Time Series Analysis

Rebecca Sela

September 28, 2005

1 Preliminaries

1.1 A Review of Trigonometry and Complex Numbers

Definition Let $X_t = R\cos(\omega t + \phi)$. Then, ϕ is the *phase* of the wave (where it is at time 0), the *period* of the wave is $2\pi/\omega$ and the *frequency* is $\omega/2\pi$, which is the reciprocal of the period.

Some facts about sines and cosines:

- $\cos(A+B) = \cos(A)\cos(B) - \sin(A)\sin(B)$
- $\sin(A+B) = \sin(A)\cos(B) + \cos(A)\sin(B)$
- Cosine is an even function: $\cos(-\theta) = \cos(\theta)$
- Sine is an odd function: $\sin(-\theta) = -\sin(\theta)$
- $\cos(\omega t - \pi/2) = \sin(\omega t)$
- $R\cos(\omega t + \phi) = R\sin(\phi)\cos(\omega t) + R\cos(\phi)\sin(\omega t)$, and any wave form can be written as a linear combination of sines and cosines. Similarly, $A\sin(\omega t) + B\cos(\omega t)$ can be converted to the form $R\cos(\omega t + \phi)$.

According to Euler's formula, $\exp(i\omega) = e^{i\omega} = \cos(\omega) + i\sin(\omega)$. The conjugate of a complex number, $z = A + Bi$ is $\overline{z} = A - Bi$. In particular, the conjugate of a complex exponential is $\overline{e^{i\omega}} = e^{-i\omega}$. Also, multiplication of complex exponentials is like the multiplication of normal exponentials, so that $\exp(i\omega)\exp(i\lambda) = \exp(i(\omega+\lambda))$. This implies all of the trigonometric addition rules.

1.2 Cumulative Distribution Functions and Probability Measures

Suppose F is a cumulative distribution function. Then we know that F is a non-decreasing, right-continuous function with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. If F is continuous and differentiable, then the derivative is the corresponding

probability measure. If F is a step function, then it is the cumulative distribution function of a discrete random variable. If F has jumps but is not a step function, then it is a mixture of a discrete random variable and a continuous random variable. A cumulative distribution function can be used to determine the measure of sets.

Given a cumulative distribution function, F, the Lebesgue integral with respect to F is $\int g(x)dF(x)$. If F is differentiable with derivative f, then $\int g(x)dF(x) = \int g(x)f(x)dx$. If F is a step function with jumps at the points x_i, then $\int g(x)dF(x) = \sum g(x_i)\mu(\{x_i\})$. In general, the integral can be constructed using indicator functions to create simple functions that approximate arbitrary functions.

The *Lebesgue decomposition* of a measure, μ is $\mu = \mu_{ac} + \mu_d + \mu_s$, where the μ_{ac} is an absolutely continuous measure, μ_d is a discrete measure, and μ_s is everything else (which will not come up in simple examples).

2 The Spectrum

Definition A time series $\{X_t\}_{t=-\infty}^{\infty}$ is *weakly stationary* if:

- $E(X_t) = \mu$ is finite and does not depend on t.
- $Var(X_t) = \sigma^2$ is finite and does not depend on t.
- $Cov(X_t, X_{t-u})$ depends only on $|u|$.

When it is not otherwise stated, we assume that $\{X_t\}$ is a zero-mean process.

Definition Let $\{X_t\}$ be a stationary, mean-zero time series. Then the *theoretical (population) autocovariance* at lag r is $c_r = Cov(X_t, X_{t-r}) = E(X_t X_{t-r})$. The *population autocorrelation* at lag r is $\rho_r E(X_t X_{t-r})/E(X_t^2)$. Note that autocorrelations and autocovariances are symmetric about 0.

Definition An infinite sequence, $\{b_r\}_{t=-\infty}^{\infty}$ is *nonnegative definite* (or *positive semidefinite*) if for any sequence, $\{a_k\}_{t=-\infty}^{\infty}$, $\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} a_r b_{r-s} a_s \geq 0$.

Theorem 2.1 *The theoretical autocovariance sequence is nonnegative definite.*

Proof

$$\begin{aligned}
0 &\leq Var(\sum_{r=-\infty}^{\infty} a_r X_{t-r}) \\
&= E((\sum_{r=-\infty}^{\infty} a_r X_{t-r})^2) \\
&= E((\sum_{r=-\infty}^{\infty} a_r X_{t-r})(\sum_{s=-\infty}^{\infty} a_s X_{t-s})) \\
&= \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} a_r E(X_{t-r}X_{t-s})a_s \\
&= \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} a_r c_{r-s} a_s
\end{aligned}$$

Definition The *spectral distribution function* for a time series with theoretical autocovariance sequence, $\{c_r\}$, is a non-decreasing, right-continuous spectral distribution function, $F(\omega)$, defined on $[-\pi, \pi]$ such that $c_r = \int_{-\pi}^{\pi} \exp(ir\omega) dF(\omega)$. If F is differentiable with derivative $f(\omega)$, then $f(\omega)$ is called the *spectral density* or the *spectrum*, and we may write $c_r = \int_{-\pi}^{\pi} \exp(ir\omega) f(\omega)\, d\omega$.

Theorem 2.2 Herglotz's Theorem *For every nonnegative definite sequence, $\{c_r\}$, there exists such a non-decreasing, right-continuous spectral distribution function, $F(\omega)$.*

If a process has an exact cycle, then there is a jump in $F(\omega)$ at that frequency (and the spectrum does not exist, or would seem to have an infinite peak at that point), and a process that is the sum of countably many cycles would have a spectral distribution function that is a step function. A "pseudo-cyclical process" would have the mass of the spectrum centered around the approximate cycle. The spectrum of a persistent process is concentrated around the low frequencies and tends to slope downward.

If X_t and Y_t are independent with spectra $f(\lambda)$ and $g(\lambda)$, then the spectrum of $X_t + Y_t$ is $f(\lambda) + g(\lambda)$.

Theorem 2.3 Bochner's Theorem *In all cases, $c_r = \int e^{i\lambda r} F(d\lambda)$. If there is a spectral density, $f(\omega) = \frac{1}{2\pi}\sum_{r=-\infty}^{\infty} c_r \exp(-ir\omega)$.*

In particular, $Var(X_t) = c_0 = \int_{-\pi}^{\pi} f(\omega)\, d\omega$. Thus, $dF(\lambda) = f(\lambda)d\lambda$ is the contribution to the variance from the frequency λ.

The spectrum of white noise with variance σ^2 is $f(\omega) = \frac{\sigma^2}{2\pi}$.

(Note that all of this holds for continuous time processes as well, except that c_r can have $r \in R$ and all integrals have limits of $(-\infty, \infty)$ instead of $(-\pi, \pi)$.)

We may use the Lebesgue decomposition to decompose the spectrum into discrete and continuous parts. There are corresponding decompositions for the time series (into pure cycles and everything else) and the spectral measure.

Theorem 2.4 Riemann-Lebsgue Lemma. *If* $\{X_t\}$ *has a spectral density, then* $c_r = \int e^{ir\lambda} f(\lambda) d\lambda \to 0$ *as* $r \to \infty$.

2.1 The spectral representation

Definition A process $X(t)$ is *strictly stationary* if $P(X(t_1) \in S_1, ..., X(t_n) \in S_n) = P(X(t_1 + \tau) \in S_1, ..., X(t_n + \tau) \in S_n)$ for all $\tau, t_1, ..., t_n$.

Definition A vector space is an *inner product space* if there exists an inner product, (x, y), such that:

- $(x, y) = \overline{(y, x)}$
- $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$
- $(x, x) = 0$ if $x = 0$ and $(x, x) > 0$ otherwise

The *norm* of an inner product space is $\sqrt{(x, x)}$. If the vector space is a complete metric space under this norm, then it is a *Hilbert space*.

Definition Let $\{x_1, ..., x_n\}$ be an orthonormal basis of a complex vector space. If $z = \beta_1 x_1 + ... + \beta_n x_n$, with $\beta_i = (z, x_i)$, then $\beta_1, ..., \beta_n$ are the *Fourier coefficients* of z.

Definition The Hilbert space, $L_2(P)$, is the space of mean-zero, complex-valued random variables on a probability space (Ω, S, P) such that $E(X^2) < \infty$, with inner product $(X, Y) = E(X\bar{Y})$. Notice that the squared norm of $X \in L_2(P)$ is the variance.

Definition Given a stochastic process, $\{X_t\}$ on (Ω, S, P), we define $M_X \subset L_2(P)$ as the set of linear combinations of the $\{X_t\}$ for all t. (That is, the individual variables $\{X_1, X_2, ...\}$ are the basis.)

Definition Let F be the spectral distribution of a weakly stationary stochastic process $\{X_t\}$. The Hilbert space, $L_2(F)$, is the space of complex-values functions, $g(\lambda)$ such that $\int |g(\lambda)| dF(\lambda) < \infty$. The inner product is defined as $(g(\lambda), h(\lambda))_F = \int g(\lambda)\overline{h(\lambda)} dF(\lambda)$.

We define a one-to-one mapping between M_X and $L_2(F)$ by mapping X_t to $\exp(i\lambda t)$, for $t = 1, 2, 3....$ This is a mapping of bases, and may be extended linearly to the rest of the elements. This is an isometric isomorphism, because:

$$\begin{aligned}
(X_t, X_u)_P &= E(X_t \overline{X_u}) \\
&= C(t - u) \\
&= \int_{-\infty}^{\infty} \exp(i\lambda(t - u)) dF(\lambda) \\
&= \int_{-\infty}^{\infty} \exp(i\lambda t)\overline{\exp(i\lambda u)} dF(\lambda) \\
&= (\exp(i\lambda t), \exp(i\lambda u))_F
\end{aligned}$$

Definition We define a random variable called *the random measure*, $Z(A)$ (in M_X) as the image of the indicator function, $I(A)$ (in $L_2(F)$), under the transformation above. Taking the limit, $dZ(\lambda)$ is the image of the indicator function of a point, $I(\{\lambda\})$.

Some properties of the random measure:

- If $A \cap B = \emptyset$ then $E(Z(A)\overline{Z(B)}) = 0$; as a corollary, $E(dZ(\lambda)\overline{dZ(\omega)}) = 0$ if $\lambda \neq \omega$.
- $E(|Z(A)|^2) = \int_A dF(\lambda) = F(A)$ and $E(dZ(\lambda)\overline{dZ(\lambda)}) = dF(\lambda)$.
- If X_t has mean zero, then $E(dZ(\lambda)) = 0$.

Definition The *spectral representation* of a mean-zero weakly stationary process, $\{X_t\}$, is $X_t = \int e^{i\lambda t} dZ(\lambda)$.

$dZ(\lambda)$ describes the phase and amplitude for the wave at frequency λ.

2.2 Sample Analogues

Definition Given a time series of length n, we define the *Fourier frequencies* as $\omega_j = \frac{2\pi j}{n}$ for $j = 0, ..., n-1$.

Note that $\sum_{t=0}^{n-1} \exp(-i\omega_j t) = 0$ if $j \neq 0 \pmod n$. As a corollary, $\sum_{t=0}^{n-1} \cos(\omega_j t) = \sum_{t=0}^{n-1} \sin(\omega_j t) = 0$ if $j \neq 0$.

Definition Consider a time series, $\{X_0, ..., X_{n-1}\}$. The *Fourier decomposition* is

$$X_t = A_0 + A_1 \cos(\omega_1 t) + B_1 \sin(\omega_1 t) + ... + A_{\frac{n}{2}-1} \cos(\omega_{\frac{n}{2}-1} t) + B_{\frac{n}{2}-1} \sin(\omega_{\frac{n}{2}-1} t) + A_{\frac{n}{2}} (-1)^t$$

(where the last term is only there if n is even), where $A_0 = \overline{X}$, $A_{\frac{n}{2}} = \frac{1}{n} \sum_{t=0}^{n-1} X_t (-1)^t$, and for $0 < j < n/2$, $A_j = \frac{2}{n} \sum_{t=0}^{n-1} X_t \cos(\omega_j t)$ and $B_j = \frac{2}{n} \sum_{t=0}^{n-1} X_t \sin(\omega_j t)$.

We may think of $\{\exp(i\omega_j t)\}_{t=0}^{n-1}$ as a vectors in C^n for each j. This set of vectors forms an orthonormal basis of the space.

Definition The *discrete Fourier transform* (DFT) of the series $\{X_t\}_{t=0}^{n-1}$ is given by $\{J_j\}_{j=0}^{n-1}$ where

$$J_j = \frac{1}{n} \sum_{t=0}^{n-1} X_t \exp(-i\omega_j t)$$

Notice that $J_{n-j} = \overline{J_j}$, so all the information in the series is contained in the first half of the coordinates. We may also define $J(\omega) = \frac{1}{n} \sum_{t=0}^{n-1} X_t \exp(-i\omega t)$ for non-Fourier frequencies.

Definition The *inverse Fourier transform* is given by

$$X_t = \sum_{j=0}^{n-1} J_j \exp(i\omega_j t)$$

Theorem 2.5 *The inverse Fourier transform of the discrete Fourier transform recovers the original series.*

Proof

$$\begin{aligned}
\sum_{j=0}^{n-1} J_j \exp(i\omega_j t) &= \sum_{j=0}^{n-1} (\frac{1}{n} \sum_{u=0}^{n-1} X_u \exp(-i\omega_j u)) \exp(i\omega_j t) \\
&= \sum_{u=0}^{n-1} X_u (\frac{1}{n} \sum_{j=0}^{n-1} \exp(i\omega_j (t-u))) \\
&= \sum_{u=0}^{n-1} X_u 1(u=t) \\
&= X_t
\end{aligned}$$

Definition The *periodogram* is the set of random variables $\{I_j\}_{j=0}^{n-1}$, where $I_j = \frac{n}{2\pi}|J_j|^2 = \frac{n}{2\pi} J_j \overline{J_j}$. It is also a name for the plot of these random variables. Note that we only need to plot $j = 1, ..., n/2$.

$|J_j|$ is the strength of the ω_j frequency in the data (if $j < n/2$). Small values of j correspond to low frequencies (that is, long periods, and therefore smoother data). Peaks in the periodogram suggest cycles or pseudo-cycles.

Because we observe only a finite number of $\{X_t\}$, we are not able to observe all possible frequencies. In particular, if the true frequency (relative to the sample size n) is larger than π (that is, half the sample period, which is called the *folding frequency*), then it will be *aliased* to a lower frequency. For example, a frequency ω with $\pi \leq \omega \leq 2\pi$ will appear as the frequency $\omega' = 2\pi - \omega$.

Notice that $\sum_{t=0}^{n-1} |X_t|^2 = n \sum_{j=0}^{n-1} |J_j|^2$, and each $|J_j|^2$ is the contribution of the j^{th} Fourier frequency to the variance of the time series.

Definition The *sample autocovariance* at lag r (for $|r| < n$) is $\hat{c}_r = \frac{1}{n} \sum_{t=|r|}^{n-1} X_t X_{t-|r|}$.

Theorem 2.6 *We may write* $I(\omega) = \frac{1}{2\pi} \sum_{|r|<n} \hat{c}_r \exp(-ir\omega) = \frac{1}{2\pi}(\hat{c}_0 + 2\sum_{r=1}^{n-1} \hat{c}_r \cos(r\omega))$ *and* $\hat{c}_r = \int_{-\pi}^{\pi} I(\omega) \exp(ir\omega)\, d\omega$.

Proof We know that $I(\omega) = \frac{1}{2\pi n} \sum_{t=0}^{n-1} \sum_{u=0}^{n-1} X_t X_u \exp(-i(t-u)\omega)$. Let $v = t - u$. We may split the summation into two parts: first, $v = 0, ..., n-1$ and $u = 0, ..., n-1-v$, and second, $v = -(n-1), ..., -1$ and $u = -v, ..., n-1$.

Then, the summation becomes:

$$\begin{aligned} I(\omega) &= \frac{1}{2\pi n}\sum_{v=-(n-1)}^{-1}(\sum_{u=-v}^{n-1} X_{v+u}X_u)\exp(-iv\omega) + \frac{1}{2\pi n}\sum_{v=0}^{n-1}(\sum_{u=0}^{n-1-v} X_{v+u}X_u)\exp(-iv\omega) \\ &= \frac{1}{2\pi}\sum_{v=0}^{n-1}\hat{c}_v\exp(-iv\omega) + \frac{1}{2\pi}\sum_{v=-(n-1)}^{-1}\hat{c}_{-v}\exp(-iv\omega) \\ &= \frac{1}{2\pi}\sum_{|r|<n}\hat{c}_r\exp(-ir\omega) \end{aligned}$$

Second,

$$\begin{aligned} \int_{-\pi}^{\pi} I(\omega)\exp(ir\omega)d\omega &= \int_{-\pi}^{\pi}(\frac{1}{2\pi}\sum_{|s|<n}\hat{c}_s\exp(-is\omega))\exp(ir\omega)d\omega \\ &= \sum_{|s|<n}\hat{c}_s\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp(i\omega(r-s))d\omega \\ &= \hat{c}_r\frac{1}{2\pi}\int_{-\pi}^{\pi}1d\omega \\ &= \hat{c}_r \end{aligned}$$

(The second-to-last step follows because $\int_{-\pi}^{\pi}\exp(i\omega(r-s))d\omega = 0$ unless $r = s$.)
∎

Note that we only calculate the values of the periodogram at Fourier frequencies, which means that we cannot integrate over them to calculate autocovariances. However, we may use the *interpolation formula* to calculate autocovariances as well.

Theorem 2.7 *Suppose $\{X_t\}$ is of length n. Append n (or more) zeroes to the end of the series to create the new series $\{X_t'\}$, and let $I(\omega_j')$ for $j = 1, ..., 2n$ be the periodogram of the new series. Then, we may calculate $\hat{c}_r = \frac{2\pi}{2n}\sum_{j=0}^{2n-1} I(\omega_j')\exp(ir\omega_j')$.*

Definition A series is *ergodic* if time averages equal sample averages. (This might fail if there is a one-time random variable that affects each realization of the series instead of individual points.)

Theorem 2.8 *If a series is Gaussian and has a spectral density, then it is ergodic.*

Theorem 2.9 *For an ergodic stationary series, $nVar(\bar{X}) \to 2\pi f(0)$. Also, $nVar(\bar{X}) = 2\pi E(I_0)$. (This gives a consistent estimate of the variance of the mean, even in the presence of autocorrelation and heteroskedasticity.)*

If $f(0) = 0$, then the variance of the mean grows more slowly than n. If $f(0) = \infty$, then the variance of the mean grows faster than n.

2.3 Linear Filters and Spectra

Definition Let $\{y_t\}$ be any time series. A *linear filter* is a set of weights, $\{g_r, g_{r+1}, ..., g_s\}$, where we define

$$z_t = \sum_{u=r}^{s} g_u y_{t-u}$$

Note that z_t is a local weighted average of y_t, and that $\{z_t\}$ is a convolution of the sequence $\{y_t\}$ with $\{g_u\}$.

Equivalently, L is a *linear filter* if $L : \{X_t\} \to L(\{X_t\})$ such that L is:

- Scale preserving: $L(\alpha\{X_t\}) = \alpha L(\{X_t\})$
- Superposable: $L(\{X_t\} + \{Y_t\}) = L(\{X_t\}) + L(\{Y_t\})$
- Time invariant: If $L(\{X_t\}) = \{Y_t\}$ then $L(\{X_{t+h}\}) = \{Y_{t+h}\}$

If a time series is a sum of signal and noise, we hope to choose weights so that they cause the noise to cancel out, leaving only the signal.

Definition The *transfer function* of a filter, $\{g_r, g_{r+1}, ..., g_s\}$, is given by

$$G(\omega) = \sum_{u=r}^{s} g_u \exp(-i\omega u)$$

Then, if $\{z_t\}$ is the filtered version of $\{y_t\}$, $J_z(\omega) = G(\omega)J_y(\omega)$. More generally, if L is any linear filter, then there exists some $B(\lambda)$ such that $L(e^{i\lambda t}) = B(\lambda)e^{i\lambda t}$. $B(\lambda)$ is called the *transfer function*. $|B(\lambda)|$ is called the *gain function*.

The transfer function shows how the filter affects the amplitude at each frequency. (A linear filter always converts a wave at frequency ω to another wave at frequency ω, but may affect the phase and amplitude.) If we wish to annihilate a certain frequency (in seasonal adjustment, for example), we must choose a filter such that the transfer function is 0 at that frequency.

Notice that, if $\{\alpha_\lambda\}$ does not depend on t, then $L(\sum_\lambda \alpha_\lambda e^{i\lambda t}) = \sum_\lambda \alpha_\lambda L(e^{i\lambda t})$. Then, if $X_t = \sum_\lambda e^{i\lambda t}$, $L(\{X_t\}) = \sum_\lambda \alpha_\lambda B(\lambda)e^{i\lambda t} = \sum_\lambda \alpha_\lambda |B(\lambda)|e^{i\theta_\lambda}e^{i\lambda t}$. If $\{X_t\}$ is weakly stationary, then $L(X_t) = \int e^{i\lambda t}B(\lambda)dZ_X(\lambda)$, and the spectral distribution of the filtered sequence is $|B(\lambda)|^2 dF_X(\lambda)$, provided that the integral of this is finite (this ensures that the resulting series has finite variance).

Definition For a linear filter, the condition $\int |B(\lambda)|^2 dF_X(\lambda) < \infty$ is called the *matching condition*.

The inverse, L^* of a filter, L, should satisfy $L^*(L(X_t)) = X_t$. This implies that the transfer function of the inverse filter is $\frac{1}{B(\lambda)}$. This means that the transfer function of any invertible filter must be non-zero for all λ. In addition, the inverse filter must satisfy the matching condition, $\int |\frac{1}{B(\lambda)}|^2 dF_X(\lambda) < \infty$. There is not always an inverse filter.

2.4 The Periodogram of Noise

Suppose $X_t \sim Normal(0,1)$ are independent and identically distributed (this is called *Gaussian white noise*). Then, J_j and $I(\omega_j)$ are random variables. We may write:

$$\begin{aligned} I(\omega) &= \frac{n}{2\pi}|J(\omega)|^2 \\ &= \frac{1}{2\pi n}|\sum_{t=0}^{n-1} X_t \exp(-i\omega t)|^2 \\ &= \frac{1}{2\pi n}((\sum_{t=0}^{n-1} X_t \cos(\omega t))^2 + (\sum_{t=0}^{n-1} X_t \sin(\omega t))^2) \end{aligned}$$

If $k \neq j$ are both strictly between 0 and π, then

$$\begin{pmatrix} \sum_{t=0}^{n-1} X_t \cos(\omega_j t) \\ \sum_{t=0}^{n-1} X_t \sin(\omega_j t) \\ \sum_{t=0}^{n-1} X_t \cos(\omega_k t) \\ \sum_{t=0}^{n-1} X_t \sin(\omega_k t) \end{pmatrix} \sim Normal(0, \frac{n}{2} I)$$

Thus, each $I(\omega_j)$ is the sum of the squares of two independent normals, and $I(\omega_j) \sim \frac{1}{2\pi} Exponential(\lambda = 1)$ are distributed independent and identically. Since this distribution is right-skewed, periodograms will tend to have peaks, even if the time series is just noise. Also, because the distribution of $I(\omega_j)$ does not depend on n, the periodogram is not consistent. (However, more frequencies are estimated as the sample size increases.) If we have non-Gaussian white noise, sums will only be uncorrelated, not independent, but everything will hold asymptotically.

To test for whether a time series is significantly different from white noise, we use *Fisher's Test*. Let m be the number of frequencies of interest (most likely, all $n/2$ of them). Let

$$G_m = \frac{\max(I(\omega_j))_{j=1}^m}{\sum_{j=1}^m I(\omega_j)}$$

Asymptotically, under the null hypothesis of white noise,

$$P(mG_m \leq x + \ln(m)) \approx \exp(-e^{-x})$$

2.5 Leakage and Data Windows

The discrete Fourier transform of a sequence of ones (called "the boxcar") is

$$J(\omega) = \frac{1}{n}\sum_{t=0}^{n-1} \exp(-i\omega t) = \exp(-i(n-1)\omega/2)\frac{\sin(n\omega/2)}{n\sin(\omega/2)}$$

The *Dirichlet Kernel* is $D_n(\omega) = \frac{\sin(n\omega/2)}{n\sin(\omega/2)}$. Notice that $D_n(0) = 1$ and $D_n(\omega_j) = 0$ for $j \neq 0$. However, between Fourier frequencies, $D_n(\omega)$ is non-zero (though smaller than one). These non-zero parts are called the *sidelobes*.

Suppose $x_t = \exp(i\omega t)$. Then,

$$\begin{aligned} J_j &= \frac{1}{n}\sum_{t=0}^{n-1}\exp(i(\omega-\omega_j)t) \\ &= D_n(\omega-\omega_j)\exp(\frac{1}{2}i(n-1)(\omega-\omega_j)) \end{aligned}$$

If ω is a Fourier frequency, then J_j is one for that frequency and zero for all others. However, if ω is not a Fourier frequency, J_j will be non-zero for many j. This is called *leakage*. In this simple case, J_j will be largest when ω_j is closest to ω, and values will decay proportionally to $|\omega - \omega_j|$.

A *data window* is a set of constants, $\{w_t\}_{t=0}^{n-1}$. We *taper* the data according to the data window by analyzing $z_t = w_t x_t$. One common taper is the *cosine bell*:

$$\begin{aligned} w_t &= \frac{1}{2}(1-\cos(2\pi(t+\frac{1}{2})/n)) \\ J_w(\omega) &= \exp(-i\omega\frac{n-1}{2})(\frac{1}{4}D_n(\omega-\frac{2\pi}{n})+\frac{1}{2}D_n(\omega)+\frac{1}{4}D_n(\omega+\frac{2\pi}{n})) \end{aligned}$$

The second term of $J_w(\omega)$ is called the *Hanned* version of $D_n(\omega)$. The Hanned version of $D_n(\omega)$ is approximately zero for all $\omega \neq 0$, and the sidelobes decay more rapidly, decreasing the leakage. A *split cosine bell* tapers just the ends of the data, which helps reduce leakage, but is not as dramatic an improvement:

$$w_t = \begin{cases} \frac{1}{2}(1-\cos(\frac{1}{m}\pi(t+0.5))) & t = 0,...,m-1 \\ 1 & t = m,...,n-m-1 \\ \frac{1}{2}(1-\cos(\frac{1}{m}\pi(n-t-0.5))) & t = n-m,...,n-1 \end{cases}$$

2.6 Estimating the Spectrum

We use the periodogram to estimate the true spectrum. Asymptotically, whenever $0 \leq \omega_j \leq \pi/2$, $I(\omega_j) \sim f(\omega_j)\cdot\frac{1}{2}\chi^2_2$, and $E(I(\omega_j)) \approx f(\omega_j)$. The ordinates are approximately independent. Notice that this estimate is inconsistent, but more frequencies are estimated each time; therefore, we may use weighted averages to try to get a better estimate. Knowing the approximate distribution allows us to put confidence bands on the periodogram (if we have an idea of what $f(\omega)$ is).

Definition The *discrete periodogram average* is a way to do smoothing, defined by:

$$\hat{f}(\omega_j) = \sum_{k=-m}^{m} g_k I_{j-k}$$

where m is a fixed fraction of the sample size n or a function of n (such as $n^{4/5}$).

If f is relatively smooth, then $E(\hat{f}(\omega_j)) = \sum_{k=-m}^{m} g_k f(\omega_{j-k}) \approx f(\omega_j) \sum_{k=-m}^{m} g_k$, so we generally choose $\sum_{k=-m}^{m} g_k = 1$ so that the estimator is asymptotically unbiased. In addition, $Var(\hat{f}(\omega_j)) \approx f(\omega_j)^2 \sum g_k^2$.

Definition *Lag-weights (Blackman and Tukey) estimation* is defined by:

$$\hat{f}(\omega) = \frac{1}{2\pi} \sum_{|r|<m} w_r \hat{c}_r \exp(ir\omega)$$

where w_r are *lag weights* with $w_0 = 1$ and $w_r = w_{-r}$ (their shape is called the *lag window* and m is the *lag number*). The Fourier transform of the weights, $W(\lambda) = \frac{1}{2\pi} \sum_r w_r \exp(ir\lambda)$, is called the *spectral window*.

Using the relationship between the sample autocorrelations and the periodogram, we find out that lag-weights estimation is an integral convolution of the periodogram:

$$\begin{aligned} \hat{f}(\omega) &= \frac{1}{2\pi} \sum_{|r|<m} w_r \hat{c}_r \exp(ir\omega) \\ &= \frac{1}{2\pi} \sum_{|r|<m} w_r \left(\int_{-\pi}^{\pi} I(\lambda) \exp(-ir\lambda) d\lambda \right) \exp(ir\omega) \\ &= \int_{-\pi}^{\pi} W(\omega - \lambda) I(\lambda) d\lambda \end{aligned}$$

Of the two methods, the former tends to be noisier, while the latter has sidelobes and therefore leakage. Both of these estimates are non-parametric; if we assume that f has a functional form, such as that of an $AR(p)$ process, then we may be able to get a tighter estimate.

Under the null hypothesis that a time series $\{X_t\}$ is white noise, $Var(\bar{X}_m) = \frac{\sigma^2}{m}$, for any m. We may estimate $\hat{Var}(\bar{X}_m)$ by finding the sample variance of $\frac{1}{m}(X_1 + ... + X_m), \frac{1}{m}(X_{m+1} + ... + X_{2m}), ...$ (This is called the *Variance Ratio Test.*) This can be extended to making a periodogram based on each block of length m and using the average to estimate the spectral density. This method is equivalent to using the Bartlett lag-weights, $w_r = 1 - \frac{|r|}{m}$, and $\hat{Var}(\bar{X}_m)$ is the Barlett estimate of $2\pi f(0)$.

2.7 Fast Fourier Transforms

Computing Fourier Transforms (and other related things) in the naive way takes $O(n^2)$ steps. However, a Fast Fourier Transform allows calculation in (a best case of) $O(n \log n)$ steps. We consider the two-factor case, in which we wish to find the Fourier Transform of a time series, $\{X_t\}$ of length $n = n_1 n_2$:

- Convert the time series into an $n_1 \times n_2$ array, Y, where $Y(t_1, t_2) = X(t_1 n_2 + t_2)$; that is, we put the time series along the rows.

- Calculate the discrete Fourier transform of each column (using this method recursively, if desired), and replace each entry by the $J_j W_n^{j_1 t_2}$, where $W_n = \exp(-\frac{2\pi i}{n})$ is called the *twiddle factor*.
- Calculate the discrete Fourier transform of each row.
- The overall discrete Fourier Transform is read down the columns; that is, $J_{j_2 n_1 + j_1}$ is in the (j_1, j_2) position.

3 ARMA Models

Definition A time series, $\{\epsilon_t\}$, is *white noise* if it is uncorrelated at all lags, has mean zero, and constant variance σ^2. A process is white noise if and only if it has a spectral density, $f_\epsilon(\lambda) = \frac{\sigma^2}{2\pi}$.

Definition The *backshift operator*, B, is defined by $Bx_t = x_{t-1}$. The *differencing operator*, Δ, is defined by $\Delta = 1 - B$.

Definition A time series, $\{X_t\}$ is a *moving average process* if it can be written as $X_t = \sum_{j=-\infty}^{\infty} a_j \epsilon_{t-j}$, with $\sum_{j=-\infty}^{\infty} a_j^2 < \infty$. It is a *one-sided moving average process* if it can be written as $x_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j} = \sum_{j=0}^{\infty} a_j B^j \epsilon_t$.

Definition A time series, $\{X_t\}$, is an *autoregression* of order p, or $AR(p)$, if there exists a white noise process, $\{\epsilon_t\}$, with variance σ^2 and constants $b_1, ..., b_p$ such that

$$X_t + b_1 X_{t-1} + ... + b_p X_{t-p} = \epsilon_t$$

and $E(X_s \epsilon_t) = 0$ whenever $s < t$. $\{X_t\}$ is an autoregression of infinite order, $AR(\infty)$, if

$$\sum_{k=0}^{\infty} b_k X_{t-k} = \epsilon_t$$

with $b_0 = 1$, $\sum_{k=0}^{\infty} b_k^2 < \infty$, and $E(X_s \epsilon_t) = 0$ when $s < t$.

Definition An $MA(q)$ model, $x_t = \sum_{j=0}^{q} a_j \epsilon_{t-j}$, if *invertible* if all the roots of $Q(z) = 1 + a_1 z + ... + a_q z^q$ lie outside the unit circle. Equivalently, an $MA(q)$ model is invertible if it can be written as an $AR(\infty)$ model.

Since the transfer function from white noise to the MA process, $X_t = \sum_{j=-\infty}^{\infty} a_j \epsilon_{t-j}$, is $B(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}$, the spectral representation is $X_t = \int_{-\pi}^{\pi} e^{i\lambda t} B(\lambda) dZ_\epsilon(\lambda)$, and the spectral density is $f_X(\lambda) = |B(\lambda)|^2 f_\epsilon(\lambda) = \frac{\sigma^2}{2\pi} |\sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}|^2$. The spectral density of a non-invertible model is 0 at frequency 0.

Theorem 3.1 *Any weakly stationary time series with a continuous spectrum can be represented as a (possibly two-sided) moving average process.*

Proof Suppose Y_t has a strictly positive spectral density, f_Y. Let $\epsilon_t = \int_{-\infty}^{\infty} e^{i\lambda t} A(\lambda) dZ_Y(\lambda)$, where $A(\lambda) = \frac{\sigma}{\sqrt{2\pi f_Y(\lambda)}}$. Then, $\int_{-\pi}^{\pi} |A(\lambda)|^2 f_Y(\lambda) = \sigma^2$ which satisfies the matching condition. Furthermore, $f_\epsilon(\lambda) = |A(\lambda)|^2 f_Y(\lambda) = \frac{\sigma^2}{2\pi}$, and ϵ_t is white noise. Thus, $Y_t = \int_{-\pi}^{\pi} e^{i\lambda t} \frac{1}{A(\lambda)} dZ_\epsilon(\lambda)$. We may write $\frac{1}{A(\lambda)} = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}$. Then, we have $Y_t = \sum_{j=-\infty}^{\infty} a_j \epsilon_{t-j}$. ∎

For an autoregression, we have a transfer function, $B(\lambda) = \sum_{k=0}^{\infty} b_k e^{-i\lambda k}$ from X_t to ϵ_t, so that $dZ_\epsilon(\lambda) = B(\lambda) dZ_X(\lambda)$, or $dZ_X(\lambda) = \frac{1}{B(\lambda)} dZ_\epsilon(\lambda)$, and $f_X(\lambda) = \frac{\sigma^2}{2\pi |\sum b_k e^{-i\lambda k}|^2}$.

Theorem 3.2 *An AR(p) process exists and is weakly stationary if $\beta(z) = 1 + b_1 z + ... + b_q z^q$ has all its roots outside the unit circle.*

Proof Note that $B(\lambda) = \beta(e^{-i\lambda}) = \sum_{k=0}^{p} b_k e^{-i\lambda k}$ is the transfer function from X_t to white noise. Suppose all the roots of $\beta(z)$ lie outside the unit circle. Since $e^{-i\lambda}$ traces out the unit circle, $B(\lambda) \neq 0$ and is in fact bounded away from zero, so that $|B(\lambda)| > \epsilon$ for all λ. Then, $Var(X_t) = \frac{\sigma^2}{2\pi} \int_{-\pi}^{\pi} \frac{1}{|B(\lambda)|^2} d\lambda < \frac{\sigma^2}{2\pi} \int_{-\pi}^{\pi} \frac{1}{\epsilon^2} d\lambda < \infty$. By the definition of the linear filter, this process satisfies the difference equation describing the original process. By complex analysis, it turns out that $\frac{1}{B(\lambda)}$ is analytic inside a circle of radius $1 + \epsilon/2$ and therefore can be written as $\frac{1}{|\beta(z)|} = a_0 + a_1 z + a_2 z^2 + ...$ and $\frac{1}{B(\lambda)} = \sum_{j=0}^{\infty} a_j e^{-i\lambda j}$, which gives an expression of X_t in terms of only past and present shocks. Since future shocks are uncorrelated with past shocks, $E(X_t \epsilon_s) = 0$ when $t < s$, and this process exists. ∎

If any roots are inside the unit circle, the process is explosive. If there is a root on the unit circle (and none inside), the process has a "unit root". Note that a process with roots inside the unit circle may be able to be written as a weakly stationary process in terms of future shocks.

Parameters of the autoregression can be estimated using maximum likelihood:

$$L(\theta) = (2\pi)^{-n/2} |\Sigma_\theta|^{-1/2} \exp(-x^T \Sigma_\theta^{-1} x/2)$$

where $\Sigma_\theta(i,j) = c_{i-j}$ is the autocovariance implied by the parameters (this matrix is Toeplitz, since it has the same value along each diagonal). However, estimation can be hard because of the matrix inversion, and the result may not be stationary. Other methods include the Burg method and the Yule-Walker equations:

$$c_r = \sum_{k=1}^{p} a_k c_{r-k}$$

(The Yule-Walker equations are equivalent to using PACF's to calculate coefficients; since the PACF's are zero after the p^{th} PACF, so are the true coefficients.)

Combining results from autoregressive and moving average models, we find that the spectral representation and density of the $ARMA(p,q)$ process $X_t + b_1 X_{t-1} + ... + b_p X_{t-p} = \epsilon_t + a_1 \epsilon_{t-1} + ... + a_q \epsilon_{t-q}$ are

$$\begin{aligned} dZ_X(\lambda) &= \frac{\sum_{j=0}^{q} a_j e^{-i\lambda j}}{\sum_{k=0}^{p} b_k e^{-i\lambda k}} dZ_\epsilon(\lambda) \\ f_X(\lambda) &= |\frac{\sum_{j=0}^{q} a_j e^{-i\lambda j}}{\sum_{k=0}^{p} b_k e^{-i\lambda k}}|^2 \frac{\sigma^2}{2\pi} \end{aligned}$$

Using the backshift operator, an ARMA process can be written as $\phi(B)x_t = \theta(B)\epsilon_t$, the MA($\infty$) representation is $X_t = \frac{\theta(B)}{\phi(B)}\epsilon_t$, and the spectral density is $\frac{\sigma^2}{2\pi}|\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}|^2$.

Definition The *dynamic range* of a process is defined as $\frac{\max f(\lambda)}{\min f(\lambda)}$.

With a fraction in the spectral representation of an ARMA process, the dynamic range can be made quite large. Note that peaks and poles in the middle spectrum will make the process pseudo-cyclical ("seasonal long memory", or a process with slowly-changing seasonality) while a spectral density that tends to infinity at 0 leads to long memory in the process.

In estimation, we must choose p and q (this is model selection). Three different information criteria may be used; the (p,q) that minimizes the information criterion is the one chosen.

$$\begin{aligned} AIC &= -2\log(likelihood) + 2(\#parameters) \\ AIC_C &= -2\log(likelihood) + 2(m+1)\frac{n}{n-m-2} \\ BIC &= -2\log(likelihood) + (\log(n))(\#parameters) \end{aligned}$$

(In the case of these models, $-2\log(likelihood) = n\log(\hat{\sigma}^2)$.)

4 Long Memory Processes

4.1 Differencing and Unit Roots

Definition A process, $\{X_t\}$, is a *random walk* if $X_t = X_{t-1} + \epsilon_t$. Equivalently, the difference of a random walk is white noise.

We may use the Dickey-Fuller test to test whether $\rho = 1$ in the equation $X_t = \rho X_{t-1} + \epsilon_t$. To do this:

1. Regress X_t on X_{t-1}.
2. Construct $\hat{\tau}_\mu = \frac{\hat{\rho}-1}{se(\hat{\rho})}$.
3. Under the null hypothesis that $\rho = 1$, $\hat{\tau}_\mu$ has a Dickey-Fuller distribution (which is non-normal!).

To test for a unit root with additional lags of x, we use a Augmented Dickey-Fuller test. (The Said-Dickey test also allows lags of ϵ, but this is ugly.)

4.2 Long Memory Processes

Definition We say that $f(x) \sim g(x)$ as $x \to K$ if $f(x)/g(x) \to 1$ as $x \to K$.

Definition A weakly stationary process, $\{X_t\}$, has *long memory* with memory parameter $-0.5 < d < 0.5, d \neq 0$ if the spectral density $f(\lambda) \sim k\lambda^{-2d}$ as $\lambda \to 0^+$. Such processes are also called $I(d)$. (If $d = 0$, we say that $\{X_t\}$ has *short memory.*)

Though the process will no longer be weakly stationary with $|d| \geq 0.5$ and therefore the spectral density will not exist, this generalizes to a "pseudo-spectral density" for all $-1 \leq d \leq 1$. Then, we have:

- $d = 0$: Usual ARMA models
- $-1 < d \leq -1/2$: Stationary, mean-reverting processes that are not invertible.
- $-1/2 < d < 1/2$: Stationary and invertible processes
- $1/2 \leq d < 1$: Non-stationary but mean-reverting processes
- $d = 1$: Non-stationary and not mean-reverting

(A process is *mean-reverting* if the effect of a shock will disappear eventually.) The difference of a long memory process with memory parameter d has a memory parameter $d - 1$. In general, higher memory parameters correspond to smoother series.

In the presence of long memory,

- Autocovariances decay as $c_r \sim k|r|^{2d-1}$ as $|r| \to \infty$. This is called *hyperbolic decay.* In this case, the autcovariances are not summable if $d > 0$ (this also shows that $f(0) = \infty$ in this case).
- Forecasts tend to revert to the unconditional mean slowly.

Note that long memory affects low frequencies and long autocovariances – that means that long memory is a long-term effect.

Theorem 4.1 *$Var(\overline{X}_n) \sim k_1 n^{2d-1}$ as $n \to \infty$, if $d > -1$. (This means that long memory processes do not obey the Central Limit Theorem, and, when $d > 0$, confidence intervals are too narrow.)*

Proof Assume that $E(X_t) = 0$ and $d > 0$. Then,

$$\begin{aligned}
I(0) &= \frac{1}{2\pi n}|\sum_{t=0}^{n-1} X_t|^2 \\
&= \frac{n}{2\pi}\overline{X}^2 \\
&= \frac{1}{2\pi}\sum_{|r|<n} \hat{c}_r \\
Var(\overline{X}) &= E(\overline{X}^2) = E(\frac{1}{n}\sum_{|r|<n} \hat{c}_r) \\
&= \frac{1}{n}\sum_{|r|<n} E(\hat{c}_r) \\
&= \frac{1}{n}\sum_{|r|<n} (1 - \frac{|r|}{n})c_r \\
&\sim \frac{1}{n}\sum_{|r|<n} (1 - \frac{|r|}{n})kr^{2d-1} \\
&= \frac{2k}{n}\sum_{r=1}^{n}(1 - \frac{|r|}{n})r^{2d-1} \\
&= n^{2d-1}\frac{2k}{n}\sum_{r=1}^{n-1}(1 - \frac{r}{n})(\frac{r}{n})^{2d-1} \\
&\approx n^{2d-1}(2k)\int_0^1 (1-x)x^{2d-1}dx \\
&\propto n^{2d-1}
\end{aligned}$$

∎

In estimation, negative values of d require tapering to achieve good estimates – there tends to be leakage near the 0 frequency. (In general, if $f(\lambda)$ has a large range, then tapering helps control leakage.)

4.3 ARFIMA(0,d,0) Models

Definition The *Gamma function* is defined by $\Gamma(p) = \int_0^\infty x^{p-1}e^x dx$ for $p > 0$. For $p \leq 0$, we define $\Gamma(p+1) = p\Gamma(p)$; with this extension, $\Gamma(p)$ is defined everywhere except the negative integers. $\Gamma(p+1) = p!$ for positive integers.

Definition We define $\Delta^d = (1-B)^d$ by

$$\Delta^d = (1-B)^d = \sum_{j=0}^{\infty}(-1)^j \binom{d}{j} B^j = \sum_{j=0}^{\infty} \pi_j B^j$$

where $\binom{d}{j} = \frac{1}{j!}d(d-1)...(d-j+1)$ and $\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(d)}$.

Definition A time series, $\{X_t\}$ is $ARFIMA(0,d,0)$ if $\Delta^d X_t = \epsilon_t$ where ϵ_t is white noise. That is, $\{X_t\}$ is fractionally integrated white noise.

This is an $AR(\infty)$ representation: $\sum_{j=0}^{\infty} \pi_j X_{t-j} = \epsilon_t$. Stirling's Formula states that as $p \to \infty$, $\Gamma(p) \sim \sqrt{2\pi}e^{-p+1}(p-1)^{p-\frac{1}{2}}$. This means that as $j \to \infty$, $\pi_j \sim j^{d-1}/\Gamma(-d)$, which is hyperbolic decay. (In fact, if $d > 0$, the π_j always decay faster than j^{-d-1}, and if $d < 0$, the π_j decay more slowly.)

Similarly, writing $X_t = \Delta^{-d}\epsilon_t = \psi_j \epsilon_{t-j}$, where $\psi_j = (-1)^j\binom{-d}{j} \sim \frac{j^{d-1}}{\Gamma(d)}$. Notice that $\sum_{j=0}^{\infty} \psi_j^2 \leq c\sum j^{2d-2} < \infty$ if $d < 0.5$, and this process is weakly stationary if $d \in (-0.5, 0.5)$.

The spectral representation and spectral density of an $ARFIMA(0,d,0)$ process are:

$$\begin{aligned} X_t &= (1-B)^{-d}\epsilon_t \\ &= \int e^{it\lambda}(1-e^{-i\lambda})^{-d}dZ_\epsilon(\lambda) \\ f_x(\lambda) &= \frac{\sigma^2}{2\pi}|1-e^{-i\lambda}|^{-2d} \\ &= \frac{\sigma^2}{2\pi}|e^{i\lambda/2}-e^{-i\lambda/2}|^{-2d}|e^{-i\lambda/2}|^{-2d} \\ &= \frac{\sigma^2}{2\pi}|2\sin(\lambda/2)|^{-2d}(1) \\ &\sim \frac{\sigma^2}{2\pi}\lambda^{-2d} \end{aligned}$$

as $\lambda \to 0$ since $\sin x \approx x$ as $x \to 0$. This shows that the long memory parameter of this process really is d.

Notice that the optimal linear forecasts for these models depend on all past values; for practical purposes, we must truncate. In addition, there is no simple updating formula for forecasts, which can make forecasting computationally intensive.

4.4 ARFIMA(p,d,q)

Definition A zero mean process, $\{X_t\}_{t=-\infty}^{\infty}$ in discrete time obeys a *fractional ARIMA(p,d,q)* model if $\phi(B)\Delta^d X_t = \theta(B)\epsilon_t$, where ϵ_t is white noise, $\phi(B)$ is a lag polynomial of order p, $\theta(B)$ is a lag polynomial of order q, and $0 < |d| < 0.5$.

This process is stationary and invertible if all the roots of $\theta(B)$ and $\phi(B)$ lie outside the unit circle. Because polynomials commute, this is equivalent to:

- $\Delta^d X_t$ is $ARMA(p,q)$.
- $\phi(B)X_t = \theta(B)(\Delta^{-d}\epsilon_t)$, and X_t is an $ARMA(p,q)$ process driven by fractionally integrated white noise ($ARFIMA(0,d,0)$).

The $MA(\infty)$ and $AR(\infty)$ representations can be found by solving:

$$\begin{aligned}\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}(1-e^{-i\lambda})^{-d} &= \sum_{k=0}^{\infty}\psi_k e^{-ik\lambda} \\ \frac{\phi(e^{-i\lambda})}{\theta(e^{-i\lambda})}(1-e^{-i\lambda})^{d} &= \sum_{k=0}^{\infty}\pi_k e^{-ij\lambda}\end{aligned}$$

These allow us to calculate the best linear forecast.

The spectral representation and spectral density are:

$$\begin{aligned}X_t &= \int_{-\pi}^{\pi} e^{it\lambda}\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}(1-e^{-i\lambda})^{-d}dZ_\epsilon(\lambda) \\ f_X(\lambda) &= \frac{\sigma_\epsilon^2}{2\pi}\left|\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}\right|^2|1-e^{-i\lambda}|^2 \\ &= \frac{\sigma_\epsilon^2}{2\pi}\left|\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}\right|^2|2\sin(\lambda/2)|^{-2d} \\ &\sim \frac{\sigma_\epsilon^2}{2\pi}\left|\frac{\theta(1)}{\phi(1)}\right|^2|\lambda|^{-2d}\end{aligned}$$

as $\lambda \to 0$, and this process has long memory. The choice of d affects the low frequency components, while $\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}$ affects the higher frequency components. With the proper choice of ARMA parameters, there can be peaks elsewhere in the spectral density with a sort of power law decay on either side of them (even with different powers), corresponding to slowly evolving seasonal behavior.

Note that forecasting still requires the infinite past. If you want to avoid the truncation, solving the Yule-Walker equations allows you to calculate the best linear forecast based on up to n observations.

4.5 Parameter estimation

4.5.1 Semi-parametric long memory estimation

Suppose $f(\lambda) = |1-e^{-i\lambda}|^{-2d}f^*(\lambda)$, where $|d| < 0.5$ and $f^*(\lambda)$ is a short memory process; that is, $f^*(\lambda)$ is positive, finite, and continuous in a neighborhood of 0. If we only wish to estimate the long memory parameter, then $f^*(\lambda)$ is a nuisance parameter, and is also subject to misspecification (which will bias estimates of d).

If we assume that $\frac{I_j}{f_j} \sim Exponential$ are independent and identically distributed, then $\log(\frac{I_j}{f_j})$ has mean $-C = -0.577216$ (negative Euler's constant) and standard deviation $\frac{\pi^2}{6}$. Let $\epsilon_j = \log(\frac{I_j}{f_j}) + C$. Then, the ϵ_j are independent and identically distributed with mean 0, and we have:

$$log(I_j) = (\log(f_j^*) - C) - 2d\log|1-e^{-i\omega_j}| + \epsilon_j$$

If we estimate only using $j = 1, ..., M$ (so that estimation is mean-invariant), where $M \to \infty$ and $\frac{M}{n} \to 0$, then we assume that $\log(f_j^*) \approx \log(f_0^*)$ for all $j = 1, ..., M$. In addition, notice that $|1 - e^{-i\omega_j}| = |2\sin(\frac{\omega_j}{2})|$. Then, with this choice of M, we have:

$$\log I_j \approx Constant - 2d \log |2 \sin \frac{\omega_j}{2}| + \epsilon_j$$

The estimate from this regression, $\hat{d}$, is called the *Geweke-Porter-Hudak (GPH) estimator*. Furthermore, $|2\sin(\frac{\omega_j}{2})| \approx |\omega_j| \propto |j|$, and we may also use the regression:

$$\log I_j \approx Constant - 2d \log |\omega_j| + \epsilon_j$$

Both of these estimators are robust against misspecification of the short-memory part of the process.

Theorem 4.2 *Suppose $M \to \infty$ and $\frac{M}{n} \to 0$. Then, $\sqrt{M}(\hat{d}_{GPH} - d)$ converges in distribution to $Normal(0, \frac{\pi^2}{24})$. Thus, for large samples, $E(\hat{d}_{GPH}) = d$ and $Var(\hat{d}_{GPH}) = \frac{\pi^2}{24M}$.*

Note that holding M fixed at some constant leads to bias (and isn't consistent). In general, choosing $M \propto n^{4/5}$ is optimal; the constant depends on $f^*(\lambda)$. If there is additional noise in the periodogram (such as in the stochastic volatility model), the exponent should be closer to 0.

Even though the spectral density is not defined above $d > 0.5$, we may define a non-integrable "pseudo-spectrum" of the same form for higher d. If $d > 0.75$, tapering and a correction to the standard errors is necessary, but the method holds for $d < 1.5$.

Note that this method is semi-parametric. This means that the specific model (aside from long memory) doesn't matter, but it also means that the estimates are not as tight as if the correct parametric model were used.

4.5.2 Parametric Long Memory Estimation

For an $ARFIMA(p, d, q)$ model (with p, q known), maximum likelihood estimation can be used to estimate $\Theta(d, \theta(B), \phi(B), \sigma_\epsilon^2)$ simultaneously, if we assume that the white noise is Gaussian. ("Quasi-Gaussian Maximum Likelihood Estimation" makes this assumption and tries to show that the results apply even if the Gaussian assumption is relaxed.) Despite the odd asymptotic behavior of the variance of the mean, $\sqrt{n}(\hat{\Theta} - \Theta_0) \to_D Normal(0, \Sigma_\Theta)$. However, we cannot estimate the mean as one of the parameters and have this hold. This method involves inverting an $n \times n$ matrix, so it can be slow. It also fails when $d \geq 1/2$, since the covariance matrix no longer exists.

Definition The *Whittle estimator* is given by:

$$-2\log(likelihood) \approx \sum_{j=1}^{n/2} (\log(f_\Theta(\omega_j)) + \frac{I(\omega_j)}{f_\Theta(\omega_j)}$$

where $I(\omega_j)$ is the periodogram of the data (and does not depend on the parameter estimates) and $f_\Theta(\omega_j) = |\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}|^2 |e^{i\lambda} - e^{-i\lambda}|^{-2d}$.

Theorem 4.3 *The Whittle estimator is asymptotically equal to the maximum likelihood estimator.*

Proof Let $X = (X_0, ..., X_{n-1})^T$ where $\{X_t\}$ is Gaussian with zero mean, has spectral density $f_\Theta(\omega)$, and covariance matrix $\Sigma_{n,\Theta} = E(XX^T)$. Then we have the likelihood:

$$-2l(\Theta) = n\log(2\pi) + \log|\Sigma_{n,\Theta}| + X^T\Sigma_{n,\Theta}X$$

$\Sigma_{n,\Theta}$ is Toeplitz, because the lag r covariance is constant along each diagonal (by stationarity). As $n \to \infty$, the eigenvectors of a Toeplitz matrix are approximately $v_j = \frac{1}{\sqrt{n}}\{\exp(-it\omega_j)\}_{t=0}^{n-1}$ for $j = 0, ..., n-1$. This is an orthonormal basis. The corresponding eigenvalues are $2\pi f_\Theta(\omega_j)$. Let V be the matrix with the v_j's down the columns. Let Λ be the matrix with $2\pi f_\Theta(\omega_j)$ down the diagonal. Then V is a unitary matrix, and $\Sigma_{n,\Theta} \approx V\Lambda V^*$. We may rewrite the log likelihood as:

$$-2l(\Theta) = m\log(2\pi) + \log(|det(\Lambda|)) + (X^TV\Lambda^{-1/2})(X^TV\Lambda^{-1/2})^*$$

Note that X^TV is a row vector containing the DFT's, and $det(\Lambda) = \prod_{i=0}^{n-1} 2\pi f_\Theta(\omega_j)$, so that we find that:

$$-2l(\Theta) = 2n\log(2\pi) + \sum_{j=0}^{n-1}\log(f_\Theta(\omega_j)) + \sum_{j=0}^{n-1} I_j/f_\Theta(\omega_j)$$

We leave off $j = 0$ since $f_\Theta(0)$ is infinite for long memory time series, and this also makes the estimator invariant to the mean. ∎

The Whittle estimator also works from $d \geq 1/2$, through the use of a pseudo-spectral density.

Both of these estimators will be biased for d if p and q are wrong (this is one reason that the semiparametric method is used instead). We may also use the local Whittle estimator to estimate d (ignoring the short memory part): $-2\log(likelihood) \approx \sum_{j=1}^{m}(\log(f_\Theta(\omega_j)) + \frac{I(\omega_j)}{f_\Theta(\omega_j)})$, where $m/n \to 0$ but $m \to \infty$. Using this estimator, $\sqrt{m}(\hat{d}_{LW} - d_0) \to_D Normal(0, 0.25)$ (as opposed to the log periodogram estimator, which has variance $\frac{\pi^2}{24}$).

4.6 Continuous Time Long Memory

Definition A Gaussian stochastic process, $V_H(t)$, is *fractional Brownian motion* with *Hurst index* H if $0 < H < 1$ and, for all t_1, t_2, $Var(V_H(t_2) - V_H(t_1)) \propto |t_2 - t_1|^{2H}$.

This is regular Brownian motion if $H = 0.5$.

Let $X_t = V_H(t) - V_h(t-1)$; these are called *increments*. Note that the process $\{X_t\}$ is strictly stationary, and

$$Var(\bar{X}) = Var(\frac{1}{n}(V_h(n) - V_h(0))) \propto n^{2H-2}$$

so that $\{X_t\}$ is a long-memory process with memory parameter $d_{diff} = H - 1/2$. (The memory parameter for the levels is $d_{levels} = H + 1/2$, so the levels of fractional Brownian motion are non-stationary.)

The "derivative" (limit of the changes) of fractional Brownian motion is called *fractional Gaussian noise* and is a continuous time analog of $ARFIMA(0, d, 0)$.

For any $r > 0$, $V_H(rt)$ has the same distribution as $r^H V_H(t)$. Thus, the process does not change (except for the scale) as we zoom in. In fact, this process is a fractal with dimension $D = \log N^{2-H} / \log N = 2 - H$, which is between 1 and 2.

The pseudo-spectral density of this process is $f_H(\omega) \propto |\omega|^{-2H+1} = |\omega|^{-2d_{levels}}$ on $(0, \infty)$, and the continuous-time periodogram ($\int_0^1 x(t)e^{-it\omega}dt$) is still a good estimate of the pseudo-spectral density. If we only observe the continuous time process in discrete time, the the periodogram will not have as clear a shape because of aliasing.

5 Linear Prediction

Let $\{X_t\}_{t=-\infty}^{\infty}$ be a mean zero weakly stationary process with a spectral density, f. For a given lead time, ν, we want to find a linear combination, $\hat{X}_{t+\nu}$ of $X_t, X_{t-1}, ...$ to minimize the forecasting error, $E((X_{t+\nu} - \hat{X}_{t+\nu})^2)$. Let M_t^X be the subspace of M^X generated by $X_t, X_{t-1}, ...$ (this is called the *linear past* of the time series). Minimizing the squared forecast error is equivalent to minimizing the (squared) distance between $X_{t+\nu}$ and this subspace; that is, $\hat{X}_{t+\nu}$ is the orthogonal projection of $X_{t+\nu}$ onto this subspace. That means that:

- $\hat{X}_{t+\nu} \in M_t^X$
- $X_{t+\nu} - \hat{X}_{t+\nu} \perp M_t^X$, which means that $E((X_{t+\nu} - \hat{X}_{t+\nu})Y) = 0$ for all $Y \in M_t^X$

Since the forecast is a linear combination of present and past values, $\hat{X}_{t+\nu} = \sum_{k=0}^{\infty} d_k^{(\nu)} X_{t-k}$, this defines a linear filter and a transfer function, $D(\lambda) = \sum_{k=0}^{\infty} d_k^{(\nu)} e^{-i\lambda k}$.

Theorem 5.1 Wold's Theorem *Any weakly stationary process, $\{X_t\}$, which is not perfectly linearly predictable can be written as $X_t = U_t + V_t$ where:*

- *$\{U_t\}$ and $\{V_t\}$ are uncorrelated.*
- *$\{U_t\}$ has a one-sided moving average representation, $U_t = \sum_{k=0}^{\infty} a_k \epsilon_{t-k}$ with $a_0 = 1$ and $M_t^{\epsilon} = M_t^U$*

- $\{V_t\}$ *is perfectly predictable; that is,* $M_t^V = M_s^V$ *for all* s, t.

Definition Given a spectral density, f, for $\{X_t\}$, a *spectral factorization* is an expression of the form $f(\lambda) = \frac{\sigma^2}{2\pi}|A(\lambda)|^2$, where $A(\lambda) = \sum_{k=0}^{\infty} a_k \exp(-i\lambda k)$ is a one-sided linear filter mapping a white noise process to $\{X_t\}$.

If we do not require that $A(\lambda)$ is one-sided, then it is not unique (for example, $\overline{A(\lambda)}$ works as well). If $|A(\lambda)| > 0$, then we may use the reciprocal of this filter to find an $AR(\infty)$ representation of $\{X_t\}$ as well.

Theorem 5.2 Kolgomorov's Formula *If* $\{X_t\}$ *can be written as a one-sided* $MA(\infty)$ *process, then the variance of this white noise process (which equals the variance of the one-step-ahead forecast error) is given by*

$$\sigma^2 = 2\pi \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(f_X(\lambda))d\lambda\right)$$

Theorem 5.3 *We may find a spectral factorization if and only if* $\sigma^2 = 2\pi \exp(\frac{1}{2\pi}\int_{\pi}^{\pi} \ln(f_X(\lambda))d\lambda) > 0$.

$\sigma^2 = 0$ when the integral is $-\infty$. This may occur if $f(\lambda) = 0$ for a range of λ; if this occurs, the part of the process corresponding to that range is perfectly predictable. The integral may be $-\infty$ in other cases as well.

6 Non-linear models and prediction

Two random variables, X and Y are *uncorrelated* if Y cannot be predicted by a linear function of X. X and Y are *independent* if no function of Y can be predicted by any function of X. Independent implies uncorrelated. For multivariate normal random variables, uncorrelated implies independent.

The optimal (not-necessarily-linear) forecast is given by the conditional mean, $E(X_{t+h}|X_t, \epsilon_t, X_{t-1}, \epsilon_{t-1}, ...) = E(X_{t+h}|\Psi_t)$.

Definition A process, $\{e_t\}$ is *strict white noise* if $E(e_t) = 0$, $Var(e_t) = \sigma^2 > 0$, and the e_t are independent and identically distributed.

Definition A time series is *linear* if it may be written as $X_t = \sum_{k=0}^{\infty} a_k e_{t-k}$, where e_t is strict white noise.

Notice that any weakly stationary process has an $MA(\infty)$ representation, but only linear processes have $MA(\infty)$ representations involving strict white noise. Non-linear time series may have optimal forecasts that are not linear combinations of past errors and observations; however, the best forecasts of the level of a non-linear time series may still be linear (forecasts of other functions of the time series might not be linear, though).

Definition $\{\epsilon_t\}$ is a *martingale difference* if $E(\epsilon_{t+h}|\epsilon_t, \epsilon_{t-1}, ...) = 0$ for all $h > 0$.

Theorem 6.1 *A martingale difference sequence (with constant variance) is white noise.*

Theorem 6.2 *The best possible forecast equals the best linear forecast if and only if the white noise sequence is a martingale difference.*

Definition The time series $\{X_t\}$ is a *martingale* if $E(X_t|X_{t-1},...) = X_{t-1}$.

Some non-linear models include:

- Bilinear models: $X_t + \sum_{k=1}^{p} a_k X_{t-k} = e_t + \sum_{k=1}^{q} b_k e_{t-k} + \sum_{i=1}^{q}\sum_{j=1}^{p} c_{ij} e_{t-i} X_{t-j}$
- Threshold autoregression: $X_t = \begin{cases} a^{(1)} X_{t-1} + e_t^{(1)}, & X_{t-1} < d \\ a^{(2)} X_{t-1} + e_t^{(2)}, & X_{t-1} \geq d \end{cases}$
- General non-linear autoregression: $X_t = \lambda(X_{t-1}) + \epsilon_t$ where $\lambda()$ is a non-linear function.

7 Volatility Models

7.1 ARCH and GARCH Models

Definition A white noise series, $\{\epsilon_t\}$ is $GARCH(p,q)$ if:

$$\begin{aligned} \epsilon_t|\Psi_{t-1} &\sim Normal(0, h_t) \\ h_t &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \end{aligned}$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$, and $\sum \alpha_i + \sum \beta_j < 1$ (for weak stationarity). Equivalently, we may write $\epsilon_t = \sqrt{h_t} e_t$, where e_t is standard Gaussian white noise. If $p = 0$, this is called an $ARCH(q)$ model.

In this model, the variance of future shocks increases if past shocks were large (this is *observation-driven* since the present variance depends on past observations of shocks). Since the conditional variances are predictable from past observations, this is a non-linear model, but the $\{\epsilon_t\}$ are martingale differences. Notice that volatility is more persistent in a GARCH model than in an ARCH model.

Note that this is a model only for white noise. We may use this as the white noise for an ARMA process. In this case, the point forecasts are the same as with any other ARMA model, but the one-step-ahead confidence interval is now $\hat{Z}_{t+1} \pm z_{\alpha/2}\sqrt{h_{t+1}}$.

If $\{e_t\}$ is strict white noise, then the asymptotic standard errors for sample autocorrelations are $1/\sqrt{n}$. Suppose $\epsilon_t \sim ARCH(q)$ and $E(\epsilon_t^4) < \infty$. Then, the sample autocorrelations, $\hat{\rho}_r = \hat{c}_r/\hat{c}_0$, are asymptotically independent with $Var(\hat{\rho}_r) \approx \frac{1}{n}(1 + \gamma_r(1 - \sum_{i=1}^{q} \alpha_i^2)/\omega^2)$, where $\gamma_r = Cov(\epsilon_t^2, \epsilon_{t-r}^2)$. In particular,

for an $ARCH(1)$ process, the asymptotical variance is $Var(\hat{\rho}_r) \approx \frac{1}{n}(1+\frac{2\alpha_1^r}{1-3\alpha_1^2})$, which means that the standard errors are bigger than one would calculate for white noise. In the $ARCH(1)$ case, $E(\epsilon_t^4) < \infty$ if and only if $\alpha_1 < \frac{1}{\sqrt{3}}$. If the fourth moment is infinite, then the variance of squared returns is infinite, and the ACF's converge in distribution, not to a number.

We may estimate the coefficients of an $ARCH(q)$ model using the Yule-Walker equations. We may also use maximum likelihood estimation for general GARCH models:

$$\begin{aligned} L(\epsilon_1, ..., \epsilon_n|\Psi_0, \theta) &= f(\epsilon_1|\Psi_0, \theta)f(\epsilon_2|\Psi_1, \theta)...f(\epsilon_n|\Psi_{n-1}, \theta) \\ &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi h_t}} \exp(-\frac{1}{2h_t}\epsilon_t^2) \\ -2\log L(\theta) &= \sum_{t=1}^{n} \log(h_t) + \sum_{t=1}^{n} \epsilon_t^2/h_t^2 \end{aligned}$$

(We cut off the initial q observed errors and initialize h_0 in order to be able to estimate this.)

Theorem 7.1 Engle's Criterion *An $ARCH(q)$ process, $\{\epsilon_t\}$ is weakly stationary if and only if $P(Z) = 1 - \alpha_1 Z - ... - \alpha_q Z^q$ has all its roots outside the unit circle.*

Theorem 7.2 Milhoj's Criterion *If $\sum_{i=1}^{q} \alpha_i < 1$, the corresponding $ARCH(q)$ process is weakly stationary.*

If $\alpha_i \geq 0$ for all i (as they are in an ARCH process), these two conditions are equivalent.

To deal with long memory in volatility, we may use FIGARCH, where $h_t = \omega + \sum_{k=1}^{\infty} \alpha_k \epsilon_{t-k}^2$, where the α_k are the $AR(\infty)$ coefficients of an $ARFIMA(1, d, 0)$ model. However, because of the restrictions on the ARCH parameters (and the importance of the fourth moment), no FIGARCH process with both long memory and a finite unconditional variance has been shown to exist.

7.2 Stochastic Volatility Models

Definition A *stochastic volatility model* models returns as $\epsilon_t = e^{h_t/2} e_t$, where h_t is any stationary Gaussian process (which is not observation-driven) and the e_t are strict white noise.

To actually estimate this, we must put more structure on h_t (since it is a latent process and cannot be observed directly). Often, part of the structure is that h_t and e_t are contemporaneously independent. In this case, $\log(\epsilon_t^2) = h_t + log(e_t^2)$, where $log(e_t^2)$ is just extra error. Also by the independence of h_t and e_t, the spectral density of ϵ_t^2 is a constant plus the spectral density of h_t.

To get long memory in volatility, h_t may be a latent long memory process (which does not depend on the observed shocks). In this case, we may still

regress the logarithm of the periodogram of squared returns on the log of the index to estimate d, but there will be more noise in the estimation.

7.3 Realized Volatility

We may measure realized volatility using high frequency data. The realized volatility for a day is the sum of squared returns over all the short intervals in a day. (This assumes no intra-day effects.) Using this realized volatility, one may measure the properties of volatility; empirically, realized volatility seems to have a long memory parameter of $d = 0.4$.

This also gives a simple forecasting method: $\hat{\sigma}_{t+1} = (1-B)^{\hat{d}}\sigma_t$. (This should be checked for additional ARMA structure, though.)

At a very high frequency, stock prices are jumps caused by individual trades (this is the difference between "trading time" and "clock time"). *Autoregressive conditional duration* studies the relationship between price changes and the duration between trades; more activity may make volatility higher.

8 Chaos and Fractals

A simple chaos model for a time series defines a time series by $X_t = f(X_{t-1})$, where f is a non-random (and non-linear) function. In this case, a time series is completely determined by f and its initial condition, X_0. For many f, even rounding error can lead to chaotic effects, so these time series can seem random. (Adding a random component into the function or choosing a random initial value can add even more randomness.)

Definition For a given function, f, the *Lyapunov exponent* is defined as $\lambda = \lim_{n\to\infty} \frac{1}{n} \log |\frac{d}{dx} f^n(x)|$, where $f^n(x) = f \circ f \circ ... \circ f(x)$.

Given two initial conditions, x_0 and x_0', the distance initially grows as $(x_t - x_t') \sim (x_0 - x_0')e^{\lambda t}$, and paths diverge exponentially fast.

Definition Suppose an initial value, x_0, is chosen from randomly from a probability distribution, F. If the probability distribution of $f(x_0)$ equals the probability distribution of x_0, then F is called an *invariant distribution*.

Definition Given a path in the plane, suppose we cover it with boxes with sides of length $1/N$. Let $N(L)$ be the number of boxes of side L required to cover the path. The *dimension* of a curve is given by $D = \lim_{N\to\infty} \frac{1}{N} \log N(1/N)$.

Definition A *fractal* is a set in space with fractional dimension.

9 Other types of Spectra

9.1 Cross-Spectra

Definition Consider two jointly weakly stationary time series, $X_t = \int e^{i\lambda t} dZ_X(\lambda)$ and $Y_t = \int e^{i\lambda t} dZ_Y(\lambda)$. The *cross-spectrum*, $f_{XY}(\lambda)$ is given by:

$$f_{XY}(\lambda)d\lambda = E(dZ_X(\lambda)\overline{dZ_Y(\lambda)})$$

As before, there is a relationship between autocorrelations and spectra:

$$f_{XY}(\lambda) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} c_{XY,r} \exp(-i\lambda r)$$

where $c_{XY,r} = E(X_t Y_{t-r})$. Note that the cross-spectrum may be complex-valued.

Definition The *coherence* of two time series is given by:

$$r_{XY}(\lambda) = \frac{|f_{XY}(\lambda)|}{\sqrt{f_{XX}(\lambda) f_{YY}(\lambda)}}$$

This is the correlation at frequency λ. The corresponding *phase*, θ_{XY}, measures the difference in the angles of $f_{XY}(\lambda)$ and $\sqrt{f_{XX}(\lambda) f_{YY}(\lambda)}$.

Definition The *cross-periodogram* is given by:

$$I_{XY}(\lambda) = \frac{n}{2\pi} J_X(\lambda)\overline{J_Y(\lambda)} = \frac{1}{2\pi} \sum_{|r|<n} \hat{c}_{XY,r} \exp(-ir\lambda)$$

As before, the values of the cross-periodogram at different frequencies are roughly independent and we estimate the cross-spectrum by smoothing the cross-periodogram.

9.2 Bispectrum

Definition Let $\{X_t\}$ be a strictly stationary process with zero mean. We define the third order moment function by $C(r,s) = E(X_t X_{t+r} X_{t+s})$. We define the *bispectrum*, $f_X(\lambda_1, \lambda_2)$, as the contribution of the frequencies λ_1, λ_2 to the third moment of X_t. That is,

$$C(r,s) = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{ir\lambda_1} e^{is\lambda_2} f(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2$$

As before, we may invert this to find that $f(\lambda_1, \lambda_2) = \frac{1}{(2\pi)^2} \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} C(r,s) e^{-ir\lambda_1} e^{-is\lambda_2}$. Note that $C(0,0) = E(X_t^3)$ is the skewness. The bispectrum does not have to

be positive or real-valued. We may use this to find something about the spectral representation for strictly stationary time series:

$$\begin{aligned} C(r,s) &= E((\int_{-\pi}^{\pi} e^{i\lambda_1 t} dZ(\lambda_1))(\int_{-\pi}^{\pi} e^{i\lambda_2 t} dZ(\lambda_2))(\int_{-\pi}^{\pi} e^{i\lambda_3 t} dZ(\lambda_3))) \\ &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{it(\lambda_1+\lambda_2+\lambda_3)} E(dZ(\lambda_1)dZ(\lambda_2)dZ(\lambda_3)) \end{aligned}$$

Since the third moment cannot depend on time for a strictly stationary series, this means that $E(dZ(\lambda_1)dZ(\lambda_2)dZ(\lambda_3)) = 0$ unless $\lambda_1+\lambda_2+\lambda_3 = 0$. This means we may write $C(r,s) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{i\lambda_1 r} e^{i\lambda_2 s} E(dZ(\lambda_1)dZ(\lambda_2)dZ(-\lambda_1-\lambda_2))$.

To estimate the bispectrum, we may use $J(\omega_j)J(\omega_k)J(-\omega_j-\omega_k)$. Since all the higher order moments of a Gaussian random variable are 0, the bispectrum (and all higher versions of this) must be 0 for a Gaussian process. For a linear process, the bispectrum is constant (and equal to $\frac{\mu_3^2}{2\pi(\sigma_\epsilon^2)^3}$, where μ_3 is the skewness). Let $T(\lambda_1,\lambda_2) = \frac{|f(\lambda_1,\lambda_2)|^2}{|f(\lambda_1)f(\lambda_2)f(\lambda_3)|}$; comparing an estimate of this quantity to the values above gives a test for linearity and Gaussianity. (This may be extended to spectra based on even higher moments, which can also be used for tests of linearity and Gaussianity.)

10 Cointegration

Definition We say that a process is *integrated of order d*, or $I(d)$, if its k^{th} difference has spectral density $f(\lambda) \sim C|\lambda|^{-2(d-k)}$ as $\lambda \to 0$, where k is any non-negative integer with $d-k \leq \frac{1}{2}$.

Definition Suppose we have two time series, $\{X_t\}$ and $\{Y_t\}$, each of which is $I(d)$. Suppose there is some β such that $U_t = Y_t - \beta X_t$ is $I(d_U)$ with $d_U < d$. Then, we say that X_t and Y_t are *fractionally cointegrated*. If $d = 1$ and $d_U = 0$, then we say that X_t and Y_t are *classically cointegrated*. The *degree of correlation* is $d - d_U$.

The series U_t represents deviations from an equilibrium. With fractional cointegration, reversion to the equilibrium will be slower than with classical cointegration.

Testing for classical cointegration (ignoring the possibility of fractional cointegration):

- Test that $\{X_t\}$ and $\{Y_t\}$ have unit roots, using a Dickey-Fuller Test.
- Regress $\{Y_t\}$ on $\{X_t\}$ using ordinary least squares, and find $\hat{\beta}$.
- Test whether $\{Y_t - \hat{\beta}X_t\}$ has a unit root, using a Dickey-Fuller test.

Note that the Dickey-Fuller test will reject the null hypothesis of a unit root even if U_t has long memory with a long memory parameter between 0 and 1.

Testing for fractional cointegration:

- Difference X_t and Y_t enough times to remove any possible polynomial trends.
- Test that the long memory parameters of X_t and Y_t are equal.
- Regress the discrete Fourier transforms of the two series on each other: $J_{Y,j} = \beta J_{X,j}$. To exclude short-memory movements (and possibly a constant term) from the regression, use only $j = 1, ..., M$. Tapering the data first also helps.
- Test that the long memory parameter of $Y_t - \hat{\beta} X_t$ is lower than the original long memory parameters.