1 Chapter 1: Introduction to main concepts

1.1 Introduction to time series analysis

The focus of this course is on the analysis, modeling and prediction of data observed in a sequential order. Standard inferential methods no longer work in this setup without further modifications since the data and innovations no longer can be viewed as independent. This requires a refined limit theory to allow for statements on asymptotic properties of estimators and test statistics. One possible conceptual solution to the dependence problem is to try to decompose the observed dependent data into independent innovations. These mechanisms are usually refereed as time series models.

1.2 Stochastic Processes

To formalize the discussion of main concepts, we start by defining what we mean by a time series. The mathematical theory on time series is based on an abstract probability space $(\Omega, \mathcal{F}, P)$.

Here Ω is the sample space of elementary events, $\mathcal{F}$ is a sigma-algebra defined on the sample space and P is a probability measure on Ω. We also introduce an index T that has an ordering relation defined on it. For example, for $t_1, t_2 \in T$, $t_1 \leq t_2$ or $t_1 < t_2$. Typically $T = R$, $T = Z^+$ or $T = Z$.

Definition A stochastic process is a family of random variables $\{X(t, w), t \in T, w \in \Omega\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$.

In particular, for a fixed w, $X(t, \cdot)$ is a function from T into R. This is called a realization of the stochastic process. On the other hand, for a fixed t, $X(\cdot, w)$ is a function from Ω into R. The definition contains continuous and discrete time processes.

In reality we have just one realization of the stochastic process, i.e. just for one given w, and we typically observe only finite number of points, i.e. T is finite.

1.3 Stationary Stochastic Processes

Assume that we have a process $\{X_t, t = 0, \pm 1, \pm 2; \ldots\}$. There are two widely used definitions of stationarity.

Definition *Strict* or *Strong Stationarity* is said to hold if, for any positive integer k, and k time points $t_1, \ldots, t_k$ and any integer lag h, we have that the vectors

$$(X_{t1}, X_{t2}, \ldots, X_{tk}) \tag{1.1}$$

and

$$(X_{t1+h}, X_{t2+h}, \ldots, X_{tk+h}) \tag{1.2}$$

have the same joint distribution.

Definition *Weak* or *Wide-sense* or *Second-order Stationarity* is said to hold if all the variances of the process are finite and we have

$$\begin{aligned} E(X_t) &= \mu, \\ Cov(X_t, X_{t+h}) &= \gamma_h, \end{aligned} \tag{1.3}$$

the quantities μ and λ_h being independent of t.

Provided the variances are finite, it is clear that a strictly stationary process must also be second-order stationary. However, the converse is not necessarily true. It is easy to construct examples of random variables which agree in the first- and second-order moment but not to any higher order.

However, in the case of a Gaussian process, the two concepts are equivalent. A process $\{X_t, t = 0, \pm1, \pm2, \ldots\}$ is said to be Gaussian if, for any k time points $t_1, \ldots, t_k$, the joint distribution of $X_{t1}, \ldots, X_{tk}$ has a multivariate normal distribution. This distribution is, of course, completely determined once its mean vector and covariance matrix have been specified. Therefore it follows that if a Gaussian process satisfies the second-order stationarity condition, it must also be strictly stationary.

Definition A special case of a weakly stationary process is white noise. We say that ϵ_t is white noise, or $WN(0, \sigma^2)$, if $E(\epsilon_t) = 0$ and $\mathrm{Cov}(\epsilon_\mathrm{i}, \epsilon_\mathrm{j}) = \sigma^2\delta_\mathrm{ij}$, where δ_{ij} is the Kronecker symbol that is equal to 1 if $i = j$ and 0 otherwise.

Definition A process $\{X_t, t = 0, \pm1, \pm2, \ldots\}$ is said to be linear if it has a representation of the form

$$X_t = \mu + \sum_{r=-\infty}^{\infty} c_r \epsilon_{t-r} \tag{1.4}$$

where μ is a common mean, $\{c_r\}$ is a sequence of absolute summable deterministic constants and $\{\epsilon_t\}$ are independent and identically distributed random variables with mean 0.

If $c_r = 0$ for all $r < 0$ it is said to be causal (i.e. in this case the process at time t does not depend on on future, as yet unobserved, values of ϵ_t). A Gaussian process necessarily has a representation as a linear process with normal $\{\epsilon_t\}$, but we may also want to consider non-Gaussian linear processes. The AR, MA and ARMA classes with i.i.d. innovations are all special cases of causal linear processes.

Remark. If ϵ_t is a martingale difference, then the process 1.4 is sometimes called a *weakly linear.*

1.4 Autocovariances, autocorrelations and spectral representations

For a weakly stationary process of mean 0, the autocovariance function is given by

$$\gamma_k = E\left\{X_t X_{t+k}\right\}. \tag{1.5}$$

It follows from the definition of weak stationarity that this does not depend on t. Also, note that $\gamma_{-k} = \gamma_k$ for all k.

The *autocorrelation function* is

$$\rho_k = \frac{\gamma_k}{\gamma_0}, \quad k = 0, \pm 1, \pm 2, \ldots \tag{1.6}$$

For any sequence of autocovariances $\{\gamma_k\}$ generated by a stationary process, there exists a function F such that

$$\gamma_k = \int\limits_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda) \tag{1.7}$$

where F is the unique function on $[-\pi, \pi]$ satisfying

(i) $F(-\pi) = 0$,

(ii) F is non-decreasing and right-continuous,

(iii) F has increments symmetric about 0, meaning that for any $0 \leq a < b \leq \pi$ we have

$$F(b) - F(a) = F(-a) - F(-b). \tag{1.8}$$

Then F is called the spectral distribution function, so called because it has many of the properties of a probability distribution function except that $F(\pi) = \gamma_0 = \mathrm{Var(X_t)}$, which is not necessarily always 1. Note that the integral 1.7 is a Stieltjes integral reflecting the fact that F may have discontinuities.

Definition If F is everywhere continuous and differentiable, with $f(\lambda) = dF(\lambda)/d\lambda$, then f is called the *spectral density function.*

Hence,1.7 may be simplified to

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda. \tag{1.9}$$

If $\sum |\gamma_k| < \infty$ then it can be shown that f always exists and is given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\lambda k} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{\infty} \gamma_k \cos(\lambda k). \tag{1.10}$$

The interpretation of F is that, for any $0 \leq \lambda_1 < \lambda_2 \leq \pi$, $F(\lambda_2) - F(\lambda_1)$ measures the contribution to the total variability of the process within the frequency range $\lambda_1 < \lambda \leq \lambda_2$.

Examples:

1. White noise: suppose $\lambda_0 = \sigma^2 > 0$ but $\lambda_k = 0$ for all $k \neq 0$.

   In this case it is immediately seen that

   $$f(\lambda) = \frac{\sigma^2}{2\pi} \text{ for all } \lambda, \tag{1.11}$$

   which is independent of λ. The converse also holds, i.e. a process is white noise if and only if its spectral density is constant.

2. Consider the process

   $$X_t = \cos(wt + U) \tag{1.12}$$

   where U is a random variable uniformly distributed on $(-\pi, \pi)$ and $0 \leq \omega \leq \pi$. We can easily calculate

   $$\begin{aligned} E(X_t) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega t + u)\, du = 0, \\ E(X_t X_{t+k}) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega t + u) \cos(\omega t + \omega k + u)\, du \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \{\cos(\omega k) + \cos(2\omega t + \omega k + 2u)\}\, du \\ &= \frac{\cos(\omega k)}{2}. \end{aligned} \tag{1.13}$$

Thus we see that $\{X_t\}$ is a stationary process. To find the spectral representation, we want to represent the autocovariance in the form

$$\gamma_k = \frac{\cos(\omega k)}{2} = \int_{(-\pi,\pi]} \cos(\lambda k) dF(\lambda). \tag{1.14}$$

A suitable F is one that takes jumps of $1/4$ at $\pm\omega$, i.e.

$$F(\lambda) = \begin{cases} 0 & \text{if } -\pi \leq \lambda < -\omega, \\ 1/4 & \text{if } -\omega \leq \lambda < \omega, \\ 1/2 & \text{if } \omega \leq \lambda \leq \pi. \end{cases}$$

By the uniqueness of F, this is the spectral distribution function in this case. Thus, a spectral distribution function which has discontinuities at $\pm\omega$ and is elsewhere flat, corresponds to a single sinusoid which is perfectly predictable once one observation in the series is known. Note that $\sum |\gamma_k| = \infty$ in this case.

An obvious extension is the case when F is flat except for k discontinuities at $\omega_1, \omega_2 \ldots, \omega_k$, where $0 \leq \omega_1 < \omega_2 < \ldots < \omega_k \leq \pi$. This corresponds to a process of the form

$$X_t = \sum_{j=1}^{k} a_j \cos(\omega_j t + U_j) \tag{1.15}$$

for some deterministic constants a_j and independent uniformly distributed on $(-\pi, \pi)$ random variables U_j.

The restriction to $0 \leq \omega_i \leq \pi$ in this example is in fact no restriction at all. In fact, suppose we have a process $X_t = \cos(\Phi t + U)$ for some Φ. Then we can expresss $\Phi = N\pi + \omega$ for some integer N, $\omega \in [0, \pi)$. If N is even then $\cos(\Phi t + U) = \cos(\omega t + U)$ for any integer t, so the frequencies ω and Φ are indistinguishable, or to use the conventional terminology in this situation, they are aliases. If N is odd, then $\cos(\Phi t + U) = \cos(\omega t - \pi t + U) = \cos(\pi t - \omega t - U)$. But U has the same distribution as $-U$ so in this case the frequency Φ is aliased to $\pi - \omega$. Thus, any frequency Φ is aliased to some frequency in the interval $[0, \pi]$.

The same comment obviously applies to 1.15 in which any k frequencies Φ_j may be aliased to frequencies ω_j in the interval $[0, \pi]$.

3. The AR(1) process

$$X_t = \phi_1 X_{t-1} + \epsilon_t, \tag{1.16}$$

in which $\{\epsilon_t\}$ is an uncorrelated sequence of random variables with mean 0 and common mean σ_ϵ^2 satisfies the relation

$$\text{Var}\{X_t\} = \phi_1^2 \text{Var}\{X_{t-1}\} + \sigma_\epsilon^2 \tag{1.17}$$

so under stationarity, in which $\text{Var}\{X_t\} = \gamma_0 = \sigma_X^2$ independently of t, we have

$$\sigma_X^2 = \frac{1}{1-\phi_1^2}\,\sigma_\epsilon^2. \tag{1.18}$$

Note that for 1.18 to make sense we require $|\phi_1| < 1$. This is the *stationarity condition* for a causal AR(1) process: without this condition the process tends to grow forever and so does not have a stationary distribution. (If $\phi_1 = \pm 1$ then the process is a random walk, which is a recurrent process but does not have a stationary distribution.) We shall see later that all AR processes require some condition of this nature.

Now for the model 1.16 satisfying the stationarity condition $|\phi_1| < 1$, and for $k > 0$, we have

$$\begin{aligned} \gamma_k &= E\{X_t X_{t-k}\} && (1.19) \\ &= \phi_1 E\{X_{t-1} X_{t-k}\} + E\{\epsilon_t X_{t-k}\} \\ &= \phi_1 \gamma_{k-1}. \end{aligned}$$

Since we also have $\gamma_{-k} = \gamma_k$ we deduce

$$\gamma_k = \phi_1^{|k|} \gamma_0, \qquad -\infty < k < \infty. \tag{1.20}$$

Direct application of 1.10 then leads to

$$f(\lambda) = \frac{\gamma_0\,(1-\phi_1^2)}{\pi\,(1 - 2\phi_1 \cos\lambda + \phi_1^2)} = \frac{\sigma_\epsilon^2}{\pi\,(1 - 2\phi_1 \cos\lambda + \phi_1^2)}. \tag{1.21}$$

Fig. 1 shows plots of $f(\lambda)$ for $\phi_1 = \pm 0.5$. In the case $\phi_1 > 0$, the power is concentrated at low frequencies, i.e. corresponding to gradual long-range fluctuations. For $\phi_1 < 0$ the power is concentrated at high frequencies, which reflects the fact that such a process tends to oscillate.

1.5 The Wold decomposition

In general, it is possible to write the spectral distribution function F in the form

$$F = F_1 + F_2, \tag{1.22}$$

where F_1 is absolutely continuous and F_2 is a purely discontinuous spectral distribution function. This corresponds to a decomposition of the process

$$X_t = U_t + V_t \tag{1.23}$$

into uncorrelated processes U and V in which U has spectral d.f. F_1 and V has spectral d.f. F_2.

As we saw in 1.15, a purely discontinuous spectral distribution function with finitely many jumps corresponds to a mixture of sinusoids, which is a purely deterministic and predictable process. The general result is given by the following:

Theorem. Suppose $F = F1 + F2$ as in 1.22 and suppose

$$\int_{-\pi}^{\pi} \log F_1'(\lambda) d\lambda > -\infty. \tag{1.24}$$

Then the decomposition of a stationary time series X into uncorrelated processes U and V as in 1.23 exists, and moreover, we have

(i) $U_t = \sum_{r=0}^{\infty} c_r \epsilon_{t-r}$ with $\{\epsilon_r\}$ uncorrelated random variables of mean 0 and common variance; without loss of generality we may take $c_0 = 1$, and we also require $\sum c_r^2 < \infty$,

(ii) V is a deterministic process, i.e. if we know V_s for all $s < t$ then we can predict V_t perfectly.

The sum in (i) is defined in the mean squared sense, i.e.

$$\lim_{R\to\infty} E\left\{\left(U_t - \sum_{r=0}^{R} c_r \epsilon_{t-r}\right)^2\right\} = 0. \tag{1.25}$$

1.6 Non-Negative Definiteness

The natural question to ask is:

if we are given an arbitrary sequence $\{\gamma_k,\ k \geq 0\}$, under what conditions is it the autocovariance function of some stationary process?

It is easily seen that there must be some restrictions, because for any finite sequence of constants $\{c_1, \ldots, c_T\}$ we have

$$\begin{aligned} 0 \leq \text{ Var } \left(\sum_{t=1}^{T} c_t X_t \right) &= \sum_{s=1}^{T} \sum_{t=1}^{T} c_s c_t \text{Cov} \left(X_s, X_t \right) \\ &= \sum_{s=1}^{T} \sum_{t=1}^{T} c_s c_t \gamma_{|t-s|}. \end{aligned} \tag{1.26}$$

If 1.26 holds for all sequences $\{c_1, \ldots, c_T\}$ then we say that $\{\gamma_k\}$ is *non-negative definite.* If the last expression is strictly positive except when $c_1 = \ldots = c_T = 0$, then it is *positive definite.* It turns out that non-negative definiteness is a necessary and sufficient condition for $\{\gamma_k\}$ to be the autocovariance function of some stationary process. (This result is known as Bochner's Theorem.)

How can we check non-negative definiteness of a given $\{\gamma_k\}$ sequence? A sufficient (and in fact necessary) condition is that the spectral density function defined by 1.10 should be non-negative for all λ. Under this condition we have

$$\begin{aligned} \sum_s \sum_t c_s c_t \gamma_{|s-t|} &= \sum_s \sum_t c_s c_t \int_{-\pi}^{\pi} e^{i(s-t)\lambda} f(\lambda) d\lambda \\ &= \int_{-\pi}^{\pi} \left\{ \sum_s \sum_t c_s c_t e^{i(s-t)\lambda} \right\} f(\lambda) d\lambda \\ &= \int_{-\pi}^{\pi} \left| \sum_t c_t e^{it\lambda} \right|^2 f(\lambda) d\lambda \\ &\geq 0. \end{aligned} \tag{1.27}$$

Thus, a very general method of constructing autocovariance functions is to take an arbitrary non-negative f and transform it via 1.9.

1.7 Estimating autocovariances and spectral densities

Suppose we have data $\{X_1, \ldots, X_t\}$. The usual estimate of γ_k for $k > 0$ is given by

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} \left(X_t - \bar{X} \right) \left(X_{t+k} - \bar{X} \right), \tag{1.28}$$

where $\bar{X}$ is the sample mean

$$\bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t. \tag{1.29}$$

Corresponding to this, we have estimates of the autocorrelations $\rho_k = \gamma_k/\gamma_0$, given by

$$r_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}. \tag{1.30}$$

In 1.28, it might seem more natural to divide by $T - k$ (rather than T) because the sum is taken over $t - k$ terms, but this is not usually done, for two reasons: (a) using the definition 1.28 ensures that the sample autocovariances are non-negative definite, which is evidently a desirable property for them to have, (b) the estimate of r_k in 1.30 often has smaller mean squared error if defined in this way, than it would in the alternative way with the divisor T in 1.28 replaced by $T - k$.

One reason for calculating and plotting the autocovariances or autocorrelations is the following easily verified fact: if $\{X_t\}$ is MA(q), then $\gamma_k = 0$ for $|k| > q$ and so plots of $\{\hat{\gamma}_k\}$ should show a sharp drop to near 0 after the q-th coefficient. This is therefore a diagnostic for an MA(q) process. The corresponding diagnostic for an AP(p) process is based on a different quantity, known as the partial *autocorrelation* function.

Definition The partial autocorrelation function (pacf) of lag k is based on the least-squares regression of X_t on $X_{t-k}, \ldots, X_{t-1}$. Formally, this is based on postulating the model

$$X_t = \sum_{j=1}^{p} a_{j,k} X_{t-j} + \epsilon_t, \quad t > k, \tag{1.31}$$

with ϵ_t independent X_t on $X_{t-k}, \ldots, X_{t-1}$. Least squares estimates of $\{a_{j,k},\ j = 1, \ldots, k\}$, obtained by minimization of

$$\sigma_k^2 = \frac{1}{T} \sum_{t=k+1}^{T} \left(X_t - \sum_{j=1}^{p} a_{j,k} X_{t-j} \right)^2, \tag{1.32}$$

which is (almost) equivalent to solving the equations

$$\hat{\gamma}_\ell = \sum_{j=1}^{k} a_{j,k} \hat{\gamma}_{|j-\ell|}, \quad 1 \le \ell \le k, \tag{1.33}$$

and then calculating the mean sum of squares by substituting in 1.32. In practice, these coefficients may be calculated recursively in k from the *Levinson-Durbin recursion*:

$$\begin{aligned} a_{k,k} &= \frac{\hat{\gamma}_k - \sum_{j=1}^{k-1} a_{j,k-1} \hat{\gamma}_{j-k}}{\sigma_{k-1}^2}, \\ a_{k,k} &= a_{j,k-1} - a_{k,k} a_{k-j,k-1}, \quad 1 \le j \le k-1, \\ \sigma_k^2 &= \sigma_{k-1}^2 \left(1 - a_{k,k}^2\right). \end{aligned} \tag{1.34}$$

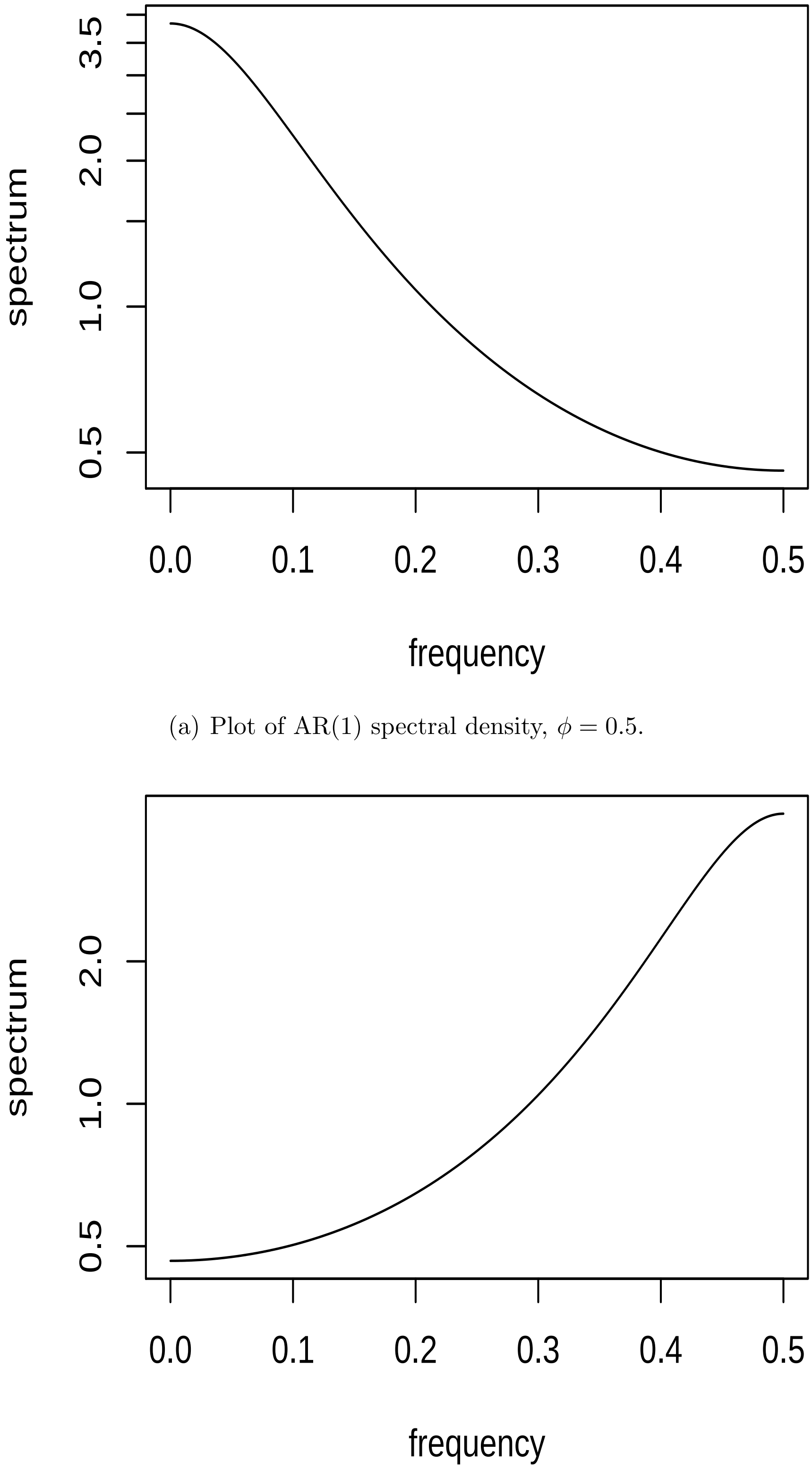
An obvious measure of how much the k-th order regression improves on that of order $k-1$ is the drop in mean squared residual error, and 1.34 shows that this is determined by the coefficient $a_{k,k}$. We therefore call $a_{k,k}$ the k-th order sample partial autocorrelation coefficient. The corresponding population autocorrelation coefficient is obtained from the same sequence of equations but with γ_k replacing $\hat{\gamma}_k$.

The most important property of pacf is the following: if the true process is AR(p), then the population partial autocorrelations of order $k > p$ are all 0, and therefore we would expect the sample partial autocorrelations to drop off sharply after lag p.

Definition For spectral densities, the simplest estimate is given by the *periodogram*

$$I_T(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} X_t e^{i\lambda t} \right|^2. \tag{1.35}$$

It will be shown later that the periodogram for fixed λ is an almost unbiased estimator of $f(\lambda)$, provided the underlying process is stationary and its spectral density exists. However, the sample periodogram is too rough to be a good estimator for most practical purposes. Various operations on the periodogram, in particular *smoothing* and *tapering*, will be introduced, in order to improve on the raw periodogram as a spectral density estimator.



(a) Plot of AR(1) spectral density, $\phi = 0.5$.

(b) Plot of AR(1) spectral density, $\phi = -0.5$.

Figure 1.1: Plot of AR(1) spectral density.