Time Series

Hilary Term 2002

Dr. Gesine Reinert

Outline

1. **The nature of time series**
   Types of data, examples, objectives, informal analysis, overview of techniques for time series analysis

2. **Stationary models**
   Weak and strong stationarity, some time-domain models, analysis in the frequency domain, state-space models, continuous-time models

3. **Statistical analysis**
   Precision of mean and sample autocorrelations, maximum-likelihood fitting, frequency domain

4. **Some more advanced topics**
   Multiple time series, nonlinear models, chaos

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. Much further reading is strongly recommended.

Organization of the class

This class will take place Tuesdays 9-10 and Thursdays 10-11, starting week 5, ending week 8, in the Department of Statistics. There will be a practical class on

Tuesday, week 7, 2:15–4:30, in the Computer lab in the department

and there will be an examples class on

Wednesday, week 7, 12-1, in the seminar room in the department.

Recommended reading

1. P.J. Brockwell and R.A. Davis (1996). *Introduction to Time Series and Forecasting.* Springer.

2. P.J. Brockwell and R.A. Davis (1991). *Time Series. Theory and Methods.* Second edition. Springer.

3. P.J. Diggle (1990). *Time Series.* Clarendon Press.

4. A.C. Harvey (1993). *Time Series Models.* Second edition. Harrister Wheatsheaf.

5. M.B. Priestley (1982). *Spectral Analysis and Time Series.* Academic Press.

6. R.L. Smith (2001) *Time Series.* At `http://www.stat.unc.edu/faculty/rs/s133/s133.html`

7. R.H. Shumway and D.S. Stoffer (2000). *Time Series Analysis and its Applications.* Springer.

1 The nature of time series

Often, observations are recorded at different times, usually but not necessarily at equally spaced time points. The dependence among these observations is of particular interest in time series analysis.

1.1 Types of data

To illustrate different types of data, consider the some examples.

1. Speech data. In Shumway and Stoffer (2000) there is a .1 second (1000 point) sample of recorded speech of the phrase $aaaa \ldots hhhh$. The signal is highly repetitive and has rather regular periodicities.

2. Total ozone series in Arosa, Switzerland.
The total ozone series of Arosa is the longest ozone series in the world. The measurements began in 1926; since 1988, the Swiss Meteorological Institute is responsible for operational measurements at Arosa. The total ozone amount is given in DU (Dobson units). The time series first fluctuates about a nearly constant value, then it seems to be decreasing. See
`http://www.lapeth.ethz.ch/doc/chemie/tpeter/totozon.html`.

3. Wolf sunspot series.
In 1848, R. Wolf devised a daily method of estimating solar activity by counting the number of individual spots and groups of spots on the face of the sun. To compute his sunspot number, he added 10 times the number of groups to the total count of individual spots. Today, Wolf sunspot counts continue, since no other index of the sun's activity reaches into the past as far. On average, the spot count takes about 4.8 years to rise from a minimum to a maximum, and another 6.2 years to fall to a minimum once again; see
`http://www.ece.ogi.edu/ ericwan/DATA/sunspots94.dat`.

4. EEG sleep data. This is a discrete-valued time series. In connection with a study on the effects of prenatal exposure to alcohol, the per-minute sleep state of infants is recorded by an EEG. Sleep state is categorized per minute, into one of six possible states: quiet sleep - trace alternant (1), quiet sleep - high voltage (2), transitional sleep (3), active sleep - low voltage (4), active sleep - high voltage (5), awake (6); see Shumway and Stoffer (2000).

5. A financial time series. Standard & Poor's Fund Services provide the S&P 500 index, reflecting stock market behaviour. In the S&P 500 series from 1960 to 16th October 1987, one might be more interested in the extreme value behaviour than in the average behaviour. From *A.J. McNeil. On Extremes and Crashes. RISK, January 1998: page 99.*

6. El Niño and fish population. An environmental series called the Southern Oscillation Index (SOI) and associated recruitment (number of new fish) for a period of 453 months ranging over 1950–1987 can be found in Shumway and Stoffer (2000). The SOI index measures changes in air pressure related to sea surface temperatures in the central pacific.About every 3-7 years there is a warming effect, called El Niño. This is an example for using two time series; here the question is to assess the effect of El Niño on fish population. Both series display repeating cycles; would these be related?

7. Gaussian white noise.
This is an erratic series; it is a collection of uncorrelated mean zero random variables with same, finite variance. See Shumway and Stoffer (2000), Diggle (1990).

The simplest form of data is a longish series of continuous measurements at equally spaced time points. Distinguish between aggregated values and point values. Also possible are discrete data, or mixtures between discrete and continuous. Of special interest are spatial and spatial-temporal time series. In medical statistics, one might want to analyze a number of possibly short series (corresponding to different individuals) of similar structure; such problems are also refereed to as *longitudinal data analysis*.

1.2 Objectives

The main objectives in time series analysis are

- Analysis and interpretation: find a model to describe the time dependence in the data. Can the model be interpreted?
- Forecasting; Given a finite sample from the series, forecast the next value, or the next several values
- Estimation of derived quantities, or signal extraction from observations of signal plus noise

- Control: How to adjust various control parameters to make the series fit closer to a target
- Adjustment: For example, in a linear model the errors could form a time series of correlated observations; one would then want to adjust the estimated variances to allow for this serial correlation.

1.3 Informal analysis

Suppose we observe a time series $(y_t)_{t \in \mathcal{T}}$, where $\mathcal{T}$ is some subset of the real numbers. Firstly, inspect the data for broad features, such as periodicity and trend, and for possible anomalies, such as outliers, missing observations, and discontinuities corresponding to changes of instrumentation or definition.

Typically one is then interested in some of the following features.

Long-term structure? These include drifts in mean, change of variability, etc. Plot means, etc, of non-overlapping blocks of data. For elimination of trend in order to study local structure

- take residuals from smooth, e.g. polynomial or other (spline, e.g.), fit (regression). Example: wool price data, Diggle (1990)
- take first or higher order differences (example: wool price data, Diggle (1990)
- take residuals from a suitable smoother, for example by a moving average operation. Example: female deaths in the UK attributed to bronchitis, emphysema, and asthma; Diggle (1990).

Periodic structure? Form a two-way table of, for example, years times month (or quarters) and examine marginal means.

Local structure? If necessary, eliminate trend etc. Plot y_{t+h} versus y_t for $h = 1$ and perhaps for $h = 2, 3, \ldots$. If appropriate linearity is present, compute correlations r_h and plot them against h; this is the *sample autocorrelation function (acf)*. Are there any oscillations? Or does it appear to be effectively random? Example: Wolf sunspot numbers, see Brockwell and Davis (1991), p. 32. These are *time-domain methods*.

Spectral analysis? Use a series expansion; for data $y_1, y_2, \ldots, y_n$ write

$$\begin{aligned} y_t &= \bar{y} + \sum_{p=1}^{\lfloor \frac{n}{2} \rfloor} \{ c_p cos(\omega_p t) + s_p sin(\omega_p t) \}, \\ \omega_p &= 2\pi \frac{p}{n}, \quad p = 1, \ldots, \lfloor \frac{n}{2} \rfloor. \end{aligned}$$

The phases are then $tan \phi_p = \frac{s_p}{c_p}$, and the squared amplitudes (powers) are $a_p^2 = c_p^2 + s_p^2$. Plot a scaling constant times a_p^2 versus ω_p; this plot is called a *periodogram*. Sometimes we are interested in the possibility of very high outlying values, in other cases it is the broad trend with ω_p that is of concern. Then some smoothing would again be in place. Example: SOI-data, Shumway and Stoffer (2000), p. 241.

In preliminary analysis of a long series it is often a good idea to split the data into, say, three or four sections and initially look at these separately.

1.4 An overview on techniques of time series analysis

Assume now that we sample at equally spaced time points, with a spacing of 1 time unit. Suppose our observations come from a stationary series $(Y_t)_{t=0,\pm 1, \pm 2, \ldots}$, i.e. all trends and other non-random effects have been removed. Three main examples are as follows.

AR(p). An *autoregressive process of order p (AR(p)-process)* is given by

$$Y_t = \sum_{r=1}^{p} \phi_r Y_{t-r} + \epsilon_t,$$

where $\phi_1, \ldots, \phi_p$ are fixed coefficients, and $(\epsilon_t)_t$ are independent errors (disturbances) with mean zero and variance σ^2: *white noise*. The name *white noise* stems from an analogy with white light, where all possible periodic oscillations are present with equal strength. Often the errors are assumed to be Gaussian; then $(\epsilon_t)_t$ is called *Gaussian white noise*.

MA(q). A *moving average process of order q (MA(q))* is given by

$$Y_t = \sum_{s=0}^{q} \theta_s \epsilon_{t-s},$$

where $\theta_0 = 1, \theta_1, \ldots, \theta_q$ are fixed coefficients, and $(\epsilon_t)_t$ is white noise.

ARMA(p,q). An *ARMA(p,q)-process* is given by

$$Y_t = \sum_{r=1}^{p} \phi_r Y_{t-r} + \sum_{s=0}^{q} \theta_s \epsilon_{t-s}, \tag{1}$$

where $\phi_1, \ldots, \phi_p$ and $\theta_0 = 1, \theta_1, \ldots, \theta_q$ are fixed coefficients, and $(\epsilon_t)_t$ is white noise.

Sometimes these models are written in short using the *back-shift operator* B,

$$By_t = y_{t-1}.$$

Define for complex numbers z, the *autoregressive polynomial* Φ,

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \tag{2}$$

with $\phi_p \neq 0$, and the *moving average polynomial* Θ,

$$\Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q, \tag{3}$$

with $\theta_q \neq 0$. With the notation

$$B^r y_t = y_{t-r},$$

an ARMA(p,q)-process is given by the solution of the *ARMA equations*

$$\Phi(B) Y_t = \Theta(B) \epsilon_t. \tag{4}$$

ARIMA(p,d,q). If the process is not stationary, then we would try to take differences and investigate whether the process is stationary. The *difference operator* is given by

$$\nabla y_t = y_t - y_{t-1};$$

so that

$$\nabla^2 y_t = \nabla(\nabla y_t) = y_t - 2y_{t-1} + y_{t-2}$$

and so on. If

$$Y_t = \nabla^d Z_t,$$

is an ARMA(p,q)-process, then $(Z_t)_{t=0,\pm 1,\pm 2,\ldots}$ is called an *integrated autoregressive moving average process*, in short, *ARIMA(p,d, q)*.

State-space approach. Sometimes a useful model for the data is of the form

$$\begin{aligned} Y_t &= S_t + \zeta_t \\ S_t &= S_{t-1} + \epsilon_t, \end{aligned} \tag{5}$$

where $(S_t)_t$ are unobservable states of the system. Here $(\zeta_t)_t$ are independent or uncorrelated, and $(\epsilon_t)_t$ is white noise. We will see later that the model (5) is equivalent to an ARMA(p,q)-model. However, more general state space models go much beyond the ARMA framework; models can be much more general than (5), for example

$$\begin{aligned} Y_t &= h_t(S_t, \zeta_t) \\ S_t &= g_t(S_{t-1}, \epsilon_t), \end{aligned} \tag{6}$$

where h_t and g_t are known functions.

A typical tool for analyzing these models are *Kalman filters*, to be encountered in a later section.

Further methods include

- nonlinear models
- models for irregularly spaced data
- continuous time models
- generalized linear models with dependent errors
- Bayesian analysis
- long-range dependence models, often based on fractals: use the difference operator ∇^d, where d is a fractional number.

Further reading

1. G.E. Box and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control.* Holden-Day.

2. P.J. Diggle, K.-L. Yang, and S.L. Zeger (Ed.) (1994). *The Analysis of Longitudinal Data.* Oxford Statistical Science.

2 Stationary models

Assume that we consider data $(y_1, y_2, \ldots, y_n)$ that is a realization of a random vector $(Y_1, Y_2, \ldots, Y_n)$,. That is, we consider an often largely hypothetical ensemble of repetitions of the data. Often we even assume that $(y_1, y_2, \ldots, y_n)$ is one finite realization of a sample of size n from a stochastic process $(Y_t)_{t=0,\pm1,\pm2,\ldots}$.

For many statistical purposes, the process $(Y_t)_{t=0,\pm1,\pm2,\ldots}$ should exhibit some redundancy. Either $(Y_t)_{t=0,\pm1,\pm2,\ldots}$ is itself stationary, or it can be reduced to some stationary process. A general model is

$$Y_t = m_t + s_t + Z_t, \quad t = 0, \pm1, \pm2, \ldots,$$

where m_t is a deterministic trend, s_t is a seasonal effect, and $(Z_t)_{t=0,\pm1,\pm2,\ldots}$ is a stationary stochastic process with mean zero. Real-life data are seldom stationary, but often they can be reduced to approximate stationarity by

- removing deterministic trend
- removing seasonal effects
- differencing
- transformations of the data.

2.1 Definitions

A process $(Y_t)_{t=0,\pm1,\pm2,\ldots}$ is called *second-order stationary* or *weak sense stationary* if there are $\mu, (\gamma_h)_h$ such that

$$\begin{aligned} \mathbf{E}Y_t &= \mu \qquad (7) \\ Cov(Y_t, Y_{t+h}) &= \gamma_h \quad \text{for all } t, h. \end{aligned}$$

In particular, $Var(Y_t) = \gamma_0$ is constant (and finite). Often we will call a second-order stationary processes just a *stationary* process.

A process $(Y_t)_{t=0,\pm1,\pm2,\ldots}$ is *strictly stationary* or *strong sense stationary* if for all $\tau, s, t_1, \ldots, t_s$, the two vectors

$$(Y_{t_1}, \ldots, Y_{t_s}) \quad \text{and} \quad (Y_{t_1+\tau}, \ldots, Y_{t_s+\tau})$$

have the same distribution.

A process $(Y_t)_{t=0,\pm1,\pm2,\ldots}$ is *Gaussian* if the joint distribution of any subset of values is multivariate normal. This distribution is completely determined by its mean vector and covariance matrix. Hence, if a Gaussian process satisfies the second-order stationary condition, it must also be strictly stationary.

For a stationary process, the function γ_h as a function of h is called the *autocovariance function*, and

$$\rho_h = \frac{\gamma_h}{\gamma_0}$$

is called the *autocorrelation function* (acf). When necessary, define $\gamma_{-h} = \gamma_h$. Thus the acf describes the second-order properties of the time series.

Studies of processes based on γ_h or on similar higher-moment properties are said to be in the *time domain*, to be contrasted later with properties based on a Fourier series-like analysis, which are said to be in the *frequency domain*.

2.2 Some time-domain models

Suppose $(Y_t)_{t=0,\pm1,\pm2,\ldots}$ is a stationary process with $\mu = 0$, where μ is given in (7.

Example 1: White noise. A set of uncorrelated random variables of zero mean and finite variance is stationary with $\gamma_h = 0$ for $h \neq 0$.

Example 2: AR(1). Suppose

$$\begin{aligned} Y_t &= \phi Y_{t-1} + \epsilon_t, \quad t > 0 \\ Y_0 &= Z_0. \end{aligned}$$

One also says that the system is forced by the innovation ϵ_t. Here, $(\epsilon_t)_t$ is white noise with $Var(\epsilon_t) = \sigma_\epsilon^2$, and ϵ_t is uncorrelated with Y_t and indeed with all previous values of Y. Backward recursion gives

$$Y_t = \phi^k Y_{t-k} + \sum_{j=0}^{k-1} \phi^j \epsilon_{t-j}.$$

1. If $|\phi| > 1$ then the only stationary solution would be

$$Y_t = -\sum_{j=1}^{\infty} \frac{1}{\phi^j} \epsilon_{t+j}.$$

   This solution depends on the future. We do not study this solution here.

2. If $|\phi| = 1$ then the process is a random walk with $Var(Y_t) = \sigma_\epsilon^2 t$.

3. Provided that $|\phi| < 1$ and that the variance of Y_t is bounded, we can represent an AR(1) model by

$$Y_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} \quad \text{(almost surely)}, \tag{8}$$

   regarded as starting in the remote past

4. If $|\phi| < 1$ and if Z_0 is replaced by a random variable having the stationary distribution of the process, with mean zero and variance $\frac{\sigma_\epsilon^2}{1-\phi^2}$, then stationarity can be verified. Usually in a representation of the type (8) we shall omit the *almost surely* from now on.

Properties can now be obtained from the infinite series, or using the following type of argument. Due to stationarity, Y_{t-1} and Y_t have the same variance σ_Y^2, so

$$\sigma_Y^2 = \phi^2 \sigma_Y^2 + \sigma_\epsilon^2,$$

or

$$\sigma_Y^2 = \frac{\sigma_\epsilon^2}{1 - \phi^2}.$$

Similarly,

$$\gamma_{h+1} = \mathbf{E} Y_t Y_{t+h+1} = \mathbf{E} Y_t (\phi Y_{t+h} + \epsilon_{t+h+1}) = \phi \gamma_h,$$

so

$$\rho_h = \phi^h.$$

Thus, if $0 < \phi < 1$ then the observations at two consecutive times are positively correlated. If $-1 < \phi < 0$, then the observations at two consecutive times are negatively correlated, whereas the observations two time steps apart are positively correlated.

When a process does not depend on the future, such as AR(1) when $|\phi| < 1$, the process is called *causal*. More generally, an ARMA process $(Y_t)_t$ is causal if it can be represented as

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

Example 3: MA(1). Suppose

$$Y_t \;=\; \epsilon_t + \theta \epsilon_{t-1}, \quad t = 0, \pm 1, \pm 2, \ldots.$$

Then $\gamma_0 = (1+\theta^2)\sigma_\epsilon^2$, $\gamma_1 \neq 0$, and $\gamma_h = 0$ for $h > 1$. Thus the process is stationary.

In general, MA(q) is stationary for any choice of $\theta_1, \ldots, \theta_q$. The acf of an MA($q$) is given by

$$\rho_h \;=\; \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1+\theta_1^2+\cdots+\theta_q^2}, & 1 \leq h \leq q \\ 0, & h > q. \end{cases}$$

Thus the acf vanishes for $h > q$. This can be used as a diagnostics for an MA process.

In particular, the acf of $MA(1)$ is given by

$$\rho_h \;=\; \begin{cases} \frac{\theta}{1+\theta^2}, & h = 1 \\ 0, & h > 1. \end{cases}$$

Thus two MA(1) processes defined by θ and by $\frac{1}{\theta}$ are identical for all practical purposes. This leads to the *identifiability condition*: Recall the *MA polynomial* $\Theta(z)$ given in (3). Then the identifiability condition states that $\Theta(z) \neq 0$ for all z such that $|z| \leq 1$,that is, all zeros of $\Theta(z)$ lie outside the unit circle.

The results below will usually assume that an MA(1) process is identifiable. However, most of them can be extended to include the case that $|\theta| = 1$, if the notion of invertibility is generalised to assume only that $\epsilon_t \in span\{Y_s, -\infty < s \leq t\}$.

For the above MA(1) process this implies that $|\theta| < 1$.

Note that we can invert the roles of Y and ϵ in the MA(1)-process and write

$$\epsilon_t = -\theta\epsilon_{t-1} + Y_t.$$

If $|\theta| < 1$ then we have the infinite AR representation of the model:

$$\epsilon_t = \sum_{j=0}^{\infty}(-\theta)^j Y_{t-j}.$$

Such a process is called an *invertible* process. (Thus an invertible MA(1) process can be represented by a causal AR(∞)-process.)

Similarly recall the definition (2) of the *AR polynomial* $\Phi(z)$.

Example. Consider the process

$$Y_t - \frac{1}{2}Y_{t-1} = \epsilon_t - \frac{1}{2}\epsilon_{t-1}.$$

At first glance this looks like an ARMA(1,1)-process. However, in operator form, we obtain

$$\left(1 - \frac{1}{2}B\right) Y_t = \left(1 - \frac{1}{2}B\right) \epsilon_t,$$

that is, a solution of the above equation is $Y_t = \epsilon_t$, so Y_t is white noise. Therefore we typically assume that the AR polynomial $\Phi(z)$ and the MA polynomial $\Theta(z)$ have no common factors.

From the ARMA equations, heuristically we might write

$$Y_t = \Phi(B)^{-1}\Theta(B)\epsilon_t$$

for causality, and

$$\epsilon_t = \Theta(B)^{-1}\Phi(B)Y_t$$

for invertibility. However, it is not obvious that these operators would be invertible.

Theorem 1 *Suppose an ARMA(p,q)-process has AR polynomial $\Phi(z)$ and MA polynomial $\Theta(z)$, where $\Phi(z)$ and $\Theta(z)$ have no common factors. Then the ARMA(p,q)-process is causal only if the roots of $\Phi(z)$ lie outside the unit circle. Then*

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \tag{9}$$

where the coefficients are given by solving

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)}, \quad |z| \leq 1.$$

Theorem 2 *Suppose an ARMA(p,q)-process has AR polynomial $\Phi(z)$ and MA polynomial $\Theta(z)$, where $\Phi(z)$ and $\Theta(z)$ have no common factors. Then the ARMA(p,q)-process is invertible only if the roots of $\Theta(z)$ lie outside the unit circle. Then*

$$\epsilon_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j},$$

where the coefficients are given by solving

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\Phi(z)}{\Theta(z)}, \quad |z| \leq 1.$$

For the proofs, see Brockwell & Davis (1991), pp. 85-87.

- If the ARMA(p,q)-process $(Y_t)_t$ is such that the polynomials $\Phi(z)$ and $\Theta(z)$ have common factors, then there are two possibilities.

  1. None of the common zeros lie on the unit circles, in which case $(Y_t)_t$ is the unique stationary solution of the ARMA equations with no common zeroes, obtained by cancelling the common factors of $\Phi(z)$ and $\Theta(z)$.
  2. At least one of the common zeros lies one the unit circle, in which case the ARMA equations may have more than one stationary solution. (See Brockwell & Davis (1991), Problem 3.24.)

- If $\Phi(z)$ and $\Theta(z)$ have no common factors and if $\Phi(z)$ has a zero lying on the unit circle, then there is no stationary solution to the ARMA equations (4) .

For a causal ARMA(p,q)-process we have, from (9), that

$$\gamma_h = Cov(Y_t, Y_{t+h}) = \sigma_\epsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \quad h \geq 0.$$

Either we could solve for the ψ's in $\Psi(z)\Phi(z) = \Theta(z)$ by comparing coefficients, or we could use that

$$E(Y_t \Phi(B) Y_{t+h}) = E(Y_t \Theta(B) \epsilon_{t+h}),$$

giving

$$\gamma_h = \sum_{j=1}^{p} \phi_j \gamma_{h-j} - \sigma_\epsilon^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}, \quad h \geq 0.$$

This leads to the difference equations

$$\begin{aligned}
\gamma_h - \sum_{j=1}^{p} \phi_j \gamma_{h-j} &= 0, \quad h \geq max(p, q+1) \\
\gamma_h - \sum_{j=1}^{p} \phi_j \gamma_{h-j} &= \sigma_\epsilon^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}, \quad 0 \leq h < max(p, q+1).
\end{aligned}$$

Example. To calculate the acf of an ARMA(1,1)-process

$$Y_t = \phi Y_{t-1} + \theta \epsilon_{t-1} + \epsilon_t$$

with $|\phi| < 1, |\theta| < 1$, we have

$$\gamma_h - \phi \gamma_{h-1} = 0, \quad h = 2, 3, \ldots$$

so

$$\gamma_h = c\phi^h$$

for some constant c. To determine c, note that

$$\begin{aligned}
\gamma_0 &= \phi\gamma_1 + \sigma_\epsilon^2 (1 + \theta\phi + \theta^2) \\
\gamma_1 &= \phi\gamma_0 + \sigma_\epsilon^2 \theta.
\end{aligned}$$

Solving this system gives

$$\begin{aligned}
\gamma_0 &= \sigma_\epsilon^2 \frac{1+2\theta\phi+\theta^2}{1-\phi^2} \\
\gamma = 1 &= \sigma_\epsilon^2 \frac{(1+\theta\phi)(\theta+\phi)}{1-\phi^2}
\end{aligned}$$

and, as $\gamma_1 = \phi$,

$$c = \frac{\gamma_1}{\phi},$$

giving

$$\gamma_h = \sigma_\epsilon^2 \frac{(1+\theta\phi)(\theta+\phi)}{1+2\theta\phi+\theta^2}\phi^{h-1}, \quad h \geq 0,$$

and

$$\rho_h = \frac{(1+\theta\phi)(\theta+\phi)}{1+2\theta\phi+\theta^2}\phi^{h-1}, \quad h \geq 1. \tag{10}$$

Example. For an AR(p) process, (10) translates into the *Yule Walker equations*

$$\rho_h = \sum_{l=1}^{p} \phi_l \rho_{h-l}.$$

Let $z_1, \ldots, z_r$ denote the roots of $\Phi(z)$, with multiplicities $m_1, \ldots, m_r$. Then (see Shumway & Stoffer. pp.109-110) we obtain

$$\rho_h = z_1^{-h} P_1(h) + z_2^{-h} P_2(h) + \cdots + z_r^{-h} P_r(h), \quad h \geq p,$$

where $P_j(h)$ is a polynomial of degree $m_j - 1$.

- If all roots are real, then ρ_h dampens to zero geometrically fast, as, from causality, all roots have absolute value larger than 1.
- The rate of decay depends on the roots that are closest to the unit circle.
- A pair of complex conjugate zeros together contribute a geometrically damped sinusoidal term.

Note that we can interpret the causal representation of an ARMA process as an MA(∞) representation.

2.3 Prediction

If we make the wholly unrealistic assumption that the parameters in the process are known, then the AR-representation specifies the optimal mean square predictor one step ahead, and by modest extension of the argument, k steps ahead, as follows. Denote the k-step ahead predictor given the observations $y_1, \ldots, y_n$ by

$$y_{n+k}^k = E(Y_{n+k} | Y_1 = y_1, \ldots, Y_n = y_n).$$

Consider the one-step ahead predictor y_{n+1}^1, the best linear predictor. This predictor is chosen to minimize the mean-square error,

$$P_{n+1}^n = E(Y_{n+1} - Y_{n+1}^1)^2.$$

It can be shown (Shumway & Stoffer, p. 115) that the best linear predictor is found by solving the *prediction equations*

$$\begin{aligned} E(Y_{n+1} - Y_{n+1}^1) &= 0 \\ E((Y_{n+1} - Y_{n+1}^1)Y_k) &= 0, \quad k = 1, \ldots, n. \end{aligned} \tag{11}$$

Thus y_{n+1}^1 will be a linear combination of $y_1, \ldots, y_n$; write

$$y_{n+1}^1 = \phi_{n1} y_n + \phi_{n2} y_{n-1} + \cdots + \phi_{nn} y_1.$$

Then the prediction equations (11) yield that

$$E((Y_{n+1} - \sum_{j=1}^n \phi_{nj} Y_{n+1-j}) Y_{n+1-k}) = 0, \quad k = 1, \ldots, n$$

giving that

$$\sum_{j=1}^n \phi_{nj} \gamma_{k-j} = \gamma_k, \quad k = 1, \ldots, n. \tag{12}$$

We can write (12) in matrix form: Let

$$\begin{aligned} \Gamma_n &= (\gamma_{k-1})_{j,k=1,\ldots,n} \\ \underline{\gamma}_n &= (\gamma_1, \ldots, \gamma_n)^T \\ \underline{\phi}_n &= (\phi_{n1}, \ldots, \phi_{nn})^T, \end{aligned}$$

then (12) translates into

$$\Gamma_n \underline{\phi}_n = \underline{\gamma}_n.$$

If $\sigma_\epsilon^2 > 0$ and if $\gamma_h \to 0$ for $h \to \infty$, then Γ_n is invertible, and

$$\underline{\phi}_n = \Gamma_n^{-1} \underline{\gamma}_n.$$

This matrix equation can be solved recursively using the *Levinson-Durbin recursion.* Start with $\phi_{00} = 0, P_1^0 = \gamma_0$. For $n \geq 1$,

$$\begin{aligned} \phi_{nn} &= \frac{\rho_n - \sum_{k=1}^{n-1} \phi_{n-1,k}\rho_{n-k}}{1 - \sum_{k=1}^{n-1} \phi_{n-1,k}\rho_k} \\ P_{n+1}^n &= P_n^{n-1}(1 - \phi_{nn}^2), \end{aligned} \tag{13}$$

and, for $n \geq 2$,

$$\phi_{nk} = \phi_{n-1,k} - \phi_{nn}\phi_{n-1,n-k}, \quad k = 1, \ldots, n-1.$$

Thus the first steps of the recursion give

$$\begin{aligned} \phi_{00} &= 0, \quad P_1^0 = \gamma_0 \\ \phi_{11} &= \rho_1, \quad P_2^1 = \gamma_0(1 - \phi_{11}^2) = \gamma_0(1 - \rho_1^2) \\ \phi_{22} &= \frac{\rho_2 - \phi_{11}\rho_1}{1 - \phi_{11}\rho_1} = \frac{\rho_2 - \rho_1^2}{-\rho_1^2} \\ \phi_{21} &= \phi_{11} - \phi_{22}\phi_{11} = \rho_1(1 - \phi_{22}) \\ P_3^2 &- \gamma_0(1 - \phi_{11}^2)(1 - \phi_{22}^2) \\ \phi_{33} &= \frac{\rho_3 0 - \phi_{21}\rho_2 - \phi_{22}\rho_1}{1 - \phi_{21}\rho_1 - \phi_{22}\rho_2}. \end{aligned}$$

Example. For an AR(2)-process we have

$$\begin{aligned} \rho_1 &= \frac{\phi_1}{1 - \phi_2} \\ \rho_h - \phi_1\rho_{h-1} - \phi_2\rho_{h-2} &= 0, \quad h \geq 2, \end{aligned}$$

hence we obtain

$$\begin{aligned} \phi_{11} &= \frac{\phi_1}{1 - \phi_2} \\ \phi_{22} &= \phi_2 \\ \phi_{21} &= \phi_1 \\ \phi_{33} &= 0. \end{aligned}$$

We obtain $Y_2^1 = \rho_1 Y_1$, which is what we would have expected from linear regression. Moreover, $Y_3^2 = \phi_1 Y_1 + \phi_2 Y_2$, from the model, and similarly all predictions for higher values will use the AR(2) equation directly.

Note that similarly to forecasting one can also backcast. Let Y_0^{h-1} be the minimum mean square linear predictor of Y_0 based on $\{Y_1, \dots, Y_{h-1}\}$. This leads to the same prediction equations as for forecasting, see Shumway & Stoffer (2000), p.124.

2.4 The partial autocorrelation function

One reason for calculating the acf is that if $(Y_t)_t$ is MA(q) then $\rho_h = 0$ for $|h| > q$, so plots of the sample acf should show a sharp drop to near zero after the qth coefficient; this is a diagnostic tool for an MA(q)-process. A corresponding tool for AR(p) is given by the *partial autocorrelation function* (pacf); the pacf of lag h is just ϕ_{hh}, given by (13). Note that

$$\begin{aligned} \phi_{11} &= corr(Y_1, Y_0) = \rho_1 \\ \phi_{hh} &= corr(Y_h - Y_h^{h-1}, Y_0 - Y_0^{h-1}), \quad h \geq 2. \end{aligned}$$

Thus ϕ_{nn} is the correlation between Y_t and Y_{t+h} with the linear effect of $\{Y_{t+1}, \dots, Y_{t+h-1}\}$, on each, removed. Another way of defining the pacf is (see Shumway & Stoffer (2000), p.111),

$$\begin{aligned} \phi_{11} &= Corr(Y_1, Y_0) = \rho_1 \\ \phi_{hh} &= Corr(Y_h - Y_h^{h-1}, Y_0 - Y_0^{h-1}). \end{aligned}$$

Example. Causal AR(1). The prediction of Y_2 based on Y_1 will be a linear function, $Y_2^1 = \alpha Y_1$. The mean-square error is

$$\begin{aligned} E(Y_2 = Y_2^1)^2 &= E(Y_2 - \alpha Y_1)^2 \\ &= \gamma_0 - 2\alpha\gamma_1 + \alpha^2\gamma_0, \end{aligned}$$

hence

$$\alpha = \frac{\gamma_1}{\gamma_0} = \rho_1 = \phi,$$

and

$$\phi_{22} = Corr(Y_2 - \phi Y_1, Y_0 - \phi Y_1) = \gamma_2 - 2\phi\gamma_1 + \phi^2\gamma_0 = 0.$$

Moreover $\phi_{hh} = 0$ for all $h > 1$.

Example. Causal AR(p). Here, for $h > p$, we have

$$Y_h^{h-1} = \sum_{j=1}^{p} \phi_j Y_{h-j},$$

hence

$$\phi_{hh} = Corr(\epsilon_h, Y_h = Y_h^{h-1}) = 0, \quad h > p.$$

This can be used as a diagnostic tool for an AR(p)-process.

Other examples.

1. MA(1): Here $\phi_{hh} = \frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1}}, \quad h \geq 1.$
2. MA(q): The pacf will never cut off.

Remarks.

1. Processes for which ρ_h decays like $h^{-(1+\delta)}$ for some $0 < \delta < 1$ are said to have *long-range dependence*.
2. First or higher order differences can be modelled by an ARMA(p,q) model, leading to ARIMA(p,d,q) model. Seasonal models. usually of a somewhat artificial kind, can be produced by differencing s terms apart. The family of models resulting from these observations are called *Box-Jenkins models*. They are flexible but typically wholly empirical.
3. There are general theorems expressing a broad class of stationary processes in autoregressive and in moving average form.
4. Processes of AR form in which the innovations are not merely uncorrelated but i.i.d. are called *linear processes.*
5. A linear process is time-reversible if and only if it is Gaussian.
6. The Gaussian AR(1) is a Markov process.

2.5 Fitting ARIMA models using the Box-Jenkins Approach

The classical Box-Jenkins approach to fitting ARIMA models can be decomposed into *identification, estimation,* and *verification.*

Identification

First we need to assess whether the observations come indeed from a stationary process. For this purpose, the acf should decay to zero fairly rapidly. If this is not the case, (repeated) differencing would be in place, until the acf would decay to zero fairly rapidly. If differencing seems to increase the variance, the model might be over-differenced. Note that in models with long-range dependence the acf would not decay to zero rapidly; typically, the autocorrelations tend to zero hyperbolically, that is, $\rho_h \sim h^{-\alpha}$, with $\alpha > 0$; in this case, so-called *fractionally differenced* ARIMA models would be in place. So differencing might not always be successful.

Once the series is accepted as stationary, the next step is initial identification of p and q. For this we use the acf and the pacf. An MA(q) series is identified from the property that all values of the acf after the qth are negligible, whereas an AR(p) series is identified from the property that all values of the pacf after the pth are negligible. To determine whether values of the acf, or the pacf, are negligible, use as a very rough approximation that acf and pacf have a standard deviation of around $\frac{1}{\sqrt{n}}$. As a rule of thumb, $\pm\frac{2}{\sqrt{n}}$ would give very approximate 95 % confidence bounds. In S-PLUS these are shown as dotted lines.

Note that for diagnostics of an ARMA(p,q) process, these two criteria cannot be combined easily, due to the contribution of the MA-part in the pacf of the process, and due to the contribution of the AR-part in the acf of the process.

Estimation: Yule-Walker estimators

For AR(p) models, the method of moments is very useful. From $Y_t = \sum_{j=1}^p \phi_j Y_{t-j} + \epsilon_t$, $(\epsilon_t)_t$ being i.i.d. WN$(0, \sigma_\epsilon^2)$, we obtain the *Yule-Walker equations*

$$\gamma_h = \sum_{i=1}^p \phi_j \gamma_{|j-h|}, \tag{14}$$

for $h > 0$. For $1 \leq h \leq p$, these are p equations in p unknowns $\phi_1, \ldots, \phi_p$, hence (14) can be solved. Indeed, a Durbin-Levinson type recursion can be used again. Note that in practice we will use $\hat{\gamma}_h, h = 1, \ldots, p$ instead of $\gamma_1, \ldots, \gamma_p$.

In matrix notation, put $\Gamma_p = \{\gamma_{i-j}\}_{i,j=1}^{\infty}$, and $\underline{\gamma}_p = (\gamma_1, \ldots, \gamma_p)^T$. Then the *Yule-Walker estimators* $\underline{\hat{\phi}}$ and $\hat{\sigma_\epsilon^2}$ of $\underline{\phi}$ and σ_ϵ^2 are given by the solutions of

$$\begin{aligned} \hat{\Gamma}_p \underline{\hat{\phi}} &= \underline{\hat{\gamma}}_p \\ \hat{\sigma_\epsilon^2} &= \underline{\hat{\gamma}}_0 - \underline{\hat{\phi}}^T \underline{\hat{\gamma}}_p. \end{aligned}$$

This gives

$$\begin{aligned} \underline{\hat{\phi}} &= \hat{\Gamma}_p^{-1} \underline{\hat{\gamma}}_p \\ \hat{\sigma_\epsilon^2} &= \underline{\hat{\gamma}}_0 - \underline{\hat{\phi}}^T \underline{\hat{\gamma}}_p. \end{aligned}$$

If $(Y_t)_t$ is a causal AR(p) process with i.i.d. WN$(0, \sigma_\epsilon^2)$, then (see Brockwell and Davis (1991), p.241)

$$\sqrt{n} \left(\underline{\hat{\phi}} - \underline{\phi} \right)$$

is approximately $\mathcal{MVN}(\underline{0}, \sigma_\epsilon^2 \Gamma_p^{-1})$-distributed, and $\hat{\sigma_\epsilon^2}$ converges to σ_ϵ^2 in probability.

The Yule-Walker estimator $\underline{\hat{\phi}}$ 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 $m > p$,

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

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

The residual variance

$$\hat{\sigma}_p^2 = \frac{1}{n} \sum_{t=p+1}^{n} \left(Y_t - \sum_{j=1}^{p} \hat{\phi}_j Y_{t-j} \right)^2$$

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

$$-2 \log L = n \log(\sigma_p^2),$$

then this can be used either for likelihood ratio tests, or by minimizing the $AIC = -2\log L + 2p$. Note that this should not be applied in a totally undiscriminatory way, as one would still like to be able to interpret the results.

Estimation: Maximum likelihood estimators

For general ARMA(p,q) models, the Yule-Walker estimator is not optimal. If a parametric model for the white noise is assumed, then maximum likelihood estimation can be used. Mostly this relies on the *prediction error decomposition*: Use similar ideas as in the state-space formulation of an ARMA(p,q). The $Y_1, Y_2, \ldots, Y_n$ have joint density

$$f(Y_1, Y_2, \ldots, Y_n) \quad = \quad f(Y_1) \prod_{t=2}^{n} f(Y_t | Y_s, 1 \leq s \leq t-1).$$

Assume Gaussian WN, and, as for the Kalman filter, that

$$\mathcal{L}(Y_t | Y_s, 1 \leq s \leq t-1) = \mathcal{N}(\hat{Y}_t, P_t^{t-1})$$

and

$$\mathcal{L}(Y_1) = \mathcal{N}(\hat{Y}_1, P_1^0).$$

Then for the log likelihood we obtain

$$-2\log L \quad = \quad \sum_{t=1}^{n} \left\{ \log(2\pi) + \log P_t^{t-1} + \frac{(Y_t - \hat{Y}_t)^2}{P_t^{t-1}} \right\}. \tag{15}$$

Thus the log likelihood is written in terms of the *innovations* $\epsilon_t = Y_t - \hat{Y}_t$. These innovations are independent Gaussian mean zero, hence this formulation is more amenable to analysis. Here, $\hat{Y}_t$ and P_t^{t-1} are functions of the unknown parameter

$$\Theta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$$

and (15) can be minimized with respect to Θ, giving the maximum likelihood estimator $\hat{\Theta}$. In general, numerical approximations to the mle will be needed here.

The second derivative of $-2\log L$, evaluated at the mle $\hat{\Theta}$, is the observed information matrix, and its inverse is an approximation of the variance-covariance matrix of the estimator. Asymptotic normality holds, also for non-Gaussian but "regular" white noise, see Brockwell and Davis (1991), p.386. Thus we obtain standard errors for the parameters $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$.

In practice, for AR(p), for example, $\hat{Y}_t = \sum_{j=1}^{p} \hat{\phi}_j Y_{t-j}$ for $t > p$ is obtainable, but it is not so obvious how to obtain the corresponding quantity for $t \leq p$; it requires some consideration of the stationary distribution of the process. A similar argument holds for ARMA(p,q) models. Hence the exact likelihood function is often replaced by a conditional likelihood function. One conditions on the first m values of the series, where $m \geq p$ is small. The conditional likelihood is then given by

$$-2 \log L_m = \sum_{t=m+1}^{n} \left\{ \log(2\pi) + \log P_t^{t-1} + \frac{(Y_t - \hat{Y}_t)^2}{P_t^{t-1}} \right\}.$$

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

Verification

Two main techniques for model verification are

1. Overfitting: Add extra parameters to the model and use a likelihood ratio test to see that these are not significant
2. Residual analysis: Calculate residuals from the fitted model and plot their acf, pacf, spectral density estimates, etc. to check whether they are consistent with white noise.

Another possibility is the *Box-Pierce test* (*portmanteau test*) based on

$$Q = n \sum_{h=1}^{K} r_h^2$$

where $K > p + d + q$ but much smaller than n. Here, r_h is the hth sample autocorrelation of the residual series. If the model is correct, then, asymptotically, Q follows a Chi-square distribution with $K - (p + d + q)$ degrees of freedom. This procedure appears also in S-PLUS. We would reject the model at level α if

$$Q > \chi^2_{1-\alpha}(K - p - d - q).$$

An improved test is the *Box-Ljung procedure*:

$$\tilde{Q} = n(n+2)\sum_{h=1}^{K} \frac{r_h^2}{n-h};$$

the distribution of $\tilde{Q}$ is closer to a Chi-square distribution with $K-(p+d+q)$ degrees of freedom if the model is correct. A problem here is the rather arbitrary choice of K.

For ARIMA(p, d, q) models, the main steps of identification, estimation and verification are the same as above. Generalizations of ARIMA(p, d, q) include

- seasonal ARIMA (SARIMA); in examining autocorrelations, particular attention must be paid to the values at or near multiples of the period
- ARMAX (ARMA with exogenous process)
- fractionally differenced ARIMA, where the fractional difference operator is defined as

$$\nabla^d = (1-B)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} B^j.$$

2.6 Analysis in the frequency domain

The idea here is to express the regularity of a series in terms of periodic variations of the underlying phenomenon. Thus it heavily relies on Fourier methods.

The spectral theory for stationary processes is based on the following fact. For any sequence $\{\gamma_h\}_h$ of autocovariances generated by a stationary process, there exists a function F such that

$$\gamma_h = \int_{(-\pi,\pi]} e^{ih\lambda} dF(\lambda), \tag{16}$$

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 around zero: for $0 \leq a < b \leq \pi$,

$$F(b) - F(a) = F(-a) - F(-b).$$

The function F is called the *spectral distribution function* (see Smith (2001)). Note that this is not necessarily a probability distribution function, as $F(\pi) = 1$ is not required. The interpretation is that, for $0 \leq a < b \leq \pi$, $F(b) - F(a)$ measures the contribution to the total variability of the process within the frequency range $(a, b]$.

If F is everywhere continuous and differentiable, then

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

is called the *spectral density function*, and (16) becomes

$$\gamma_h = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda. \tag{17}$$

It $\sum_{h\geq 0} |\gamma_h| < \infty$, then it can be shown that f always exists, and is given by

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

We observe that, due to symmetry, $f(\lambda) = f(-\lambda)$.

The spectrum and the acf contain equivalent information concerning the underlying sequence. However, the spectrum has a more tangible interpretation in terms of the inherent tendency for realizations of $\{Y_t\}_t$ to exhibit cyclic variations about the mean.

Note that some textbooks define the spectral density function in a different form; Shumway & Stoffer (2000) use

$$\gamma_h = \int_{-1/2}^{1/2} e^{2\pi i\nu h} f(\nu) d\nu.$$

This is just the same integral as above, using a change of variable.

Sometimes it is also convenient to consider the *normalized* spectral density function

$$f^*(\lambda) = \frac{f(\lambda)}{\sigma_\epsilon^2} = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \rho_h e^{i\lambda h},$$

see also the Exercise sheet.

Example. $\mathrm{WN}(0, \sigma^2)$, $\sigma^2 > 0$. Here, $\gamma_0 = \sigma^2$, $\gamma_h = 0$ for $h \neq 0$, and

$$f(\lambda) = \frac{\sigma^2}{2\pi},$$

independent of λ. The spectral distribution function is uniform. The converse also holds: A process is WN if and only if its spectral density is constant. Indeed, a quantile-quantile plot for the spectral distribution function versus the uniform can be used to test for white noise, see Brockwell & Davis (1991), p. 223.

Example. AR(1). Here $\gamma_0 = \frac{\sigma_\epsilon^2}{1-\phi^2}$, and $\gamma_h = \phi^h \gamma_0$, so

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

$$
\begin{aligned}
&= \frac{\gamma_0(1-\phi^2)}{2\pi(1-2\phi\cos(\lambda)+\phi^2)} \\
&= \frac{\sigma_\epsilon^2}{2\pi(1-2\phi\cos(\lambda)+\phi^2)},
\end{aligned}
$$

where we used that $e^{-i\lambda} + e^{i\lambda} = 2\cos(\lambda)$. Plotting the spectral density $f(\lambda)$, we see that in the case $\phi > 0$ the spectral density $f(\lambda)$ is a decreasing function of λ, the power is concentrated at low frequencies, corresponding to graduate long-range fluctuation. For $\phi < 0$ the spectral density $f(\lambda)$ increases, the power is concentrated at high frequencies; such a process tends to oscillate.

The spectral density for an ARMA(p,q)-process is again related the the AR and to the MA polynomials. See Brockwell and Davis (1991), p.123, for the following theorem.

Theorem 3 *Let $\Phi(B)Y_t = \Theta(B)\epsilon_t$ describe an ARMA(p,q)-process, where $\{\epsilon_t\}_t \sim WN(0,\sigma_\epsilon^2)$. Suppose that $\Phi(z)$ and $\Theta(z)$ have no common factors, and Φ has no zeros on the unit circle. Then $\{Y_t\}$ has spectral density*

$$
f(\lambda) = \frac{{\sigma_\epsilon}^2}{2\pi}\frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}.
$$

Example. AR(1). Here $\Phi(z) = 1 - \phi z$, and $\Theta(z) = 1$, so, for $-\pi \leq \lambda < \pi$,

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

as calculated before.

Example. MA(1). Here $\Phi(z) = 1, \Theta(z) = 1 + \theta z$, and we obtain, for $-\pi \leq \lambda < \pi$,

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

Plotting the spectral density $f(\lambda)$, we 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.

Working with the discrete Fourier transform offers advantages, in particular there are fast algorithms available (FFTs). For estimating the spectrum, the periodogram has some nice properties, which we will see later.

Note: Similarly, complex-valued processes can be analyzed, using the definition of covariance that, for X and Y mean zero,

$$Cov(X, Y) = E(X\bar{Y}),$$

where $\bar{Y}$ is the complex conjugate of Y. Then

$$f_c(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_h e^{i\lambda h}$$

and

$$\gamma_h = \int_{-\pi}^{\pi} e^{ih\lambda} f_c(\lambda) d\lambda.$$

Time-invariant linear filters

A different view point on time series is provided by time-invariant linear filters. In order to extract signals from a time series, the distribution of power or variance can be modified by making a linear transformation. We say that the process $\{Z_t\}_{t=0,\pm1,\pm2,\ldots}$ is obtained from the process $\{Y_t\}_{t=0,\pm1,\pm2,\ldots}$ by application of the *linear filter*

$$\mathcal{C} = \{c_{t,k}, t, k = 0, \pm1, \pm2, \ldots\}$$

if

$$Z_t = \sum_{k=-\infty}^{\infty} c_{t,k} Y_k, \quad t = 0, \pm1, \pm2, \ldots.$$

The filter is *time-invariant* if $c_{t,k} = a_{t-k}$ depends only on $t-k$. Then

$$Z_t = \sum_{k=-\infty}^{\infty} a_k Y_{t-k}, \quad t = 0, \pm 1, \pm 2, \ldots.$$

Z_t is also called the *output*, and $\{Y_t\}_t$ is called the *input*. The Fourier transform

$$A(\nu) = \sum_{t=-\infty}^{\infty} a_t e^{-2\pi i \nu t}$$

is called the *frequency response function*.

It is straightforward (see Shumway and Stoffer (2000), p. 228) to calculate that the spectral density of $\{Z_t\}_t$ is related to the spectral density of $\{Y_t\}_t$ by

$$f_Z(\nu) = |A(\nu)|^2 f_Y(\nu).$$

The spectral density multiplier $|A(\nu)|^2$ is called the *transfer function*.

Example. A *first difference filter*

$$Z_t = \nabla Y_t = Y_t - Y_{t-1}.$$

Here $a_0 = 1, a_1 = -1, a_r = 0$ for $r \neq 0, 1$, and

$$\begin{aligned} A(\nu) &= 1 - e^{-2\pi i \nu} \\ |A(\nu)|^2 &= 2(1 - \cos(2\pi\nu)). \end{aligned}$$

Thus this filter is large for higher frequencies, and small for lower frequencies; it will enhance higher frequencies. Such a filter is called a *high-pass filter*.

Example. A *centred moving average filter* is given by

$$Z_t = \sum_{k=-p}^{p} a_k Y_{t-k}.$$

Then

$$A(\nu) = 2\sum_{k=1}^{p} a_k \cos(2\pi\nu k) + 1.$$

The filter will enhance lower frequencies; it is a *low-pass filter*.

Example. A causal ARMA(p,q) process $Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ can be interpreted as obtained from $\{\epsilon_t\}_t$ by application of the time-invariant linear filter $\{\psi_j, j = 0, 1, 2, \ldots\}$.

2.7 State-space models

State-space models assume that the observations $\{Y_t\}_t$ are incomplete and noisy functions of some underlying unobservable process $\{X_t\}_t$, called the *state process*, which is assumed to have a simple Markovian dynamics. The general state space model (see Künsch (2001)) is described by

1. $X_0, X_1, X_2, \ldots$ is a Markov chain
2. Conditionally on $\{X_t\}_t$, the Y_t's are independent, and Y_t depends on X_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.

A prominent role is played by the linear state space model

$$\begin{aligned} X_t &= G_t X_{t-1} + v_t && (18) \\ Y_t &= H_t X_t + w_t, && (19) \end{aligned}$$

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,

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

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 $\{X_t\}_t$ in terms of the observed data points $\{Y_t\}_t$.

Example. Suppose the model

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

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 covariance V_t^2 and W_t^2, respectively. Then

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

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

The following example shows that any ARMA(p,q)-model with Gaussian WN can be formulated as a state space model.

Example. ARMA(p,q) with Gaussian WN (thus independent WN). For a causal invertible ARMA(p,q) process

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=0}^{q} \theta_j \epsilon_{t-j},$$

write

$$Y_{t|s} = E(Y_t | Y_u lu \leq s).$$

Then (see Künsch 2001)

$$Y_{t+j|t} = \sum_{i=1}^{\min(j-1,p)} \phi_i Y_{t+j-i|t} + \sum_{i=\min(j,p)}^{p} \phi_i Y_{t+j-i} + \sum_{i=0}^{q} \theta_i \epsilon_{t+j-i},$$

empty sums being taken as zero; so that for $j \geq k = \max(p, q+1)$, $Y_{t+j|t}$ is linearly dependent with $X_t := (Y_t, Y_{t+1|t}, \ldots, Y_{t+k-1|t})^T$. We thus can choose X_t as state vector. Then (19) holds with $H_t = (1, 0, \ldots, O))$ and $w_t = 0$. Also we have that $Y_{t+i|t+1} = Y_{t+i|t} + b_i \epsilon_{t+1}$ for some coefficients b_i, so that

$$Y_{t+k|t+1} = \sum_{j=1}^{p} \phi_j Y_{t+k-j|t} + b_k \epsilon_{t+1},$$

thus establishing (18). The representation of an ARMA model as a state-space model is not unique, see Brockwell and Davis (1991), pp.469-470.

Note that the above model is more flexible than ARMA(p,q). If, for example, the observation at time t is missing, then we simply put $H_t = (0, 0, \ldots, 0)^T$. However, it is difficult to deal with non-Gaussian linear state space models.

Many examples for state-space models come from engineering, and, more recently, from biology, and from mathematical finance.

Example. If P_t denotes the price of an asset at time t, then at least to a first approximation the log return

$$Y_t = \log\left(\frac{P_t}{P_{t-1}}\right)$$

has conditional mean zero given the past, but its conditional variance - called *volatility* - depends on the past. Stochastic volatility models consider the conditional variance as an exogenous random process. The easiest example is

$$\begin{aligned} X_t &= m + \phi X_{t-1} + v_t \\ Y_t &= \exp(X_t) w_t, \end{aligned}$$

see Timmer & Weigend (1997). For more general stochastic volatility models, see Shephard (1996).

Filtering, smoothing, and forecasting

The primary aims of the analysis of state space models are to produce estimators for the underlying unobserved signal X_t given the data $\mathbf{Y}^s = (Y_1, \ldots, Y_s)$ 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.

For filtering and forecasting, we use the *Kalman filter*. It is a recursive method to calculate a conditional distribution within a multivariate normal framework.

It is useful to first revise some distributional results for multivariate normal distributions. Suppose that

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \tag{20}$$

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

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

and conversely, if $Z_2 \sim \mathcal{MVN}(\mu_2, \Sigma_{22})$ and if (21) holds, then (20) holds.

Assume the model

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

with $(v_t)_t$ ind.$WN(0, V_t)$, and $(w_t)_t$ ind.$WN(0, W_t)$. Here, X_t is a vector. Put $\mathbf{Y}^t = (Y_1, Y_2, \ldots, Y_t)$ and

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

When $t_1 = t_2 = t$, we will write P_t^s for convenience. Suppose $X_0^0 = \mu$ and $P_0^0 = \sigma_0$, and that

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

Then

$$\begin{aligned}
\mathcal{L}(X_t|\mathbf{Y}^{t-1}) &= \mathcal{L}(G_t X_{t-1} + v_t|\mathbf{Y}^{t-1}) \\
&= \mathcal{MVN}(G_t X_{t-1}^{t-1}, R_t),
\end{aligned}$$

where

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

Also,

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

Let Z_1 have the conditional distribution of Y_t given $\mathbf{Y}^{t-1}$, and let Z_2 have the conditional distribution of X_t given $\mathbf{Y}^{t-1}$. Put

$$\begin{aligned}
Z_1 &= Y_t \\
Z_2 &= X_t \\
\mu_2 &= G_t X_{t-1}^{t-1} \\
V_{22} &= R_t \\
\mu_1 + \Sigma_{12}\Sigma_{11}^{-1}(Z_2 - \mu_2) &= H_t X_t \\
\Sigma_{11} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21} &= W_t.
\end{aligned}$$

Then, with (20) and (21), the conditional distribution of $(Y_r, X_t)^T$ given $\mathbf{Y}^{t-1}$ is given by

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

Apply (20) and (21) again, reversing the roles of Z_1 and Z_2, to compute that the conditional distribution of X_t given $\mathbf{Y}^{t-1}$ is multivariate normal with mean X_t^t and variance P_t, where

$$\begin{aligned} X_t^t &= G_t X_{t-1}^{t-1} + R_t H_t^T (W_t + H_t R_t H_t^T)^{-1} (Y_t - H_t G_t X_{t-1}^{t-1}) \\ P_t &= R_t - R_t H_t^T (W_t + H_t R_t H_t^T)^{-1} H_t R_t. \end{aligned} \tag{22}$$

Equations (22) and (22) are known as the *Kalman filter updating equations.* This solves the filtering problem.

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 X_{t-1}^s, P_{t-1}^s. Then

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

Recursion solves the forecasting problem.

Note that the conditional distribution of Y_{t+1} given $\mathbf{Y}^1$ is

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

This fact is the basis of the *prediction error decomposition*, useful for parameter estimation.

For smoothing we use the *Kalman smoother*. We proceed by backwards induction. Suppose that X_t^t, P_t^t are known, where P_t^t is the conditional covariance matrix of $\mathbf{Y}_t$ given $\{X_1, \dots, X_t\}$. With a similar derivation as above, for $t = n, n-1, \dots, 1$,

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

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 often it is not clear which initial distribution is appropriate.

2.8 Continuous time models

In continuous time, stochastic difference equations are replaced by stochastic differential equations, and the white noise process $(\epsilon_t)_t$ by a Brownian process $dW(t)$ with the formal properties that

$$\begin{aligned} E(dW(t)) &= 0 \\ Var(dW(t)) &= \sigma^2 dt \\ cov(dW(t), dW(s)) &= 0, \quad t \neq s. \end{aligned}$$

The simplest and most important example, the analogue in continuous time of AR(1), is the *Ornstein-Uhlenbeck process* (OU-process) with

$$dY(t) \;=\; -\rho Y(t)dt + dW(t),$$

leading to the autocorrelation function $e^{-\rho h}$. If the OU-process is sampled at equally spaced time points ℓ apart, then there results and AR(1) with $\gamma_1 = e^{-\rho\ell}$, whereas if it is observed in discrete time in aggregated form, the result is an ARMA(1,1).

Rigorous theory needs more theoretical tools for integrals with respect to $dW(t)$.

Further reading

1. P. Bloomfield (2000). *Fourier Analysis of Time Series.* Second edition, Wiley.

2. H. R. Künsch (2001). *State space and hidden Markov models.* In *Complex Stochastic Systems*, O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg, eds. Chapman and Hall.

3. N. Shephard (1996). *Statistical aspects of ARCH and stochastic volatility* In *Time Series Models with Econometric, Finance and Other Applications*, D.R. Cox, D.V. Hinkley, and O.E. Barndorff-Nielsen, eds. Chapman and Hall.

4. J. Timmer and A.S. Weigend (1997). Modeling volatility using state space models. *International Journal of Neural Systems* **8**, 385–398.

3 Statistical Analysis

3.1 Time domain

There are two broad complementary approaches to analysis, one essentially nonparametric, and one specific to a family of models. Often one starts with the first to get a broad idea of the structure of the data. That may be sufficient or it may lead to the more formal fitting of a model often out of the ARIMA family.

We here always consider the situation of n observations $y_1, \ldots, y_n$.

3.1.1 Nonparametric approach

First having removed any substantial trend that is present, we assume that the observations come from a stationary process with mean μ and autocovariance function γ_h. We estimate the mean μ by the data mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

and the variance $\sigma^2 = \gamma_0$ by

$$\hat{\gamma}_0 = \frac{1}{n} \sum_{i=1}^{n} (y_t - \bar{y})^2,$$

where the divisor n may need modifying in short series not in general to $n-1$ however.

A direct calculation gives

$$\begin{aligned} Var(\bar{Y}) &= \frac{\sigma^2}{n} + \frac{2}{n^2} \sum_h (n-h)\gamma_h \\ &\approx \frac{\sigma^2}{n} \left(1 + 2\sum_h \rho_h\right) \\ &= \frac{2}{n}\pi f(0) = \frac{\sigma^2}{n} \left(\sum_{j=-\infty}^{\infty} \psi_j \right)^2, \end{aligned}$$

assuming that $\sum_h |\rho_h| < \infty$. Note the very important point that this is often extremely different from $\frac{\sigma^2}{n}$ (which we obtain for WN). The situation is much worse for series with long-range dependence, so that $\sum_h |\rho_h| = \infty$; then convergence of variance is at a slower rate than $\frac{1}{n}$.

Suppose

$$Y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j},$$

where $\{\epsilon_t\}_t$ is WN$(0, \sigma^2)$. If $\sum |\psi_j| < \infty$ and if $\sum \psi_j \neq 0$, then (see Brockwell & Davis (1991), p.219), the scaled sample mean $\frac{\sqrt{n}}{\sigma}\bar{Y}$ is asymptotically normal with mean μ and variance $\left(\sum_{j=\infty}^{\infty} \psi_j\right)^2$.

We now estimate the lag h autocorrelation by

$$r_h = \frac{\sum_{i=1}^{n}(y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{i=1}^{n}(y_t - \bar{y})^2},$$

or, especially if h is appreciable compared with n some adaptation thereof with different divisors. In general it is probably a bad idea to go beyond $h = \frac{n}{10}$. Similarly, we estimate the lag h autocovariance by

$$\hat{\gamma}_h = \frac{1}{n}\sum_{i=1}^{n}(y_t - \bar{y})(y_{t+h} - \bar{y}).$$

If the corresponding population correlation is zero (white noise), then the sample correlation has variance $\frac{1}{n}$. Note that the presence of correlation for some, usually small, h induces correlation between the sample correlations at larger values of h, even if the population correlation is zero.

For a causal ARMA(p,q) process $Y_t = \sum_{j \geq 0} \psi_j \epsilon_{t-j}$ with $\sum |\psi_j| < \infty$, and $(\epsilon_t)_t$ being i.i.d. WN$(0, \sigma_\epsilon^2)$ such that

- either $E\epsilon_t^4 < \infty$
- or $\sum_{j \geq 0} j\psi_j^2 < \infty$

we have that (see Brockwell and Davis (1991), p. 221) for any h the vector $\underline{r} = (r_1, \dots, r_h)$ is asymptotically multivariate normal distributed with mean vector $\underline{\rho} = (\rho_1, \dots, \rho_h)$ and variance-covariance matrix $\frac{1}{n}W$, Here, $W = (w_{ij})_{i,j=1}^{h}$ is given by *Bartlett's formula*

$$\begin{aligned} w_{ij} &= \sum_{k=-\infty}^{\infty} \left\{\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho_k^2 \right. \\ &\quad \left. -2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)\right\}. \end{aligned}$$

For i.i.d. WN, for example, we obtain $w_{ij} = 1$ if $i = j$, and $w_{ij} = 0$ if $i \neq j$.

If n is small, or if the parameter values are close to the boundaries, then the convergence towards the asymptotic distribution can be very poor. In this case, bootstrap methods would be advisable. Here the specific dependence structure has to be taken into account. There are specific bootstrap methods for time series; see Bühlmann (2001) for an overview.

3.2 Fitting ARIMA models

For fitting ARIMA models, the classical Box-Jenkins approach can be used, see Subsection 2.5.

3.2.1 Estimation in state-space models

In state-space models, a maximum-likelihood approach based on the prediction error decomposition can be employed. Here, missing data can be treated as well, see Kohn and Ansley (1986). For non-Gaussian models, inference usually involves computer-intensive methods, such as MCMC and Gibbs sampling, see also Künsch (2001). An example is given by the Lake Huron data analysis in Smith (2001).

3.3 Frequency domain

3.3.1 Nonparametric approach

Interpreting the spectral density as the variance at a given frequency, we might get useful information out of the magnitude of that variance by isolating the frequency component. Hence, try to enhance the periodic component in a time series at frequency ν by correlating the series against periodic sine and cosine functions at frequency ν. The sine and cosine transforms

$$
\begin{aligned}
Y_c(\nu) &= n^{-\frac{1}{2}} \sum_{t=1}^{n} Y_t \cos(2\pi\nu t) \\
Y_s(\nu) &= n^{-\frac{1}{2}} \sum_{t=1}^{n} Y_t \sin(2\pi\nu t)
\end{aligned}
$$

should be large when the series contains the frequency ν, and should be small otherwise. Typically, we will probe at frequencies $\nu_k = \frac{k}{n}, 0 = 1, \ldots, n-1$. The corresponding frequencies $\omega_k = 2\pi \frac{k}{n}$ are called the *Fourier frequencies.*

The *(sample) periodogram*

$$I_n(\nu) = Y_c^2(\nu) + Y_s^2(\nu)$$

measures the sample variance or power at the frequency ν. We will see that the periodogram is an approximately unbiased estimator of the spectral density function. The estimated spectral density can be used to estimate the autocorrelation function, using (17).

A shorter way of describing the periodogram is via the discrete Fourier transform: write

$$\begin{aligned} Y_f(\nu) &= n^{-\frac{1}{2}} \sum_{t=1}^{n} Y_t e^{-2\pi i \nu t} \\ I_n(\nu) &= |Y_f(\nu)|^2. \end{aligned}$$

Note that the time series can be recovered from the Fourier transform via

$$Y_t = n^{-\frac{1}{2}} \sum_{t=1}^{n} Y_t e^{2\pi i \nu t}.$$

For Gaussian WN, because Y_c and Y_s are obtained by orthogonal transformation and are orthogonal to the general mean, they are themselves independently normally distributed with constant variance. In fact, if ν_n^k denotes a frequency such that $2\pi\nu_n^k$ is close to ν for any given n, then $Y_c(\nu_n^k)$ and $Y_s(\nu_n^k)$ are asymptotically i.i.d. normal with mean zero and variance $\frac{1}{2}f(\nu)$, and

$$\frac{2I_n(\nu)}{f(\nu)}$$

is asymptotically Chisquare distributed with two degrees of freedom. This result can be extended to the simultaneous estimation (Brockwell and Davis (1991), p.346; for a heuristic proof see Smith (2000), pp. 41-42).

Theorem 4 *Let $w_j = \frac{2\pi j}{n}, 1 \le j < \frac{n}{2}$ be the Fourier frequencies. Suppose that $(Y_t)_t$ is a causal ARMA(p,q) process with $\sum_{j=0}^{\infty} |\phi_j| < \infty$, and assume that the i.i.d. WN $(\epsilon_t)_t$ satisfies that $E\epsilon_t^4 < \infty$. Then the periodogram ordinates $\{I_n(w_j), 1 \le j < \frac{n}{2}\}$ are approximately independent and exponentially distributed, with means $\{f(w_j), 1 \le j < \frac{n}{2}\}$.*

Remarks.

- Theorem 4 is central for spectral estimation theory. No similar result for the sample acf is available.
- From Theorem 4, we have that for a Fourier frequency w, $I_n(w)$ is approximately exponentially distributed with mean $f(w)$, so $VarI_n(w) \approx f(w)^2$ does not converge to zero as $n \to \infty$, Thus $I_n(w)$ is NOT a consistent estimator of $f(w)$.

Note that, regardless of the size of n, the approximating Chisquare distribution will always have only two degrees of freedom. The degrees of freedom can be increased by taking a moving average, for example. In addition, the independence of the periodogram ordinates at different Fourier frequencies results in the sample periodogram, as a function of w, being extremely irregular. Also for this reason often smoothing is applied, for instance using a moving average, or more generally a smoothing kernel. Here the choice of bandwidth is nontrivial. Note that we need to adjust the spectral estimates for the effects of filtering.

Example. Suppose we use a centred moving average smoothed spectral density estimator

$$\hat{f}(\nu) = \frac{1}{L} \sum_{\ell=-\frac{L-1}{2}}^{\ell=\frac{L-1}{2}} I_n\left(\nu_k^n + \frac{\ell}{n}\right),$$

where $\nu_k = \frac{k}{n}$ is a Fourier frequency that is close to the frequency of interest ν. Then, for large n, (Shumway & Stoffer (2000), p.243)

$$\frac{2L\hat{f}(\nu_k)}{f(\nu)} \approx \chi^2_{2L},$$

that is, $\frac{2L\hat{f}(\nu_k)}{f(\nu)}$ follows approximately a chi-squared distribution with $2L$ degrees of freedom. Thus we obtain a 100(1-α)% confidence interval

$$\frac{2L\hat{f}(\nu_k)}{\chi^2_{2L}(\alpha/2)} \leq f(\nu) \leq \frac{2L\hat{f}(\nu_k)}{\chi^2_{2L}(1-\alpha/2)}.$$

Another problem with the periodogram is that of *leakage*: if I_n is evaluated at non-Fourier frequencies w there may be "side-lobes" to either side of the main peak, which may significantly distort the spectrum. In this case *tapering* is used, that is, we transform the series by putting $\tilde{y}_t = y_t h_t$, for a function h_t. A 100p% *cosine taper*, for example, is given by

$$h_t = \begin{cases} \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(t-1/2)}{m}\right)\right\}, & 1 \leq t \leq m \\ 1, & m+1 \leq t \leq n-m \\ \frac{1}{2}\left\{1 - \cos\left(\frac{\pi(n-t+1/2)}{m}\right)\right\}, & n-m \leq t \leq n, \end{cases}$$

where usually $p = \frac{2m}{n}$. S-PLUS departs from this convention by using $p = \frac{m}{n}$ instead. For more detail see, for example, Shumway & Stoffer (2000), pp.247-251.

3.4 Parametric approach

It is also possible to fit parametric models to the spectral density function. A convenient class of models is given by the *Bloomfield class*,

$$f(w) = \exp\left\{2\sum_{k=0}^{n} \theta_k \cos(kw)\right\}.$$

Any parametric spectral density function $f(\nu, \Theta)$ can be estimated using the approximate *Whittle log likelihood*

$$\log L \approx -\sum_j \log f(w_j, \Theta) - \sum_j \frac{I_n(w_j)}{f(w_j, \Theta)}.$$

Thus maximum likelihood estimation can be applied.

Note that spectral analysis assumes stationarity, so no information about trend is available. It is basically a nonparametric approach, which is good as long as n is large. Finally, note that it is not always clear whether interpretation of the underlying process in terms of cyclic patterns of variation makes sense.

4 Some more advanced topics

4.1 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.

4.2 Nonlinear models

Virtually all the above models are essentially linear. They can be extended in an important although relatively minor way by point-wise transformation of the variable, for example by taking logarithms. Formal tests for nonlinearity can be based on lagged third moments

$$C_3(r,s) = E(Y_t - \mu)(Y_{t+r} - \mu)(Y_{t+s} - \mu)$$

in the time domain, and by higher-order spectra, in particular the bispectrum

$$f_3(\nu_1, \nu_2) = \frac{1}{4\pi^2} \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} C_3(r,s) e^{-ir\nu_1 - is\nu_2},$$

in the frequency domain. For linear Gaussian models the bispectrum and the lagged third moments would vanish.

Explicitly nonlinear models can show quite different kinds of behaviour from that typical of linear models but tend to be difficult to study other than numerically. Some special cases are

- models of variance heterogeneity, ARCH and its generalizations, of much interest in finance and econometrics more generally,
- nonlinear autoregressions of the form

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

  for suitable nonlinear functions g. Note that polynomials cannot be used except for rather limited purposes because of nonstationarity,
- threshold autoregressions in which $g(Y_t)$ is piecewise linear.

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. 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. For more details see, e.g., Diks (1999).

4.3 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. Takens (1981), for example, gives a reconstruction of the asymptotic dynamics of a dynamical system from an observed deterministic time series. This is a completely different approach to time series.

Acknowledgement. Much of these notes are based on lecture notes by D.R. Cox. Helpful discussions with him and with N. Stollenwerk are gratefully acknowledged.

Further reading

1. P. B"uhlmann (2001). Bootstrap for time series. To appear in *Statistical Science.* See also `http://stat.ethz.ch/ buhlmann/bibliog.html`

2. G.E.P. Box, and G.M. Jenkins (1970). *Time Series Analysis, Forecasting, and Control.* Holden-Day.

3. C. Diks (1999). *Nonlinear Time Series Analysis.* World Scientific.

4. R. Kohn and C.F. Ansley (1986). Estimation, interpolation and prediction for ARIMA models with missing data. *J. Amer. Statist. Assoc.* **81**, 751–761.

5. H. R. Künsch (2001). *State space and hidden Markov models.* In *Complex Stochastic Systems*, O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg, eds. Chapman and Hall.

6. F. Takens (1981). Detecting strange attractors in turbulence. in *Dynamical Systems and Turbulence, Warwick 1980 (Lecture Notes in Mathematics)*, (eds D.A. Rand and L.-S. Young), vol. 898, 366–381.