1 Models for time series

1.1 Time series data

A time series is a set of statistics, usually collected at regular intervals. Time series data occur naturally in many application areas.

- economics - e.g., monthly data for unemployment, hospital admissions, etc.
- finance - e.g., daily exchange rate, a share price, etc.
- environmental - e.g., daily rainfall, air quality readings.
- medicine - e.g., ECG brain wave activity every 2^{-8} secs.

The methods of time series analysis pre-date those for general stochastic processes and Markov Chains. The aims of time series analysis are to describe and summarise time series data, fit low-dimensional models, and make forecasts.

We write our real-valued series of observations as $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$, a doubly infinite sequence of real-valued random variables indexed by $\mathbb{Z}$.

1.2 Trend, seasonality, cycles and residuals

One simple method of describing a series is that of **classical decomposition**. The notion is that the series can be decomposed into four elements:

Trend (T_t) — long term movements in the mean;

Seasonal effects (I_t) — cyclical fluctuations related to the calendar;

Cycles (C_t) — other cyclical fluctuations (such as a business cycles);

Residuals (E_t) — other random or systematic fluctuations.

The idea is to create separate models for these four elements and then combine them, either additively

$$X_t = T_t + I_t + C_t + E_t$$

or multiplicatively

$$X_t = T_t \cdot I_t \cdot C_t \cdot E_t \,.$$

1.3 Stationary processes

1. A sequence $\{X_t, t \in \mathbb{Z}\}$ is **strongly stationary** or **strictly stationary** if

$$(X_{t_1}, \ldots, X_{t_k}) \stackrel{\mathcal{D}}{=} (X_{t_1+h}, \ldots, X_{t_k+h})$$

   for all sets of time points $t_1, \ldots, t_k$ and integer h.

2. A sequence is **weakly stationary**, or **second order stationary** if

(a) $\mathbb{E}(X_t) = \mu$, and

(b) $\text{cov}(X_t, X_{t+k}) = \gamma_k$,

where μ is constant and γ_k is independent of t.

3. The sequence $\{\gamma_k,\ k \in \mathbb{Z}\}$ is called the **autocovariance function**.

4. We also define
$$\rho_k = \gamma_k/\gamma_0 = \text{corr}(X_t, X_{t+k})$$
and call $\{\rho_k, k \in \mathbb{Z}\}$ the **autocorrelation function** (ACF).

Remarks.

1. A strictly stationary process is weakly stationary.

2. If the process is Gaussian, that is $(X_{t_1}, \ldots, X_{t_k})$ is multivariate normal, for all $t_1, \ldots, t_k$, then weak stationarity implies strong stationarity.

3. $\gamma_0 = \text{var}(X_t) > 0$, assuming X_t is genuinely random.

4. By symmetry, $\gamma_k = \gamma_{-k}$, for all k.

1.4 Autoregressive processes

The **autoregressive process** of order p is denoted AR(p), and defined by

$$X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \epsilon_t \tag{1.1}$$

where $\phi_1, \ldots, \phi_r$ are fixed constants and $\{\epsilon_t\}$ is a sequence of independent (or uncorrelated) random variables with mean 0 and variance σ^2.

The AR(1) process is defined by

$$X_t = \phi_1 X_{t-1} + \epsilon_t\,. \tag{1.2}$$

To find its autocovariance function we make successive substitutions, to get

$$X_t = \epsilon_t + \phi_1(\epsilon_{t-1} + \phi_1(\epsilon_{t-2} + \cdots)) = \epsilon_t + \phi_1\epsilon_{t-1} + \phi_1^2\epsilon_{t-2} + \cdots$$

The fact that $\{X_t\}$ is second order stationary follows from the observation that $\mathbb{E}(X_t) = 0$ and that the autocovariance function can be calculated as follows:

$$\gamma_0 = \mathbb{E}\left(\epsilon_t + \phi_1\epsilon_{t-1} + \phi_1^2\epsilon_{t-2} + \cdots\right)^2 = \left(1 + \phi_1^2 + \phi_1^4 + \cdots\right)\sigma^2 = \frac{\sigma^2}{1-\phi_1^2}$$

$$\gamma_k = \mathbb{E}\left(\sum_{r=0}^{\infty} \phi_1^r \epsilon_{t-r} \sum_{s=0}^{\infty} \phi_1^s \epsilon_{t+k-s}\right) = \frac{\sigma^2\phi_1^k}{1-\phi_1^2}\,.$$

There is an easier way to obtain these results. Multiply equation (1.2) by X_{t-k} and take the expected value, to give

$$\mathbb{E}(X_t X_{t-k}) = \mathbb{E}(\phi_1 X_{t-1} X_{t-k}) + \mathbb{E}(\epsilon_t X_{t-k}) \,.$$

Thus $\gamma_k = \phi_1 \gamma_{k-1}$, $k = 1, 2, \ldots$

Similarly, squaring (1.2) and taking the expected value gives

$$\mathbb{E}(X_t^2) = \phi_1 \mathbb{E}(X_{t-1}^2) + 2\phi_1 \mathbb{E}(X_{t-1}\epsilon_t) + \mathbb{E}(\epsilon_t^2) = \phi_1^2 \mathbb{E}(X_{t-1}^2) + 0 + \sigma^2$$

and so $\gamma_0 = \sigma^2/(1 - \phi_1^2)$.

More generally, the AR(p) process is defined as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t \,. \tag{1.3}$$

Again, the autocorrelation function can be found by multiplying (1.3) by X_{t-k}, taking the expected value and dividing by γ_0, thus producing the **Yule-Walker equations**

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}, \quad k = 1, 2, \ldots$$

These are linear recurrence relations, with general solution of the form

$$\rho_k = C_1 \omega_1^{|k|} + \cdots + C_p \omega_p^{|k|} \,,$$

where $\omega_1, \ldots, \omega_p$ are the roots of

$$\omega^p - \phi_1 \omega^{p-1} - \phi_2 \omega^{p-2} - \cdots - \phi_p = 0$$

and $C_1, \ldots, C_p$ are determined by $\rho_0 = 1$ and the equations for $k = 1, \ldots, p-1$. It is natural to require $\gamma_k \to 0$ as $k \to \infty$, in which case the roots must lie inside the unit circle, that is, $|\omega_i| < 1$. Thus there is a restriction on the values of $\phi_1, \ldots, \phi_p$ that can be chosen.

1.5 Moving average processes

The **moving average process** of order q is denoted MA(q) and defined by

$$X_t = \sum_{s=0}^{q} \theta_s \epsilon_{t-s} \tag{1.4}$$

where $\theta_1, \ldots, \theta_q$ are fixed constants, $\theta_0 = 1$, and $\{\epsilon_t\}$ is a sequence of independent (or uncorrelated) random variables with mean 0 and variance σ^2.

It is clear from the definition that this is second order stationary and that

$$\gamma_k = \begin{cases} 0, & |k| > q \\ \sigma^2 \sum_{s=0}^{q-|k|} \theta_s \theta_{s+k}, & |k| \le q \end{cases}$$

We remark that two moving average processes can have the same autocorrelation function. For example,

$$X_t = \epsilon_t + \theta\epsilon_{t-1} \quad \text{and} \quad X_t = \epsilon_t + (1/\theta)\epsilon_{t-1}$$

both have $\rho_1 = \theta/(1+\theta^2)$, $\rho_k = 0$, $|k| > 1$. However, the first gives

$$\epsilon_t = X_t - \theta\epsilon_{t-1} = X_t - \theta(X_{t-1} - \theta\epsilon_{t-2}) = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \cdots$$

This is only valid for $|\theta| < 1$, a so-called **invertible process**. No two invertible processes have the same autocorrelation function.

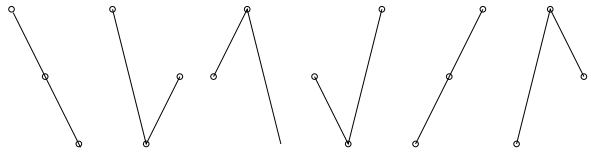
1.6 White noise

The sequence $\{\epsilon_t\}$, consisting of independent (or uncorrelated) random variables with mean 0 and variance σ^2 is called **white noise** (for reasons that will become clear later.) It is a second order stationary series with $\gamma_0 = \sigma^2$ and $\gamma_k = 0$, $k \neq 0$.

1.7 The turning point test

We may wish to test whether a series can be considered to be white noise, or whether a more complicated model is required. In later chapters we shall consider various ways to do this, for example, we might estimate the autocovariance function, say $\{\hat{\gamma}_k\}$, and observe whether or not $\hat{\gamma}_k$ is near zero for all $k > 0$.

However, a very simple diagnostic is the **turning point test**, which examines a series $\{X_t\}$ to test whether it is purely random. The idea is that if $\{X_t\}$ is purely random then three successive values are equally likely to occur in any of the six possible orders.



In four cases there is a turning point in the middle. Thus in a series of n points we might expect $(2/3)(n-2)$ turning points.

In fact, it can be shown that for large n, the number of turning points should be distributed as about $N(2n/3, 8n/45)$. We reject (at the 5% level) the hypothesis that the series is unsystematic if the number of turning points lies outside the range $2n/3 \pm 1.96\sqrt{8n/45}$.

2 Models of stationary processes

2.1 Purely indeterministic processes

Suppose $\{X_t\}$ is a second order stationary process, with mean 0. Its **autocovariance function** is

$$\gamma_k = \mathbb{E}(X_t X_{t+k}) = \mathrm{cov}(X_t, X_{t+k}), \quad k \in \mathbb{Z}.$$

1. As $\{X_t\}$ is stationary, γ_k does not depend on t.

2. A process is said to be **purely-indeterministic** if the regression of X_t on $X_{t-q}, X_{t-q-1}, \ldots$ has explanatory power tending to 0 as $q \to \infty$. That is, the residual variance tends to $\mathrm{var}(X_t)$.

   An important theorem due to Wold (1938) states that every purely-indeterministic second order stationary process $\{X_t\}$ can be written in the form

   $$X_t = \mu + \theta_0 Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots$$

   where $\{Z_t\}$ is a sequence of uncorrelated random variables.

3. A **Gaussian process** is one for which $X_{t_1}, \ldots, X_{t_n}$ has a joint normal distribution for all $t_1, \ldots, t_n$. No two distinct Gaussian processes have the same autocovariance function.

2.2 ARMA processes

The **autoregressive moving average process**, ARMA(p, q), is defined by

$$X_t - \sum_{r=1}^{p} \phi_r X_{t-r} = \sum_{s=0}^{q} \theta_s \epsilon_{t-s}$$

where again $\{\epsilon_t\}$ is white noise. This process is stationary for appropriate ϕ, θ.

Example 2.1
Consider the **state space model**

$$\begin{aligned} X_t &= \phi X_{t-1} + \epsilon_t \,, \\ Y_t &= X_t + \eta_t \,. \end{aligned}$$

Suppose $\{X_t\}$ is unobserved, $\{Y_t\}$ is observed and $\{\epsilon_t\}$ and $\{\eta_t\}$ are independent white noise sequences. Note that $\{X_t\}$ is AR(1). We can write

$$\begin{aligned} \xi_t &= Y_t - \phi Y_{t-1} \\ &= (X_t + \eta_t) - \phi(X_{t-1} + \eta_{t-1}) \\ &= (X_t - \phi X_{t-1}) + (\eta_t - \phi \eta_{t-1}) \\ &= \epsilon_t + \eta_t - \phi \eta_{t-1} \end{aligned}$$

Now ξ_t is stationary and $\text{cov}(\xi_t, \xi_{t+k}) = 0$, $k \geq 2$. As such, ξ_t can be modelled as a MA(1) process and $\{Y_t\}$ as ARMA$(1,1)$.

2.3 ARIMA processes

If the original process $\{Y_t\}$ is not stationary, we can look at the first order difference process

$$X_t = \nabla Y_t = Y_t - Y_{t-1}$$

or the second order differences

$$X_t = \nabla^2 Y_t = \nabla(\nabla Y)_t = Y_t - 2Y_{t-1} + Y_{t-2}$$

and so on. If we ever find that the differenced process is a stationary process we can look for a ARMA model of that.

The process $\{Y_t\}$ is said to be an **autoregressive integrated moving average process**, ARIMA(p,d,q), if $X_t = \nabla^d Y_t$ is an ARMA(p,q) process.

AR, MA, ARMA and ARIMA processes can be used to model many time series. A key tool in identifying a model is an estimate of the autocovariance function.

2.4 Estimation of the autocovariance function

Suppose we have data $(X_1, \ldots, X_T)$ from a stationary time series. We can estimate

- the mean by $\bar{X} = (1/T)\sum_1^T X_t$,
- the autocovariance by $c_k = \hat{\gamma}_k = (1/T)\sum_{t=k+1}^T (X_t - \bar{X})(X_{t-k} - \bar{X})$, and
- the autocorrelation by $r_k = \hat{\rho}_k = \hat{\gamma}_k/\hat{\gamma}_0$.

The plot of r_k against k is known as the **correlogram**. If it is known that μ is 0 there is no need to correct for the mean and γ_k can be estimated by

$$\hat{\gamma}_k = (1/T)\sum_{t=k+1}^T X_t X_{t-k}\,.$$

Notice that in defining $\hat{\gamma}_k$ we divide by T rather than by $(T-k)$. When T is large relative to k it does not much matter which divisor we use. However, for mathematical simplicity and other reasons there are advantages in dividing by T.

Suppose the stationary process $\{X_t\}$ has autocovariance function $\{\gamma_k\}$. Then

$$\text{var}\left(\sum_{t=1}^T a_t X_t\right) = \sum_{t=1}^T \sum_{s=1}^T a_t a_s \,\text{cov}(X_t, X_s) = \sum_{t=1}^T \sum_{s=1}^T a_t a_s \gamma_{|t-s|} \geq 0.$$

A sequence $\{\gamma_k\}$ for which this holds for every $T \geq 1$ and set of constants $(a_1, \ldots, a_T)$ is called a **nonnegative definite sequence**. The following theorem states that $\{\gamma_k\}$ is a valid autocovariance function if and only if it is nonnegative definite.

THEOREM 2.2 (Blochner) The following are equivalent.

1. There exists a stationary sequence with autocovariance function $\{\gamma_k\}$.
2. $\{\gamma_k\}$ is nonnegative definite.
3. The spectral density function,

$$f(\omega) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{ik\omega} = \frac{1}{\pi}\gamma_0 + \frac{2}{\pi} \sum_{k=1}^{\infty} \gamma_k \cos(\omega k)\,,$$

   is positive if it exists.

Dividing by T rather than by $(T-k)$ in the definition of $\hat{\gamma}_k$

- ensures that $\{\hat{\gamma}_k\}$ is nonnegative definite (and thus that it could be the autocovariance function of a stationary process), and
- can reduce the L^2-error of r_k.

2.5 Identifying a MA(q) process

In a later lecture we consider the problem of identifying an ARMA or ARIMA model for a given time series. A key tool in doing this is the correlogram.

The MA(q) process X_t has $\rho_k = 0$ for all k, $|k| > q$. So a diagnostic for MA(q) is that $|r_k|$ drops to near zero beyond some threshold.

2.6 Identifying an AR(p) process

The AR(p) process has ρ_k decaying exponentially. This can be difficult to recognise in the correlogram. Suppose we have a process X_t which we believe is AR(k) with

$$X_t = \sum_{j=1}^{k} \phi_{j,k} X_{t-j} + \epsilon_t$$

with ϵ_t independent of $X_1, \ldots, X_{t-1}$.

Given the data $X_1, \ldots, X_T$, the least squares estimates of $(\phi_{1,k}, \ldots, \phi_{k,k})$ are obtained by minimizing

$$\frac{1}{T} \sum_{t=k+1}^{T} \left(X_t - \sum_{j=1}^{k} \phi_{j,k} X_{t-j} \right)^2 .$$

This is approximately equivalent to solving equations similar to the Yule-Walker equations,

$$\hat{\gamma}_j = \sum_{\ell=1}^{k} \hat{\phi}_{\ell,k} \hat{\gamma}_{|j-\ell|}, \quad j = 1, \ldots, k$$

These can be solved by the **Levinson-Durbin recursion**:

Step 0. $\sigma_0^2 := \hat{\gamma}_0, \quad \hat{\phi}_{1,1} = \hat{\gamma}_1/\hat{\gamma}_0, \quad k := 0$

Step 1. Repeat until $\hat{\phi}_{k,k}$ near 0:

$$k := k + 1$$

$$\hat{\phi}_{k,k} := \left(\hat{\gamma}_k - \sum_{j=1}^{k-1} \hat{\phi}_{j,k-1}\hat{\gamma}_{k-j}\right) \Bigg/ \sigma_{k-1}^2$$

$$\hat{\phi}_{j,k} := \hat{\phi}_{j,k-1} - \hat{\phi}_{k,k}\hat{\phi}_{k-j,k-1}, \text{ for } j = 1, \ldots, k-1$$

$$\sigma_k^2 := \sigma_{k-1}^2(1 - \hat{\phi}_{k,k}^2)$$

We test whether the order k fit is an improvement over the order $k-1$ fit by looking to see if $\hat{\phi}_{k,k}$ is far from zero.

The statistic $\hat{\phi}_{k,k}$ is called the kth **sample partial autocorrelation coefficient** (PACF). If the process X_t is genuinely AR(p) then the population PACF, $\phi_{k,k}$, is exactly zero for all $k > p$. Thus a diagnostic for AR(p) is that the sample PACFs are close to zero for $k > p$.

2.7 Distributions of the ACF and PACF

Both the sample ACF and PACF are approximately normally distributed about their population values, and have standard deviation of about $1/\sqrt{T}$, where T is the length of the series. A rule of thumb it that ρ_k is negligible (and similarly $\phi_{k,k}$) if r_k (similarly $\hat{\phi}_{k,k}$) lies between $\pm 2/\sqrt{T}$. (2 is an approximation to 1.96. Recall that if $Z_1, \ldots, Z_n \sim N(\mu, 1)$, a test of size 0.05 of the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ rejects H_0 if and only if $\bar{Z}$ lies outside $\pm 1.96/\sqrt{n}$).

Care is needed in applying this rule of thumb. It is important to realize that the sample autocorrelations, $r_1, r_2, \ldots$, (and sample partial autocorrelations, $\hat{\phi}_{1,1}, \hat{\phi}_{2,2}, \ldots$) are not independently distributed. The probability that any one r_k should lie outside $\pm 2/\sqrt{T}$ depends on the values of the other r_k.

A 'portmanteau' test of white noise (due to Box & Pierce and Ljung & Box) can be based on the fact that approximately

$$Q'_m = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}r_k^2 \sim \chi_m^2 \,.$$

The sensitivity of the test to departure from white noise depends on the choice of m. If the true model is ARMA(p, q) then greatest power is obtained (rejection of the white noise hypothesis is most probable) when m is about $p + q$.

3 Spectral methods

3.1 The discrete Fourier transform

If $h(t)$ is defined for integers t, the discrete Fourier transform of h is

$$H(\omega) = \sum_{t=-\infty}^{\infty} h(t)e^{-i\omega t}, \quad -\pi \leq \omega \leq \pi$$

The inverse transform is

$$h(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega t} H(\omega)\, d\omega\,.$$

If $h(t)$ is real-valued, and an even function such that $h(t) = h(-t)$, then

$$H(\omega) = h(0) + 2\sum_{t=1}^{\infty} h(t)\cos(\omega t)$$

and

$$h(t) = \frac{1}{\pi} \int_{0}^{\pi} \cos(\omega t) H(\omega)\, d\omega\,.$$

3.2 The spectral density

The Wiener-Khintchine theorem states that for any real-valued stationary process there exists a **spectral distribution function**, $F(\cdot)$, which is nondecreasing and right continuous on $[0, \pi]$ such that $F(0) = 0$, $F(\pi) = \gamma_0$ and

$$\gamma_k = \int_{0}^{\pi} \cos(\omega k)\, dF(\omega)\,.$$

The integral is a Lebesgue-Stieltges integral and is defined even if F has discontinuities. Informally, $F(\omega_2) - F(\omega_1)$ is the contribution to the variance of the series made by frequencies in the range (ω_1, ω_2).

$F(\cdot)$ can have jump discontinuities, but always can be decomposed as

$$F(\omega) = F_1(\omega) + F_2(\omega)$$

where $F_1(\cdot)$ is a nondecreasing continuous function and $F_2(\cdot)$ is a nondecreasing step function. This is a decomposition of the series into a purely indeterministic component and a deterministic component.

Suppose the process is purely indeterministic, (which happens if and only if $\sum_k |\gamma_k| < \infty$). In this case $F(\cdot)$ is a nondecreasing continuous function, and differentiable at all points (except possibly on a set of measure zero). Its derivative $f(\omega) = F'(\omega)$ exists, and is called the **spectral density function**. Apart from a

multiplication by $1/\pi$ it is simply the discrete Fourier transform of the autocovariance function and is given by

$$f(\omega) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-ik\omega} = \frac{1}{\pi}\gamma_0 + \frac{2}{\pi} \sum_{k=1}^{\infty} \gamma_k \cos(\omega k)\,,$$

with inverse

$$\gamma_k = \int_0^{\pi} \cos(\omega k) f(\omega)\, d\omega\,.$$

Note. Some authors define the spectral distribution function on $[-\pi, \pi]$; the use of negative frequencies makes the interpretation of the spectral distribution less intuitive and leads to a difference of a factor of 2 in the definition of the spectra density. Notice, however, that if f is defined as above and extended to negative frequencies, $f(-\omega) = f(\omega)$, then we can write

$$\gamma_k = \int_{-\pi}^{\pi} \tfrac{1}{2} e^{i\omega k} f(\omega)\, d\omega\,.$$

Example 3.1

(a) Suppose $\{X_t\}$ is i.i.d., $\gamma_0 = \mathrm{var}(X_t) = \sigma^2 > 0$ and $\gamma_k = 0$, $k \geq 1$. Then $f(\omega) = \sigma^2/\pi$. The fact that the spectral density is flat means that all frequencies are equally present accounts for our calling this sequence **white noise**.

(b) As an example of a process which is not purely indeterministic, consider $X_t = \cos(\omega_0 t + U)$ where ω_0 is a value in $[0, \pi]$ and $U \sim U[-\pi, \pi]$. The process has zero mean, since

$$\mathbb{E}(X_t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega_0 t + u)\, du = 0$$

and autocovariance

$$\begin{aligned}
\gamma_k &= \mathbb{E}(X_t, X_{t+k}) \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\omega_0 t + u) \cos(\omega_0 t + \omega_0 k + u) du \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \tfrac{1}{2} \left[\cos(\omega_0 k) + \cos(2\omega_0 t + \omega_0 k + 2u)\right] du \\
&= \frac{1}{2\pi} \tfrac{1}{2} [2\pi \cos(\omega_0 k) + 0] \\
&= \frac{1}{2} \cos(\omega_0 k)\,.
\end{aligned}$$

Hence X_t is second order stationary and we have

$$\gamma_k = \frac{1}{2}\cos(\omega_0 k), \quad F(\omega) = \frac{1}{2} I_{[\omega \geq \omega_0]} \quad \text{and} \quad f(\omega) = \frac{1}{2}\delta_{\omega_0}(\omega)\,.$$

Note that F is a nondecreasing step function.

More generally, the spectral density

$$f(\omega) = \sum_{j=1}^{n} \frac{1}{2} a_j \delta_{\omega_j}(\omega)$$

corresponds to the process $X_t = \sum_{j=1}^{n} a_j \cos(\omega_j t + U_j)$ where $\omega_j \in [0, \pi]$ and $U_1, \ldots, U_n$ are i.i.d. $U[-\pi, \pi]$.

(c) The MA(1) process, $X_t = \theta_1 \epsilon_{t-1} + \epsilon_t$, where $\{\epsilon_t\}$ is white noise. Recall $\gamma_0 = (1 + \theta_1^2)\sigma^2$, $\gamma_1 = \theta_1 \sigma^2$, and $\gamma_k = 0$, $k > 1$. Thus

$$f(\omega) = \frac{\sigma^2(1 + 2\theta_1 \cos\omega + \theta_1^2)}{\pi}.$$

(d) The AR(1) process, $X_t = \phi_1 X_{t-1} + \epsilon_t$, where $\{\epsilon_t\}$ is white noise. Recall

$$\mathrm{var}(X_t) = \phi_1^2 \,\mathrm{var}(X_{t-1}) + \sigma^2 \implies \gamma_0 = \phi_1^2 \gamma_0 + \sigma^2 \implies \gamma_0 = \sigma^2/(1 - \phi_1^2)$$

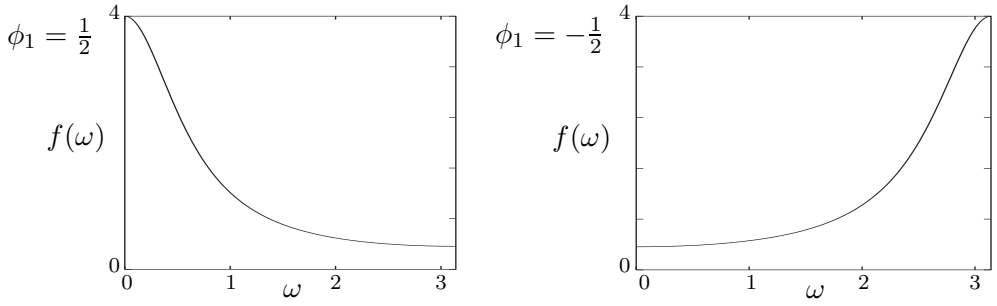
where we need $|\phi_1| < 1$ for X_t stationary. Also,

$$\gamma_k = \mathrm{cov}(X_t, X_{t-k}) = \mathrm{cov}(\phi_1 X_{t-1} + \epsilon_t, X_{t-k}) = \phi_1 \gamma_{k-1}.$$

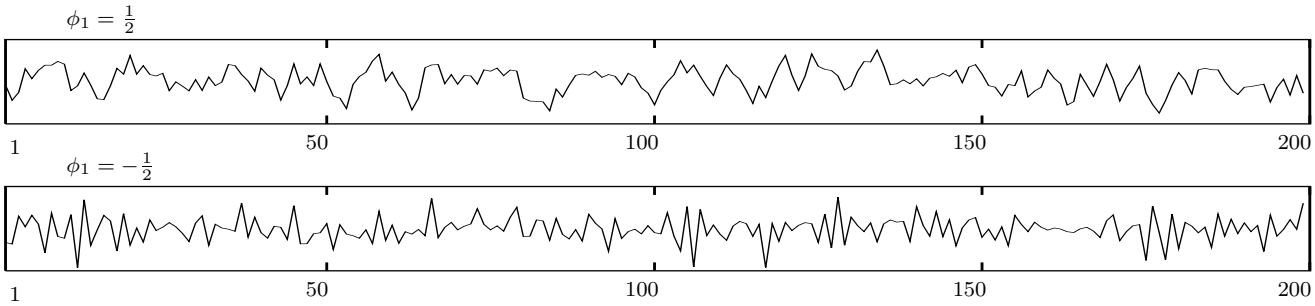
So $\gamma_k = \phi_1^{|k|} \gamma_0$, $k \in \mathbb{Z}$. Thus

$$\begin{aligned} f(\omega) &= \frac{\gamma_0}{\pi} + \frac{2}{\pi} \sum_{k=1}^{\infty} \phi_1^k \gamma_0 \cos(k\omega) = \frac{\gamma_0}{\pi} \left\{ 1 + \sum_{k=1}^{\infty} \phi_1^k \left[e^{i\omega k} + e^{-i\omega k} \right] \right\} \\ &= \frac{\gamma_0}{\pi} \left\{ 1 + \frac{\phi_1 e^{i\omega}}{1 - \phi_1 e^{i\omega}} + \frac{\phi_1 e^{-i\omega}}{1 - \phi_1 e^{-i\omega}} \right\} = \frac{\gamma_0}{\pi} \frac{1 - \phi_1^2}{1 - 2\phi_1 \cos\omega + \phi_1^2} \\ &= \frac{\sigma^2}{\pi(1 - 2\phi_1 \cos\omega + \phi_1^2)}. \end{aligned}$$

Note that $\phi > 0$ has power at low frequency, whereas $\phi < 0$ has power at high frequency.



Plots above are the spectral densities for AR(1) processes in which $\{\epsilon_t\}$ is Gaussian white noise, with $\sigma^2/\pi = 1$. Samples for 200 data points are shown below.



3.3 Analysing the effects of smoothing

Let $\{a_s\}$ be a sequence of real numbers. A **linear filter** of $\{X_t\}$ is

$$Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s} \, .$$

In Chapter 5 we show that the spectral density of $\{Y_t\}$ is given by

$$f_Y(\omega) = |a(\omega)|^2 \, f_X(\omega) \, ,$$

where $a(z)$ is the **transfer function**

$$a(\omega) = \sum_{s=-\infty}^{\infty} a_s e^{i\omega s} \, .$$

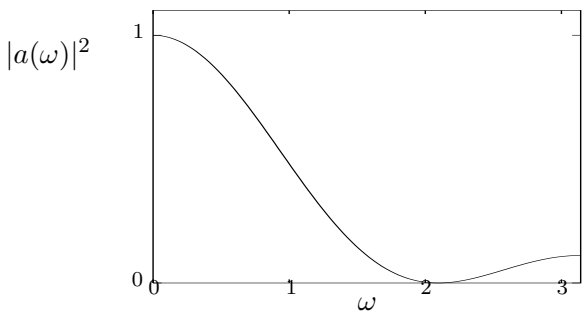
This result can be used to explore the effect of smoothing a series.

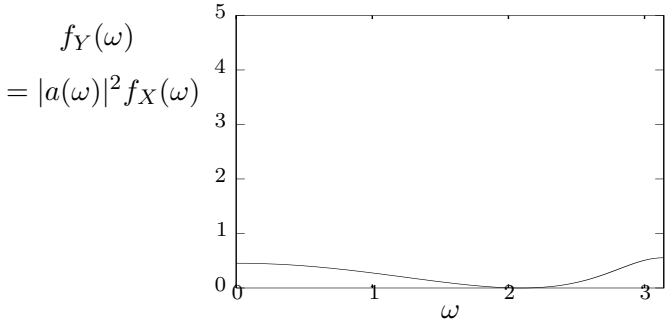
Example 3.2
Suppose the AR(1) series above, with $\phi_1 = -0.5$, is smoothed by a moving average on three points, so that smoothed series is

$$Y_t = \tfrac{1}{3}[X_{t+1} + X_t + X_{t-1}] \, .$$

Then $|a(\omega)|^2 = |\frac{1}{3}e^{-i\omega} + \frac{1}{3} + \frac{1}{3}e^{i\omega}|^2 = \frac{1}{9}(1 + 2\cos\omega)^2$.

Notice that $\gamma_X(0) = 4\pi/3$, $\gamma_Y(0) = 2\pi/9$, so $\{Y_t\}$ has 1/6 the variance of $\{X_t\}$. Moreover, all components of frequency $\omega = 2\pi/3$ (i.e., period 3) are eliminated in the smoothed series.





4 Estimation of the spectrum

4.1 The periodogram

Suppose we have $T = 2m + 1$ observations of a time series, $y_1, \ldots, y_T$. Define the **Fourier frequencies**, $\omega_j = 2\pi j/T$, $j = 1, \ldots, m$, and consider the regression model

$$y_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j \cos(\omega_j t) + \sum_{j=1}^{m} \beta_j \sin(\omega_j t)\,,$$

which can be written as a general linear model, $Y = X\theta + \epsilon$, where

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}, \quad X = \begin{pmatrix} 1 & c_{11} & s_{11} & \cdots & c_{m1} & s_{m1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & c_{1T} & s_{1T} & \cdots & c_{mT} & s_{mT} \end{pmatrix}, \quad \theta = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_m \\ \beta_m \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix},$$

$$c_{jt} = \cos(\omega_j t), \qquad s_{jt} = \sin(\omega_j t)\,.$$

The least squares estimates in this model are given by

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y\,.$$

Note that

$$\sum_{t=1}^{T} e^{i\omega_j t} = \frac{e^{i\omega_j}(1 - e^{i\omega_j T})}{1 - e^{i\omega_j}} = 0$$

$$\implies \quad \sum_{t=1}^{T} c_{jt} + i \sum_{t=1}^{T} s_{jt} = 0 \quad \implies \quad \sum_{t=1}^{T} c_{jt} = \sum_{t=1}^{T} s_{jt} = 0$$

and

$$\sum_{t=1}^{T} c_{jt} s_{jt} = \tfrac{1}{2} \sum_{t=1}^{T} \sin(2\omega_j t) = 0\,,$$

$$\sum_{t=1}^{T} c_{jt}^2 = \tfrac{1}{2} \sum_{t=1}^{T} \{1 + \cos(2\omega_j t)\} = T/2\,,$$

$$\sum_{t=1}^{T} s_{jt}^2 = \tfrac{1}{2} \sum_{t=1}^{T} \{1 - \cos(2\omega_j t)\} = T/2\,,$$

$$\sum_{t=1}^{T} c_{jt} s_{kt} = \sum_{t=1}^{T} c_{jt} c_{kt} = \sum_{t=1}^{T} s_{jt} s_{kt} = 0, \quad j \neq k\,.$$

Using these, we have

$$\hat{\theta} = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} = \begin{pmatrix} T & 0 & \cdots & 0 \\ 0 & T/2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T/2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_t y_t \\ \sum_t c_{1t} y_t \\ \vdots \\ \sum_t s_{mt} y_t \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (2/T) \sum_t c_{1t} y_t \\ \vdots \\ (2/T) \sum_t s_{mt} y_t \end{pmatrix}$$

and the regression sum of squares is

$$\hat{Y}^\top \hat{Y} = Y^\top X (X^\top X)^{-1} X^\top Y = T\bar{y}^2 + \sum_{j=1}^{m} \frac{2}{T} \left[\left\{ \sum_{t=1}^{T} c_{jt} y_t \right\}^2 + \left\{ \sum_{t=1}^{T} s_{jt} y_t \right\}^2 \right] .$$

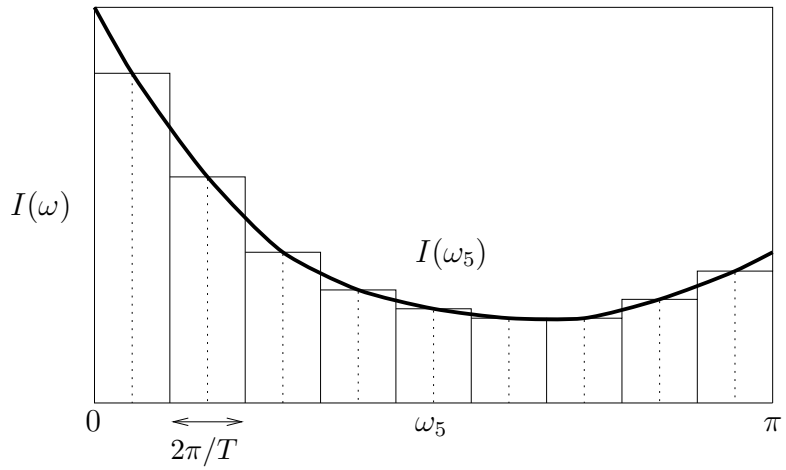
Since we are fitting T unknown parameters to T data points, the model fits with no residual error, i.e., $\hat{Y} = Y$. Hence

$$\sum_{t=1}^{T} (y_t - \bar{y})^2 = \sum_{j=1}^{m} \frac{2}{T} \left[\left\{ \sum_{t=1}^{T} c_{jt} y_t \right\}^2 + \left\{ \sum_{t=1}^{T} s_{jt} y_t \right\}^2 \right] .$$

This motivates definition of the **periodogram** as

$$I(\omega) = \frac{1}{\pi T} \left[\left\{ \sum_{t=1}^{T} y_t \cos(\omega t) \right\}^2 + \left\{ \sum_{t=1}^{T} y_t \sin(\omega t) \right\}^2 \right] .$$

A factor of $(1/2\pi)$ has been introduced into this definition so that the sample variance, $\hat{\gamma}_0 = (1/T) \sum_{t=1}^{T} (y_t - \bar{y})^2$, equates to the sum of the areas of m rectangles, whose heights are $I(\omega_1), \ldots, I(\omega_m)$, whose widths are $2\pi/T$, and whose bases are centred at $\omega_1, \ldots, \omega_m$. I.e., $\hat{\gamma}_0 = (2\pi/T) \sum_{j=1}^{m} I(\omega_j)$. These rectangles approximate the area under the curve $I(\omega)$, $0 \leq \omega \leq \pi$.



Using the fact that $\sum_{t=1}^T c_{jt} = \sum_{t=1}^T s_{jt} = 0$, we can write

$$\begin{aligned}
\pi T I(\omega_j) &= \left\{\sum_{t=1}^T y_t \cos(\omega_j t)\right\}^2 + \left\{\sum_{t=1}^T y_t \sin(\omega_j t)\right\}^2 \\
&= \left\{\sum_{t=1}^T (y_t - \bar{y}) \cos(\omega_j t)\right\}^2 + \left\{\sum_{t=1}^T (y_t - \bar{y}) \sin(\omega_j t)\right\}^2 \\
&= \left|\sum_{t=1}^T (y_t - \bar{y}) e^{i\omega_j t}\right|^2 \\
&= \sum_{t=1}^T (y_t - \bar{y}) e^{i\omega_j t} \sum_{s=1}^T (y_s - \bar{y}) e^{-i\omega_j s} \\
&= \sum_{t=1}^T (y_t - \bar{y})^2 + 2\sum_{k=1}^{T-1} \sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y}) \cos(\omega_j k)\,.
\end{aligned}$$

Hence

$$I(\omega_j) = \frac{1}{\pi}\hat{\gamma}_0 + \frac{2}{\pi}\sum_{k=1}^{T-1} \hat{\gamma}_k \cos(\omega_j k)\,.$$

$I(\omega)$ is therefore a sample version of the spectral density $f(\omega)$.

4.2 Distribution of spectral estimates

If the process is stationary and the spectral density exists then $I(\omega)$ is an almost unbiased estimator of $f(\omega)$, but it is a rather poor estimator without some smoothing.

Suppose $\{y_t\}$ is Gaussian white noise, i.e., $y_1, \ldots, y_T$ are iid $N(0, \sigma^2)$. Then for any Fourier frequency $\omega = 2\pi j/T$,

$$I(\omega) = \frac{1}{\pi T}\left[A(\omega)^2 + B(\omega)^2\right]\,, \tag{4.1}$$

where

$$A(\omega) = \sum_{t=1}^T y_t \cos(\omega t)\,, \qquad B(\omega) = \sum_{t=1}^T y_t \sin(\omega t)\,. \tag{4.2}$$

Clearly $A(\omega)$ and $B(\omega)$ have zero means, and

$$\begin{aligned}
\text{var}[A(\omega)] &= \sigma^2 \sum_{t=1}^T \cos^2(\omega t) = T\sigma^2/2\,, \\
\text{var}[B(\omega)] &= \sigma^2 \sum_{t=1}^T \sin^2(\omega t) = T\sigma^2/2\,,
\end{aligned}$$

$$\mathrm{cov}[A(\omega), B(\omega)] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{T} y_t y_s \cos(\omega t)\sin(\omega s)\right] = \sigma^2 \sum_{t=1}^{T} \cos(\omega t)\sin(\omega t) = 0\,.$$

Hence $A(\omega)\sqrt{2/T\sigma^2}$ and $B(\omega)\sqrt{2/T\sigma^2}$ are independently distributed as $N(0,1)$, and $2\left[A(\omega)^2 + B(\omega)^2\right]/(T\sigma^2)$ is distributed as χ^2_2. This gives $I(\omega) \sim (\sigma^2/\pi)\chi^2_2/2$. Thus we see that $I(w)$ is an unbiased estimator of the spectrum, $f(\omega) = \sigma^2/\pi$, but it is not consistent, since $\mathrm{var}[I(\omega)] = \sigma^4/\pi^2$ does not tend to 0 as $T \to \infty$. This is perhaps surprising, but is explained by the fact that as T increases we are attempting to estimate $I(\omega)$ for an increasing number of Fourier frequencies, with the consequence that the precision of each estimate does not change.

By a similar argument, we can show that for any two Fourier frequencies, ω_j and ω_k the estimates $I(\omega_j)$ and $I(\omega_k)$ are statistically independent. These conclusions hold more generally.

THEOREM 4.1 Let $\{Y_t\}$ be a stationary Gaussian process with spectrum $f(\omega)$. Let $I(\cdot)$ be the periodogram based on samples $Y_1, \ldots, Y_T$, and let $\omega_j = 2\pi j/T$, $j < T/2$, be a Fourier frequency. Then in the limit as $T \to \infty$,

(a) $I(\omega_j) \sim f(\omega_j)\chi^2_2/2$.

(b) $I(\omega_j)$ and $I(\omega_k)$ are independent for $j \neq k$.

Assuming that the underlying spectrum is smooth, $f(\omega)$ is nearly constant over a small range of ω. This motivates use of an estimator for the spectrum of

$$\hat{f}(\omega_j) = \frac{1}{2p+1}\sum_{\ell=-p}^{p} I(\omega_{j+\ell})\,.$$

Then $\hat{f}(\omega_j) \sim f(\omega_j)\chi^2_{2(2p+1)}/[2(2p+1)]$, which has variance $f(\omega)^2/(2p+1)$. The idea is to let $p \to \infty$ as $T \to \infty$.

4.3 The fast Fourier transform

$I(\omega_j)$ can be calculated from (4.1)–(4.2), or from

$$I(\omega_j) = \frac{1}{\pi T}\left|\sum_{t=1}^{T} y_t e^{i\omega_j t}\right|^2 .$$

Either way, this requires of order T multiplications. Hence to calculate the complete periodogram, i.e., $I(\omega_1), \ldots, I(\omega_m)$, requires of order T^2 multiplications. Computation effort can be reduced significantly by use of the **fast Fourier transform**, which computes $I(\omega_1), \ldots, I(\omega_m)$ using only order $T\log_2 T$ multiplications.

5 Linear filters

5.1 The Filter Theorem

A linear filter of one random sequence $\{X_t\}$ into another sequence $\{Y_t\}$ is

$$Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}\,. \tag{5.1}$$

THEOREM 5.1 (the filter theorem) Suppose X_t is a stationary time series with spectral density $f_X(\omega)$. Let $\{a_t\}$ be a sequence of real numbers such that $\sum_{t=-\infty}^{\infty} |a_t| < \infty$. Then the process $Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}$ is a stationary time series with spectral density function

$$f_Y(\omega) = \left|A(e^{i\omega})\right|^2 f_X(\omega) = |a(\omega)|^2 f_X(\omega)\,,$$

where $A(z)$ is the **filter generating function**

$$A(z) = \sum_{s=-\infty}^{\infty} a_s z^s, \quad |z| \le 1\,.$$

and $a(\omega) = A(e^{i\omega})$ is the **transfer function** of the linear filter.

Proof.

$$\begin{aligned}
\operatorname{cov}(Y_t, Y_{t+k}) &= \sum_{r\in\mathbb{Z}}\sum_{s\in\mathbb{Z}} a_r a_s \operatorname{cov}(X_{t-r}, X_{t+k-s}) \\
&= \sum_{r,s\in\mathbb{Z}} a_r a_s \gamma_{k+r-s} \\
&= \sum_{r,s\in\mathbb{Z}} a_r a_s \int_{-\pi}^{\pi} \tfrac{1}{2} e^{i\omega(k+r-s)} f_X(\omega) d\omega \\
&= \int_{-\pi}^{\pi} A(e^{i\omega}) A(e^{-i\omega}) \tfrac{1}{2} e^{i\omega k} f_X(\omega) d\omega \\
&= \int_{-\pi}^{\pi} \tfrac{1}{2} e^{i\omega k} \left|A(e^{i\omega})\right|^2 f_X(\omega) d\omega \\
&= \int_{-\pi}^{\pi} \tfrac{1}{2} e^{i\omega k} f_Y(\omega) d\omega\,.
\end{aligned}$$

Thus $f_Y(\omega)$ is the spectral density for Y and Y is stationary. ■

5.2 Application to autoregressive processes

Let us use the notation B for the **backshift operator**

$$B^0 = I, \quad (B^0 X)_t = X_t, \quad (BX)_t = X_{t-1}, \quad (B^2 X)_t = X_{t-2}, \quad \ldots$$

Then the AR(p) process can be written as

$$(I - \textstyle\sum_{r=1}^{p} \phi_r B^r)\, X = \epsilon$$

or $\phi(B)X = \epsilon$, where ϕ is the function

$$\phi(z) = 1 - \textstyle\sum_{r=1}^{p} \phi_r z^r \,.$$

By the filter theorem, $f_\epsilon(\omega) = |\phi\left(e^{i\omega}\right)|^2 f_X(\omega)\big)$, so since $f_\epsilon(\omega) = \sigma^2/\pi$,

$$f_X(\omega) = \frac{\sigma^2}{\pi |\phi(e^{i\omega})|^2} \,. \tag{5.2}$$

As $f_X(\omega) = (1/\pi) \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k}$, we can calculate the autocovariances by expanding $f_X(\omega)$ as a power series in $e^{i\omega}$. For this to work, the zeros of $\phi(z)$ must lie outside the unit circle in $\mathbb{C}$. This is the stationarity condition for the AR(p) process.

Example 5.2
For the AR(1) process, $X_t - \phi_1 X_{t-1} = \epsilon_t$, we have $\phi(z) = 1 - \phi_1 z$, with its zero at $z = 1/\phi_1$. The stationarity condition is $|\phi_1| < 1$. Using (5.2) we find

$$f_X(\omega) = \frac{\sigma^2}{\pi |1 - \phi e^{i\omega}|^2} = \frac{\sigma^2}{\pi(1 - 2\phi \cos\omega + \phi^2)} \,,$$

which is what we found by other another method in Example 3.1(c). To find the autocovariances we can write, taking $z = e^{i\omega}$,

$$\begin{aligned} \frac{1}{|\phi_1(z)|^2} = \frac{1}{\phi_1(z)\phi_1(1/z)} = \frac{1}{(1-\phi_1 z)(1-\phi_1/z)} &= \sum_{r=0}^{\infty} \phi_1^r z^r \sum_{s=0}^{\infty} \phi_1^s z^{-s} \\ = \sum_{k=-\infty}^{\infty} z^k (\phi_1^{|k|}(1 + \phi_1^2 + \phi_1^4 + \cdots)) &= \sum_{k=-\infty}^{\infty} \frac{z^k \phi_1^{|k|}}{1 - \phi_1^2} \\ \implies f_X(\omega) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \frac{\sigma^2 \phi_1^{|k|}}{1-\phi_1^2} e^{i\omega k} & \end{aligned}$$

and so $\gamma_k = \sigma^2 \phi_1^{|k|}/(1 - \phi_1^2)$ as we saw before.

In general, it is often easier to calculate the spectral density function first, using filters, and then deduce the autocovariance function from it.

5.3 Application to moving average processes

The MA(q) process $X_t = \epsilon_t + \sum_{s=1}^{q} \theta_s \epsilon_{t-s}$ can be written as

$$X = \theta(B)\epsilon$$

where $\theta(z) = \sum_{s=0}^{q} \theta_s B^s$. By the filter theorem, $f_X(\omega) = |\theta(e^{i\omega})|^2 (\sigma^2/\pi)$.

EXAMPLE 5.3
For the MA(1), $X_t = \epsilon_t + \theta_1 \epsilon_{t-1}$, $\theta(z) = 1 + \theta_1 z$ and

$$f_X(\omega) = \frac{\sigma^2}{\pi}\left(1 + 2\theta_1 \cos\omega + \theta_1^2\right) .$$

As above, we can obtain the autocovariance function by expressing $f_X(\omega)$ as a power series in $e^{i\omega}$. We have

$$f_X(\omega) = \frac{\sigma^2}{\pi}\left(\theta_1 e^{-i\omega} + (1+\theta_1^2) + \theta_1 e^{i\omega}\right) = \frac{\sigma^2}{\pi}\theta(e^{i\omega})\theta(e^{-i\omega})$$

So $\gamma_0 = \sigma^2(1+\theta_1^2)$, $\gamma_1 = \theta_1\sigma^2$, $\gamma_2 = 0$, $|k| > 1$.

As we remarked in Section 1.5, the autocovariance function of a MA(1) process with parameters (σ^2, θ_1) is identical to one with parameters $(\theta_1^2\sigma^2, \theta_1^{-1})$. That is,

$$\gamma_0^* = \theta_1^2\sigma^2(1 + 1/\theta_1^2) = \sigma^2(1+\theta_1^2) = \gamma_0$$
$$\rho_1^* = \theta_1^{-1}/(1+\theta_1^{-2}) = \theta_1/(1+\theta_1^2) = \rho_1 .$$

In general, the MA(q) process can be written as $X = \theta(B)\epsilon$, where

$$\theta(z) = \sum_{k=0}^{q} \theta_k z^k = \prod_{k=1}^{q} (\omega_k - z) .$$

So the autocovariance generating function is

$$g(z) = \sum_{k=-q}^{q} \gamma_k z^k = \sigma^2\theta(z)\theta(z^{-1}) = \sigma^2 \prod_{k=1}^{q} (\omega_k - z)(\omega_k - z^{-1}) . \tag{5.3}$$

Note that $(\omega_k - z)(\omega_k - z^{-1}) = \omega_k^2(\omega_k^{-1} - z)(\omega_k^{-1} - z^{-1})$. So $g(z)$ is unchanged in (5.3) if (for any k such that ω_k is real) we replace ω_k by ω_k^{-1} and multiply σ^2 by ω_k^2. Thus (if all roots of $\theta(z) = 0$ are real) there can be 2^q different $MA(q)$ processes with the same autocovariance function. For **identifiability**, we assume that all the roots of $\theta(z)$ lie outside the unit circle in $\mathbb{C}$. This is equivalent to the invertibility condition, that ϵ_t can be written as a convergent power series in $\{X_t, X_{t-1}, \ldots\}$.

5.4 The general linear process

A special case of (5.1) is the **general linear process**,

$$Y_t = \sum_{s=0}^{\infty} a_s X_{t-s} ,$$

where $\{X_t\}$ is white noise. This has

$$\mathrm{cov}(Y_t, Y_{t+k}) = \sigma^2 \sum_{s=0}^{\infty} a_s a_{s+k} \le \sigma^2 \sum_{s=0}^{\infty} a_s^2 ,$$

where the inequality is an equality when $k = 0$. Thus $\{Y_t\}$ is stationary if and only if $\sum_{s=0}^{\infty} a_s^2 < \infty$. In practice the general linear model is useful when the a_s are expressible in terms of a finite number of parameters which can be estimated. A rich class of such models are the ARMA models.

5.5 Filters and ARMA processes

The ARMA(p, q) model can be written as $\phi(B)X = \theta(B)\epsilon$. Thus

$$|\phi(e^{i\omega})|^2 f_X(\omega) = |\theta(e^{i\omega})|^2 \frac{\sigma^2}{\pi} \quad \Longrightarrow \quad f_X(\omega) = \left|\frac{\theta(e^{i\omega})}{\phi(e^{i\omega})}\right|^2 \frac{\sigma^2}{\pi}\,.$$

This is subject to the conditions that

- the zeros of ϕ lie outside the unit circle in $\mathbb{C}$ for stationarity.
- the zeros of θ lie outside the unit circle in $\mathbb{C}$ for identifiability.
- $\phi(z)$ and $\theta(z)$ have no common roots.

If there were a common root, say $1/\alpha$, so that $(I - \alpha B)\phi_1(B)X = (I - \alpha B)\theta_1(B)\epsilon$, then we could multiply both sides by $\sum_{n=0}^{\infty} \alpha^n B^n$ and deduce $\phi_1(B)X = \theta_1(B)\epsilon$, and thus that a more economical ARMA$(p-1, q-1)$ model suffices.

5.6 Calculating autocovariances in ARMA models

As above, the filter theorem can assist in calculating the autocovariances of a model. These can be compared with autocovariances estimated from the data. For example, an ARMA$(1, 2)$ has

$$\phi(z) = 1 - \phi z, \quad \theta(z) = 1 + \theta_1 z + \theta_2 z^2, \quad \text{where } |\phi| < 1.$$

Then $X = C(B)\epsilon$, where

$$C(z) = \theta(z)/\phi(z) = \left(1 + \theta_1 z + \theta_2 z^2\right) \sum_{n=0}^{\infty} \phi^n z^n = \sum_{n=0}^{\infty} c_n z^n\,,$$

with $c_0 = 1$, $c_1 = \phi + \theta_1$, and

$$c_n = \phi^n + \phi^{n-1}\theta_1 + \phi^{n-1}\theta_2 = \phi^{n-2}\left(\phi^2 + \phi\theta_1 + \theta^2\right), \quad n \geq 2.$$

So $X_t = \sum_{n=0}^{\infty} c_n \epsilon_{t-n}$ and we can compute covariances as

$$\gamma_k = \operatorname{cov}(X_t, X_{t+k}) = \sum_{n,m=0}^{\infty} c_n c_m \operatorname{cov}(\epsilon_{t-n}, \epsilon_{t+k-m}) = \sum_{n=0}^{\infty} c_n c_{n+k} \sigma^2\,.$$

For example, $\gamma_k = \phi\gamma_{k-1}$, $k \geq 3$. As a test of whether the model is ARMA$(1, 2)$ we might look to see if the sample autocovariances decay geometrically, for $k \geq 2$,

6 Estimation of trend and seasonality

6.1 Moving averages

Consider a decomposition into trend, seasonal, cyclic and residual components.

$$X_t = T_t + I_t + C_t + E_t \,.$$

Thus far we have been concerned with modelling $\{E_t\}$. We have also seen that the periodogram can be useful for recognising the presence of $\{C_t\}$.

We can estimate trend using a **symmetric moving average**,

$$\hat{T}_t = \sum_{s=-k}^{k} a_s X_{t+s} \,,$$

where $a_s = a_{-s}$. In this case the transfer function is real-valued.

The choice of moving averages requires care. For example, we might try to estimate the trend with

$$\hat{T}_t = \tfrac{1}{3}\left(X_{t-1} + X_t + X_{t+1}\right) .$$

But suppose $X_t = T_t + \epsilon_t$, where trend is the quadratic $T_t = a + bt + ct^2$. Then

$$\hat{T}_t = T_t + \tfrac{2}{3}c + \tfrac{1}{3}(\epsilon_{t-1} + \epsilon_t + \epsilon_{t+1}) \,,$$

so $\mathbb{E}\hat{T}_t = \mathbb{E}X_t + \frac{2}{3}c$ and thus $\hat{T}$ is a biased estimator of the trend.

This problem is avoided if we estimate trend by fitting a polynomial of sufficient degree, e.g., to find a cubic that best fits seven successive points we minimize

$$\sum_{t=-3}^{3} \left(X_t - b_0 - b_1 t - b_2 t^2 - b_3 t^3\right)^2 .$$

So

$$\begin{array}{rcccccc}
\sum X_t & = & 7\hat{b}_0 & & + & 28\hat{b}_2 & \\
\sum tX_t & = & & 28\hat{b}_1 & & & + \quad 196\hat{b}_3 \\
\sum t^2 X_t & = & 28\hat{b}_0 & & + & 196\hat{b}_2 & \\
\sum t^3 X_t & = & & 196\hat{b}_1 & & & + \quad 1588\hat{b}_3
\end{array}$$

Then

$$\begin{aligned}
\hat{b}_0 &= \tfrac{1}{21}\left(7\textstyle\sum X_t - \sum t^2 X_t\right) \\
&= \tfrac{1}{21}\left(-2X_{-3} + 3X_{-2} + 6X_{-1} + 7X_0 + 6X_1 + 3X_2 - 2X_3\right) .
\end{aligned}$$

We estimate the trend at time 0 by $\hat{T}_0 = \hat{b}_0$, and similarly,

$$\hat{T}_t = \tfrac{1}{21}\left(-2X_{t-3} + 3X_{t-2} + 6X_{t-1} + 7X_t + 6X_{t+1} + 3X_{t+2} - 2X_{t+3}\right) .$$

A notation for this moving average is $\frac{1}{21}[-2,3,6,7,6,3,-2]$. Note that the weights sum to 1. In general, we can fit a polynomial of degree q to $2q+1$ points by applying a symmetric moving average. (We fit to an odd number of points so that the midpoint of fitted range coincides with a point in time at which data is measured.)

A value for q can be identified using the **variate difference method**: if $\{X_t\}$ is indeed a polynomial of degree q, plus residual error $\{\epsilon_t\}$, then the trend in $\Delta^r X_t$ is a polynomial of degree $q-r$ and

$$\Delta^q X_t = \text{constant} + \Delta^q \epsilon_t = \text{constant} + \epsilon_t - \binom{q}{1}\epsilon_{t-1} + \binom{q}{2}\epsilon_{t-2} - \cdots + (-1)^q \epsilon_{t-q}\,.$$

The variance of $\Delta^q X_t$ is therefore

$$\text{var}(\Delta^q \epsilon_t) = \left[1 + \binom{q}{1}^2 + \binom{q}{2}^2 + \cdots + 1\right]\sigma^2 = \binom{2q}{q}\sigma^2\,,$$

where the simplification in the final line comes from looking at the coefficient of z^q in expansions of both sides of

$$(1+z)^q(1+z)^q = (1+z)^{2q}\,.$$

Define $V_r = \text{var}(\Delta^r X_t)/\binom{2r}{r}$. The fact that the plot of V_r against r should flatten out at $r \geq q$ can be used to identify q.

6.2 Centred moving averages

If there is a seasonal component then a **centred-moving average** is useful. Suppose data is measured quarterly, then applying twice the moving average $\frac{1}{4}[1,1,1,1]$ is equivalent to applying once the moving average $\frac{1}{8}[1,2,2,2,1]$. Notice that this so-called **centred average of fours** weights each quarter equally. Thus if $X_t = I_t + \epsilon_t$, where I_t has period 4, and $I_1 + I_2 + I_3 + I_4 = 0$, then $\hat{T}_t$ has no seasonal component. Similarly, if data were monthly we use a centred average of 12s, that is, $\frac{1}{24}[1,2,2,2,2,2,2,2,2,2,2,1]$.

6.3 The Slutzky-Yule effect

To remove both trend and seasonal components we might successively apply a number of moving averages, one or more to remove trend and another to remove seasonal effects. This is the procedure followed by some standard forecasting packages.

However, there is a danger that application of successive moving averages can introduce spurious effects. The **Slutzky-Yule effect** is concerned with the fact that a moving average repeatedly applied to a purely random series can introduce artificial cycles. Slutzky (1927) showed that some trade cycles of the nineteenth century were no more than artifacts of moving averages that had been used to smooth the data.

To illustrate this idea, suppose the moving average $\frac{1}{6}[-1, 2, 4, 2, -1]$ is applied k times to a white noise series. This moving average has transfer function, $a(\omega) = \frac{1}{6}(4 + 4\cos\omega - 2\cos 2\omega)$, which is maximal at $\omega = \pi/3$. The smoothed series has a spectral density, say $f_k(\omega)$, proportional to $a(\omega)^{2k}$, and hence for $\omega \neq \pi/3$, $f_k(\omega)/f_k(\pi/3) \to 0$ as $k \to \infty$. Thus in the limit the smoothed series is a periodic wave with period 6.

6.4 Exponential smoothing

Single exponential smoothing

Suppose the mean level of a series drifts slowly over time. A naive one-step-ahead forecast is $X_t(1) = X_t$. However, we might let all past observations play a part in the forecast, but give greater weights to those that are more recent. Choose weights to decrease exponentially and let

$$X_t(1) = \frac{1-\omega}{1-\omega^t}\left(X_t + \omega X_{t-1} + \omega^2 X_{t-2} + \cdots + \omega^{t-1} X_1\right),$$

where $0 < \omega < 1$. Define S_t as the right hand side of the above as $t \to \infty$, i.e.,

$$S_t = (1-\omega)\sum_{s=0}^{\infty} \omega^s X_{t-s}\,.$$

S_t can serve as a one-step-ahead forecast, $X_t(1)$. S_t is known as **simple exponential smoothing**. Let $\alpha = 1 - \omega$. Simple algebra gives

$$\begin{aligned} S_t &= \alpha X_t + (1-\alpha) S_{t-1} \\ X_t(1) &= X_{t-1}(1) + \alpha[X_t - X_{t-1}(1)]\,. \end{aligned}$$

This shows that the one-step-ahead forecast at time t is the one-step-ahead forecast at time $t-1$, modified by α times the forecasting error incurred at time $t-1$.

To get things started we might set S_0 equal to the average of the first few data points. We can play around with α, choosing it to minimize the mean square forecasting error. In practice, α in the range 0.25–0.5 usually works well.

Double exponential smoothing

Suppose the series is approximately linear, but with a slowly varying trend. If it were true that $X_t = b_0 + b_1 t + \epsilon_t$, then

$$\begin{aligned} S_t &= (1-\omega)\sum_{s=0}^{\infty} \omega^s \left(b_0 + b_1(t-s) + \epsilon_t\right) \\ &= b_0 + b_1 t - b_1(1-\omega)\sum_{s=0}^{\infty} \omega^s s + b_1(1-\omega)\sum_{s=0}^{\infty} \omega^s \epsilon_{t-s}\,, \end{aligned}$$

and hence

$$\mathbb{E}S_t = b_0 + b_1 t - b_1\omega/(1-\omega) = \mathbb{E}X_{t+1} - b_1/(1-\omega)\,.$$

Thus the forecast has a bias of $-b_1/(1-\omega)$. To eliminate this bias let $S_t^1 = S_t$ be the first smoothing, and $S_t^2 = \alpha S_t^1 + (1-\alpha)S_{t-1}^2$ be the simple exponential smoothing of S_t^1. Then

$$\mathbb{E}S_t^2 = \mathbb{E}S_t^1 - b_1\omega/(1-\omega) = \mathbb{E}X_t - 2b_1\omega/(1-\omega)\,,$$
$$\mathbb{E}(2S_t^1 - S_t^2) = b_0 + b_1 t, \qquad \mathbb{E}(S_t^1 - S_t^2) = b_1(1-\alpha)/\alpha\,.$$

This suggests the estimates $\hat{b}_0 + \hat{b}_1 t = 2S_t^1 - S_t^2$ and $\hat{b}_1 = \alpha(S_t^1 - S_t^2)/(1-\alpha)$. The forecasting equation is then

$$X_t(s) = \hat{b}_0 + \hat{b}_1(t+s) = (2S_t^1 - S_t^2) + s\alpha(S_t^1 - S_t^2)/(1-\alpha)\,.$$

As with single exponential smoothing we can experiment with choices of α and find S_0^1 and S_0^2 by fitting a regression line, $X_t = \hat{\beta}_0 + \hat{\beta}_1 t$, to the first few points of the series and solving

$$S_0^1 = \hat{\beta}_0 - (1-\alpha)\hat{\beta}_1/\alpha, \qquad S_0^2 = \hat{\beta}_0 - 2(1-\alpha)\hat{\beta}_1/\alpha\,.$$

6.5 Calculation of seasonal indices

Suppose data is quarterly and we want to fit an additive model. Let $\hat{I}_1$ be the average of $X_1, X_5, X_9, \ldots$, let $\hat{I}_2$ be the average of $X_2, X_6, X_{10}, \ldots$, and so on for $\hat{I}_3$ and $\hat{I}_4$. The cumulative seasonal effects over the course of year should cancel, so that if $X_t = a + I_t$, then $X_t + X_{t+1} + X_{t+2} + X_{t+3} = 4a$. To ensure this we take our final estimates of the seasonal indices as $I_t^* = \hat{I}_t - \frac{1}{4}(\hat{I}_1 + \cdots + \hat{I}_4)$.

If the model is multiplicative and $X_t = aI_t$, we again wish to see the cumulative effects over a year cancel, so that $X_t + X_{t+1} + X_{t+2} + X_{t+3} = 4a$. This means that we should take $I_t^* = \hat{I}_t - \frac{1}{4}(\hat{I}_1 + \cdots + \hat{I}_4) + 1$, adjusting so the mean of $I_1^*, I_2^*, I_3^*, I_4^*$ is 1.

When both trend and seasonality are to be extracted a two-stage procedure is recommended:

(a) Make a first estimate of trend, say $\hat{T}_t^1$.

Subtract this from $\{X_t\}$ and calculate first estimates of the seasonal indices, say I_t^1, from $X_t - \hat{T}_t^1$.

The first estimate of the deseasonalized series is $Y_t^1 = X_t - I_t^1$.

(b) Make a second estimate of the trend by smoothing Y_t^1, say $\hat{T}_t^2$.

Subtract this from $\{X_t\}$ and calculate second estimates of the seasonal indices, say I_t^2, from $X_t - \hat{T}_t^2$.

The second estimate of the deseasonalized series is $Y_t^2 = X_t - I_t^2$.

7 Fitting ARIMA models

7.1 The Box-Jenkins procedure

A general ARIMA(p, d, q) model is $\phi(B)\nabla(B)^d X = \theta(B)\epsilon$, where $\nabla(B) = I - B$.

The **Box-Jenkins** procedure is concerned with fitting an ARIMA model to data. It has three parts: **identification**, **estimation**, and **verification**.

7.2 Identification

The data may require pre-processing to make it stationary. To achieve stationarity we may do any of the following.

- Look at it.
- Re-scale it (for instance, by a logarithmic or exponential transform.)
- Remove deterministic components.
- Difference it. That is, take $\nabla(B)^d X$ until stationary. In practice $d = 1, 2$ should suffice.

We recognise stationarity by the observation that the autocorrelations decay to zero exponentially fast.

Once the series is stationary, we can try to fit an ARMA(p, q) model. We consider the correlogram $r_k = \hat{\gamma}_k/\hat{\gamma}_0$ and the partial autocorrelations $\hat{\phi}_{k,k}$. We have already made the following observations.

- An MA(q) process has negligible ACF after the qth term.
- An AR(p) process has negligible PACF after the pth term.

As we have noted, very approximately, both the sample ACF and PACF have standard deviation of around $1/\sqrt{T}$, where T is the length of the series. A rule of thumb is that ACF and PACF values are negligible when they lie between $\pm 2/\sqrt{T}$. An ARMA(p, q) process has kth order sample ACF and PACF decaying geometrically for $k > \max(p, q)$.

7.3 Estimation

AR processes

To fit a pure AR(p), i.e., $X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \epsilon_t$ we can use the **Yule-Walker equations** $\gamma_k = \sum_{r=1}^{p} \phi_r \gamma_{|k-r|}$. We fit ϕ by solving $\hat{\gamma}_k = \sum_{1}^{p} \phi_r \hat{\gamma}_{|k-r|}$, $k = 1, \ldots, p$. These can be solved by a **Levinson-Durbin recursion**, (similar to that used to solve for partial autocorrelations in Section 2.6). This recursion also gives the estimated

residual variance $\hat{\sigma}_p^2$, and helps in choice of p through the approximate log likelihood $-2\log L \simeq T\log(\hat{\sigma}_p^2)$.

Another popular way to choose p is by minimizing **Akaike's AIC** (*an information criterion*), defined as AIC $= -2\log L + 2k$, where k is the number of parameters estimated, (in the above case p). As motivation, suppose that in a general modelling context we attempt to fit a model with parameterised likelihood function $f(X \mid \theta)$, $\theta \in \Theta$, and this includes the true model for some $\theta_0 \in \Theta$. Let $X = (X_1, \ldots, X_n)$ be a vector of n independent samples and let $\hat{\theta}(X)$ be the maximum likelihood estimator of θ. Suppose Y is a further independent sample. Then

$$-2n\mathbb{E}_Y\mathbb{E}_X \log f\left(Y \mid \hat{\theta}(X)\right) = -2\mathbb{E}_X \log f\left(X \mid \hat{\theta}(X)\right) + 2k + O\left(1/\sqrt{n}\right) ,$$

where $k = |\Theta|$. The left hand side is $2n$ times the conditional entropy of Y given $\hat{\theta}(X)$, i.e., the average number of bits required to specify Y given $\hat{\theta}(X)$. The right hand side is approximately the AIC and this is to be minimized over a set of models, say $(f_1, \Theta_1), \ldots, (f_m, \Theta_m)$.

ARMA processes

Generally, we use the maximum likelihood estimators, or at least squares numerical approximations to the MLEs. The essential idea is prediction error decomposition. We can factorize the joint density of $(X_1, \ldots, X_T)$ as

$$f(X_1, \ldots, X_T) = f(X_1) \prod_{t=2}^{T} f(X_t \mid X_1, \ldots, X_{t-1}) .$$

Suppose the conditional distribution of X_t given $(X_1, \ldots, X_{t-1})$ is normal with mean $\hat{X}_t$ and variance P_{t-1}, and suppose also that X_1 is normal $N(\hat{X}_1, P_0)$. Here $\hat{X}_t$ and P_{t-1} are functions of the unknown parameters $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ and the data.

The log likelihood is

$$-2\log L = -2\log f = \sum_{t=1}^{T} \left[\log(2\pi) + \log P_{t-1} + \frac{(X_t - \hat{X}_t)^2}{P_{t-1}} \right] .$$

We can minimize this with respect to $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ to fit ARMA(p, q).

Additionally, the second derivative matrix of $-\log L$ (at the MLE) is the observed information matrix, whose inverse is an approximation to the variance-covariance matrix of the estimators.

In practice, fitting ARMA(p, q) the log likelihood $(-2\log L)$ is modified to sum only over the range $\{m+1, \ldots, T\}$, where m is small.

Example 7.1
For AR(p), take $m = p$ so $\hat{X}_t = \sum_{r=1}^{p} \phi_r X_{t-r}$, $t \geq m+1$, $P_{t-1} = \sigma_\epsilon^2$.

Note. When using this approximation to compare models with different numbers of parameters we should always use the same m.

Again we might choose p and q by minimizing the AIC of $-2\log L + 2k$, where $k = p + q$ is the total number of parameters in the model.

7.4 Verification

The third stage in the Box-Jenkins algorithm is to check whether the model fits the data. There are several tools we may use.

- Overfitting. Add extra parameters to the model and use likelihood ratio test or t-test to check that they are not significant.
- Residuals analysis. Calculate the residuals from the model and plot them. The autocorrelation functions, ACFs, PACFs, spectral densities, estimates, etc., and confirm that they are consistent with white noise.

7.5 Tests for white noise

Tests for white noise include the following.

(a) The turning point test (explained in Lecture 1) compares the number of peaks and troughs to the number that would be expected for a white noise series.

(b) The **Box–Pierce test** is based on the statistic

$$Q_m = T\sum_{k=1}^{m} r_k^2\,,$$

where r_k is the kth sample autocorrelation coefficient of the residual series, and $p + q < m \ll T$. It is called a 'portmanteau test', because it is based on the all-inclusive statistic. If the model is correct then $Q_m \sim \chi^2_{m-p-q}$ approximately.

In fact, r_k has variance $(T-k)/(T(T+2))$, and a somewhat more powerful test uses the Ljung-Box statistic quoted in Section 2.7,

$$Q'_m = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}r_k^2\,,$$

where again, $Q'_m \sim \chi^2_{m-p-q}$ approximately.

(c) Another test for white noise can be constructed from the periodogram. Recall that $I(\omega_j) \sim (\sigma^2/\pi)\chi^2_2/2$ and that $I(\omega_1), \ldots, I(\omega_m)$ are mutually independent.

Define $C_j = \sum_{k=1}^{j} I(\omega_k)$ and $U_j = C_j/C_m$. Recall that χ^2_2 is the same as the exponential distribution and that if $Y_1, \ldots, Y_m$ are i.i.d. exponential random variables,

then $(Y_1 + \cdots + Y_j)/(Y_1 + \cdots + Y_m)$, $j = 1, \ldots, m-1$, have the distribution of an ordered sample of $m-1$ uniform random variables drawn from $[0,1]$. Hence under the hypothesis that $\{X_t\}$ is Gaussian white noise U_j, $j = 1, \ldots, m-1$ have the distribution of an ordered sample of $m-1$ uniform random variables on $[0,1]$. The standard test for this is the Kolomogorov-Smirnov test, which uses as a test statistic, D, defined as the maximum difference between the theoretical distribution function for $U[0,1]$, $F(u) = u$, and the empirical distribution $\hat{F}(u) = \{\#(U_j \leq u)\}/(m-1)$. Percentage points for D can be found in tables.

7.6 Forecasting with ARMA models

Recall that $\phi(B)X = \theta(B)\epsilon$, so the power series coefficients of $C(z) = \theta(z)/\phi(z) = \sum_{r=0}^{\infty} c_r z^r$ give an expression for X_t as $X_t = \sum_{r=0}^{\infty} c_r \epsilon_{t-r}$.

But also, $\epsilon = D(B)X$, where $D(z) = \phi(z)/\theta(z) = \sum_{r=0}^{\infty} d_r z^r$ — as long as the zeros of θ lie strictly outside the unit circle and thus $\epsilon_t = \sum_{r=0}^{\infty} d_r X_{t-r}$.

The advantage of the representation above is that given $(\ldots, X_{t-1}, X_t)$ we can calculate values for $(\ldots, \epsilon_{t-1}, \epsilon_t)$ and so can forecast X_{t+1}.

In general, if we want to forecast X_{T+k} from $(\ldots, X_{T-1}, X_T)$ we use

$$\hat{X}_{T,k} = \sum_{r=k}^{\infty} c_r \epsilon_{T+k-r} \ = \sum_{r=0}^{\infty} c_{k+r} \epsilon_{T-r}\,,$$

which has the least mean squared error over all linear combinations of $(\ldots, \epsilon_{T-1}, \epsilon_T)$. In fact,

$$\mathbb{E}\left((\hat{X}_{T,k} - X_{T+k})^2\right) = \sigma_\epsilon^2 \sum_{r=0}^{k-1} c_r^2\,.$$

In practice, there is an alternative recursive approach. Define

$$\hat{X}_{T,k} = \begin{cases} X_{T+k}, & -(T-1) \leq k \leq 0\,, \\ \text{optimal predictor of } X_{T+k} \text{ given } X_1, \ldots, X_T, & 1 \leq k\,. \end{cases}$$

We have the recursive relation

$$\hat{X}_{T,k} = \sum_{r=1}^{p} \phi_r \hat{X}_{T,k-r} + \hat{\epsilon}_{T+k} + \sum_{s=1}^{q} \theta_s \hat{\epsilon}_{T+k-s}$$

For $k = -(T-1), -(T-2), \ldots, 0$ this gives estimates of $\hat{\epsilon}_t$ for $t = 1, \ldots, T$.

For $k > 0$, this gives a forecast $\hat{X}_{T,k}$ for X_{T+k}. We take $\hat{\epsilon}_t = 0$ for $t > T$.

But this needs to be started off. We need to know $(X_t,\ t \leq 0)$ and ϵ_t, $t \leq 0$. There are two standard approaches.

1. Conditional approach: take $X_t = \epsilon_t = 0$, $t \leq 0$.

2. Backcasting: we forecast the series in the reverse direction to determine estimators of $X_0, X_{-1}, \ldots$ and $\epsilon_0, \epsilon_{-1}, \ldots$.

8 State space models

8.1 Models with unobserved states

State space models are an alternative formulation of time series with a number of advantages for forecasting.

1. All ARMA models can be written as state space models.

2. Nonstationary models (e.g., ARMA with time varying coefficients) are also state space models.

3. Multivariate time series can be handled more easily.

4. State space models are consistent with Bayesian methods.

In general, the model consists of

$$\begin{array}{ll} \text{observed data:} & X_t = F_t S_t + v_t \\ \text{unobserved state:} & S_t = G_t S_{t-1} + w_t \\ \text{observation noise:} & v_t \sim N(0, V_t) \\ \text{state noise:} & w_t \sim N(0, W_t) \end{array}$$

where v_t, w_t are independent and F_t, G_t are known matrices — often time dependent (e.g., because of seasonality).

Example 8.1
$X_t = S_t + v_t$, $S_t = \phi S_{t-1} + w_t$. Define $Y_t = X_t - \phi X_{t-1} = (S_t + v_t) - \phi(S_{t-1} + v_{t-1}) = w_t + v_t - \phi v_{t-1}$. The autocorrelations of $\{y_t\}$ are zero at all lags greater than 1. So $\{Y_t\}$ is MA(1) and thus $\{X_t\}$ is ARMA(1, 1).

Example 8.2
The general ARMA(p, q) model $X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \sum_{s=0}^{q} \theta_s \epsilon_{t-s}$ is a state space model. We write $X_t = F_t S_t$, where

$$F_t = (\phi_1, \phi_2, \cdots, \phi_p, 1, \theta_1, \cdots, \theta_q), \quad S_t = \begin{pmatrix} X_{t-1} \\ \vdots \\ X_{t-p} \\ \epsilon_t \\ \vdots \\ \epsilon_{t-q} \end{pmatrix} \in \mathbb{R}^{p+q+1}$$

with $v_t = 0$, $V_t = 0$. $S_t = G_t S_{t-1} + w_t$.

$$S_t = \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ X_{t-3} \\ \vdots \\ X_{t-p-1} \\ \epsilon_t \\ \epsilon_{t-1} \\ \epsilon_{t-2} \\ \vdots \\ \epsilon_{t-q} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_p & 1 & \theta_1 & \theta_2 & \cdots & \theta_{q-1} & \theta_q \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \vdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-2} \\ X_{t-3} \\ \vdots \\ X_{t-p-1} \\ \epsilon_{t-1} \\ \epsilon_{t-2} \\ \epsilon_{t-3} \\ \vdots \\ \epsilon_{t-q} \\ \epsilon_{t-q-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \epsilon_t \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}.$$

8.2 The Kalman filter

Given observed data $X_1, \ldots, X_t$ we want to find the conditional distribution of S_t and a forecast of X_{t+1}.

Recall the following multivariate normal fact: If

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right) \tag{8.1}$$

then

$$(Y_1 \mid Y_2) \sim N\left(\mu_1 + A_{12}A_{22}^{-1}(Y_2 - \mu_2), A_{11} - A_{12}A_{22}^{-1}A_{21}\right). \tag{8.2}$$

Conversely, if $(Y_1 \mid Y_2)$ satisfies (8.2), and $Y_2 \sim N(\mu_2, A_{22})$ then the joint distribution is as in (8.1).

Now let $\mathcal{F}_{t-1} = (X_1, \ldots, X_{t-1})$ and suppose we know that $(S_{t-1} \mid \mathcal{F}_{t-1}) \sim N\left(\hat{S}_{t-1}, P_{t-1}\right)$. Then

$$S_t = G_t S_{t-1} + w_t\,,$$

so

$$(S_t \mid \mathcal{F}_{t-1}) \sim N\left(G_t \hat{S}_{t-1}, G_t P_{t-1} G_t^\top + W_t\right),$$

and also $(X_t \mid S_t, \mathcal{F}_{t-1}) \sim N(F_t S_t, V_t)$.

Put $Y_1 = X_t$ and $Y_2 = S_t$. Let $R_t = G_t P_{t-1} G_t^\top + W_t$. Taking all variables conditional on $\mathcal{F}_{t-1}$ we can use the converse of the multivariate normal fact and identify

$$\mu_2 = G_t \hat{S}_{t-1} \quad \text{and} \quad A_{22} = R_t\,.$$

Since S_t is a random variable,

$$\mu_1 + A_{12}A_{22}^{-1}(S_t - \mu_2) = F_t S_t \Longrightarrow A_{12} = F_t R_t \quad \text{and} \quad \mu_1 = F_t \mu_2\,.$$

Also

$$A_{11} - A_{12}A_{22}^{-1}A_{21} = V_t \Longrightarrow A_{11} = V_t + F_t R_t R_t^{-1} R_t^\top F_t^\top = V_t + F_t R_t F_t^\top .$$

What this says is that

$$\left.\begin{pmatrix} X_t \\ S_t \end{pmatrix}\right|_{\mathcal{F}_{t-1}} = N\left(\begin{pmatrix} F_t G_t \hat{S}_{t-1} \\ G_t \hat{S}_{t-1} \end{pmatrix}, \begin{pmatrix} V_t + F_t R_t F_t^\top & F_t R_t \\ R_t^\top F_t^\top & R_t \end{pmatrix}\right) .$$

Now apply the multivariate normal fact directly to get $(S_t \mid X_t, \mathcal{F}_{t-1}) = (S_t \mid \mathcal{F}_t) \sim N(\hat{S}_t, P_t)$, where

$$\begin{aligned} \hat{S}_t &= G_t \hat{S}_{t-1} + R_t F_t^\top \left(V_t + F_t R_t F_t^\top\right)^{-1} \left(X_t - F_t G_t \hat{S}_{t-1}\right) \\ P_t &= R_t - R_t F_t^\top \left(V_t + F_t R_t F_t^\top\right)^{-1} F_t R_t \end{aligned}$$

These are the **Kalman filter updating equations**.

Note the form of the right hand side of the expression for $\hat{S}_t$. If contains the term $G_t \hat{S}_{t-1}$, which is simply what we would predict if it were known that $S_{t-1} = \hat{S}_{t-1}$, plus a term that depends on the observed error in forecasting X_t, i.e., $\left(X_t - F_t G_t \hat{S}_{t-1}\right)$. This is similar to the forecast updating expression for simple exponential smoothing in Section 6.4.

All we need to start updating the estimates are the initial values $\hat{S}_0$ and P_0. Three ways are commonly used.

1. Use a Bayesian prior distribution.

2. If F, G, V, W are independent of t the process is stationary. We could use the stationary distribution of S to start.

3. Choosing $S_0 = 0$, $P_0 = kI$ (k large) reflects prior ignorance.

8.3 Prediction

Suppose we want to predict the X_{T+k} given $(X_1, \ldots, X_T)$. We already have

$$(X_{T+1} \mid X_1, \ldots, X_T) \sim N\left(F_{T+1} G_{T+1} S_t, V_{T+1} + F_{T+1} R_{T+1} F_{T+1}^\top\right)$$

which solves the problem for the case $k = 1$. By induction we can show that

$$(S_{T+k} \mid X_1, \ldots, X_T) \sim N\left(\hat{S}_{T+k}, P_{T+k}\right)$$

where

$$\begin{aligned}\hat{S}_{T,0} &= \hat{S}_T \\ P_{T,0} &= P_T \\ \hat{S}_{T,k} &= G_{T+k}\hat{S}_{T,k-1} \\ P_{T,k} &= G_{T+k}P_{T,k-1}G_{T+k}^\top + W_{T+k}\end{aligned}$$

and hence that $(X_{T+k} \mid X_1, \ldots, X_T) \sim N\left(F_{T+k}\hat{S}_{T,k}, V_{T+k} + F_{T+k}P_{T,k}F_{T+k}^\top\right)$.

8.4 Parameter estimation revisited

In practice, of course, we may not know the matrices F_t, G_t, V_t, W_t. For example, in ARMA(p, q) they will depend on the parameters $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, σ^2, which we may not know.

We saw that when performing prediction error decomposition that we needed to calculate the distribution of $(X_t \mid X_1, \ldots, X_{t-1})$. This we have now done.

Example 8.3
Consider the state space model

$$\begin{aligned}&\text{observed data} && X_t = S_t + v_t\,, \\ &\text{unobserved state} && S_t = S_{t-1} + w_t\,,\end{aligned}$$

where v_t, w_t are independent errors, $v_t \sim N(0, V)$ and $w_t \sim N(0, W)$.

Then we have $F_t = 1$, $G_t = 1$, $V_t = V$, $W_t = W$. $R_t = P_{t-1} + W$. So if $(S_{t-1} \mid X_1, \ldots, X_{t-1}) \sim N\left(\hat{S}_{t-1}, P_{t-1}\right)$ then $(S_t \mid X_1, \ldots, X_t) \sim N\left(\hat{S}_t, P_t\right)$, where

$$\hat{S}_t = \hat{S}_{t-1} + R_t(V + R_t)^{-1}(X_t - \hat{S}_{t-1})$$

$$P_t = R_t - \frac{R_t^2}{V + R_t} = \frac{VR_t}{V + R_t} = \frac{V(P_{t-1} + W)}{V + P_{t-1} + W}\,.$$

Asymptotically, $P_t \to P$, where P is the positive root of $P^2 + WP - WV = 0$ and $\hat{S}_t$ behaves like $\hat{S}_t = (1 - \alpha)\sum_{r=0}^{\infty} \alpha^r X_{t-r}$, where $\alpha = V/(V + W + P)$. Note that this is simple exponential smoothing.

Equally, we can predict S_{T+k} given $(X_1, \ldots, X_T)$ as $N\left(\hat{S}_{T,k}, P_{T,k}\right)$ where

$$\begin{aligned}\hat{S}_{T,0} &= S_t\,, \\ P_{T,0} &= P_T\,, \\ \hat{S}_{T,k} &= \hat{S}_T\,, \\ P_{T,k} &= P_T + kW\,.\end{aligned}$$

So $(X_{T+k} \mid X_1, \ldots, X_T) \sim N\left(\hat{S}_T, V + P_T + kW\right)$.