Time Series

HILARY TERM 2010

PROF. GESINE REINERT
http://www.stats.ox.ac.uk/~reinert

Overview

- Chapter 1: *What are time series?* Types of data, examples, objectives. Definitions, stationarity and autocovariances.
- Chapter 2: *Models of stationary processes.* Linear processes. Autoregressive, moving average models, ARMA processes, the Backshift operator. Differencing, ARIMA processes. Second-order properties. Autocorrelation and partial autocorrelation function. Tests on sample autorcorrelations.
- Chapter 3: *Statistical Analyis.* Fitting ARIMA models: The Box-Jenkins approach. Model identification, estimation, verification. Analysis in the frequency domain. Spectrum, periodogram, smoothing, filters.
- Chapter 4: *State space models.* Linear models. Kalman filters.
- Chapter 5: *Nonlinear models.* ARCH and stochastic volatility models. Chaos.

Relevant books

1. P.J. Brockwell and R.A. Davis (2002). *Introduction to Time Series and Forecasting*. Springer.
2. P.J. Brockwell and R.A. Davis (1991). *Time Series: Theory and methods*. Springer.
3. P. Diggle (1990). *Time Series*. Clarendon Press.
4. R.H. Shumway and D.S. Stoffer (2006). *Time Series Analysis and Its Applications. With R Examples. 2nd edition.* Springer.
5. R.L. Smith (2001) *Time Series.* At http://www.stat.unc.edu/faculty/rs/s133/tsnotes.pdf

6. W.N. Venables and B.D. Ripley (2002). *Modern Applied Statistics with S*. Springer.

Lectures take place Mondays 11-12 and Thursdays 10-11, weeks 1-4, plus Wednesday Week 1 at 11, and **not** Thursday Week 3 at 10. There will be two problem sheets, and two Practical classes Friday of Week 2 and Friday of Week 4 and there will be two Examples classes Tuesday 10-11 of Weeks 3 and 5. The Practical in Week 4 will be assessed. Your marker for the problem sheets is Yang Wu; the work is due Friday of Weeks 2 and 4 at 5 pm.

While the examples class will cover problems from the problem sheet, there may not be enough time to cover all the problems. You will benefit most from the examples class if you (attempt to) solve the problems on the sheet ahead of the examples class.

Lecture notes are published at `http://www.stats.ox.ac.uk/~reinert/timeseries/timeseries.htm`. The notes may cover more material than the lectures. The notes may be updated throughout the lecture course.

Time series analysis is a very complex topic, far beyond what could be covered in an 8-hour class. Hence the goal of the class is to give a brief overview of the basics in time series analysis. Further reading is recommended.

1 What are Time Series?

Many statistical methods relate to data which are independent, or at least uncorrelated. There are many practical situations where data might be correlated. This is particularly so where repeated observations on a given system are made sequentially in time.

Data gathered sequentially in time are called a *time series*.

Examples

Here are some examples in which time series arise:

- Economics and Finance
- Environmental Modelling
- Meteorology and Hydrology

- Demographics
- Medicine
- Engineering
- Quality Control

The simplest form of data is a long-ish series of continuous measurements at equally spaced time points.

That is

- observations are made at distinct points in time, these time points being equally spaced
- and, the observations may take values from a continuous distribution.

The above setup could be easily generalised: for example, the times of observation need not be equally spaced in time, the observations may only take values from a discrete distribution, ...

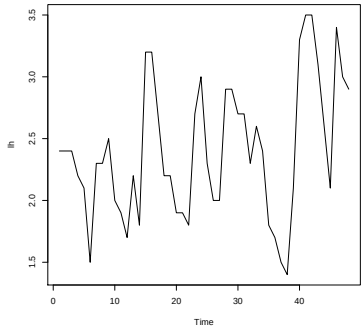
If we repeatedly observe a given system at regular time intervals, it is very likely that the observations we make will be correlated. So we cannot assume that the data constitute a random sample. The time-order in which the observations are made is vital.

Objectives of time series analysis:

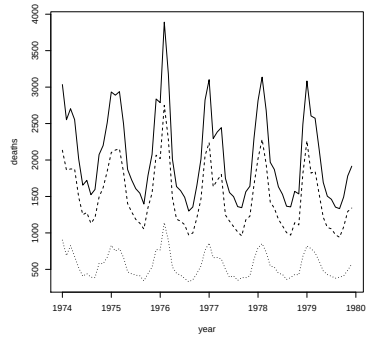
- description - summary statistics, graphs
- analysis and interpretation - find a model to describe the time dependence in the data, can we interpret the model?
- forecasting or prediction - given a sample from the series, forecast the next value, or the next few values
- control - adjust various control parameters to make the series fit closer to a target
- adjustment - in a linear model the errors could form a time series of correlated observations, and we might want to adjust estimated variances to allow for this

2 Examples: from Venables and Ripley, data from Diggle (1990)

`lh`: a series of 48 observations at 10-minute intervals on luteinizing hormone levels for a human female



`deaths`: monthly deaths in the UK from a set of common lung diseases for the years 1974 to 1979



dotted series = males, dashed = females, solid line = total
(We will not split the series into males and females from now on.)

1.1 Definitions

Assume that the series X_t runs throughout time, that is $(X_t)_{t=0,\pm1,\pm2,\ldots}$, but is only observed at times $t = 1, \ldots, n$.

So we observe $(X_1, \ldots, X_n)$. Theoretical properties refer to the underlying process $(X_t)_{t\in\mathbb{Z}}$.

The notations X_t and $X(t)$ are interchangeable.

The theory for time series is based on the assumption of 'second-order stationarity'. Real-life data are often not stationary: e.g. they exhibit a linear trend over time, or they have a seasonal effect. So the assumptions of stationarity below apply after any trends/seasonal effects have been removed. (We will look at the issues of trends/seasonal effects later.)

1.2 Stationarity and autocovariances

The process is called *weakly stationary* or *second-order stationary* if for all integers t, τ

$$E(X_t) = \mu$$
$$\mathrm{cov}(X_{t+\tau}, X_\tau) = \gamma_t$$

where μ is constant and γ_t does not depend on τ.

The process is *strictly stationary* or *strongly stationary* if

$$(X_{t_1}, \ldots, X_{t_k}) \quad \text{and} \quad (X_{t_1+\tau}, \ldots, X_{t_k+\tau})$$

have the same distribution for all sets of time points $t_1, \ldots, t_k$ and all integers τ.

Notice that a process that is strictly stationary is automatically weakly stationary. The converse of this is not true in general.

However, if the process is Gaussian, that is if $(X_{t_1}, \ldots, X_{t_k})$ has a multivariate normal distribution for all $t_1, \ldots, t_k$, then weak stationarity does imply strong stationarity.

Note that $\mathrm{var}(X_t) = \gamma_0$ and, by stationarity, $\gamma_{-t} = \gamma_t$.

The sequence (γ_t) is called the *autocovariance function.*

The *autocorrelation function* (acf) (ρ_t) is given by

$$\rho_t = \mathrm{corr}(X_{t+\tau}, X_\tau) = \frac{\gamma_t}{\gamma_0}.$$

The acf describes the second-order properties of the time series.

We estimate γ_t by c_t, and ρ_t by r_t, where

$$c_t = \frac{1}{n} \sum_{s=\max(1,1-t)}^{\min(n-t,n)} [X_{s+t} - \overline{X}][X_s - \overline{X}] \qquad \text{and} \qquad r_t = \frac{c_t}{c_0}.$$

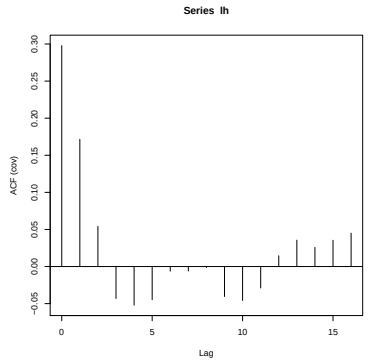
- For $t > 0$, the covariance $\mathrm{cov}(X_{t+\tau}, X_\tau)$ is estimated from the $n - t$ observed pairs
$$(X_{t+1}, X_1), \ \ldots \ , (X_n, X_{n-t}).$$
If we take the usual covariance of these pairs, we would be using different estimates of the mean and variances for each of the subseries $(X_{t+1}, \ldots, X_n)$ and $(X_1, \ldots, X_{n-t})$, whereas under the stationarity assumption these have the same mean and variance. So we use $\overline{X}$ (twice) in the above formula.

A plot of r_t against t is called the *correlogram*.

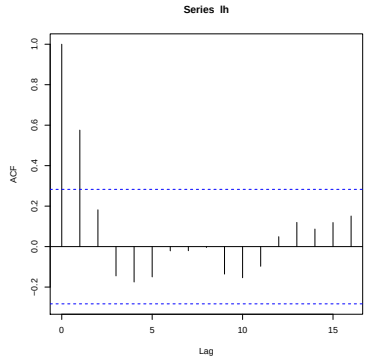
A series (X_t) is said to be *lagged* if its time axis is shifted: shifting by τ lags gives the series $(X_{t-\tau})$.

So r_t is the estimated autocorrelation at lag t; it is also called the *sample autocorrelation function*.

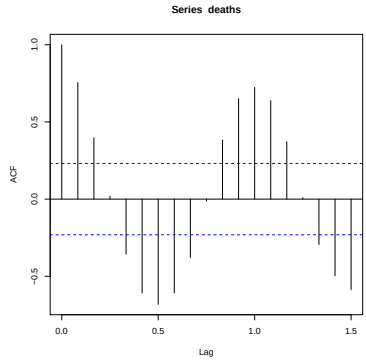
`lh`: autocovariance function



`lh`: autocorrelation function



`deaths`: autocorrelation function



2 Models of stationary processes

Assume we have a time series without trends or seasonal effects. That is, if necessary, any trends or seasonal effects have already been removed from the series.

How might we construct a linear model for a time series with autocorrelation?

Linear processes

The process (X_t) is called a *linear process* if it has a representation of the form

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

where μ is a common mean, $\{c_r\}$ is a sequence of fixed constants and $\{\epsilon_t\}$ are independent random variables with mean 0 and common variance.

We assume $\sum c_r^2 < \infty$ to ensure that the variance of X_t is finite.

If the $\{\epsilon_t\}$ are identically distributed, then such a orocess is strictly stationary. If $c_r = 0$ for $r < 0$ it is said to be *causal*, i.e. the process at time t does not depend on the future, as yet unobserved, values of ϵ_t.

The AR, MA and ARMA processes that we are now going to define are all special cases of causal linear processes.

2.1 Autoregressive processes

Assume that a current value of the series is linearly dependent upon its previous value, with some error. Then we could have the linear relationship

$$X_t = \alpha X_{t-1} + \epsilon_t$$

where ϵ_t is a *white noise* time series. [That is, the ϵ_t are a sequence of uncorrelated random variables (possibly normally distributed, but not necessarily normal) with mean 0 and variance σ^2.]

This model is called an *autoregressive* (AR) model, since X is regressed on itself. Here the lag of the autoregression is 1.

More generally we could have an autoregressive model of order p, an AR(p) model, defined by

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t.$$

At first sight, the AR(1) process

$$X_t = \alpha X_{t-1} + \epsilon_t$$

is not in the linear form $X_t = \mu + \sum c_r \epsilon_{t-r}$. However note that

$$\begin{aligned} X_t &= \alpha X_{t-1} + \epsilon_t \\ &= \epsilon_t + \alpha(\epsilon_{t-1} + \alpha X_{t-2}) \\ &= \epsilon_t + \alpha\epsilon_{t-1} + \alpha^2\epsilon_{t-2} + \cdots + \alpha^{k-1}\epsilon_{t-k+1} + \alpha^k X_{t-k} \\ &= \epsilon_t + \alpha\epsilon_{t-1} + \alpha^2\epsilon_{t-2} + \cdots \end{aligned}$$

which is in linear form.

If ϵ_t has variance σ^2, then from independence we have that

$$Var(X_t) = \sigma^2 + \alpha^2\sigma^2 + \cdots + \alpha^{2(k-1)}\sigma^2 + \alpha^{2k}Var(X_{t-k}).$$

The sum converges as we assume finite variance.

But the sum converges only if $|\alpha| < 1$. Thus $|\alpha| < 1$ is a requirement for the AR(1) process to be stationary.

We shall calculate the acf later.

2.2 Moving average processes

Another possibility is to assume that the current value of the series is a weighted sum of past white noise terms, so for example that

$$X_t = \epsilon_t + \beta\epsilon_{t-1}.$$

Such a model is called a *moving average* (MA) model, since X is expressed as a weighted average of past values of the white noise series.

Here the lag of the moving average is 1. We can think of the white noise series as being *innovations* or *shocks*: new stochastically uncorrelated information which appears at each time step, which is combined with other innovations (or shocks) to provide the observable series X.

More generally we could have a moving average model of order q, an MA(q) model, defined by

$$X_t = \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}.$$

If ϵ_t has variance σ^2, then from independence we have that

$$Var(X_t) = \sigma^2 + \sum_{j=1}^{q} \beta_j^2 \sigma^2.$$

We shall calculate the acf later.

2.3 ARMA processes

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

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=0}^{q} \beta_j \epsilon_{t-j}$$

where $\beta_0 = 1$.

A slightly more general definition of an ARMA process incorporates a non-zero mean value μ, and can be obtained by replacing X_t by $X_t - \mu$ and X_{t-i} by $X_{t-i} - \mu$ above.

From its definition we see that an MA(q) process is second-order stationary for any $\beta_1, \ldots, \beta_q$.

However the AR(p) and ARMA(p, q) models do not necessarily define second-order stationary time series.

For example, we have already seen that for an AR(1) model we need the condition $|\alpha| < 1$. This is the *stationarity condition* for an AR(1) process. All AR processes require a condition of this type.

Define, for any complex number z, the *autoregressive polynomial*

$$\phi_\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p.$$

Then the *stationarity condition* for an AR(p) process is:

all the zeros of the function $\phi_\alpha(z)$ lie outside the unit circle in the complex plane.

This is exactly the condition that is needed on $\{\alpha_1, \ldots, \alpha_p\}$ to ensure that the process is well-defined and stationary (see *Brockwell and Davis 1991), pp. 85-87.*

2.4 The backshift operator

Define the *backshift operator* B by

$$BX_t = X_{t-1}, \quad B^2 X_t = B(BX_t) = X_{t-2}, \quad \ldots$$

We include the identity operator $IX_t = B^0 X_t = X_t$.

Using this notation we can write the AR(p) process $X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t$ as

$$\left(I - \sum_{i=1}^{p} \alpha_i B^i \right) X_t = \epsilon_t$$

or even more concisely

$$\phi_\alpha(B)X = \epsilon.$$

Recall that an MA(q) process is $X_t = \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}$.
Define, for any complex number z, the *moving average polynomial*

$$\phi_\beta(z) = 1 + \beta_1 z + \cdots + \beta_q z^q.$$

Then, in operator notation, the MA(q) process can be written

$$X_t = \left(I + \sum_{j=1}^{q} \beta_j B^j \right) \epsilon_t$$

or

$$X = \phi_\beta(B)\epsilon.$$

For an MA(q) process we have already noted that there is no need for a stationarity condition on the coefficients β_j, but there is a different difficulty requiring some restriction on the coefficients.

Consider the MA(1) process

$$X_t = \epsilon_t + \beta\epsilon_{t-1}.$$

As ϵ_t has mean zero and variance σ^2, we can calculate the autocovariances to be

$$\begin{aligned}
\gamma_0 &= Var(X_0) = (1+\beta^2)\sigma^2 \\
\gamma_1 &= Cov(X_0, X_1) \\
&= Cov(\epsilon_0, \epsilon_1) + Cov(\epsilon_0, \beta\epsilon_0) + Cov(\beta\epsilon_{-1}, \epsilon_1) + Cov(\beta\epsilon_{-1}, \beta\epsilon_0) \\
&= Cov(\epsilon_0, \beta\epsilon_0) \\
&= \beta\sigma^2, \\
\gamma_k &= 0, \quad k \geqslant 2.
\end{aligned}$$

So the autocorrelations are

$$\rho_0 = 1, \quad \rho_1 = \frac{\beta}{1+\beta^2}, \quad \rho_k = 0 \quad k \geqslant 2.$$

Now consider the identical process but with β replaced by $1/\beta$. From above we can see that the autocorrelation function is unchanged by this transformation: the two processes defined by β and $1/\beta$ cannot be distinguished.

It is customary to impose the following *identifiability condition*:

all the zeros of the function $\phi_\beta(z)$ lie outside the unit circle in the complex plane.

The ARMA(p, q) process

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=0}^{q} \beta_j \epsilon_{t-j}$$

where $\beta_0 = 1$, can be written

$$\phi_\alpha(B)X = \phi_\beta(B)\epsilon.$$

The conditions required are

1. the stationarity condition on $\{\alpha_1, \ldots, \alpha_p\}$
2. the identifiability condition on $\{\beta_1, \ldots, \beta_q\}$
3. an additional identifiability condition: $\phi_\alpha(z)$ and $\phi_\beta(z)$ have no common roots.

Condition 3 is to avoid having an ARMA(p, q) model which can, in fact, be expressed as a lower order model, say as an ARMA($p-1, q-1$) model.

2.5 Differencing

The *difference operator* ∇ is given by

$$\nabla X_t = X_t - X_{t-1}$$

These differences form a new time series ∇X (of length $n-1$ if the original series had length n). Similarly

$$\nabla^2 X_t = \nabla(\nabla X_t) = X_t - 2X_{t-1} + X_{t-2}$$

and so on.

If our original time series is not stationary, we can look at the first order difference process ∇X, or second order differences $\nabla^2 X$, and so on. If we find that a differenced process is a stationary process, we can look for an ARMA model of that differenced process.

In practice if differencing is used, usually $d = 1$, or maybe $d = 2$, is enough.

2.6 ARIMA processes

The process X_t is said to be an *autoregressive integrated moving average process* ARIMA(p, d, q) if its dth difference $\nabla^d X$ is an ARMA(p, q) process.

An ARIMA(p, d, q) model can be written

$$\phi_\alpha(B)\nabla^d X = \phi_\beta(B)\epsilon$$

or

$$\phi_\alpha(B)(I - B)^d X = \phi_\beta(B)\epsilon.$$

2.7 Second order properties of MA(q)

For the MA(q) process $X_t = \sum_{j=0}^{q} \beta_j \epsilon_{t-j}$, where $\beta_0 = 1$, it is clear that $E(X_t) = 0$ for all t.

Hence, for $k > 0$, the autocovariance function is

$$\begin{aligned}
\gamma_k &= E(X_t X_{t-k}) \\
&= E\left[(\sum_{j=0}^{q} \beta_j \epsilon_{t-j})(\sum_{i=0}^{q} \beta_i \epsilon_{t-k-i}) \right] \\
&= \sum_{j=0}^{q} \sum_{i=0}^{q} \beta_j \beta_i E(\epsilon_{t-j}\epsilon_{t-k-i}).
\end{aligned}$$

Since the ϵ_t sequence is white noise, $E(\epsilon_{t-j}\epsilon_{t-k-i}) = 0$ unless $j = i + k$.

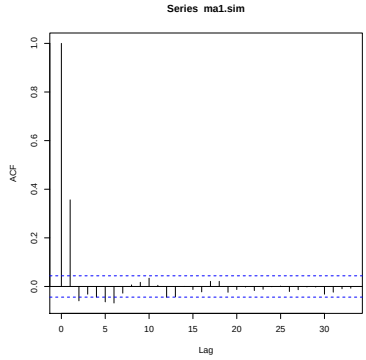
Hence the only non-zero terms in the sum are of the form $\sigma^2 \beta_i \beta_{i+k}$ and we have

$$\gamma_k = \begin{cases} \sigma^2 \sum_{i=0}^{q-|k|} \beta_i \beta_{i+|k|} & |k| \leqslant q \\ 0 & |k| > q \end{cases}$$

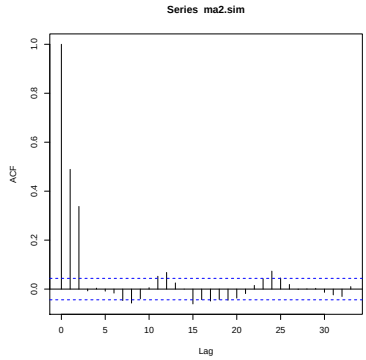
and the acf is obtained via $\rho_k = \gamma_k / \gamma_0$.

In particular notice that the acf if zero for $|k| > q$. This 'cut-off' in the acf after lag q is a characteristic property of the MA process and can be used in identifying the order of an MA process.

`Simulation`: MA(1) with $\beta = 0.5$



`Simulation`: MA(2) with $\beta_1 = \beta_2 = 0.5$



To identify an MA(q) process:

We have already seen that for an MA(q) time series, all values of the acf beyond lag q are zero: i.e. $\rho_k = 0$ for $k > q$.

So plots of the acf should show a sharp drop to near zero after the qth coefficient. This is therefore a diagnostic for an MA(q) process.

2.8 Second order properties of AR(p)

Consider the AR(p) process

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t.$$

For this model $E(X_t) = 0$ (why?).

Hence multiplying both sides of the above equation by X_{t-k} and taking expectations gives

$$\gamma_k = \sum_{i=1}^{p} \alpha_i \gamma_{k-i}, \qquad k > 0.$$

In terms of the autocorrelations $\rho_k = \gamma_k / \gamma_0$

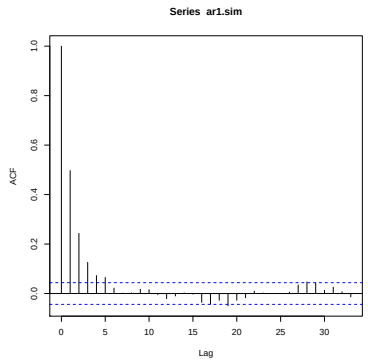
$$\rho_k = \sum_{i=1}^{p} \alpha_i \rho_{k-i}, \qquad k > 0$$

These are the *Yule-Walker* equations.

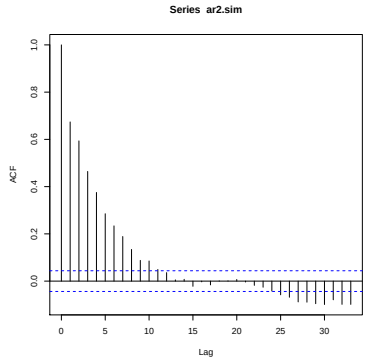
The population autocorrelations ρ_k are thus found by solving the Yule-Walker equations: these autocorrelations are generally all non-zero.

Our present interest in the Yule-Walker equations is that we could use them to calculate the ρ_k if we knew the α_i. However later we will be interested in using them to infer the values of α_i corresponding to an observed set of sample autocorrelation coefficients.

`Simulation`: AR(1) with $\alpha = 0.5$



`Simulation`: AR(2) with $\alpha_1 = 0.5, \alpha_2 = 0.25$



To identify an AR(p) process:

The AR(p) process has ρ_k decaying smoothly as k increases, which can be difficult to recognize in a plot of the acf.

Instead, the corresponding diagnostic for an AR(p) process is based on a quantity known as the *partial autocorrelation function* (pacf).

The partial autocorrelation at lag k is the correlation between X_t and X_{t-k} after regression on $X_{t-1}, \ldots, X_{t-k+1}$.

To construct these partial autocorrelations we successively fit autoregressive processes of order $1, 2, \ldots$ and, at each stage, define the partial autocorrelation coefficient a_k to be the estimate of the final autoregressive coefficient: so a_k is the estimate of α_k in an AR(k) process. If the underlying process is AR(p), then $\alpha_k = 0$ for $k > p$, so a plot of the pacf should show a cutoff after lag p.

The simplest way to construct the pacf is via the sample analogues of the Yule-Walker equations for an AR(p)

$$\rho_k = \sum_{i=1}^{p} \alpha_i \rho_{|k-i|} \qquad k = 1, \ldots, p$$

The sample analogue of these equations replaces ρ_k by its sample value r_k:

$$r_k = \sum_{i=1}^{p} a_{i,p} r_{|k-i|} \qquad k = 1, \ldots, p$$

where we write $a_{i,p}$ to emphasize that we are estimating the autoregressive coefficients $\alpha_1, \ldots, \alpha_p$ on the assumption that the underlying process is autoregressive of order p.

So we have p equations in the unknowns $a_{1,p}, \ldots, a_{p,p}$, which could be solved, and the pth partial autocorrelation coefficient is $a_{p,p}$.

Calculating the pacf

In practice the pacf is found as follows.

Consider the regression of X_t on $X_{t-1}, \ldots, X_{t-k}$, that is the model

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

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

Given data $X_1, \ldots, X_n$, least squares estimates of $\{a_{1,k}, \ldots, a_{k,k}\}$ are obtained by minimising

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

These $a_{j,k}$ coefficients can be found recursively in k for $k = 0, 1, 2, \ldots$.

For $k = 0$: $\sigma_0^2 = c_0$; $a_{0,0} = 0$, and $a_{1,1} = \rho(1)$.

And then, given the $a_{j,k-1}$ values, the $a_{j,k}$ values are given by

$$a_{k,k} = \frac{\rho_k - \sum_{j=1}^{k-1} a_{j,k-1} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} a_{j,k-1} \rho_j}$$

$$a_{j,k} = a_{j,k-1} - a_{k,k} a_{k-j,k-1} \qquad j = 1, \ldots, k-1$$

and then

$$\sigma_k^2 = \sigma_{k-1}^2 (1 - a_{k,k}^2).$$

This recursive method is the *Levinson-Durbin* recursion.

The $a_{k,k}$ value is the kth sample *partial correlation coefficient*.

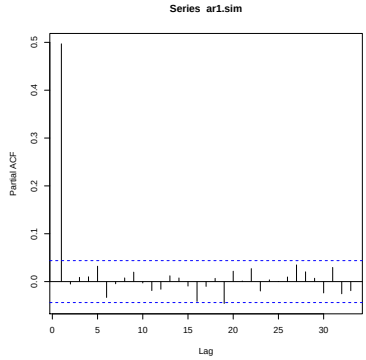
In the case of a Gaussian process, we have the interpretation that

$$a_{k,k} = \mathrm{corr}(X_t, X_{t-k} \mid X_{t-1}, \ldots, X_{t-k+1}).$$

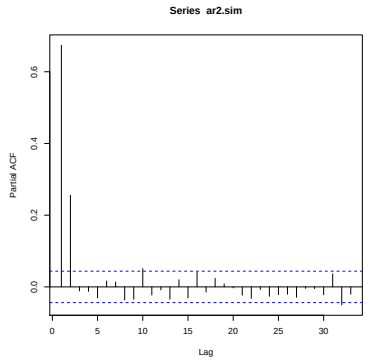
If the process X_t is genuinely an AR(p) process, then $a_{k,k} = 0$ for $k > p$.

So a plot of the pacf should show a sharp drop to near zero after lag p, and this is a diagnostic for identifying an AR(p).

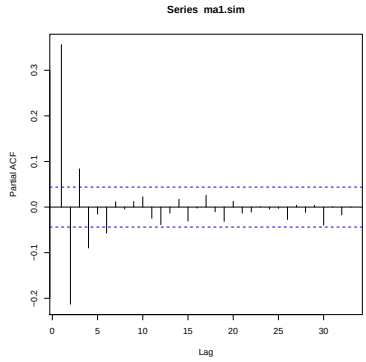
`Simulation`: AR(1) with $\alpha = 0.5$



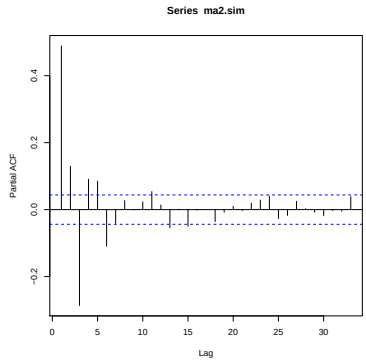
`Simulation`: AR(2) with $\alpha_1 = 0.5, \alpha_2 = 0.25$



`Simulation`: MA(1) with $\beta = 0.5$



`Simulation`: MA(2) with $\beta_1 = \beta_2 = 0.5$



Tests on sample autocorrelations

To determine whether the values of the acf, or the pacf, are negligible, we can use the approximation that they each have a standard deviation of around $1/\sqrt{n}$.

So this would give $\pm 2/\sqrt{n}$ as approximate confidence bounds (2 is an approximation to 1.96). In R these are shown as blue dotted lines.

Values outside the range $\pm 2/\sqrt{n}$ can be regarded as significant at about the 5% level. But if a large number of r_k values, say, are calculated it is likely that some will exceed this threshold even if the underlying time series is a white noise sequence.

Interpretation is also complicated by the fact that the r_k are not independently distributed. The probability of any one r_k lying outside $\pm 2/\sqrt{n}$ depends on the values of the other r_k.

3 Statistical Analysis

3.1 Fitting ARIMA models: The Box-Jenkins approach

The Box-Jenkins approach to fitting ARIMA models can be divided into three parts:

- Identification;
- Estimation;
- Verification.

3.1.1 Identification

This refers to initial preprocessing of the data to make it stationary, and choosing plausible values of p and q (which can of course be adjusted as model fitting progresses).

To assess whether the data come from a stationary process we can

- look at the data: e.g. a time plot as we looked at for the `lh` series;
- consider transforming it (e.g. by taking logs;)
- consider if we need to difference the series to make it stationary.

For stationarity the acf should decay to zero fairly rapidly. If this is not true, then try differencing the series, and maybe a second time if necessary. (In practice it is rare to go beyond $d = 2$ stages of differencing.)

The next step is initial identification of p and q. For this we use the acf and the pacf, recalling that

- for an MA(q) series, the acf is zero beyond lag q;
- for an AR(p) series, the pacf is zero beyond lag p.

We can use plots of the acf/pacf and the approximate $\pm 2/\sqrt{n}$ confidence bounds.

3.1.2 Estimation: AR processes

For the AR(p) process

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t$$

we have the Yule-Walker equations $\rho_k = \sum_{i=1}^{p} \alpha_i \rho_{|i-k|}$, for $k > 0$.

We fit the parameters $\alpha_1, \ldots, \alpha_p$ by solving

$$r_k = \sum_{i=1}^{p} \alpha_i r_{|i-k|}, \qquad k = 1, \ldots, p$$

These are p equations for the p unknowns $\alpha_1, \ldots, \alpha_p$ which, as before, can be solved using a Levinson-Durbin recursion.

The Levinson-Durbin recursion gives the residual variance

$$\widehat{\sigma}_p^2 = \frac{1}{n} \sum_{t=p+1}^{n} \left(X_t - \sum_{j=1}^{p} \widehat{\alpha}_j X_{t-j} \right)^2 .$$

This can be used to guide our selection of the appropriate order p. Define an approximate log likelihood by

$$-2 \log L = n \log(\widehat{\sigma}_p^2).$$

Then this can be used for likelihood ratio tests.

Alternatively, p can be chosen by minimising AIC where

$$AIC = -2 \log L + 2k$$

and $k = p$ is the number of unknown parameters in the model.

If $(X_t)_t$ is a causal AR(p) process with i.i.d. WN$(0, \sigma_\epsilon^2)$, then (see Brockwell and Davis (1991), p.241) then the Yule-Walker estimator $\underline{\hat{\alpha}}$ is optimal with respect to the normal distribution.

Moreover (Brockwell and Davis (1991), p.241) for the pacf of a causal AR(p) process we have that, for $m > p$,

$$\sqrt{n} \hat{\alpha}_{mm}$$

is asymptotically standard normal. However, the elements of the vector $\hat{\alpha}_m = (\hat{\alpha}_{1m}, \ldots, \hat{\alpha}_{mm})$ are in general not asymptotically uncorrelated.

3.1.3 Estimation: ARMA processes

Now we consider an ARMA(p, q) process. If we assume a parametric model for the white noise – this parametric model will be that of Gaussian white noise – we can use maximum likelihood.

We rely on the *prediction error decomposition*. That is, $X_1, \ldots, X_n$ have joint density

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

Suppose the conditional distribution of X_t given $X_1, \ldots, X_{t-1}$ is normal with mean $\widehat{X}_t$ and variance P_{t-1}, and suppose that $X_1 \sim N(\widehat{X}_1, P_0)$. (This is as for the *Kalman filter* – see later.)

Then for the log likelihood we obtain

$$-2 \log L = \sum_{t=1}^{n} \left\{ \log(2\pi) + \log P_{t-1} + \frac{(X_t - \widehat{X}_t)^2}{P_{t-1}} \right\}.$$

Here $\widehat{X}_t$ and P_{t-1} are functions of the parameters $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$, and so maximum likelihood estimators can be found (numerically) by minimising $-2 \log L$ with respect to these parameters.

The matrix of second derivatives of $-2 \log L$, evaluated at the mle, is the observed information matrix, and its inverse is an approximation to the covariance matrix of the estimators. Hence we can obtain approximate standard errors for the parameters from this matrix.

In practice, for AR(p) for example, the calculation is often simplified if we condition on the first m values of the series for some small m. That is, we use a conditional likelihood, and so the sum in the expression for $-2 \log L$ is taken over $t = m + 1$ to n.

For an AR(p) we would use some small value of m, $m \geqslant p$.

When comparing models with different numbers of parameters, it is important to use the same value of m, in particular when minimising AIC $= -2 \log L + 2(p + q)$. In R this corresponds to keeping `n.cond` in the `arima` command fixed when comparing the AIC of several models.

3.1.4 Verification

The third step is to check whether the model fits the data.

Two main techniques for model verification are

- Overfitting: add extra parameters to the model and use likelihood ratio or t tests to check that they are not significant.
- Residual analysis: calculate residuals from the fitted model and plot their acf, pacf, 'spectral density estimates', etc, to check that they are consistent with white noise.

3.1.5 Portmanteau test of white noise

A useful test for the residuals is the Box-Pierce portmanteau test. This is based on

$$Q = n \sum_{k=1}^{K} r_k^2$$

where $K > p + q$ but much smaller than n, and r_k is the acf of the residual series. If the model is correct then, approximately,

$$Q \sim \chi^2_{K-p-q}$$

so we can base a test on this: we would reject the model at level α if $Q > \chi^2_{K-p-q}(1-\alpha)$.

An improved test is the Box-Ljung procedure which replaces Q by

$$\tilde{Q} = n(n+2) \sum_{k=1}^{K} \frac{r_k^2}{n-k}.$$

The distribution of $\tilde{Q}$ is closer to a χ^2_{K-p-q} than that of Q.

Once we have a suitable model for the time series, we could apply it to estimate, say, a trend in a time series. For example, suppose that $x_1, \ldots, x_k$ are explanatory variables, that ϵ_t is an ARMA(p,q)-process, and that we observe a series y_t. Our null model may then be that

$$Y_t = \mu + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon_t, \quad t = 1, \ldots, T,$$

and the alternative model could be

$$Y_t = \mu + f_t(\lambda) + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon_t, \quad t = 1, \ldots, T,$$

where $f_t(\lambda)$ is a function for the trend. As ϵ_t is ARMA, we can write down the likelihoods under the two models, and then carry out a generalised likelihood ratio test to assess whether the trend is significant.

For confidence intervals, assume that all errors are independently normally distributed. Then we can estimate the covariance matrix for ϵ_t using the Yule-Walker equations; call this estimate V. Let X be the $T \times (k+2)$ design matrix. Then we estimate the covariance matrix of $(\hat{\mu}, \hat{\lambda}, \hat{\beta}_k)$ by

$$\hat{\Sigma} = (X^T(\hat{\sigma^2}V)^{-1}X)^{-1}.$$

If σ_λ is the square root of the diagonal element in $\hat{\Sigma}$ corresponding to λ, then $\hat{\lambda} \pm \sigma_\lambda t_{\alpha/2}$ is a 100 α-confidence interval for λ.

As an example, see *X.Zheng, R.E.Basher, C.S.Thompson: Trend detection in regional-mean temperature series: Maximum, minimum, mean, diurnal range and SST. In: Journal of Climate Vol. 10 Issue 2 (1997), pp. 317–326.*

3.2 Analysis in the frequency domain

We can consider representing the variability in a time series in terms of harmonic components at various frequencies. For example, a very simple model for a time series X_t exhibiting cyclic fluctuations with a known period, p say, is

$$X_t = \alpha \cos(\omega t) + \beta \sin(\omega t) + \epsilon_t$$

where ϵ_t is a white noise sequence, $\omega = 2\pi/p$ is the known frequency of the cyclic fluctuations, and α and β are parameters (which we might want to estimate).

Examining the second-order properties of a time series via autocovariances/autocorrelations is ‘analysis in the time domain’.

What we are about to look at now, examining the second-order properties by considering the frequency components of a series is ‘analysis in the frequency domain’.

3.2.1 The spectrum

Suppose we have a stationary time series X_t with autocovariances (γ_k).

For any sequence of autocovariances (γ_k) generated by a stationary process, there exists a function F such that

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda)$$

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

1. $F(-\pi) = 0$
2. F is non-decreasing and right-continuous
3. the increments of F are symmetric about zero, meaning that for $0 \leqslant a < b \leqslant \pi$,
$$F(b) - F(a) = F(-a) - F(-b).$$

The function F is called the *spectral distribution function* or *spectrum*. F has many of the properties of a probability distribution function, which helps explain its name, but $F(\pi) = 1$ is not required.

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

If F is everywhere continuous and differentiable, then

$$f(\lambda) = \frac{dF(\lambda)}{d\lambda}$$

is called the *spectral density function* and we have

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

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

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

By the symmetry of γ_k, $f(\lambda) = f(-\lambda)$.

From the mathematical point of view, the spectrum and acf contain equivalent information concerning the underlying stationary random sequence (X_t). However, the spectrum has a more tangible interpretation in terms of the inherent tendency for realizations of (X_t) to exhibit cyclic variations about the mean.

[Note that some authors put constants of 2π in different places. For example, some put a factor of $1/(2\pi)$ in the integral expression for γ_k in terms of F, f, and then they don't need a $1/(2\pi)$ factor when giving f in terms of γ_k.]

Example: WN$(0, \sigma^2)$

Here, $\gamma_0 = \sigma^2$, $\gamma_k = 0$ for $k \neq 0$, and so we have immediately

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

which is independent of λ.

The fact that the spectral density is constant means that all frequencies are equally present, and this is why the sequence is called 'white noise'. The converse also holds: i.e. a process is white noise if and only if its spectral density is constant.

Note that the frequency is measured in cycles per unit time; for example, at frequency $\frac{1}{2}$ the series makes a cycle every two time units. The number of time periods to complete a cycle is 2. In general, for frequency λ the number of time units to complete a cycle is $\frac{1}{\lambda}$.

Data which occurs at discrete time points will need at least two points to determine a cycle. Hence the highest frequency of interest is $\frac{1}{2}$.

The integral $\int_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda)$ is interpreted as a so-called Riemann-Stieltjes integral. If F is differentiable with derivative f, then

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

If F is such that

$$F(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_0 \\ a & \text{if } \lambda \geq \lambda_0 \end{cases}$$

then

$$\int_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda) = a e^{ik\lambda_0}.$$

The integral is additive; if

$$F(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_0 \\ a & \text{if } \lambda_0 \leq \lambda < \lambda_1 \\ a+b & \text{if } \lambda \geq \lambda_1 \end{cases}$$

then

$$\begin{aligned} \int_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda) &= \int_{\lambda_0}^{\lambda_1} e^{ik\lambda} dF(\lambda) + \int_{\lambda_1}^{\pi} e^{ik\lambda} dF(\lambda) \\ &= ae^{ik\lambda_0} + (a+b-a)e^{ik\lambda_1} \\ &= ae^{ik\lambda_0} + be^{ik\lambda_1}. \end{aligned}$$

Example: Consider the process

$$X_t = U_1 \sin(2\pi\lambda_0 t) + U_2 \cos(2\pi\lambda_0 t)$$

with U_1, U_2 independent, mean zero, variance σ^2 random variables. Then this process has frequency λ_0; the number of time periods for the above series to complete one cycle is exactly $\frac{1}{\lambda_0}$. We calculate

$$\begin{aligned} \gamma_h &= E\{U_1 \sin(2\pi\lambda_0 t) + U_2 \cos(2\pi\lambda_0 t)) \\ &\quad \times (U_1 \sin(2\pi\lambda_0(t+h)) + U_2 \cos(2\pi\lambda_0(t+h))\} \\ &= \sigma^2\{\sin(2\pi\lambda_0 t)\sin(2\pi\lambda_0(t+h)) + \cos(2\pi\lambda_0 t))\cos(2\pi\lambda_0(t+h))\}. \end{aligned}$$

Now we use that

$$\begin{aligned} \sin\alpha \sin\beta &= \frac{1}{2}(\cos(\alpha-\beta) - \cos(\alpha+\beta)) \\ \cos\alpha \cos\beta &= \frac{1}{2}(\cos(\alpha-\beta) + \cos(\alpha+\beta)) \end{aligned}$$

to get

$$\begin{aligned} \gamma_h &= \frac{\sigma^2}{2}\left(\cos(2\pi\lambda_0 h) - \cos(2\pi\lambda_0(2t+h))\right. \\ &\quad \left. + \cos(2\pi\lambda_0 h) - + \cos(2\pi\lambda_0(2t+h))\right) \\ &= \sigma^2 \cos(2\pi\lambda_0 h) \\ &= \frac{\sigma^2}{2}\left(e^{-2\pi i\lambda_0 h} + e^{2\pi i\lambda_0 h}\right). \end{aligned}$$

So, with $a = b = \frac{\sigma^2}{2}$, we use

$$F(\lambda) = \begin{cases} 0 & \text{if } \lambda < -\lambda_0 \\ \frac{\sigma^2}{2} & \text{if } -\lambda_0 \le \lambda < \lambda_0 \\ \sigma^2 & \text{if } \lambda \ge \lambda_0. \end{cases}$$

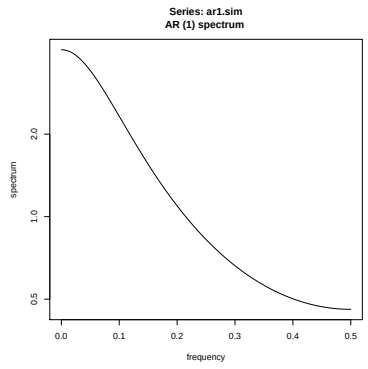
Example: AR(1): $X_t = \alpha X_{t-1} + \epsilon_t$.

Here $\gamma_0 = \sigma^2/(1-\alpha^2)$ and $\gamma_k = \alpha^{|k|}\gamma_0$ for $k \neq 0$.
So

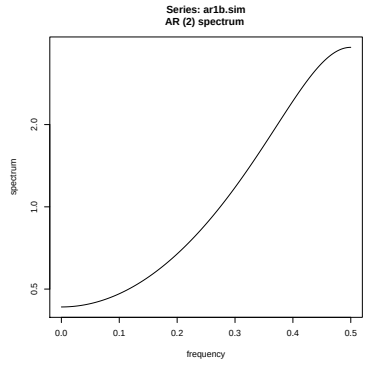
$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi}\gamma_0 \sum_{k=-\infty}^{\infty} \alpha^{|k|} e^{i\lambda k} \\ &= \frac{\gamma_0}{2\pi} + \frac{1}{2\pi}\gamma_0 \sum_{k=1}^{\infty} \alpha^k e^{i\lambda k} + \frac{1}{2\pi}\gamma_0 \sum_{k=1}^{\infty} \alpha^k e^{-i\lambda k} \\ &= \frac{\gamma_0}{2\pi}\left(1 + \frac{\alpha e^{i\lambda}}{1-\alpha e^{i\lambda}} + \frac{\alpha e^{-i\lambda}}{1-\alpha e^{-i\lambda}}\right) \\ &= \frac{\gamma_0(1-\alpha^2)}{2\pi(1-2\alpha\cos\lambda+\alpha^2)} \\ &= \frac{\sigma^2}{2\pi(1-2\alpha\cos\lambda+\alpha^2)} \end{aligned}$$

where we used $e^{-i\lambda} + e^{i\lambda} = 2\cos\lambda$.

`Simulation`: AR(1) with $\alpha = 0.5$



`Simulation`: AR(1) with $\alpha = -0.5$



Plotting the spectral density $f(\lambda)$, we see that in the case $\alpha > 0$ the spectral density $f(\lambda)$ is a decreasing function of λ: that is, the power is concentrated at low frequencies, corresponding to gradual long-range fluctuations.

For $\alpha < 0$ the spectral density $f(\lambda)$ increases as a function of λ: that is, the power is concentrated at high frequencies, which reflects the fact that such a process tends to oscillate.

ARMA(p, q) process

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=0}^{q} \beta_j \epsilon_{t-j}$$

The spectral density for an ARMA(p,q) process is related to the AR and MA polynomials $\phi_\alpha(z)$ and $\phi_\beta(z)$.

The spectral density of X_t is

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\phi_\beta(e^{-i\lambda})|^2}{|\phi_\alpha(e^{-i\lambda})|^2}.$$

Example: AR(1) Here $\phi_\alpha(z) = 1 - \alpha z$ and $\phi_\beta(z) = 1$, so, for $-\pi \leqslant \lambda < \pi$,

$$\begin{aligned} f(\lambda) &= \frac{\sigma^2}{2\pi}|1 - \alpha e^{-i\lambda}|^{-2} \\ &= \frac{\sigma^2}{2\pi}|1 - \alpha\cos\lambda + i\alpha\sin\lambda|^{-2} \\ &= \frac{\sigma^2}{2\pi}\{(1 - \alpha\cos\lambda)^2 + (\alpha\sin\lambda)^2\}^{-1} \\ &= \frac{\sigma^2}{2\pi(1 - 2\alpha\cos\lambda + \alpha^2)} \end{aligned}$$

as calculated before.

Example: MA(1)
Here $\phi_\alpha(z) = 1, \phi_\beta(z) = 1 + \theta z$, and we obtain, for $-\pi \leqslant \lambda < \pi$,

$$\begin{aligned} f(\lambda) &= \frac{\sigma_\epsilon^2}{2\pi}|1 + \theta e^{-i\lambda}|^2 \\ &= \frac{\sigma_\epsilon^2}{2\pi}(1 + 2\theta\cos(\lambda) + \theta^2). \end{aligned}$$

Plotting the spectral density $f(\lambda)$, we would see that in the case $\theta > 0$ the spectral density is large for low frequencies, small for high frequencies. This is not surprising, as we have short-range positive correlation, smoothing the series.

For $\theta < 0$ the spectral density is large around high frequencies, and small for low frequencies; the series fluctuates rapidly about its mean value. Thus, to a coarse order, the qualitative behaviour of the spectral density is similar to that of an AR(1) spectral density.

3.2.2 The Periodogram

To estimate the spectral density we use the *periodogram*.

For a frequency ω we compute the squared correlation between the time series

and the sine/cosine waves of frequency ω. The periodogram $I(\omega)$ is given by

$$I(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} e^{-i\omega t} X_t\right|^2$$

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

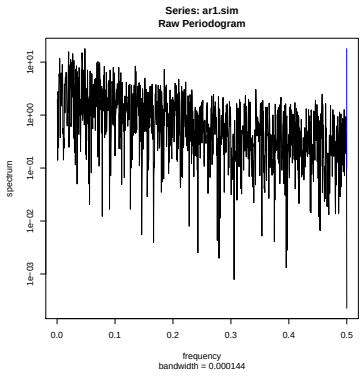
The periodogram is related to the estimated autocovariance function by

$$I(\omega) = \frac{1}{2\pi}\sum_{t=-\infty}^{\infty} c_t e^{-i\omega t} = \frac{c_0}{2\pi} + \frac{1}{\pi}\sum_{t=1}^{\infty} c_t \cos(\omega t);$$

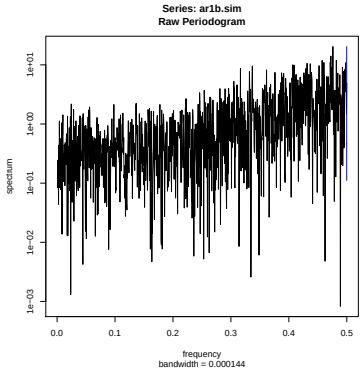
$$c_t = \int_{-\pi}^{\pi} e^{i\omega t} I(\omega) d\omega.$$

So the periodogram and the estimated autocovariance function contain the same information. For the purposes of interpretation, sometimes one will be easier to interpret, other times the other will be easier to interpret.

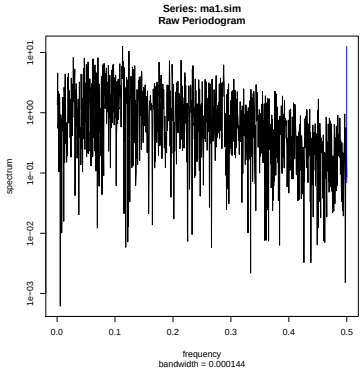
`Simulation`: AR(1) with $\alpha = 0.5$



`Simulation`: AR(1) with $\alpha = -0.5$



`Simulation`: MA(1) with $\beta = 0.5$



From asymptotic theory, at *Fourier frequencies* $\omega = \omega_j = 2\pi j/n$, $j = 1, 2, \ldots$, the periodogram ordinates $\{I(\omega_1), I(\omega_2), \ldots\}$ are approximately independent with means $\{f(\omega_1), f(\omega_2), \ldots\}$. That is for these ω

$$I(\omega) \sim f(\omega)E$$

where E is an exponential distribution with mean 1.

Note that $\mathrm{var}[I(\omega)] \approx f(\omega)^2$, which does not tend to zero as $n \to \infty$. So $I(\omega)$ is NOT a consistent estimator.

The *cumulative periodogram* $U(\omega)$ is defined by

$$U(\omega) = \sum_{0<\omega_k \leqslant \omega} I(\omega_k) \,/\, \sum_{1}^{\lfloor n/2 \rfloor} I(\omega_k).$$

This can be used to test residuals in a fitted model, for example. If we hope that our residual series is white noise, the the cumulative periodogram of the residuals should increase linearly: i.e. we can plot the cumulative periodogram (in R) and look to see if the plot is an approximate straight line.

If $X_t, t = 0, \pm 1, \pm 2, \ldots$ is Gaussian white noise, and if $\omega_k = \frac{2\pi k}{n}$ are the Fourier frequencies, $-\pi < \omega_k \leq \pi$, then the random variables

$$\frac{\sum_{k=1}^{i} I(\omega_k)}{\sum_{k=1}^{q} I(\omega_k)}, \quad r = 1, \ldots, q-1,$$

are distributed as the order statistics of $q-1$ independent random variables, each being uniformly distributed on $[0, 1]$.

As a consequence, we may apply a Kolmogorov-Smirnov test to assess whether the residuals of a time series are white noise.

Example: Brockwell & Davis (p 339, 340): Data generated by

$$X_t = \cos(\pi t/3) + \epsilon_t \qquad t = \ldots, 100$$

where $\{\epsilon_t\}$ is Gaussian white noise with variance 1. There is a peak in the periodogram at $\omega_{17} = 0.34\pi$.

In addition, the independence of the periodogram ordinates at different Fourier frequencies suggests that the sample periodogram, as a function of ω, will be extremely irregular. For this reason smoothing is often applied, for instance using a moving average, or more generally a smoothing kernel.

3.2.3 Smoothing

The idea behind smoothing is to take weighted averages over neighbouring frequencies in order to reduce the variability associated with individual periodogram values.

The main form of a smoothed esimator is given by

$$\hat{f}(\omega) = \int \frac{1}{h} K\left(\frac{\lambda - \omega}{h}\right) I(\lambda) d\lambda.$$

Here K is some *kernel function* (= a probability density function), for example a standard normal pdf, and h is the *bandwidth*.

The bandwidth h affects the degree to which this process smooths the periodogram. Small $h =$ indictes a little smoothing, large $h =$ a lot of smoothing.

In practice, the smoothed esimate $\hat{f}(\omega)$ will be evaluated by the sum

$$\begin{aligned}\hat{f}(\omega) &= \sum_j \int_{\omega_{j-1}}^{\omega_j} \frac{1}{h} K\left(\frac{\lambda - \omega}{h}\right) I(\lambda) d\lambda \\ &\approx \frac{2\pi}{n} \sum_j \frac{1}{h} K\left(\frac{\omega_j - \omega}{h}\right) I(\omega_j).\end{aligned}$$

Writing

$$g_j = \frac{2\pi}{hn} K\left(\frac{\omega_j - \omega}{h}\right)$$

we calculate that

$$E(\hat{f}(\omega)) \approx \sum_j g_j f(\omega_j)$$

and

$$Var(\hat{f}(\omega)) \approx \sum_j g_j^2 f(\omega_j)^2 \approx \frac{2\pi}{nh} f(\omega)^2 \int K(x)^2 dx$$

as well as

$$bias(\hat{f}(\omega)) \approx \frac{f''(\omega)}{2} h^2 \int x^2 K(x) dx,$$

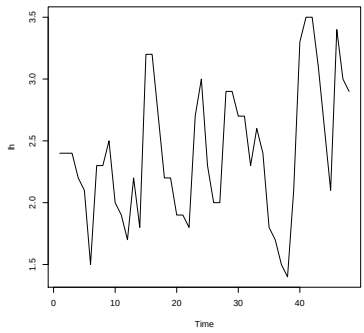
see *Venables and Ripley, p.408.* Then

$$\sqrt{2bias(\hat{f}(\omega))/f''(\omega)}$$

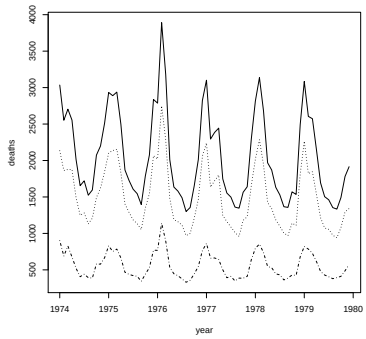
is referred to as the *bandwidth* in `R`.

As the degree of smoothing h increases, the variance decreases but the bias increases.

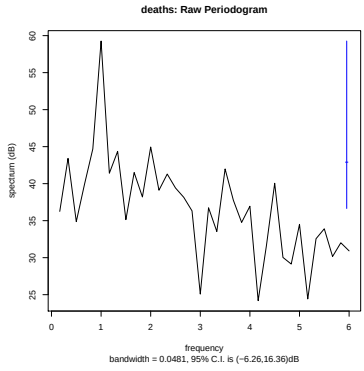
Example series: `lh`



Example series: `deaths`



`deaths`: unsmoothed periodogram



Suppose we have estimated the periodogram values $I(\omega_1), I(\omega_2), \ldots$, where $\omega_j = 2\pi j/n$, $j = 1, 2, \ldots$.

An example of a simple way to smooth is to use a moving average, and so estimate $I(\omega_j)$ by

$$\frac{1}{16}I(\omega_{j-4}) + \frac{1}{8}[I(\omega_{j-3}) + I(\omega_{j-2}) + \cdots + I(\omega_{j+3})] + \frac{1}{16}I(\omega_{j+4}).$$

Observe that the sum of the weights above (i.e. the $\frac{1}{16}$s and the $\frac{1}{8}$s) is 1.

Keeping the sum of weights equal to 1, this process could be modified by using more, or fewer, $I(\omega_k)$ values to estimate $I(\omega_j)$.

Also, this smoothing process could be repeated.

If a series is (approximately) periodic, say with frequency ω_0, then periodogram will show a peak near this frequency.

It may well also show smaller peaks at frequencies $2\omega_0, 3\omega_0, \ldots$.

The integer multiples of ω_0 are called its *harmonics*, and the secondary peaks at these high frequencies arise because the cyclic variation in the original series is non-sinusoidal. (So a situation like this warns against interpreting multiple peaks in the periodogram as indicating the presence of several distinct cyclic mechanisms in the underlying process.)

In R, smoothing is controlled by the option `spans` to the `spectrum` function.

The unsmoothed periodogram (above) was obtained via `spectrum(lh)`. The plots are on log scale, in units of *decibels*, that is, the plot is of $10 \log_{10} I(\omega)$.

The smoothed versions below are

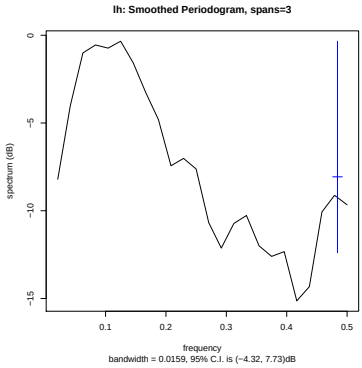
```
spectrum(lh, spans = 3)
spectrum(lh, spans = c(3,3))
spectrum(lh, spans = c(3,5))
```

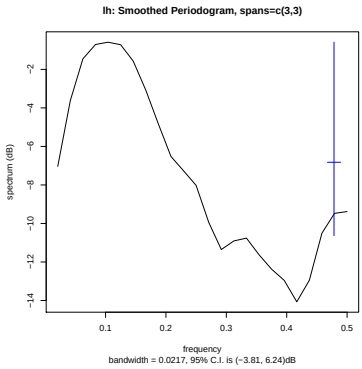
In R, the default is to use the *modified Daniell kernel*. This kernel places half the weights at the endpoints; the other half is distributed uniformly.

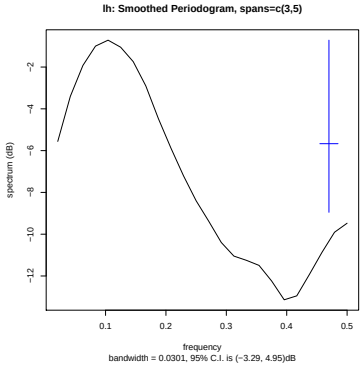
All of the examples, above and below, from Venables & Ripley.

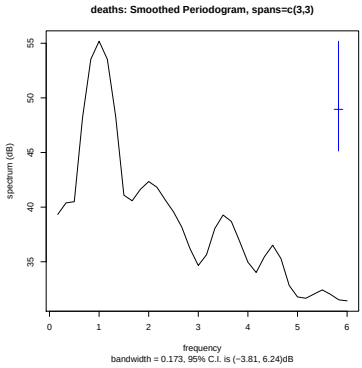
V & R advise:

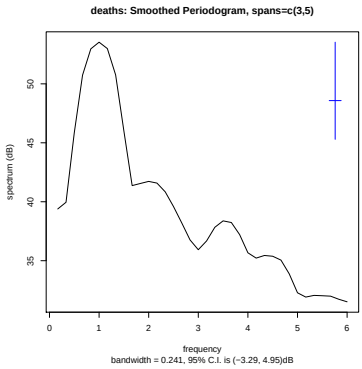
- trial and error needed to choose the spans;
- spans should be odd integers;
- use at least two, which are different, to get a smooth plot.

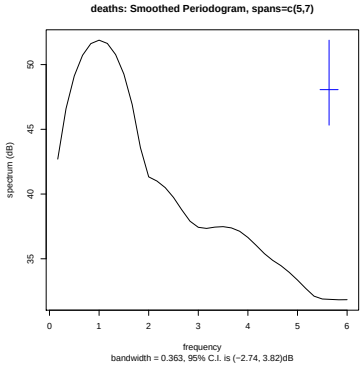

lh: Smoothed Periodogram, spans=3
spectrum (dB)
0
−5
−10
−15
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.0159, 95% C.I. is (−4.32, 7.73)dB


lh: Smoothed Periodogram, spans=c(3,3)
spectrum (dB)
−2
−4
−6
−8
−10
−12
−14
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.0217, 95% C.I. is (−3.81, 6.24)dB

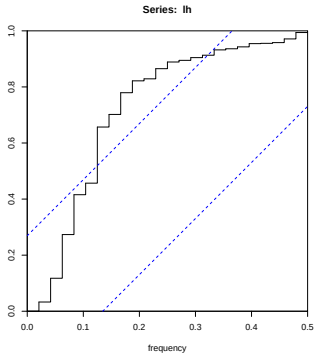

lh: Smoothed Periodogram, spans=c(3,5)
spectrum (dB)
−2
−4
−6
−8
−10
−12
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.0301, 95% C.I. is (−3.29, 4.95)dB



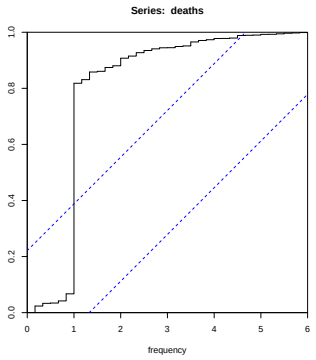




`lh`: cumulative periodogram



`deaths`: cumulative periodogram



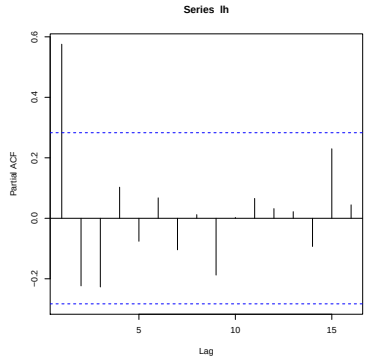
3.3 Model fitting using time and frequency domain

3.3.1 Fitting ARMA models

The value of ARMA processes lies primarily in their ability to approximate a wide range of second-order behaviour using only a small number of parameters.

Occasionally, we may be able to justify ARMA processes in terms of the basic mechanisms generating the data. But more frequently, they are used as a means of summarising a time series by a few well-chosen summary statistics: i.e. the parameters of the ARMA process.

Now consider fitting an AR model to the `lh` series. Look at the pacf:



Fit an AR(1) model:

```
lh.ar1 <- ar(lh, F, 1)
```

The fitted model is:

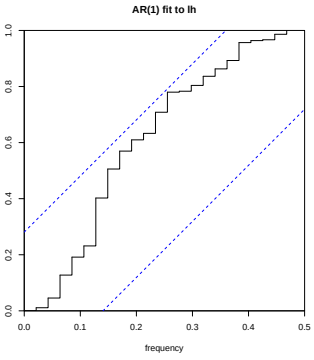
$$X_t = 0.58X_{t-1} + \epsilon_t$$

with $\sigma^2 = 0.21$.

One residual plot we could look at is

```
cpgram(lh.ar1$resid)
```

`lh`: cumulative periodogram of residuals from AR(1) model



Also try select the order of the model using AIC:

```
lh.ar <- ar(lh, order.max = 9)
lh.ar$order
lh.ar$aic
```

This selects the AR(3) model:

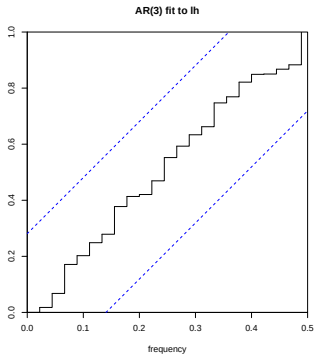
$$X_t = 0.65X_{t-1} - 0.06X_{t-2} - 0.23X_{t-3} + \epsilon_t$$

with $\sigma^2 = 0.20$.

The same order is selected when using

```
lh.ar <- ar(lh, order.max = 20)
lh.ar$order
```

`lh`: cumulative periodogram of residuals from AR(3) model



By default, `ar` fits by using the Yule-Walker equations.

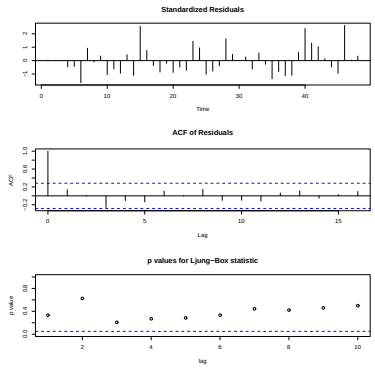
We can also use

`arima` in `library(MASS)`

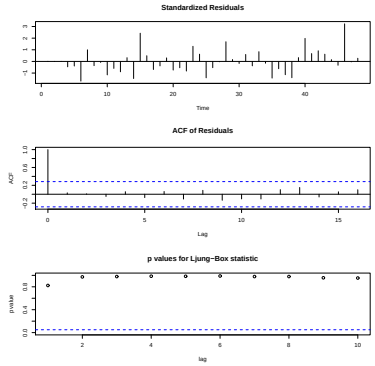
to fit these models using maximum likelihood. (Examples in Venables & Ripley, and in the practical class)

The function `tsdiag` produces diagnostic residuals plots. As mentioned in a previous lecture, the p-values from the Ljung-Box statistic are of concern if they go below 0.05 (marked with a dotted line on the plot).

`lh`: diagnostic plots from AR(1) model



`lh`: diagnostic plots from AR(3) model



3.3.2 Estimation and elimination of trend and seasonal components

The first step in the analysis of any time series is to plot the data.

If there are any apparent discontinuities, such as a sudden change of level, it may be advisable to analyse the series by first breaking it into a homogeneous segments.

We can think of a simple model of a time series as comprising

- deterministic components, i.e. trend and seasonal components
- plus a random or stochastic component which shows no informative pattern.

We might write such a *decomposition model* as the additive model

$$X_t = m_t + s_t + Z_t$$

where

$$\begin{aligned} m_t &= \text{trend component (or mean level) at time } t; \\ s_t &= \text{seasonal component at time } t; \\ Z_t &= \text{random noise component at time } t. \end{aligned}$$

Here the trend m_t is a slowly changing function of t, and if d is the number of observations in a complete cycle then $s_t = s_{t-d}$.

In some applications a multiplicative model may be appropriate

$$X_t = m_t s_t Z_t.$$

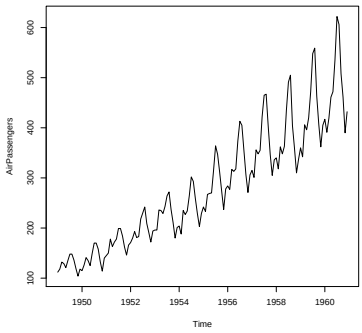
After taking logs, this becomes the previous additive model.

It is often possible to look at a time plot of the series to spot trend and seasonal behaviour. We might look for a linear trend in the first instance, though in many applications non-linear trend is also of interest and present.

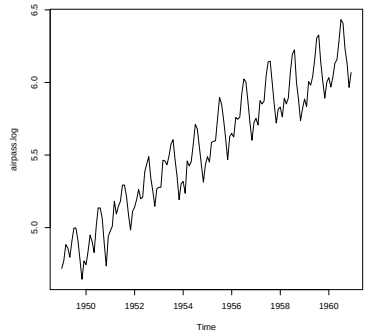
Periodic behaviour is also relatively straightforward to spot. However, if there are two or more cycles operating at different periods in a time series, then it may be difficult to detect such cycles by eye. A formal Fourier analysis can help.

The presence of both trend and seasonality together can make it more difficult to detect one or the other by eye.

Example: Box and Jenkins airline data. Monthly totals (thousands) of international airline passengers, 1949 to 1960.



```
airpass.log <- log(AirPassengers)
ts.plot(airpass.log)
```



We can aim to estimate and extract the deterministic components m_t and s_t, and hope that the residual or noise component Z_t turns out to be a stationary process. We can then try to fit an ARMA process, for example, to Z_t.

An alternative approach (Box-Jenkins) is to apply the difference operator ∇ repeatedly to the series X_t until the differenced series resembles a realization of a stationary process, and then fit an ARMA model to the suitably differenced series.

3.3.3 Elimination of trend when there is no seasonal component

The model is

$$X_t = m_t + Z_t$$

where we can assume $E(Z_t) = 0$.

1: *Fit a Parametric Relationship*

We can take m_t to be the linear trend $m_t = \alpha_0 + \alpha_1 t$, or some similar polynomial trend, and estimate m_t by minimising $\sum (X_t - m_t)^2$ with respect to α_0, α_1.

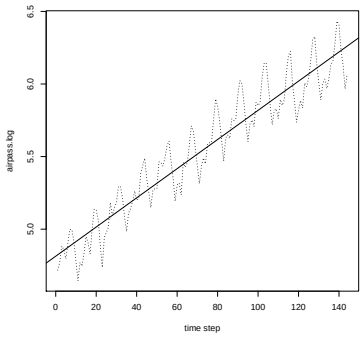
Then consider fitting stationary models to $Y_t = X_t - \widehat{m}_t$, where $\widehat{m}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 t$.

Non-linear trends are also possible of course, say $\log m_t = \alpha_0 + \alpha_1 k^t$ ($0 < k < 1$), $m_t = \alpha_0/(1 + \alpha_1 e^{-\alpha_2 t})$, ...

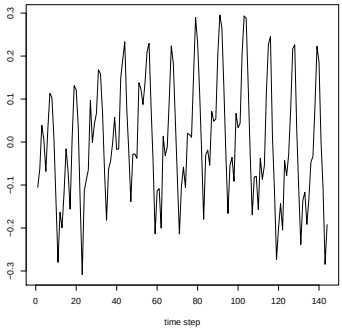
In practice, fitting a single parametric relationship to an entire time series is unrealistic, so we may fit such curves as these locally, by allowing the parameters α to vary (slowly) with time.

The resulting series $Y_t = X_t - \widehat{m}_t$ is the *detrended time series*.

Fit a linear trend:



The detrended time series:



2: *Smoothing*

If the aim is to provide an estimate of the local trend in a time series, then we can apply a *moving average*. That is, take a small sequence of the series values $X_{t-q}, \ldots, X_t, \ldots, X_{t+q}$, and compute a (weighted) average of them to obtain a smoothed series value at time t, say $\widehat{m}_t$, where

$$\widehat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}.$$

It is useful to think of $\{\widehat{m}_t\}$ as a process obtained from $\{\widehat{X}_t\}$ by application of a linear *filter* $\widehat{m}_t = \sum_{j=-\infty}^{\infty} a_j X_{t+j}$, with weights $a_j = 1/(2q+1)$, $-q \leqslant j \leqslant q$, and $a_j = 0$, $|j| > q$.

This filter is a 'low pass' filter since it takes data X_t and removes from it the rapidly fluctuating component $Y_t = X_t - \widehat{m}_t$, to leave the slowly varying estimated trend term $\widehat{m}_t$.

We should not choose q too large since, if m_t is not linear, although the filtered process will be smooth, it will not be a good estimate of m_t.

If we apply two filters in succession, for example to progressively smooth a series, we are said to be using a convolution of the filters.

By careful choice of the weights a_j, it is possible to design a filter that will not only be effective in attenuating noise from the data, but which will also allow a larger class of trend functions.

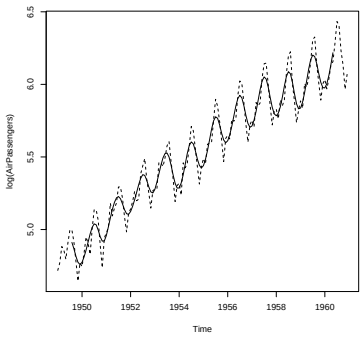
Spencer's 15-point filter has weights

$$\begin{aligned} a_j &= a_{-j} \qquad |j| \leqslant 7 \\ a_j &= 0 \qquad |j| > 7 \\ (a_0, a_1, \ldots, a_7) &= \frac{1}{320}(74, 67, 46, 21, 3, -5, -6, -3) \end{aligned}$$

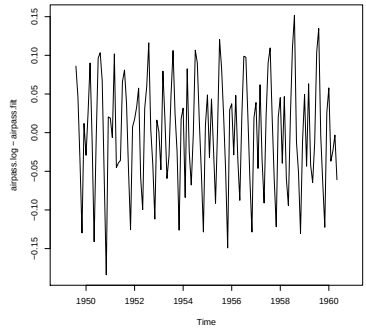
and has the property that a cubic polynomial passes through the filter undistorted.

```
spencer.wts <- c(-3,-6,-5,3,21,46,67,74,67,46,21,3,-5,-6,-3)/320
airpass.filt <- filter(airpass.log, spencer.wts)
ts.plot(airpass.log, airpass.filt, lty=c(2,1))
```

Original series and filtered series using Spencer's 15-point filter:



Detrended series via filtering:



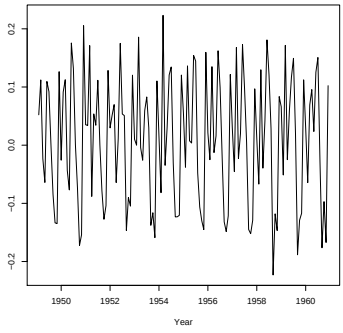
3: *Differencing*

Recall that the *difference operator* is $\nabla X_t = X_t - X_{t-1}$. Note that differencing is a special case of applying a linear filter.

We can think of differencing as a 'sample derivative'. If we start with a linear function, then differentiation yields a constant function, while if we start with a quadratic function we need to differentiate twice to get to a constant function.

Similarly, if a time series has a linear trend, differencing the series once will remove it, while if the series has a quadratic trend we would need to difference twice to remove the trend.

Detrended series via differencing:



3.4 Seasonality

After removing trend, we can remove seasonality. (Above, all detrended versions of the airline data clearly still have a seasonal component.)

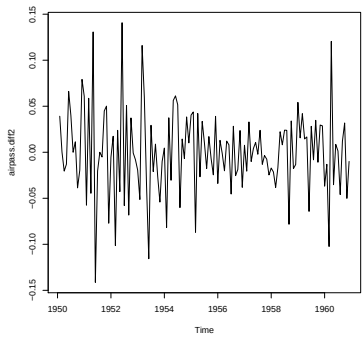
1: *Block averaging*

The simplest way to remove seasonality is to average the observations at the same point in each repetition of the cycle (for example, for monthly data average all the January values) and subtract that average from the values at those respective points in the cycle.

2: *Seasonal differencing*

The seasonal difference operator is $\nabla_s X_t = X_t - X_{t-s}$ where s is the period of the seasonal cycle. Seasonal differencing will remove seasonality in the same way that ordinary differencing will remove a polynomial trend.

```
airpass.diff<-diff(airpass.log)
airpass.diff2 <- diff(airpass.diff, lag=12)
ts.plot(airpass.diff2)
```



After differencing at lag 1 (to remove trend), then at lag 12 (to remove seasonal effects), the `log(AirPassengers)` series appears stationary.

That is, the series $\nabla\nabla_{12}X$, or equivalently the series $(1-B)(1-B^{12})X$, appears stationary.

R has a function `stl` which you can use to estimate and remove trend and seasonality using 'loess'.

`stl` is a complex function, you should consult the online documentation before you use it. The time series chapter of Venables & Ripley contains examples of how to use `stl`. As with all aspects of that chapter, it would be a good idea for you to work through the examples there.

We could now look to fit an ARMA model to $\nabla\nabla_{12}X$, or to the residual component extracted by `stl`.

Seasonal ARIMA models

Recall that X is an ARMA(p, q) process if

$$X_t - \sum_{i=1}^{p} \alpha_i X_{t-i} = \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}$$

and X is an ARIMA(p, d, q) process if $\nabla^d X$ is ARMA(p, q).

In shorthand notation, these processes are

$$\phi_\alpha(B)X = \phi_\beta(B)\epsilon \qquad \text{and} \qquad \phi_\alpha(B)\nabla^d X = \phi_\beta(B)\epsilon.$$

Suppose we have monthly observations, so that seasonal patterns repeat every $s = 12$ observations. Then we may typically expect X_t to depend on such terms as X_{t-12}, and maybe X_{t-24}, as well as $X_{t-1}, X_{t-2}, \ldots$.

A general seasonal ARIMA (SARIMA) model, is

$$\Phi_p(B)\Phi_P(B^s)Y = \Phi_q(B)\Phi_Q(B^s)\epsilon$$

where $\Phi_p, \Phi_P, \Phi_q, \Phi_Q$ are polynomials of orders p, P, q, Q and where

$$Y = (1 - B)^d(1 - B^s)^D X.$$

Here:

- s is the number of observations per season, so $s = 12$ for monthly data;
- D is the order of seasonal differencing, i.e. differencing at lag s (we were content with $D = 1$ for the air passenger data);
- d is the order of ordinary differencing (we were content with $d = 1$ for the air passenger data).

This model is often referred to as an ARIMA($(p,d,q)\times(P,D,Q)_s$) model.

Examples

1. Consider a ARIMA model of order $(1,0,0)\times(0,1,1)_{12}$.
This model can be written

$$(1-\alpha B)Y_t = (1+\beta B^{12})\epsilon_t$$

where

$$Y_t = X_t - X_{t-12}.$$

2. The 'airline model' (so named because of its relevance to the air passenger data) is a ARIMA model of order $(0,1,1)\times(0,1,1)_{12}$.
This model can be written

$$Y_t = (1+\beta_1 B)(1+\beta_2 B^{12})\epsilon_t$$

where $Y_t = \nabla\nabla_{12}X$ is the series we obtained after differencing to reach stationarity, i.e. one step of ordinary differencing, plus one step of seasonal (lag 12) differencing.

3.5 Forecasting in ARMA models

As a linear time series, under our usual assumptions on the AR-polynomial and the MA-polynomial, we can write an ARMA model as a causal model,

$$X_t = \sum_{r=0}^{\infty} c_r \epsilon_{t-r}.$$

Suppose that we are interested in forecasting X_{T+k} from observations $\{X_t, t \leq T\}$. Consider forecasts of the form

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

Then

$$\begin{aligned} X_{T+k} - \hat{X}_{T,k} &= \sum_{r=0}^{\infty} c_r \epsilon_{T+k-r} - \sum_{r=0}^{\infty} c_{r+k} \epsilon_{T-r} \\ &= \sum_{r=0}^{k-1} c_r \epsilon_{T+k-r} + \sum_{r=k}^{\infty} c_r \epsilon_{T+k-r} - \sum_{s=k}^{\infty} c_s \epsilon_{T-s+k} \\ &= \sum_{r=0}^{k-1} c_r \epsilon_{T+k-r}. \end{aligned}$$

This gives rise to the mean squared prediction error

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

Thus

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

is our theoretical optimal predictor.

Note that the mean squared prediction errors are based solely on the uncertainty of prediction; they do not take errors in model identification into account.

In practice one usually uses a recursive approach. Define $\hat{X}_{T,k}$ to be the optimal predictor of X_{T+k} given $X_1, \ldots, X_T$; for $-T+1 \leq k \leq 0$, we set $\hat{X}_{T,k} = X_{T+k}$. Then use the recursive relation

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

For $k \leq 0$ we can use thus relation to calculate $\hat{\epsilon}_t$ for $1 \leq t \leq T$. For $k > 0$ we define $\hat{\epsilon}_t = 0$ for $t > T$, to calculate the forecasts.

The difficulty is how to start off the recursion. Two standard solutions are
Either assume $X_t = \epsilon_t = 0$ for all $t \leq 0$,
or forecast the series in reverse direction to determine estimates of $X_0, X_{-1}, \ldots,$ as well as $\epsilon_0 = 0, \epsilon_{-1} = 0$, etc.
A superior approach is to recast the model in state space form and apply the Kalman filter.

4 State space models

State-space models assume that the observations $(X_t)_t$ are incomplete and noisy functions of some underlying unobservable process $(Y_t)_t$, called the *state process*, which is assumed to have a simple Markovian dynamics. The general state space model is described by

1. $Y_0, Y_1, Y_2, \ldots$ is a Markov chain
2. Conditionally on $\{Y_t\}_t$, the X_t's are independent, and X_t depends on Y_t only.

When the state variables are discrete, one usually calls this model a *hidden Markov model*; the term *state space model* is mainly used for continuous state variables.

4.1 The linear state space model

A prominent role is played by the linear state space model

$$\begin{aligned} Y_t &= G_t Y_{t-1} + v_t \quad (1) \\ X_t &= H_t Y_t + w_t, \quad (2) \end{aligned}$$

where G_t and H_t are deterministic matrices, and $(v_t)_t$ and $(w_t)_t$ are two independent white noise sequences with v_t and w_t being mean zero and having covariance matrices V_t^2 and W_t^2, respectively. The general case,

$$\begin{aligned} Y_t &= g_t(Y_{t-1}, v_t) \\ X_t &= h_t(Y_t, w_t), \end{aligned}$$

is much more flexible. Also, multivariate models are available. The typical question on state space models is the estimation or the prediction of the states $(Y_t)_t$ in terms of the observed data points $(X_t)_t$.

Example. Suppose the two-dimensional model

$$Y_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} Y_{t-1} + \begin{pmatrix} 1 \\ \beta \end{pmatrix} X_t,$$

where X_t is one-dimensional mean zero white noise. Then

$$\begin{aligned} Y_{2,t} &= \beta X_t \\ Y_{1,t} &= Y_{2,t-1} + X_t = X_t + \beta X_{t-1}, \end{aligned}$$

so we obtain an MA(1)-process.

Example. Suppose the model

$$\begin{aligned} Y_t &= \phi Y_{t-1} + v_t \\ X_t &= Y_t + w_t, \end{aligned}$$

where $(v_t)_t$ and $(w_t)_t$ are two independent white noise sequences with v_t and w_t being mean zero and having variances V^2 and W^2, respectively. Then

$$\begin{aligned} X_t - \phi X_{t-1} &= Y_t - \phi Y_{t-1} + w_t - \phi w_{t-1} \\ &= v_t + w_t - \phi w_{t-1}. \end{aligned}$$

The right-hand side shows that all correlations at lags $\geqslant 1$ are zero. Hence the right-hand side is equivalent to an MA(1) model, and thus X_t follows an ARMA(1,1)-model.

To make the connection with ARMA(1,1) more transparent, note that

$$\epsilon_t = v_t + w_t$$

gives a mean zero white noise series with variance $\sigma_\epsilon^2 = V^2 + W^2$. Thus ϵ_t has the same distribution as $\sqrt{\frac{V^2+W^2}{W^2}} w_t$. Putting

$$\beta = -\sqrt{\frac{W^2}{V^2+W^2}}\phi$$

thus gives that

$$v_t + w_t - \phi w_{t-1} = \epsilon_t + \beta \epsilon_{t-1}.$$

In fact any ARMA(p,q)-model with Gaussian WN can be formulated as a state space model. The representation of an ARMA model as a state-space model is however not unique, see Brockwell and Davis (1991), pp.469-470.

Note that the above model is more flexible than an ARMA model. If, for example, the observation at time t is missing, then we simply put $H_t = (0, 0, \ldots, 0)^T$.

4.2 Filtering, smoothing, and forecasting

The primary aims of the analysis of state space models are to produce estimators for the underlying unobserved signal Y_t given the data $\mathbf{X}^s = (X_1, \ldots, X_s)$ up to time s. When $s < t$ the problem is called *forecasting*, when $s = t$ it is called *filtering*, and when $s > t$ it is called *smoothing*. For a derivation of the results below see also Smith (2001).

We will throughout assume the white noise to be Gaussian.

In *Kalman filters made easy* by Terence Tong, at
`http://openuav.astroplanes.com/library/docs/writeup.pdf`
an analogy of the following type is given.

Suppose that you just met a new friend and you do not know how punctual your new friend will be. Based on your history, you estimate when the friend will arrive. You do not want to come too early, but also you do not want to be too late.

You arrive on time at your first meeting, while your friend arrives 30 min late. So you adapt your estimate, you will not be so early next time.

The *Kalman filter* is a method for updating parameter estimates instantly when a new observation occurs, based on the likelihood of the current data - without having to re-estimate a large number of parameters using all past data.

The Kalman filter was first developed in an engineering framework, and we shall use it for filtering and forecasting. It is a recursive method to calculate a conditional distribution within a multivariate normal framework. As it is recursive, only the estimated state from the previous time step and the current measurement are needed to compute the estimate for the current state.

The state of the filter is represented by two variables: the estimate of the state at time t; and the error covariance matrix (a measure of the estimated accuracy of the state estimate). The Kalman filter has two distinct phases: Predict and Update. The predict phase uses the state estimate from the previous timestep to produce an estimate of the state at the current timestep. In the update phase, measurement information at the current timestep is used to refine this prediction to arrive at a new, (hopefully) more accurate state estimate, again for the current timestep.

It is useful to first revise some distributional results for multivariate normal

distributions. Suppose that

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right). \tag{3}$$

Then the conditional distribution of Z_1 given $Z_2 = z_2$ is

$$\mathcal{L}(Z_1|Z_2 = z_2) = \mathcal{MVN}\left(\mu_1 + \Sigma_{12}\Sigma_{11}^{-1}(z_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21}\right) \tag{4}$$

and conversely, if $Z_2 \sim \mathcal{MVN}(\mu_2, \Sigma_{22})$ and if (4) holds, then (3) holds.

In particular, the conditional distribution of Z_1 given $Z_2 = z_2$ is again normal, and we can give its mean and its covariance matrix explicitly.

If Z_1, Z_2, Z_3 are jointly normally distributed with means μ_p and covariance matrices $\Sigma_{pq} = E[(Z_p - \mu_p)(Z_q - \mu_q)']$, for $p, q = 1, 2, 3$, and assume that $\mu_3 = 0$ and $\Sigma_{23} = 0$. Then

$$E(Z_1|Z_2, Z_3) = E(Z_1|Z_2) + \Sigma_{13}\Sigma_{33}^{-1}Z_3$$

and

$$Var(Z_1|Z_2, Z_3) = Var(Z_1|Z_2) - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{13}'.$$

To illustrate how the filter works, we first look at a one-dimensional example. Let $\mathbf{X}^{(t-1)} = \{x_1, \ldots, x_{t-1}\}$ be the set of past observations from a time series $\mathbf{X}$ which arises in the state space model

$$\begin{aligned} X_t &= Y_t + \epsilon_t \\ Y_t &= Y_{t-1} + \eta_{t-1}, \end{aligned}$$

where ϵ_t is mean-zero normal with variance σ_ϵ^2 and η_t is mean-zero normal with variance σ_η^2; all independent.

Assume that the conditional distribution of Y_t given $\mathbf{X}^{(t-1)}$ is $\mathcal{N}(a_t, P_t)$, where a_t and P_t are to be determined. Given a_t and P_t, our objective is to calculate a_{t+1} and P_{t+1} when x_t, the next observation, arrives.

Now

$$\begin{aligned} a_{t+1} &= E(Y_{t+1}|\mathbf{X}^{(t)}) \\ &= E(Y_t + \eta_t|\mathbf{X}^{(t)}) \\ &= E(Y_t|\mathbf{X}^{(t)}) \end{aligned}$$

and

$$\begin{aligned} P_{t+1} &= Var(Y_{t+1}|\mathbf{X}^{(t)}) \\ &= Var(Y_t + \eta_t|\mathbf{X}^{(t)}) \\ &= Var(Y_t|\mathbf{X}^{(t)}) + \sigma_\eta^2. \end{aligned}$$

Define $v_t = x_t - a_t$ and $F_t = Var(v_t)$. Then

$$\begin{aligned} E(v_t|\mathbf{X}^{(t-1)}) &= E(Y_t + \epsilon_t - a_t|\mathbf{X}^{(t-1)}) \\ &= a_t - a_t = 0. \end{aligned}$$

Thus $E(v_t) = E(E(v_t|\mathbf{X}^{(t-1)})) = 0$ and

$$Cov(v_t, x_j) = E(v_t x_j) = E[E(v_t|\mathbf{X}^{(t-1)})x_j] = 0,$$

and as v_t and x_j are normally distributed, they are independent for $j = 1, \ldots, t-1$. When $\mathbf{X}^{(t)}$ is fixed, $\mathbf{X}^{(t-1)}$ and x_t are fixed, so $\mathbf{X}^{(t-1)}$ and v_t are fixed, and vice versa. Thus

$$E(Y_t|\mathbf{X}^{(t)}) = E(Y_t|\mathbf{X}^{(t-1)}, v_t)$$

and

$$Var(Y_t|\mathbf{X}^{(t)}) = Var(Y_t|\mathbf{X}^{(t-1)}, v_t).$$

Now we apply the conditional mean and variance formula for multivariate normally distributed random variables:

$$\begin{aligned} E(Y_t|\mathbf{X}^{(t)}) &= E(Y_t|\mathbf{X}^{(t-1)}, v_t) \\ &= E(Y_t|\mathbf{X}^{(t-1)}) + Cov(Y_t, v_t)Var(v_t)^{-1}v_t, \end{aligned}$$

where

$$\begin{aligned} Cov(Y_t, v_t) &= E(Y_t(x_t - a_t)) \\ &= E[Y_t(Y_t + \epsilon_t - a_t)] \\ &= E[Y_t(Y_t - a_t)] \\ &= E[(Y_t - a_t)^2] + a_t E[E(Y_t - a_t|\mathbf{X}^{(t-1)})] \\ &= E[(Y_t - a_t)^2] \\ &= E[E\{(Y_t - a_t)^2|\mathbf{X}^{(t-1)}\}] \\ &= E[Var(Y_t|\mathbf{X}^{(t-1)})] \\ &= P_t, \end{aligned}$$

and

$$\begin{aligned} Var(v_t) &= F_t \\ &= Var(Y_t + \epsilon_t - a_t) \\ &= Var(Y_t|\mathbf{X}^{(t-1)}) + \sigma_\epsilon^2 \\ &= P_t + \sigma_\epsilon^2. \end{aligned}$$

Put

$$K_t = \frac{P_t}{F_t}$$

then, since $a_t = E(Y_t|\mathbf{X}^{(t-1)})$, we have

$$E(Y_t|\mathbf{X}^{(t)}) = a_t + K_t v_t.$$

Now

$$\begin{aligned} Var(Y_t|\mathbf{X}^{(t)}) &= Var(Y_t|\mathbf{X}^{(t-1)}, v_t) \\ &= Var(Y_t|\mathbf{X}^{(t-1)}) - Cov(Y_t, v_t)^2 Var(v_t)^{-1} \\ &= P_t - \frac{P_t^2}{F_t} \\ &= P_t(1 - K_t). \end{aligned}$$

Thus the rule set of relations for updating from time t to $t+1$ is

$$\begin{aligned} v_t &= x_t - a_t \quad \text{Kalman filter residual; innovation} \\ a_{t+1} &= a_t + K_t v_t \\ F_t &= P_t + \sigma_\epsilon^2 \\ P_{t+1} &= P_t(1 - K_t) + \sigma_\eta^2 \\ K_t &= \frac{P_t}{F_t}, \end{aligned}$$

for $t = 1, \ldots, n$.

Note: a_1 and P_1 are assumed to be known; we shall discuss how to initialize later.

Now consider the more general model

$$\begin{aligned} Y_t &= G_t Y_{t-1} + v_t \\ X_t &= H_t Y_t + w_t, \end{aligned}$$

with $(v_t)_t$ independent white noise $WN(0, V_t)$, and $(w_t)_t$ ind. $WN(0, W_t)$. Here, Y_t is a vector representing unknown states of the system, and X_t are the observed data. . Put $\mathbf{X}^t = (X_1, X_2, \ldots, X_t)$, the history of X up to time t, and

$$\begin{aligned} Y_t^s &= E(Y_t|\mathbf{X}^s) \\ P_{t_1,t_2}^s &= E\{(Y_{t_1} - Y_{t_1}^s)(Y_{t_2} - Y_{t_2}^s)^T\} \\ &= E\{(Y_{t_1} - Y_{t_1}^s)(Y_{t_2} - Y_{t_2}^s)^T|\mathbf{X}_s\}. \end{aligned}$$

When $t_1 = t_2 = t$, we will write P_t^s for convenience.

Suppose $Y_0^0 = \mu$ and $P_0^0 = \Sigma_0$, and that the conditional distribution of Y_{t-1} given the history $\mathbf{X}^{t-1}$ up to time $t-1$,

$$\mathcal{L}(Y_{t-1}|\mathbf{X}^{t-1}) = \mathcal{MVN}(Y_{t-1}^{t-1}, P_{t-1}).$$

Then $\mathcal{L}(Y_t|\mathbf{X}^{t-1})$ is again multivariate normal. We have that

$$\begin{aligned} E(X_t|Y_t) &= H_t Y_t \\ Var(X_t|Y_t) &= W_t. \end{aligned}$$

With

$$R_t = G_t P_{t-1} G_t^{-1} + V_t$$

the conditional distribution of $(X_t, Y_t)^T$ given $\mathbf{X}^{t-1}$ is given by

$$\mathcal{L}\left(\begin{array}{c} X_t \\ Y_t \end{array} \middle| \mathbf{X}^{t-1} \right) = \mathcal{MVN}\left(\begin{pmatrix} H_t G_t Y_{t-1}^{t-1} \\ G_t Y_{t-1}^{t-1} \end{pmatrix}, \begin{pmatrix} W_t + H_t R_t H_t^T & H_t R_t \\ R_t H_t^T & R_t \end{pmatrix} \right).$$

We can compute that the conditional distribution of Y_t given $\mathbf{X}^{t-1}$ is multivariate normal with mean Y_t^t and variance $P_t^{(t-1)}$, where

$$\begin{aligned} Y_t^t &= G_t Y_{t-1}^{t-1} + R_t H_t^T (W_t + H_t R_t H_t^T)^{-1} (X_t - H_t G_t Y_{t-1}^{t-1}) \\ P_t^{(t-1)} &= R_t - R_t H_t^T (W_t + H_t R_t H_t^T)^{-1} H_t R_t. \end{aligned}$$

These equations are known as the *Kalman filter updating equations*. This solves the filtering problem.

Have a look at the expression for Y_t^t. It contains the term $G_t Y_{t-1}^{t-1}$, which is simply what we would predict if it were known that $Y_{t-1} = Y_{t-1}^{t-1}$, plus a term which depends on the observed error in forecasting, i.e. $(X_t - H_t G_t Y_{t-1}^{t-1})$.

Note that we initialized the recursion by $X_0^0 = \mu$ and $P_0^0 = \sigma_0$. Instead one might have initialized the recursion by some prior distribution, of by an uninformative prior $X_0^0 = 0, P_0^0 = kI$, where I denotes the identity matrix.

For *forecasting*, suppose $t > s$. By induction, assume we know Y_{t-1}^s, P_{t-1}^s. Then

$$\begin{aligned} Y_t^s &= G_t Y_{t-1}^s \\ P_t^s &= G_t P_{t-1}^s G_t^T + V_t. \end{aligned}$$

Recursion solves the forecasting problem.

The `R` command `predict(arima)` uses Kalman filters for prediction; see for example the airline passenger example, with the code on the course website.

We can calculate that the conditional distribution of X_{t+1} given $\mathbf{X}^t$ is

$$\mathcal{MVN}(H_{t+1} G_{t+1} Y_{t+1}^t, H_{t+1} R_{t+1} H_{t+1}^T + W_{t+1}).$$

This fact is the basis of the *prediction error decomposition*, giving us a likelihood for parameter estimation.

For smoothing we use the *Kalman smoother*. We proceed by backwards induction. Suppose that Y_t^t, P_t^t are known, where P_t^t is the conditional covariance matrix of $\mathbf{X}_t$ given $\{Y_1, \ldots, Y_t\}$. With a similar derivation as above, for $t = n, n-1, \ldots, 1$,

$$\begin{aligned} Y_{t-1}^n &= Y_{t-1}^{t-1} + J_{t-1}(Y_t^n - Y_t^{n-1}) \\ P_{t-1}^n &= P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1}) J_{t-1}^T \end{aligned}$$

where

$$J_{t-1} = P_{t-1}^{t-1} H^T (P_t^{t-1})^{-1}.$$

Note that these procedures differ for different initial distributions, and sometimes it may not clear which initial distribution is appropriate.

See also *Kalman filters made easy* by Terence Tong, at
`http://openuav.astroplanes.com/library/docs/writeup.pdf`.

Example: Johnson & Johnson quarterly earnings per share, 1960-1980. The model is

$$\begin{aligned} X_t &= T_t + S_t + v_t, \quad \text{observed} \\ T_t &= \phi T_{t-1} + w_{t1}, \quad \text{trend} \\ S_t &= S_{t-1} + S_{t-2} + S_{t-3} + w_{t2} \quad \text{seasonal component.} \end{aligned}$$

Assume that the seasonal components sum to zero over the four quarters, in expectation. Here w_t are i.i.d. mean-zero normal vectors with covariance matrix Q, and v_t are i.i.d. mean-zero normal with covariance R.

The state vector is

$$Y_t = (T_t, S_t, S_{t-1}, S_{t-2}).$$

See *Shumway and Stoffer*, p.334-336. The initial estimates are as follows. Growth is about 3 % per year, so choose $\phi = 1.03$. The initial mean is fixed at $(0.5, 0.3, 0.2, 0.1)^t$, and the initial covariance matrix is diagonal with $\Sigma_{0,i,i} = 0.01$, for $i = 1, 2, 3, 4$. Initial state covariance values were taken as $q_{11} = 0.01, q_{22} = 0.1$ to reflect relatively low uncertainty in the trend model compared to the seasonal model. All other elements of Q are taken to be 0. We take $R = 0.04$. Iterative estimation (using the EM algorithm) yielded, after 70 iterations, $R = .0086$, $\phi = 1.035$, $q_{11} = 0.0169, q_{22} = 0.0497$, and $\mu = (.55, .21, .15, .06)$.

5 Non-linear models

Note that this chapter and the next chapter were not covered in lectures.

Financial time series, e.g. share prices, share price indices, spot interest rates, currency exchange rates, have led to many specialized models and methods.

There are two main types:

- ARCH models

- Stochastic Volatility models

ARCH = autoregressive conditionally heteroscedastic

ARCH models are models analagous to ARMA models, but with AR and MA components which act on the variances of the process as well as, or instead of, the means.

Stochastic Volatility

In stochastic volatility models there is some unobserved process known as the volatility which directly influences the variance of the observed series. That is, these have some similar characteristics to state space models.

A review of ARCH / Stochastic Volatility models is:
Shephard (1996), which is Chapter 1 of *Time Series Models* (editors: Cox, Hinkley, Barndorff-Nielsen), Chapman and Hall

Usually we consider the daily returns y_t given by

$$y_t = 100 \log \left(\frac{x_t}{x_{t-1}} \right)$$

where x_t is the price on day t.

Common features of series of this type are:

- there is a symmetric distribution about the mean
- there is little autocorrelation among the values of y_t
- there is strong autocorrelation among the values of y_t^2
- the y_t have heavy tailed distributions (i.e. heavier tails than a normal distribution)
- the variance of the process changes substantially over time

Most models of financial time series are of the general structure

$$y_t \mid z_t \sim N(\mu_t, \sigma_t^2)$$

where z_t is some set of conditioning random variables (maybe lagged values of y_t) and μ_t and σ_t^2 are functions of z_t.

An example of an ARCH model is:

$$y_t \mid z_t \sim N(0, \sigma_t^2)$$

where

$$\begin{aligned} z_t &= (y_1, \ldots, y_{t-1}) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2. \end{aligned}$$

Clearly here the variance of y_t depends on lagged values of y_t.

An example of a stochastic volatility model is

$$y_t \mid h_t \sim N(0, e^{h_t})$$

where

$$\begin{aligned} h_{t+1} &= \gamma_0 + \gamma_1 h_t + \eta_t \\ \eta_t &\sim N(0, \sigma_\eta^2) \end{aligned}$$

with the variables η_t being independent as t varies.

The state variable h_t is not observed, but could be estimated using the observations. This situation is similar to that for state space models, but it is the variance (not the mean) of y_t that depends on h_t here.

5.1 ARCH models

The simplest ARCH model, ARCH(1), is

$$y_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2$$

with $\epsilon_t \sim N(0, 1)$, and the sequence of ϵ_t variables being independent. Here $\alpha_1 > 0$ has to be satisfied to avoid negative variances. Note that the conditional distribution of Y_t given $Y_{t-1} = y_{t-1}$ is

$$\mathcal{N}(0, \alpha_0 + \alpha_1 y_{t-1}^2).$$

Hence

$$E(Y_t) = E[E(Y_t|Y_{t-1})] = 0.$$

To calculate the variance, we re-write

$$\begin{aligned} y_t^2 &= \sigma_t^2 \epsilon_t^2 \\ \alpha_0 + \alpha_1 y_{t-1}^2 &= \sigma_t^2 \end{aligned}$$

so that

$$y_t^2 - (\alpha_0 + \alpha_1 y_{t-1}^2) = \sigma_t^2 \epsilon_t^2 - \sigma_t^2,$$

or

$$y_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + v_t,$$

with

$$v_t = \sigma_t^2(\epsilon_t^2 - 1).$$

Note that $\epsilon_t^2 \sim \chi_1^2$. Now

$$\begin{aligned} E(v_t) &= E[E(v_t|Y_{t-1})] \\ &= E[\sigma_t^2 E(\epsilon_t^2 - 1)] = 0, \end{aligned}$$

and furthermore

$$\begin{aligned} Cov(v_{t+h}, v_t) &= E(v_t v_{t+h}) = E[E(v_t v_{t+h}|Y_{t+h-1})] \\ &= E[v_t E(v_t v_{t+h}|Y_{t+h-1})] = 0. \end{aligned}$$

Thus the error process v_t is uncorrelated. If the variance of v_t is finite and constant in time, and if $0 \leq \alpha_1 < 1$, then y_t^2 is a causal AR(1)-process. In particular,

$$E(Y_t^2) = Var(Y_t) = \frac{\alpha_0}{1 - \alpha_1}.$$

In order for $Var(T_t^2) < \infty$ we need $3\alpha_1^2 < 1$.

As the conditional distribution of Y_t given the past is normal and easy to write down, to estimate parameters in an ARCH(1)-model, usually conditional maximum likelihood is used. For a wide class of processes, asymptotic normality of the estimators has been proven. A practical difficulty is that the likelihood surface tends to be flat, so that even for the simplest form ARCH(1), the masimum likelihood estimates of α_0 and α_1 can be quite imprecise.

5.2 GARCH and other models

The ARCH model can be thought of as an autoregressive model in y_t^2. An obvious extension of this idea is to consider adding moving average terms as well. This generalization of ARCH is called GARCH. The simplest GARCH model is GARCH(1,1):

$$y_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

The sequence is second-order stationary if $\alpha_1 + \beta_1 < 1$.

The simplest estimation scheme for the GARCH(1,1) model uses some initial sample of observations to obtain a crude estimate of σ_t^2, and then use maximum likelihood estimation based on the prediction error decomposition.

A further extension (EGARCH, where E is for exponential) is to model the log of σ_t^2 as a function of the magnitude, and of the sign, of ϵ_{t-1}.

The R command `garch` in the `tseries` package uses the Jarque-Bera test for normality, based on sample skewness and kurtosis. For a sample $x_1, \ldots, x_n$ the test statistic is given by

$$\frac{n}{6}\left(s^2 + \frac{(\kappa - 3)^2}{4}\right)$$

with

$$s = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2\right)^{\frac{3}{2}}}$$

the sample skewness, and

$$\kappa = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2\right)^{2}}$$

the sample kurtosis. For a normal distribution, the expected skewness is 0, and the expected kurtosis is 3. To test the null hypothesis that the data come from a normal distribution, the Jarque-Bera statistic is compared to the chi-square distribution with 2 degrees of freedom.

5.3 Stochastic volatility

The basic alternative to ARCH-type models is to allow σ_t^2 to depend not on past observations but on some unobserved components.

The log-normal stochastic volatility model is

$$y_t = \exp(h_t/2), \qquad h_{t+1} = \gamma_0 + \gamma_1 h_t + \eta_t$$

where $\epsilon_t \sim N(0,1)$ and $\eta_t \sim N(0, \sigma_\eta^2)$ are independent for all t.

The process h_t is strongly stationary if and only if $|\gamma_1| < 1$, and if h_t is stationary, then so is y_t. Means, and autocorrelations can be computed.

Estimation is not straightforward any more, as $\log \epsilon_t^2$ does not have a normal distribution. Often Monte-Carlo approaches are used: see MCMC lectures!

6 Further topics

6.1 Multivariate time series

Virtually all the above discussion generalizes when a vector is observed at each point in time. In the time domain, analysis would typically use cross-correlations and vector autoregressive-moving average models. In the frequency domain, dependencies at different frequencies are analysed separately.

6.2 Threshold models

For example when considering neuron firing in the brain, neurons are stimulated but will only fire once the stimulus exceeds a threshold. Then threshold models are used;

$$Y_{t+1} = g(Y_t) + \epsilon_t,$$

where $g(Y_t)$ is piecewise linear.

6.3 More general nonlinear models

Nonlinear time series are of the form

$$Y_{t+1} = g(Y_t) + \epsilon_t, \text{ or } \quad Y_{t+1} = g(Y_t, \epsilon_t),$$

where $g(y)$ or $g(y, \epsilon)$ is nonlinear.

For nonlinear time series, the amplitude (the periodogram) does not suffice to estimate the spectral density, and the acf; instead the phase is also needed. That is, we use vectors of time-delayed observations to describe the evolution of the system. For example, suppose our time series is

$$1, 3, 6, 7, 4, 2, 4, 5, 6$$

and we want to describe it in a 3-dim space, using a delay of 1: then our vectors are

$$(1, 3, 6); (3, 6, 7); (6, 7, 4); (7, 4, 2)$$

and so on, and we can see how these vectors move around in 3-dim space.

The interplay between randomness and nonlinearity generates new effects such as coexistence of fixed points, periodic points, and chaotic attractors, and new tools have been developed for these systems. In particular, nonlinear time series analysis uses many ideas from deterministic chaos theory.

6.4 Chaos

There is a large literature centering around the idea that some simple deterministic processes generate output that is very like a realization of a stochastic process. In particular it satisfies sensitivity to the initial conditions. This is a completely different approach to time series.