and the variance reduction property of the averaged periodogram.

We now segment the data set $x[0], x[1], .., x[N-1]$ into blocks of length M according to

$$x_k[m] = x[m + kD] \quad ; \quad m = 0, 1, .., M-1 \tag{7.10}$$

where $k = 0, 1, .., L-1$. Here, kD denotes the starting point for segment number k, and L is the total number of data segments. When $D = M$ we have no overlap, and $L = K$ as in the averaged periodogram technique. However, when $D = M/2$, we obtain a 50% overlap between adjacent data segments, and we obtain $L = 2K - 1$ data segments.

In the WOSA technique, we multipy each of the data segments by a data window $w[m]$ before computing the individual periodograms. For segment number k we thus form a "modified" spectral estimate

$$\widehat{S}_{XX}^{(W)}(f;k) = \frac{\Delta t}{M\mathcal{U}}\left|\sum_{m=0}^{M-1} w[n]x_k[n]\exp(-j2\pi f m\Delta t)\right|^2 \quad ; \quad |f| < 1/(2\Delta t), \tag{7.11}$$

for $k = 0, 1, .., L-1$, where $\mathcal{U}$ is now defined as

$$\mathcal{U} = \frac{1}{M}\sum_{m=0}^{M-1} w^2[m]. \tag{7.12}$$

The resulting WOSA spectral estimator is the arithmetic average of the L modified periodograms,

$$\widehat{S}_{XX}^{(W)}(f) = \frac{1}{L}\sum_{k=0}^{L-1} \widehat{S}_{XX}^{(W)}(f;k) \quad ; \quad |f| < 1/(2\Delta t). \tag{7.13}$$

One can show that the expectation value of the WOSA estimator is given by

$$E\{\widehat{S}_{XX}^{(W)}(f)\} = S_{XX}(f) * Q(f), \tag{7.14}$$

where

$$Q(f) = \frac{\Delta t}{M\mathcal{U}} \left| \sum_{m=0}^{M-1} w[m] \exp(-j2\pi f m \Delta t) \right|^2 .$$

The variance of the WOSA depends on the data window $w[m]$ that we apply, and on the degree of overlap. *Welch* (1967) showed that for the case of a triangular window, and 50% overlap ($D = M/2$), the following approximate expression for the estimator variance applies,

$$\mathrm{var}\{\widehat{S}_{XX}^{(W)}(f)\} \simeq \frac{9}{8L} S_{XX}^2(f). \tag{7.15}$$

It is now evident that the WOSA implies a larger variance reduction that the averaged periodogram. For the case of 50% overlap and triangular window, we have $L = 2K - 1$, where K is the number of nonoverlapping segments applied in the averaged periodogram technique. Thus, the ratio of variances for the WOSA and the averaged periodogram is

$$\frac{\mathrm{var}\{\widehat{S}_{XX}^{(W)}(f)\}}{\mathrm{var}\{\widehat{S}_{XX}^{(Mper)}(f)\}} \simeq \frac{9}{16}, \tag{7.16}$$

which shows that we obtain a variance reduction close to 0.5 when applying a 50% overlap of the data segments.

In Fig. 6 we see the WOSA spectral estimate of the AR(4) data from Fig. 1. We applied $M = 100$ data points in each segment, and a 50% overlap, resulting in $L = 3$ segments. The data window was of the Hann type.

7.4 Frequency Smoothing

A simple and intuitively appealing way of reducing the variance is by smoothing of the periodogram. The simplest form of a frequency smoother weights a set of neighboring Fourier frequencies uniformly, or

$$\widehat{S}_{XX}^{(SM)}[m] = \frac{1}{2K+1} \sum_{k=-K}^{K} \widehat{S}_{XX}^{(per)}[m-k]. \tag{7.17}$$

We see that the smoothing is symmetric around the frequency m of interest. Since the periodogram at different harmonic frequencies are statistically independent, we understand that the variance is reduced by a factor $2K + 1$ with respect to the periodogram, i.e.,

$$var\left\{\widehat{S}_{XX}^{(SM)}[m]\right\} = \frac{S_{XX}^2[m]}{2K+1}. \tag{7.18}$$

It is obvious that the frequency averaging procedure may introduce a bias in the spectral estimate. A spectral maximum will e.g. be underestimated, since the spectral value at the maximum will be downweighted by the surrounding lower values due to the averaging. Also, the frequency resolution will be decreased, since we introduce correlations between frequencies.

In general, one may apply a frequency smoothing kernel of a different shape than the uniform smoothing described above. Such a general frequency smoother can be written as

$$\widehat{S}_{XX}^{(SM)}[m] = \frac{1}{2K+1} \sum_{k=-K}^{K} q[k] \widehat{S}_{XX}^{per}[m-k]. \tag{7.19}$$

where $q[k]$ is a symmetric kernel with region of support $-K \leq k \leq K$. By choosing the kernel to be more concentrated than the uniform kernel, we introduce less bias. The price we pay, however, is a lower variance reduction. The statistical analysis of the frequency smoothed periodogram for a general kernel $q[k]$ is nontrivial.

7.5 Blackman-Tukey with $M < N - 1$

So far, when discussing the Blackman-Tukey method, we assumed $M = N - 1$, or that all lags of the ACF was to be applied when evaluating Eq. (6.4). Without going into the details, we will argue that an improvement of the BT-estimator can be achieved by neglecting the higher lags, i.e., by applying $M < N - 1$. The hope is that the statistical errors involved in the PSD estimate decreases when we skip the ACF-estimates that have a large variance.

A further generalisation can be invoked by multiplying the ACF estimates by a so-called *lag-window* prior to the Fourier-transformation implied by Eq. (6.4).

Note that using $M < N - 1$ can produce invalid spectral estimates in the sense that *negative* spectral densities are obtained for some frequencies. One should therefore be *very careful* when drawing conclusions from the BT-estimator in the general case. Some authors invoke additional contraints to ensure that the lag-windowed BT-estimator yield valid non-negative spectral estimates. This puts restrictions both on the choice of lag-windows, and on M for a given lag-window.

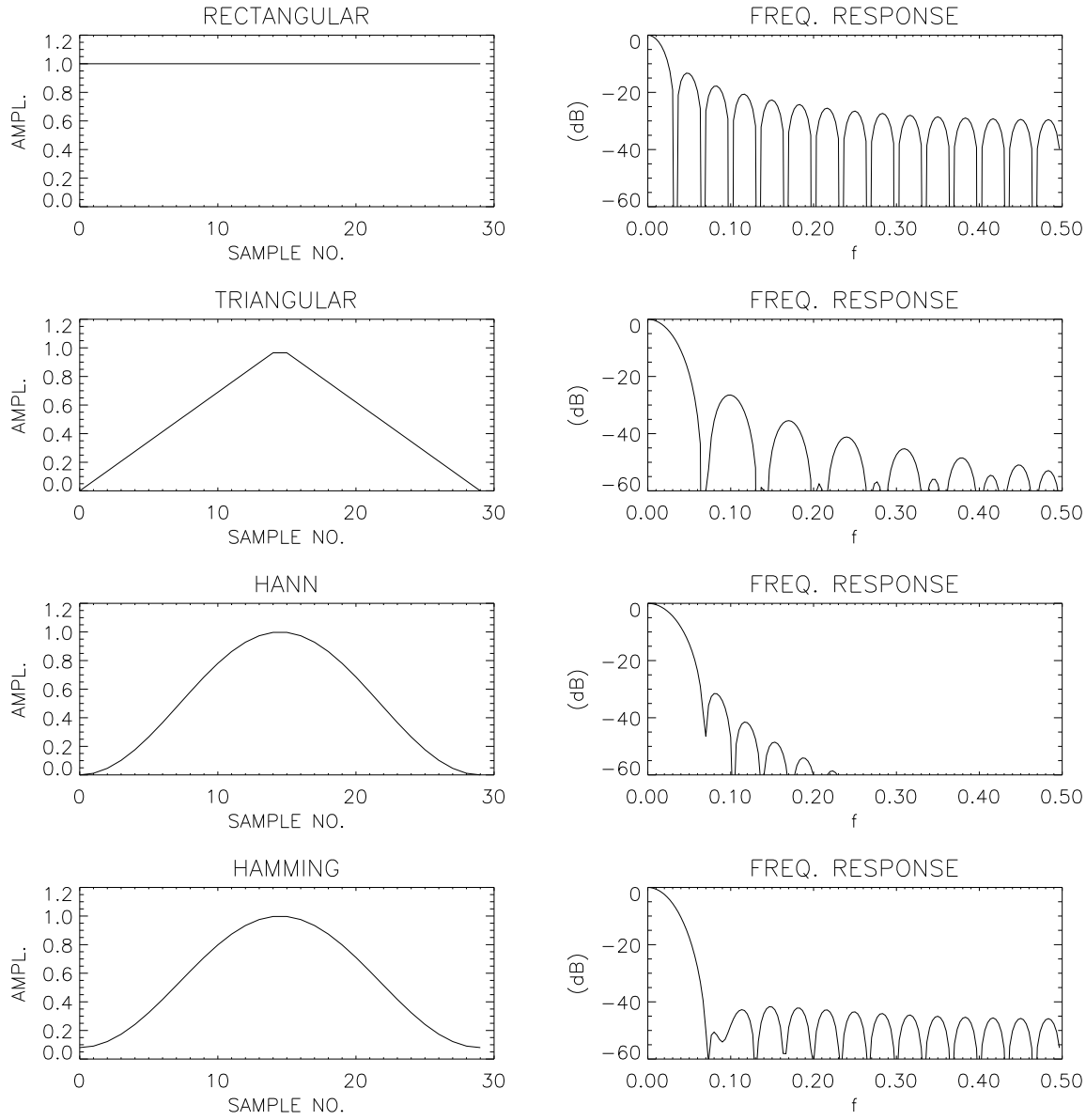


Figure 7.1: Some standard discrete time data windows (left) and their frequency responses (right) for $N = 30$. From top to bottom: rectangular, triangular, Hann and Hamming windows, respectively.

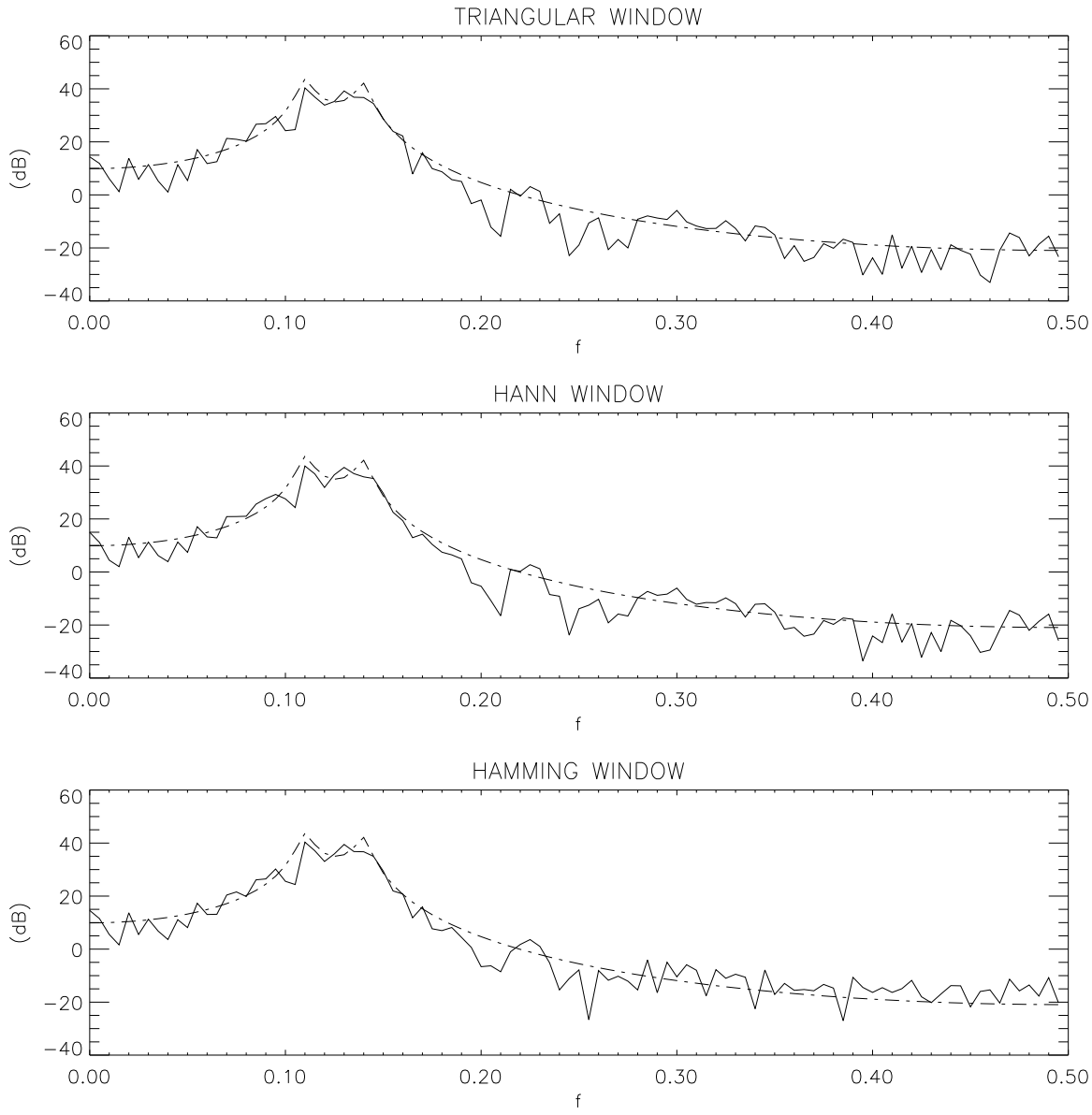


Figure 7.2: Windowed periodogram estimates using the AR(4)-data from Fig. 1. Upper panel: with triangular window, Central panel: with Hann window, and Lower panel: with Hamming window.

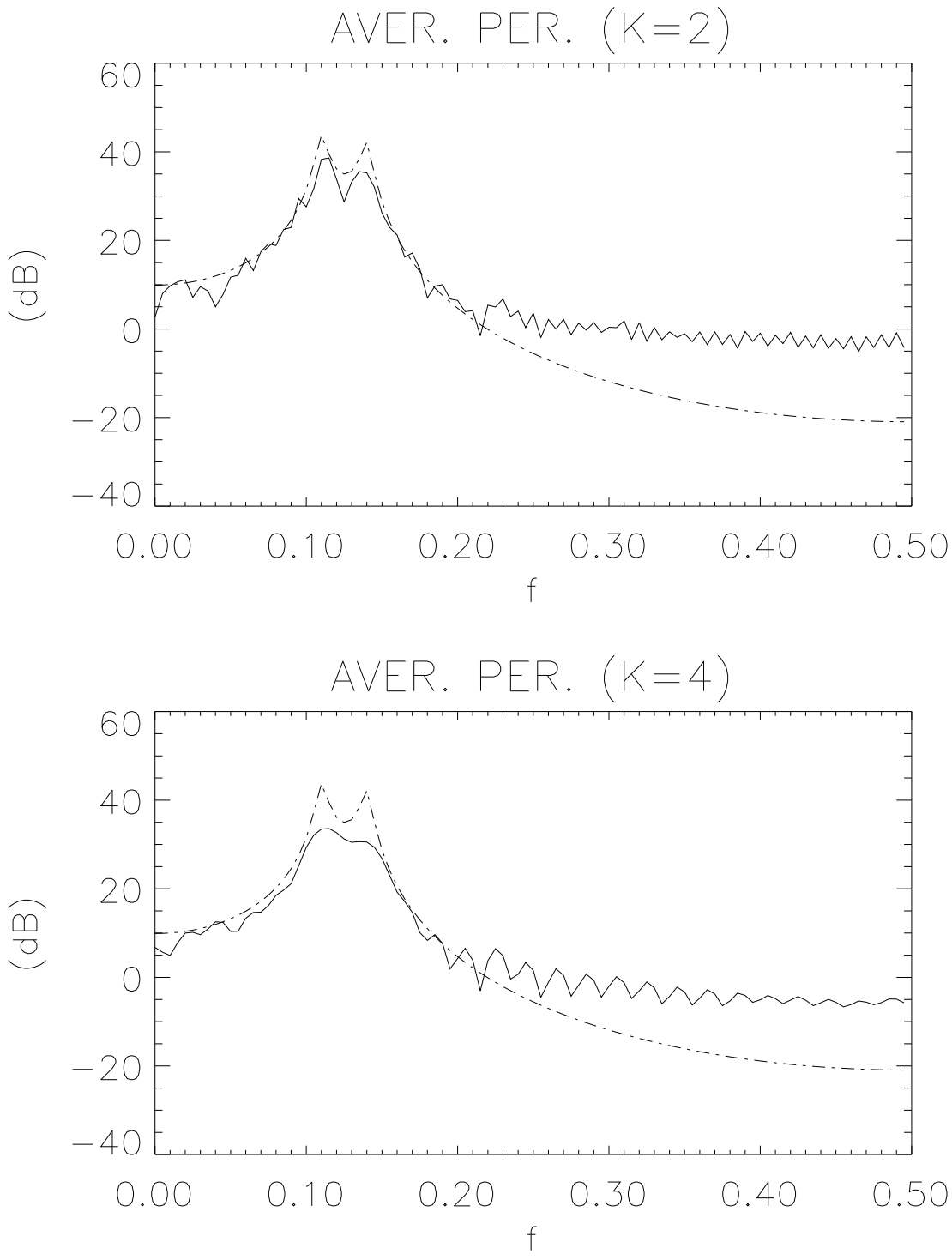


Figure 7.3: Averaged periodogram estimates of the AR(4) data from Fig. 1. Upper panel: $K = 2$ data segments each having $M = 100$ samples. Lower panel: $K = 4$ data segments with $M = 50$ samples in each. Full line: estimates, dash-dotted line: true PSD.

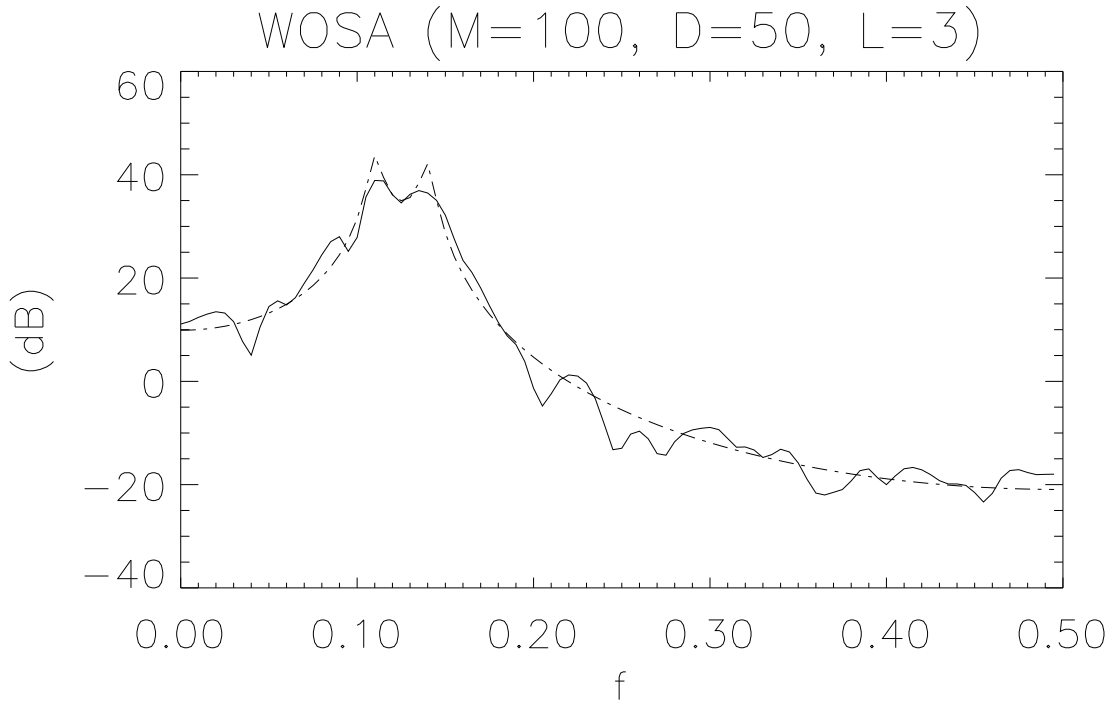


Figure 7.4: WOSA spectral estimate using 50% overlap. Same data as before.

Chapter 8

MULTITAPER SPECTRAL ESTIMATION

We will now briefly review the theory of a recent non-parametric spectral estimation method referred to as the multitaper (MT) technique proposed by *Thomson* [1982], following ideas from *Slepian* [1978]. This method combines the use of optimal data tapers, with averaging over a set of power spectral estimates. We also describe a recent important and very practical simplification of the MT method, proposed by *Riedel and Sidorenko* [1995].

8.1 Discrete Prolate Spheroidal Sequences

Thomson [1982] proposed to apply some stringent optimality criteria when selecting data tapers. He suggested to consider tapers that maximizes the "spectral concentration", or the energy contained in the mainlobe relative to the total energy of the taper. One therefore seeks the taper $v[n]$ with a discrete Fourier transform $V(f)$, that maximizes the window energy ratio

$$\lambda = \frac{\int_{-f_B}^{f_B} |V(f)|^2 \, df}{\int_{-1/2}^{1/2} |V(f)|^2 \, df} \tag{8.1}$$

where f_B is the wanted resolution half-bandwidth (a design parameter) of the taper. An ideal taper would therefore have $\lambda \simeq 1$ and f_B as small as possible (but note that $f_B > 1/N$). (Note also that we use $\Delta t = 1$ in this chapter to simplify the notation.)

Expressing $V(f)$ by its discrete Fourier transform, $V(f) = \sum_{n=0}^{N-1} v[n] \exp(-j2\pi fn)$ and maximizing the above functional with respect to $v[n]$, *Slepian* [1978] showed that the optimal taper $\mathbf{v} = [v[0], v[1], \ldots, v[N-1]]^T$ obeys the eigenvalue equation

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v} \tag{8.2}$$

where the matrix $\mathbf{A}$ has elements $[\mathbf{A}]_{nm} = \sin[2\pi f_B(n-m)]/[\pi(n-m)]$, for $n, m = 0, 1, \ldots, N-1$. Note that (8.2) is an N-dimensional eigenvector/eigenvalue problem, thus giving N eigenvector/eigenvalue pairs, $(\mathbf{v}_k, \lambda_k)$, where $k = 0, 1, \ldots, N-1$. The interpretation is thus that we obtain a *sequence* of orthogonal tapers (eigenvectors), $\mathbf{v}_k$, each with a corresponding spectral concentration measure λ_k. The first taper $\mathbf{v}_0$ has a spectral concentration

λ_0. Then, $\mathbf{v}_1$ maximizes the ratio in (8.1) subject to being orthogonal to $\mathbf{v}_0$, and with $\lambda_1 < \lambda_0$. Continuing, we can thus form up to N orthogonal tapers $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{N-1}$, with $0 < \lambda_{N-1} < \lambda_{N-2} < \cdots < \lambda_0 < 1$. Only tapers with $\lambda_k \simeq 1$ can be applied, since $\lambda_k \ll 1$ implies a large undesirable leakage.

It is usually safe to apply tapers up to order $k = 2Nf_B - 1$ [*Percival and Walden*, 1993, pp. 334–335]. It is customary to standardize the tapers such that they are orthonormal, $\mathbf{v}_k^T \mathbf{v}_{k'} = \delta_{k,k'}$, where $\delta_{k,k'}$ is the Kronecker delta, implying that $\mathcal{U} = 1/N$. The solutions $\mathbf{v}_k$ are referred to as "Discrete Prolate Spheroidal Sequences" (DPSS) [*Slepian*, 1978]. These optimal tapers are not expressible in closed form. The eigenvalue equation (8.2) must therefore be regarded as the defining equation for these tapers. It is not trivial to solve for the eigenvectors, because of inherent numerical instabilities in the problem, see e.g., [*Percival and Walden*, 1993]. When a resolution bandwidth f_B and data length N is fixed, one obtains a sequence of orthogonal tapers $\mathbf{v}_k$ that may be applied in forming a so-called "multitaper" (MT) spectral estimate. The simplest definition of an MT estimate is simply the arithmetic average of K tapered "eigenspectra"

$$\widehat{S}_{MT}(f) = \frac{1}{K} \sum_{k=0}^{K-1} \widehat{S}_{MT}^{(k)}(f) \tag{8.3}$$

where the "eigenspectrum" of order k is defined by

$$\widehat{S}_{MT}^{(k)}(f) = \left| \sum_{n=0}^{N-1} v_k[n] x[n] \exp(-j2\pi f n) \right|^2 \quad ; \quad |f| \leq 1/2 \tag{8.4}$$

where $v_k[n]$ denotes the elements of DPSS-taper of order k. Also data adaptive averaging schemes exist, see *Thomson* [1982] and *Riedel et al.* [1994]. The adaptive averaging is necessary in several applications.

The averaging of tapered spectral estimates, Eq. (8.3), leads to a decrease of the variance relative to any individual spectral estimates. An exact expression for the estimation variance is given by *Walden et al.* [1994], for stationary real-valued zero-mean Gaussian time series. Their expressions are rather complicated, but suffice it here to say that asymptotically,

$$\mathrm{var}\{\widehat{S}_{MT}(f)\} \simeq (1/K) S^2(f), \tag{8.5}$$

which is identical to the simple approximation derived by *Thomson* [1982]. One also gain control over the bias (or sidelobe leakage), since now

$$E\{\widehat{S}_{MT}(f)\} = \int_{-1/2}^{1/2} \overline{Q}(f - f') S(f') df', \tag{8.6}$$

where the so-called *total spectral window* is given by

$$\overline{Q}(f) = (1/K) \sum_{k=0}^{K-1} Q_k(f), \tag{8.7}$$

and the spectral window of order k is given by

$$Q_k(f) = \left| \sum_{n=0}^{N-1} v_k[n] \exp(-j2\pi f n) \right|^2 . \tag{8.8}$$

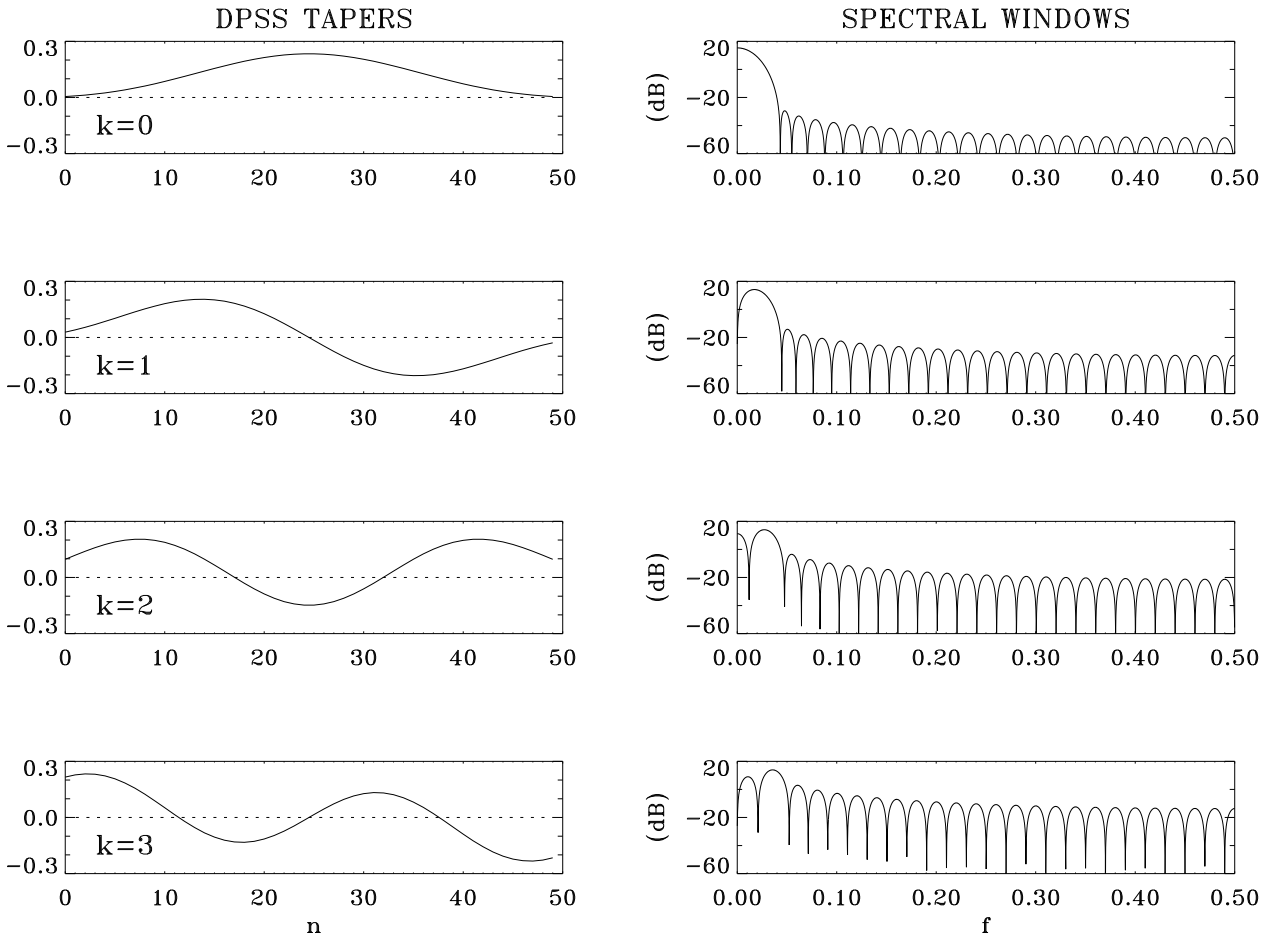


Figure 8.1: Left column: the four lowest order DPSS tapers for $N = 50$. Right column: corresponding energy spectra. From top to bottom: $k = 0, 1, 2, 3$. Design bandwith: $f_B = 1/25$

In Fig. 8, we show the first four tapers for the case $N = 50$, and $f_B = 2/N$, along with a plot of the corresponding spectral window $Q_k(f)$ in each case. Note that only the zeroth order taper has a shape similar to the classical bell shaped tapers, whereas the higher order tapers contain nodes and portions where they are negative. Further, notice that as k increases, the mainlobe-to-sidelobe ratio decreases, thus the higher order windows are more prone to leakage than the lower order ones.

Bronez [1992] compared the leakage, variance, and frequency resolution for the DPSS MT method with that of a weighted overlapped segment averaging (WOSA). He found that the MT method always performed better than the WOSA for each of the measures, when the other two measures were required to be equal for both estimation methods.

8.2 Sinusoidal tapers

Recently, a much simpler set of orthogonal tapers was proposed by *Riedel and Sidorenko* [1995]. Their tapers are approximations to a set of orthogonal tapers that minimize the local bias of the spectral estimator [*Papoulis*, 1972], $B(f) = \int_{-1/2}^{1/2} f^2 |V(f)|^2 df$. These simplified tapers are called "sinusoidal tapers", and they are expressible in closed form as [*Riedel and Sidorenko*, 1995]

$$v_k[n] = \left(\frac{2}{N+1}\right)^{1/2} \sin\left[\frac{\pi(k+1)(n+1)}{N+1}\right] \quad ; \quad k, n = 0, 1, \ldots, N-1 \tag{8.9}$$

In Fig. 9, we show the four lowest order sinusoidal tapers for $N = 50$, and their energy spectra. Riedel and Sidorenko's sinusoidal tapers will in many practical situations perform similarly to the much more complicated DPSS-tapers, as will be demonstrated by some of the results later in this paper. Note that there are no eigenvalues connected to the sinusoidal tapers, nor do we have the spectral bandwidth parameter f_B at our disposal. One can however show [*Riedel and Sidorenko*, 1995], that the energy of the sinusoidal taper of order k is mainly concentrated in the frequency range $k/[2(N+1)] \leq |f| \leq (k+2)/[2(N+1)]$. Thus, when applying a total of K sinusoidal tapers to form a multitaper spectral estimate, the resulting main lobe of the total spectral window has its energy concentrated in the range $|f| \leq (K+1)/[2(N+1)]$.

8.3 Extensions of the Multitaper Technique

The multitaper approach has been extended to signals in any number of dimensions and for any multitaper basis by *Hanssen* [1997]. By additional constraints on the shape of spectral peaks and a penalty function suppressing sidelobes, *Hansson and Salomonsson* [1997] developed a multitaper method with improved frequency resolution. A class of multitaper estimators for bispectra has been advocated and compared to classical bispectral estimators by *Birkelund and Hanssen* [1999].

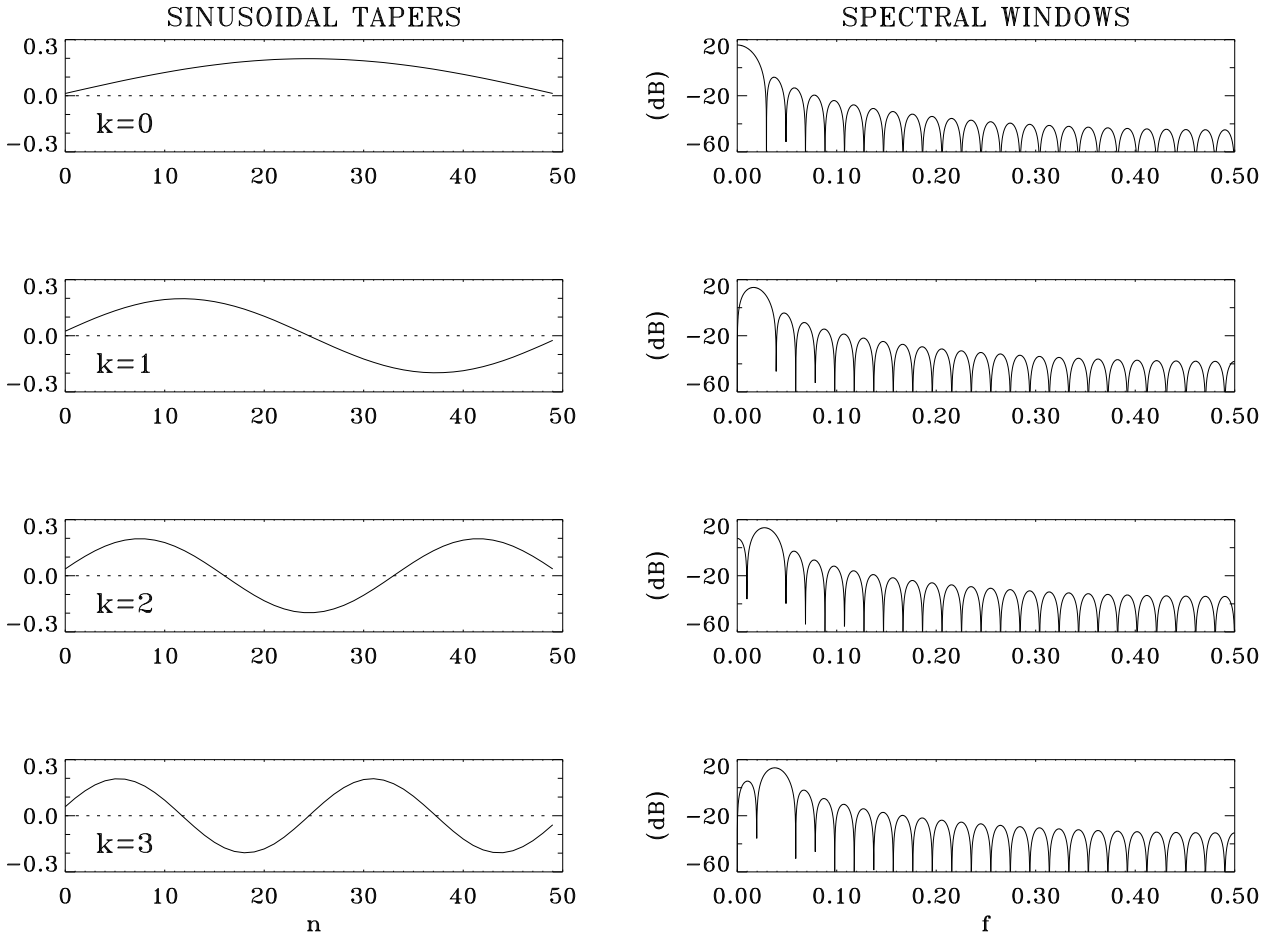


Figure 8.2: Left column: the four lowest order sinusoidal tapers for $N = 50$. Right column: corresponding energy spectra. From top to bottom: $k = 0, 1, 2, 3$.

Chapter 9

PARAMETRIC SPECTRAL ESTIMATION

The spectral estimation techniques we have examined so far, are all independent of any signal models. An important class of spectral estimators can be formed by assuming that the process has a specific underlying structure [*Priestley*, 1981; *Percival and Walden*, 1993]

9.1 Power Spectra of ARMA Processes

It is a well-known result [e.g., *Peebles*, 1993] that if we convolve a WSS stochastic process $X(t)$ with a known deterministic function $h(t)$ to obtain the output process

$$Y(t) = \int_{-\infty}^{\infty} h(\xi) X(t-\xi) d\xi, \tag{9.1}$$

then the process $Y(t)$ is WSS and it has the power spectral density

$$S_{YY}(f) = |H(f)|^2 S_{XX}(f). \tag{9.2}$$

Here, $H(f) = \mathcal{F}\{h(t)\}$ is the Fourier transform of the function $h(t)$.

The ARMA(p,q) process in Eq. (2.28) is in the form of two discrete convolutions. To apply the result from Eq. (9.2) to the ARMA-case, all we need to do is to replace the Fourier transforms by discrete Fourier transforms. By applying Eq. (9.2) separately on the left and the right hand side of Eq. (2.28), we obtain

$$|A(f)|^2 S_{XX}(f) = |B(f)|^2 S_{\varepsilon\varepsilon}(f) \quad ; \quad |f| \leq 1/(2\Delta t), \tag{9.3}$$

where

$$A(f) = \Delta t \sum_{k=0}^{p} a_k \exp(-j2\pi f k \Delta t) \tag{9.4}$$

$$B(f) = \Delta t \sum_{m=0}^{q} b_m \exp(-j2\pi f m \Delta t). \tag{9.5}$$

The driving noise $\varepsilon[n]$ is white with total average power σ^2, so the power spectral density $S_{\varepsilon\varepsilon}(f)$ in Eq. (9.3) is given by

$$S_{\varepsilon\varepsilon}(f) = \sigma^2 \Delta t \quad ; \quad |f| \leq 1/(2\Delta t). \tag{9.6}$$

Solving for $S_{XX}(f)$ from Eq. (9.3) is now easy, and we obtain the PSD of the ARMA-process as

$$\begin{aligned} S_{XX}(f) &= \sigma^2 \Delta t \frac{|B(f)|^2}{|A(f)|^2} \\ &= \sigma^2 \Delta t \frac{\left|1 + \sum_{m=1}^{q} b_m \exp(-j2\pi f m \Delta t)\right|^2}{\left|1 + \sum_{k=1}^{p} a_k \exp(-j2\pi f k \Delta t)\right|^2}. \end{aligned} \tag{9.7}$$

By examining the expression for $S_{XX}(f)$, we understand that it is possible to form power spectra of complicated shapes by assuming ARMA-processes. We notice that if a frequency f_0 exists such that $|A(f_0)|^2 \simeq 0$, then $S_{XX}(f)$ will have a pronounced maximum at $f = f_0$. Likewise, if a frequency f_1 exists such that $|B(f_1)|^2 \simeq 0$, then $S_{XX}(f)$ will have a pronounced minimum at $f = f_1$. We thus see that one can produce very complicated power spectra by adjusting the parameters in the model.

9.2 ARMA Spectral Estimation

Assume now that the signal under study has a structure that can be described by a difference equation of the form Eq. (2.28), and that we know the correct orders p and q. Then one can perform an ARMA spectral estimation by estimating the parameters $(a_1, .., a_p, b_1, .., b_q, \sigma^2)$, and subsequently substitute the estimates $(\widehat{a}_1, .., \widehat{a}_p, \widehat{b}_1, .., \widehat{b}_q, \widehat{\sigma^2})$ into Eq. (9.7) to yield

$$\widehat{S}_{XX}^{(ARMA)}(f) = \Delta t \widehat{\sigma^2} \frac{\left|1 + \sum_{m=1}^{q} \widehat{b}_m \exp(-j2\pi f m \Delta t)\right|^2}{\left|1 + \sum_{k=1}^{p} \widehat{a}_k \exp(-j2\pi f k \Delta t)\right|^2}. \tag{9.8}$$

Several methods exist to estimate the $p + q + 1$ parameters from a given data set. In practice, it is often a very difficult task to estimate the MA-parameters $b_1, .., b_q$. However, in many practical situations it is sufficient to consider the AR-part of the spectrum.

9.3 AR Spectral Estimation

If the underlying process has a structure that can be described with sufficient accuracy using only the AR-parameters, the problem simplifies significantly. In this case, the resulting spectral estimator is called an AR power spectral estimator, and it has the form [*Priestley*, 1981]

$$\widehat{S}_{XX}^{(AR)}(f) = \frac{\widehat{\sigma^2} \Delta t}{\left|1 + \sum_{k=1}^{p} \widehat{a}_k \exp(-j2\pi f k \Delta t)\right|^2}. \tag{9.9}$$

In this case, we thus need to estimate the $p+1$ parameters $(\widehat{a}_1, .., \widehat{a}_p, \widehat{\sigma^2})$.

The next important question that arises is: how do we estimate the AR-parameters $a_1, .., a_p$ and the driving noise variance σ^2? It seems obvious that the more accurately can we estimate the parameters, the more accurate spectral estimates we can expect. It is also important to notice that we do not know *a priori* whether a signal can be modeled as an AR process or not. Furthermore, the model order p is in general unknown, and it must be estimated from the data by means of some criterion (e.g., Akaike's information criterion, or a minimum average prediction error criterion, see *Percival and Walden* [1993], *Kay* [1988] or *Marple* [1987] for details).

The Yule-Walker equations

Several advanced parameter estimation techniques exist that can be applied to the problem of estimating the AR-parameters. The advanced techniques will be covered in later courses (e.g., AFys-361), so in AFys-261 we will only discuss the simplest of the estimation techniques.

We will start by showing how one may obtain the *exact* values of $a_1, .., a_p, \sigma^2$ if the *true* ACF values are known for lags $0, 1, .., p$. The AR(p) process under study can be written as

$$x[n] + a_1 x[n-1] + \cdots + a_p x[n-p] = \varepsilon[n]. \tag{9.10}$$

Multiply Eq. (9.10) by $x[n-k]$, and take the expectation value of the resulting equation to yield

$$E\left\{x[n]x[n-k] + a_1 x[n-1]x[n-k] + \ldots + a_p x[n-p]x[n-k]\right\} = E\left\{\varepsilon[n]x[n-k]\right\}. \tag{9.11}$$

From the left hand side of Eq. (9.11) we get the sum of ACF lags

$$R_{XX}[-k] + a_1 R_{XX}[-k+1] + \ldots + a_p R_{XX}[-k+p]. \tag{9.12}$$

From the right hand side of Eq. (9.11) we find by substituting the AR-process for $x[n-k]$ the relation

$$\begin{aligned} E\{\varepsilon[n]x[n-k]\} &= E\{\varepsilon[n]\left(-a_1 x[n-k-1] - \cdots - a_p x[n-k-p] + \varepsilon[n-k]\right)\} \\ &= E\{\varepsilon[n]\varepsilon[n-k]\} = \sigma^2 \delta_{k,0}, \end{aligned} \tag{9.13}$$

since $\varepsilon[n]$ is a white noise process. The result is the following equation for the ACF lags

$$R_{XX}[-k] + a_1 R_{XX}[-k+1] + \cdots + a_p R_{XX}[-k+p] = \sigma^2 \delta_{k,0}. \tag{9.14}$$

Since we have $p+1$ unknown parameters, we need to form $p+1$ equations to be able to solve for the parameters. If we evaluate Eq. (9.14) for lags $k = 1, 2, .., p$, we obtain the p equations usually known as the *Yule-Walker equations* or the *Wiener-Hopf equations*

$$\begin{array}{ccccccccc} R_{XX}[-1] & + & a_1 R_{XX}[0] & + & \cdots & + & a_p R_{XX}[p-1] & = & 0 \\ R_{XX}[-2] & + & a_1 R_{XX}[-1] & + & \cdots & + & a_p R_{XX}[p-2] & = & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ R_{XX}[-p] & + & a_1 R_{XX}[-p+1] & + & \cdots & + & a_p R_{XX}[0] & = & 0. \end{array} \tag{9.15}$$

For real valued processes the ACF is symmetric around $k = 0$, i.e., $R_{XX}[-m] = R_{XX}[m]$. Using this fact, and rewriting Eq. (9.15) by means of matrix/vector notation, we see that the Yule-Walker equations can be written in the matrix form

$$\mathbf{RA} = -\mathbf{P}, \tag{9.16}$$

where the correlation matrix $\mathbf{R}$ is defined by

$$\mathbf{R} = \begin{bmatrix} R_{XX}[0] & R_{XX}[1] & \cdots & R_{XX}[p-1] \\ R_{XX}[1] & R_{XX}[0] & \cdots & R_{XX}[p-2] \\ \vdots & \vdots & \vdots & \vdots \\ R_{XX}[p-1] & R_{XX}[p-2] & \cdots & R_{XX}[0] \end{bmatrix} \tag{9.17}$$

and the column vectors $\mathbf{A}$ and $\mathbf{P}$ are defined by

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} R_{XX}[1] \\ R_{XX}[2] \\ \vdots \\ R_{XX}[p]] \end{bmatrix}.$$

It is now straightforward to find the AR-parameters by operating on Eq. (9.16) with the inverse of the correlation matrix, so we obtain

$$\mathbf{A} = -\mathbf{R}^{-1}\mathbf{P}. \tag{9.18}$$

To obtain the solution for the variance of the driving noise, we now evaluate the $k = 0$ term of Eq. (9.14), which gives

$$\begin{aligned} \sigma^2 &= R_{XX}[0] + a_1 R_{XX}[1] + \cdots + a_p R_{XX}[p] \\ &= R_{XX}[0] + \mathbf{P}^T\mathbf{A} \\ &= R_{XX}[0] - \mathbf{P}^T\mathbf{R}^{-1}\mathbf{P}. \end{aligned} \tag{9.19}$$

In a practical situation, one would not have the true ACF values available. Instead, one would have to replace $R_{XX}[k]$ by some ACF estimates $\widehat{R}_{XX}[k]$. Notice that we only need to estimate the $p+1$ lowest ACF-lags in order to estimate the parameters describing an AR(p) process.

The AR-parameters are thus estimated as

$$\widehat{\mathbf{A}} = -\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{P}}. \tag{9.20}$$

and

$$\widehat{\sigma^2} = \widehat{R}_{XX}[0] + \widehat{\mathbf{P}}^T\widehat{\mathbf{A}}, \tag{9.21}$$

where ACF estimates are used to form all the necessary matrices and vectors.

Since this is a parameter estimation problem, it is very important that we always keep $p \ll N$, where N is the number of available samples. This will ensure that the AR parameters can be estimated with some confidence.

If the correct order of the AR-process is not known, then also p must be estimated from the available data. This is a topic which is far beyond the scope of these lectures, and in general this is one of the non-trivial tasks of a time series analyst. Note that one needs 2 AR-coefficients to represent one real valued sinusoidal component. Thus, if we know that the data set consists of N_p real periodic components, then it is a rule of the thumb to choose the order to be $p \geq 2N_p$.

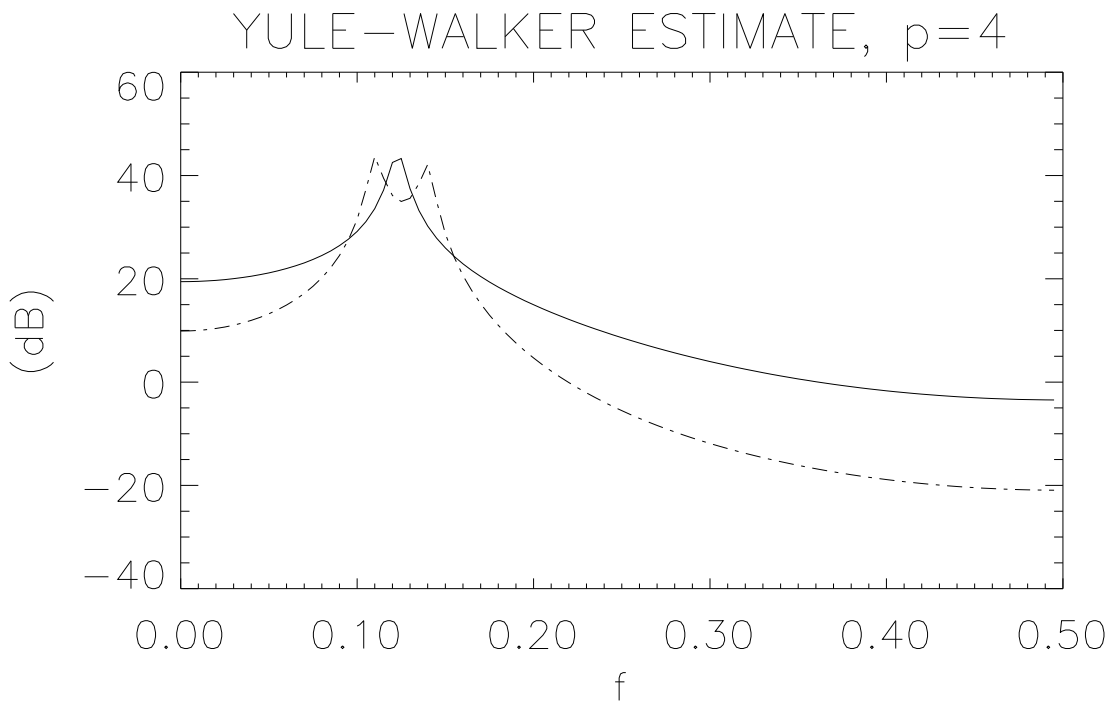


Figure 9.1: Yule-Walker spectral estimate of the AR(4) data from Fig. 1. We used $p = 4$, and the biased ACF-estimator. Full line: spectral estimate, and dash-dotted line: true PSD.

Chapter 10

PREWHITENING IN SPECTRAL ESTIMATION

As we have seen, any classical spectral estimator will suffer from spectral leakage out of spectral maxima. This is the major cause of bias in the spectral estimates. In general, the spectrum analysis becomes easier and more accurate the flatter the true spectrum is, white noise being the ideal case. If the signal is white, then it is in fact not necessary to use any data window in the spectral estimation step.

A very important improvement of classical spectral estimation techniques involves a preprocessing step that aims at reducing the *dynamic range* $R = \max\{S_{XX}(f)\} - \min\{S_{XX}(f)\}$ of the signal prior to the spectral analysis. If such a reduction is achieved, one may expect any classical spectral estimator to perform better on the preprocessed signal than on the original signal. The spectral estimate of the preprocessed signal must thereafter be corrected to take into account the part of the spectrum that we "removed" during the preprocessing. This methodology is called "prewhitening", and it is an extremely important and powerful technique for practical applications.

The standard procedure to whiten the process $x[0], \ldots, x[N-1]$ is to fit an AR-model to the signal, and to consider the difference between the signal value $x[n]$ and the linear p-step predictor $y[n] = -a_1 x[n-1] - \ldots - a_p x[n-p]$. This difference is often called the *innovation sequence* of the process, and in our case it has the form

$$r[n] = x[n] - y[n] = x[n] + a_1 x[n-1] + \cdots + a_p x[n-p]. \tag{10.1}$$

If the signal is truly AR(p), then the innovations $r[n]$ will be a white process. In general, however, $r[n]$ is not white. Hopefully, we have been able to catch some of the dynamics by the linearly predicted sequence $y[n]$, so that $r[n]$ has a smaller dynamic range than that of the original data.

Let $H(f) = 1 + \sum_{k=1}^{p} a_k \exp(-j2\pi f k \Delta t)$ denote the transfer function of the linear whitening filter. We then know that if $S_{XX}(f)$ is the PSD of the original process, then the residual process $r[n]$ has a PSD given by

$$S_{rr}(f) = |H(f)|^2 \, S_{XX}(f). \tag{10.2}$$

The method then goes as follows: (1) Fit an AR(p) model to the data $\widehat{a}_1, .., \widehat{a}_p, \widehat{\sigma^2}$. (2) Form the innovation sequence $r[n] = x[n] - y[n]$. (3) Estimate the PSD of $r[n]$ by a classical

technique of your choice. (4) Form the final PSD estimate by "postcoloring" the spectral estimate, i.e.,

$$\widehat{S}_{XX}(f) = \frac{\widehat{S}_{rr}(f)}{\left|1 + \sum_{k=1}^{p} \widehat{a}_k \exp(-j2\pi f k \Delta t)\right|^2}. \tag{10.3}$$

By examining the mathematical form of this method, it is obvious that the resulting spectral estimate will depend on a large number of different factors: (i) how well the AR(p) model captures the overall spectral shape, (ii) the model order p, (iii) the actual method we use to estimate the AR-parameters, and (iv) the estimation technique we use to estimate the spectrum of the innovations process.

The prewhitening approach in the form described here, is thus a hybrid between a classical non-parametric and a "modern" parametric method. In cases where we have enough a priori information to suggest a meaningful prewhitening filter in beforehand, one could hope to use this technique also in real-time applications. Up to now, however, this technique has mainly been applied in off-line situations.

Note that the frequency resolution degrades somewhat by using the prewhitening technique. This is because the innovation sequence $r[n]$ contains p fewer data points than the original data set N, due to the initialization of the filter. This means that the frequency resolution decreases by a factor $1 - p/N$ due to the shortening of the available data vector.

Chapter 11

ANALYSIS OF NON-STATIONARY PROCESSES

All spectral analysis methods we have discussed so far, require that the data is at least wide-sense stationary. In practical terms, this means that if we perform a spectral estimation based on two different segments of the data set, the overall shape of the two estimates should be similar. In this case, a simple time averaging procedure was a practical way of simulating the expectation operator entering all basic definitions.

Many (maybe most) real-world stochastic process are however non-stationary. This means that e.g., the total average power is a function of time, and the power spectral density may vary as time goes. The standard Fourier-techniques do not have the capability of resolving such a time variability. This is obvious when recalling the basic definition of the Fourier transform, $X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)dt$, which effectively "integrates away" the time history of the signal. We thus need to modify the standard spectral analysis techniques to also represent the time variability of the "frequency" content of the signal.

It is important to recall the importance of stationarity in data analysis: it gives the process some sort of "statistical stability" so that quantities originally defined as ensemble averages (e.g., mean values, ACFs, PSDs) can be estimated from a single realization by computing time domain averages. (Strictly speaking, the property we use in this case is ergodicity rather than stationarity, but in practical terms the two ideas are closely related.)

Non-stationarities arise in several ways when considering real-world processes. A typical case occuring in e.g. communication and radar detection, is that the observed signal is the sum of a deterministic function $s(t)$ and a (zero-mean) stochastic process $W(t)$,

$$X(t) = s(t) + W(t). \tag{11.1}$$

Among other things, this implies that we allow the mean $E[X(t)]$ to vary over time. This could include e.g., steady "growth" or "decay" of some quantity, or cyclic behaviour related to e.g. seasonal variations in the process. One way of handling such cases could be accomplished by: (1) a parametrization of the function $s(t)$, (2) estimate its parameters, (3) subtract the estimated $s(t)$, and (4) perform all subsequent analysis on the estimated residuals. Alternatively, if the trend is *linear*, i.e., $s(t) = At + B$, where A and B are (unknown) constants, then it is sufficient to perform a difference between subsequent data values in order to obtain a stationary process.

We recall that the physical interpretation of the power spectral density was that it gives

the power as a function of frequency. If we wish to preserve this interpretation also for non-stationary process, we must by some means allow the extended definition of the PSD to express "power density as a function of time".

Consider a simple non-stationary process of the form

$$X(t) = \begin{cases} X_1(t), & t \leq t_0 \\ X_2(t), & t > t_0 \end{cases} \tag{11.2}$$

where $X_1(t)$ and $X_2(t)$ are each stationary processes but with different correlation structure. Assuming that the "change point" t_0 is known, the natural way of describing the power vs. frequency properties of $X(t)$ is to introduce two spectra, one for $t \leq t_0$ and one for $t > t_0$. This is readily generalized to any partition of *local* processes valid in time intervals $t_i \leq t_{i+1}$. Going to the limit $t_{i+1} - t_i \to 0$, we are led to the notion of *continuously changing PSDs*, or *time-dependent spectra*. Note that it is unrealistic to hope for an estimation method that can give the PSD at a single point in time. However, if we assume that the process is changing reasonably smoothly over time, it may be possible to estimate some form of "average" spectrum of the process in the neighborhood of a particular time instant.

11.1 The Short-Time Fourier Transform and the Spectrogram

The so-called "Short-Time Fourier Transform" (STFT) was introduced in the electrical engineering context by Gabor in 1946. This method is the simplest of all techniques that provide a time-frequency representation of a signal. Let $w(s)$ be a (localized) window function. To obtain a local variant of the Fourier transform, we shift the window to time t, and calculate the Fourier transform of the product $x(s)w(s-t)$,

$$X_w(f,t) = \int_{-\infty}^{\infty} x(s)w(s-t)\exp(-j2\pi f s)ds. \tag{11.3}$$

We usually require that $w(s)$ is centered around $s = 0$, and that its Fourier transform $W(f)$ is centered around $f = 0$. In mathematical terms, this is equivalent to the requirements

$$\int_{-\infty}^{\infty} s|w(s)|^2 ds = 0, \qquad \int_{-\infty}^{\infty} f|W(f)|^2 df = 0. \tag{11.4}$$

The last requirement in Eq. (11.4) is of course always met when $w(s)$ is real valued.

To obtain a local analysis of the spectral properties of $x(t)$, we must require that the sliding window function $w(s)$ is "narrow" in some sense. At the same time, we would like the width of the Fourier transform $W(f)$ to be "narrow", to provide a reasonable resolution in frequency space. We now define the *effective time-duration* of the window function $w(t)$ by

$$\Delta t_e = \left[\frac{\int_{-\infty}^{\infty} t^2|w(t)|^2 dt}{\int_{-\infty}^{\infty} |w(t)|^2 dt} \right]^{1/2}, \tag{11.5}$$

and the *effective bandwidth* of $w(t)$ as

$$\Delta f_e = \left[\frac{\int_{-\infty}^{\infty} f^2 |W(f)|^2 df}{\int_{-\infty}^{\infty} |W(f)|^2 df} \right]^{1/2} . \tag{11.6}$$

Basically, two sinusoids can be discriminated only if their frequency difference is larger than Δf_e, and two pulses in time can be resolved only if they occur at time instants separated by more than Δt_e.

One cannot find a data window $w(s)$ that has an arbitrary simultaneous time- and frequency resolution. In fact, an *uncertainty* principle exists, meaning that the time-bandwidth product is limited from below,

$$\Delta t_e \Delta f_e \geq \frac{1}{2}. \tag{11.7}$$

Note the similarity with Heisenberg's uncertainty relation from quantum (wave) mechanics. *Gabor* showed that the smallest time-bandwidth product (i.e., the best simultaneous time- and frequency resolution) is obtained for a Gaussian window function. Despite its optimality, one often use other windows than the Gaussian in real applications. When applying a Gaussian window in the STFT, the transform is often referred to as a *Gabor transform*. It is customary to plot the squared modulus value of the STFT,

$$S_w(f,t) = |X_w(f,t)|^2 \tag{11.8}$$

when discussing the time-frequency behaviour of the process. This second order real valued quantity is often called the *spectrogram* of the signal. We understand that the spectrogram $S_w(f,t)$ is a measure of the *energy spectrum* in a time window centered at time t rather than a power spectrum.

If the energy of the analyzing window $w(t)$ is normalized,

$$\int_{-\infty}^{\infty} |w(t)|^2 dt = 1, \tag{11.9}$$

then it is easy to show that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S_w(f,t) df dt = \int_{-\infty}^{\infty} |x(t)|^2 dt. \tag{11.10}$$

Thus, the spectrogram preserves the signal energy.

11.1.1 Inversion of the Spectrogram

It is a widespread misunderstanding (e.g., *Haykin and van Veen* [1999]) that the spectrogram cannot be inverted to yield the signal $x(t)$ since the spectrogram is real valued and hence contains no phase information. This argument is in fact patently wrong. It is true that we have lost the phase information of the STFT by the modulus operation, but the phase of the *signal* $x(t)$ is not lost. This is because we have transformed a one-dimensional signal $x(t)$ to a two-dimensional domain. Thus, there is a high degree of redundancy in the spectrogram,

which unambigously allow us to reconstruct the time history of the underlying signal. Note however that the second order properties of the analyzing window $w(t)$ enter this discussion explicitly, so we must remove the effect of the window when inverting $S_w(f,t) = |X_w(f,t)|^2$. This puts restrictions on the class of time windows which allows for a possible inversion of the spectrogram.

To recover $x(t)$ from the spectrogram we proceed as follows. First, consider the two-dimensional Fourier transform of the spectrogram in time and frequency

$$C(\tau,\nu) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_w(f,t)\exp\left[j2\pi(\nu t + f\tau)\right] dtdf, \tag{11.11}$$

where τ and ν are the transform variables. Substituting the definition of the spectrogram eq. (11.8) into this integral, it is a simple task to rewrite $C(\tau,\nu)$ as

$$C(\tau,\nu) = A_x(\tau,\nu)A_w(\tau,-\nu), \tag{11.12}$$

where

$$A_x(\tau,\nu) = \int_{-\infty}^{\infty} x(t-\tau/2)x(t+\tau/2)\exp(j2\pi\nu t)dt \tag{11.13}$$

and

$$A_w(\tau,\nu) = \int_{-\infty}^{\infty} w(t-\tau/2)w(t+\tau/2)\exp(j2\pi\nu t)dt. \tag{11.14}$$

The functions $A_x(\tau,\nu)$ and $A_w(\tau,\nu)$ are called the *ambiguity functions* (e.g., Cohen [1995]) of $x(t)$ and $w(t)$, respectively.

From eq. (11.12) we thus see that the ambiguity function of $x(t)$ can be written as

$$A_x(\tau,\nu) = \frac{C(\tau,\nu)}{A_w(\tau,-\nu)}. \tag{11.15}$$

But from the definition of $A_x(\tau,\nu)$ we understand that the Fourier transform of eq. (11.13) yields

$$x(t-\tau/2)x(t+\tau/2) = \int_{-\infty}^{\infty} A_x(\tau,\nu)\exp(-j2\pi\nu t)d\nu. \tag{11.16}$$

If we choose the particular time $t = \tau/2$, we see that

$$x(\tau) = \frac{1}{x(0)}\int_{-\infty}^{\infty} A_x(\tau,\nu)\exp(-j\pi\nu\tau)d\nu. \tag{11.17}$$

We have thus proven that $x(t)$ is completely recoverable from the spectrogram $C(\tau,\nu)$ through the operation

$$x(t) = \gamma\int_{-\infty}^{\infty}\frac{C(t,\nu)}{A_w(t,-\nu)}\exp(-j\pi\nu t)d\nu, \tag{11.18}$$

where $\gamma = 1/x(0) \neq 0$ is a constant, and we must require that the ambiguity function of the data window $A_w(t,-\nu)$ is nonzero for all locations in the (t,ν)-plane. This will of course depend on the particular data window that is applied in the analysis.

A practical example which leads to a valid window ambiguity function is given by the Gaussian window

$$w(t) = \exp\left(-t^2/2\sigma^2\right), \tag{11.19}$$

where σ^2 is a width parameter of the Gaussian. The ambiguity function is readily found to be

$$A_w(\tau,\nu) = \sqrt{\pi\sigma^2}\exp\left(-\tau^2/4\sigma^2 - \pi^2\nu^2\sigma^2\right). \tag{11.20}$$

This function is nowhere zero, and it facilitates a particularly simple inversion integral for the spectrogram,

$$x(t) = \frac{\gamma}{\sqrt{\pi\sigma^2}}\int_{-\infty}^{\infty} C(t,\nu)\exp\left(\pi^2\nu^2\sigma^2 + t^2/4\sigma^2\right)\exp\left(-j\pi\nu t\right)d\nu. \tag{11.21}$$

At first sight, one may think that the integral in eq. (11.21) diverges asymptotically as $\nu \to \infty$ and $t \to \infty$ because of the positive exponential dependency upon these parameters. This is however wrong, since these diverging terms exactly cancel similar terms present in the spectrogram $C(t,\nu)$ introduced by the data window.

11.1.2 Simultaneous Time and Frequency Resolution

From the definition of the STFT and the spectrogram, we see that Δt_e and Δf_e are fixed when a specific window has been chosen. Thus, the time and frequency resolutions are the same same regardless of the location in the (f,t)-plane. In particular, this means that the same resolution applies when studying both low- and high-frequency phenomena, and when studying both short-time bursts, and slowly varying (quasi)-periodic signals. This lack of flexibility is a major restriction when performing a spectrogram analysis of many real-world signals.

11.2 Continuous Wavelet Transform

As noted in the previous section, the time-frequency resolution is fixed in the entire time-frequency plane for the STFT. The *wavelet transform* (WT) is a fairly recent technique that aims at removing this limitation [*Daubechies*, 1992; *Nallat*, 1998]. We wavelet transform is a method that varies the size of the analyzing window acording to which frequency (or *scale*) that is under study. When studying low frequencies (large scales) the window is wide, while when studying high frequencies (small scales), the window is narrow. Thus, this is a technique that attempts at obtaining a good compromise between time and frequency resolution simultaneously. The basic idea is that we use one single data window that is scaled according to the frequency we want to study. The time resolution is obtained by translating this common analyzing function, or *mother wavelet* as it is usually called. Note that the uncertainty principle Eq. (11.7) will still apply locally.

Now, let the effective time-duration $\Delta t_e \sim 1/\Delta f_e$ become better (i.e., become smaller) as the central frequency f increases. The simplest choice we can make is that $f\Delta t_e =$ constant, or as it is usually stated,

$$\frac{\Delta f_e}{f} = c, \tag{11.22}$$

where c is a constant. If we interpret the frequency analyzer as a filter, we see that the bandwidth of the filter increases as we try to analyze higher frequencies. This is what is often referred to as *constant-Q* filtering.

The *Continuous Wavelet Transform* (CWT) of the signal $x(t)$ is defined by the two-variable transformation [*Mallat*, 1998]

$$X_\psi(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{a}\right) dt. \tag{11.23}$$

The function $\psi(\cdot)$ is called the *mother wavelet*, and we have great freedom in choosing this function. Since the argument of the ψ-function in (11.23) must be dimensionless, we see that both transform variables a and b have the dimension of *time*. From the structure of the basic definition, we also understand that b is a *translation variable*, and a is a *scale* or *compression variable*. In general, $-\infty < b < \infty$, and a is either $-\infty < a < \infty$ or $0 \leq a < \infty$.

By defining the *wavelet function* at scale a and time b as

$$\psi_{a,b}(t) \equiv \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), \tag{11.24}$$

we see that the CWT may be understood as an *inner product* of the signal $x(t)$ and the wavelet function, since

$$X_\psi(a,b) = \int_{-\infty}^{\infty} x(t)\psi^*_{a,b}(t)dt \equiv \langle x(t), \psi_{a,b}(t)\rangle. \tag{11.25}$$

But an inner product is just a measure of the similarity between two functions, so the wavelet coefficient at scale $X_\psi(a,b)$ is a measure of how similar the signal $x(t)$ is to the analyzing wavelet located around time b, at a scale proportional to a. We thus compare the signal locally to shifted and scaled versions of the mother wavelet $\psi(t)$.

An inverse wavelet transform exists, so that $x(t)$ can be reconstructed from the wavelet coefficients $X_\psi(a,b)$. The inverse transform has the form

$$x(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X_\psi(a,b)\psi_{a,b}(t)\frac{da\ db}{a^2}. \tag{11.26}$$

The constant C_ψ depends only on the mother wavelet, and is defined by

$$C_\psi \equiv 2\pi \int_{-\infty}^{\infty} \frac{|\Psi(f)|^2}{|f|} df, \tag{11.27}$$

where $\Psi(f) = \mathcal{F}\{\psi(t)\}$ is the Fourier transform of the mother wavelet. To ensure that the inverse WT exists, we must require that the so-called *admissibility condition* is satisfied,

$$C_\psi < \infty. \tag{11.28}$$

In practical terms, this means that $\Psi(f=0) = 0$, or that

$$\int_{-\infty}^{\infty} \psi(t)dt = 0. \tag{11.29}$$

If we express $x(t)$ and $\psi(t)$ by their respective Fourier transforms, an alternative and very useful form of the WT is obtained as

$$X_\psi(a,b) = \sqrt{|a|} \int_{-\infty}^{\infty} X(f)\Psi^*(af)e^{j2\pi bf}df. \tag{11.30}$$

Note that the CWT has the apparance of an orthogonal basis. In general, however, the basis functions $\psi_{a,b}(t)$ cannot be orthogonal, since both a and b are allowed to vary continuously. If a and b are discretized in a very specific manner, one may construct an orthonormal version of the WT. Such a representation is very important for applications where an effective and sparse representation of the data is essential. The prototypical application is data compression, or filtering in the wavelet domain. Note that only a very few specific wavelets yield an orthonormal basis. The first non-trivial orthonormal wavelets were invented in Tromsø in the early 1980's by Professor Jan Strømberg at the Mathematics Department. If one in addition require that the mother wavelets have compact support, then rather bizarre functions appear. The most famous of the compact and orthonormal wavelets are the so-called *Daubechies* wavelets, named after their inventor, Ingrid Daubechies.

The CWT preserves signal energy in the sense that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |X_\psi(a,b)|^2 \frac{da\ db}{a^2} = \int_{-\infty}^{\infty} |x(t)|^2 dt. \tag{11.31}$$

The squared modulus of the CWT is therefore a proper distribution of energy in the *time-scale* plane, and it is often referred to as a *scalogram*.

REFERENCES

Birkelund, Y. and Hanssen, A., Multitaper estimators for bispectra, *Proc. IEEE Workshop on Higher-Order Statistics*, Caesarea, Israel, 207–211, 1999.

Bronez, T. P., On the performance advantage of multitaper spectral analysis, *IEEE Trans. Signal Proc.*, **40**, 2941–482, 1992.

Cohen, L., *Time-Frequency Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

Daubechies, I., *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

Hanssen, A., Multidimensional multitaper spectral estimation, *Signal Processing*, **58**, 327–332, 1997.

Hansson, M., and Salomonsson, G., A multiple window method for estimation of peaked spectra, *IEEE Trans. Signal Processing*, **45**, 778–781, 1997.

Harris, F. K., On the use of windows for harmonic analysis with the discrete Fourier transform, *Proc. IEEE*, **66**, 51–83, 1978.

Haykin, S. and Van Veen, B., *Signals and Systems*, Wiley, New York, 1999.

Kay, S. M., *Modern Spectral Estimation: Theory & Application*, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

Larsen, R. J., and Marx, M. L., *An Introduction to Mathematical Statistics and its Applications*, (2nd ed.), Prentice-Hall, Englewood Cliffs, NJ, 1986.

Mallat, S., *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

Marple, S. L. Jr., *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

Nikias, C. L. and Petropulu, A. P., *Higher-order Spectra Analysis*, Prentice-Hall, 1993.

Papoulis, A., Minimum-bias windows for high-resolution spectral estimates, *IEEE Trans. Inf. Theory*, **IT-19**, 9–12, 1973.

Peebles, P. Z., *Probability, Random Variables, and Random Signal Principles*, 3rd ed., McGraw-Hill, New York, 1993.

Percival, D. B., and A. T. Walden, *Spectral Analysis for Physical Applications: Multivariate and Conventional Univariate Techniques*, Cambridge Univ. Press, Cambridge, 1993.

Priestley, M. B., *Spectral Analysis and Time Series*, Academic Press, London, 1981.

Priestley, M. B., *Non-linear and Non-stationary Time Series Analysis*, Academic Press, London, 1988.

Proakis, J. G., C. M. Rader, F. Ling, and C. L. Nikias, *Advanced Digital Signal Processing*, Macmillan, New York, 1992.

Riedel, K. S., A. Sidorenko, and D. J. Thomson, Spectral estimation of plasma fluctuations. I. Comparison of methods, *Phys. Plasmas*, **1**, 485–500, 1994.

Riedel, K. S., and A. Sidorenko, Minimum bias multiple taper spectral estimation, *IEEE Trans. Signal Proc.*, **43**, 188–195, 1995.

Shiavi, R., *Introduction to Applied Statistical Signal Analysis*, (2nd ed.), Academic Press, 1999.

Slepian, D., Prolate spheroidal wave functions, Fourier analysis and uncertainty; V: The discrete case, *Bell Syst. Tech. J.*, **5**, 1371–1429, 1978.

Thomson, D. J., Spectrum estimation and harmonic analysis, *Proc. IEEE*, **70**, 1055–1096, 1982.

Walden, A. T., E. J. McCoy, and D. B. Percival, The variance of multitaper spectrum estimates for real Gaussian processes, *IEEE Trans. Signal Process.*, **42**, 479–482, 1994.

Welch, P. D., The use of Fast Fourier Transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.*, **15**, 70–73, 1967.