Time series analysis
Notes for the course in fall 2010

Hansruedi Künsch

Version August 2011

This script gives a brief introduction into theory and methods of time series analysis. For examples and illustrations of the concepts and methods, you should look at the *R*-demonstrations which are also on my web page.

Contents

1 Introduction

A time series is a sequence of observations of a system/phenomenon which varies irregularly in time. We interprete the irregular behavior as the effect of randomness. Observations at a fixed time point can be univariate or multivariate. The time points where observations are made can be equispaced or irregular (equispaced observations with missing values gives an intermediate situation).

The basic plot shows the observations versus time. From this plot one can usually obtain a first descriptive characterization of a time series. Things to look for are for instance whether the level is constant or changes slowly and thus exhibits a trend. In the case of a trend, one then looks if the fluctuations change with the level or remain constant. Another feature to look for is whether periodic behavior is present or absent. Periodic behavior can either have a simple explanation as a seasonal effect with known period (day, week, year), or it can be due to the internal dynamics of the system. In that case the period is typically not known a priori and both period and cyclic pattern change over time. Finally, one will look whether the behavior of the series in different time windows is essentially the same or changing with time, either by slow or by sudden changes.

Goals of time series analysis can be classified in one of the following

- Description
- Modeling
- Prediction
- Signal extraction.

1.1 Stochastic processes

A stochastic process is a mathematical model for a time series.

Stochastic process = Collection of random variables $(X_t(\omega); t \in T)$. Alternative view: Stochastic process as a random function from T to $\mathbb{R}$.

A basic distinction is between continuous and discrete equispaced time T. Models in continuous time are prefered for irregular observation points. In this course we will restrict ourselves mostly to discrete equispaced time.

In all interesting cases, there is dependence between the random variables at different times. Hence need to consider joint distributions, not only marginals. Gaussian stochastic processes have joint Gaussian distribution for any number of time points.

A stochastic process describes how different time series (when different ω's are drawn) could look like. In most cases, we observe only one realization $x_t(\omega)$ of the stochastic process (a single ω). Hence it is clear that we need additional assumptions, if we want to draw conclusions about the joint distributions (which

involves many ω's) from a single realization. The most common such assumption is stationarity.

Stationarity means the same behavior of the observed time series in different time windows. Mathematically, it is formulated as invariance of (joint) distributions when time is shifted. Stationarity justifies taking of averages (mathematically, one needs ergodicity in addition).

Simple examples of stochastic processes in discrete time:

- White noise: $X_t(.)$ i.i.d.. This model is stationary. It is not of interest by itself, but useful as building block for more complicated models. The reason for the name will become clear later.

- Harmonic oscillations plus white noise

$$X_t(\omega) = \sum_{k=1}^{K} \alpha_k \cos(\lambda_k t + \phi_k) + \varepsilon_t(\omega)$$

  where the $\varepsilon_t(.)$ are i.i.d. and the α_k, λ_k and ϕ_k are (unknown) parameters. The "signal" is periodic iff $\lambda_k/\lambda_j \in \mathbb{Q}$ for all j, k. This model is not stationary.

- Moving averages:

$$X_t(\omega) = F(\varepsilon_t(\omega), \varepsilon_{t-1}(\omega), \ldots, \varepsilon_{t-k}(\omega))$$

  where the $\varepsilon_t(.)$ are i.i.d. and F is a fixed function. This model is stationary. The special case where F is linear is often used.

- Autoregressive models: These models are defined recursively:

$$X_t(\omega) = F(X_{t-1}(\omega)) + \epsilon_t(\omega)$$

  where the $\varepsilon_t(.)$ are i.i.d. and F is a fixed function. In order to define a model, we also need an initial condition $X_{t_0}(\omega)$ which is usually assumed to be independent of all $\varepsilon_t(.)$ for $t > t_0$. This model is usually not stationary, but depending on f it can converge to a stationary model as $t \to \infty$ or $t_0 \to -\infty$. Again, the linear case $F(x) = \beta_0 + \beta_1 x$ is often used. In this case, the model is asymptotically stationary iff $|\beta_1| < 1$.

- Autoregressive conditionally heteroscedastic models

$$X_t(\omega) = F(X_{t-1}(\omega))\varepsilon_t(\omega)$$

  where the $\varepsilon_t(.)$ are i.i.d. and F is a fixed function. The same comments as in the autoregressive model apply.

The last two examples are special cases of a Markov process.

Deterministic chaotic models also show a behavior which is difficult to predict. They have been considered as an alternative model class.

Among the processes in continuous time, we only mention Gaussian processes. These are processes where all finite dimensional distributions are Gaussian. These processes are determined by the mean function $\mu(t)$ and the covariance function $C(t,s)$. Whereas the mean function is arbitrary, not every function of two arguments is a valid covariance function because the matrix $(C(t_i,t_j); 1 \leq i,j \leq n)$ must be positive definite for all n and all times $t_1,\ldots,t_n$. We will come back to this in Section 2.1.

1.2 Treatment of trend and seasonal component

Series with trends and/or seasonal component have a mean value which changes with time and thus they are not stationary. Hence one would like to be able to eliminate them so that the remaining irregular part can be modeled hopefully by a stationary process.

A seasonal component is periodic with a period which is is fixed and known in advance. We denote the number of observations in a full period by M. It is either strictly periodic or it changes very slowly in time. Time series may contain other periodic features whose period and shape are more variable and not due to a known periodic external influence like the motion of the sun or the organisation of our society in working days and weekends. In that case, the time series often can still be considered to be stationary and we should not speak about seasonality.

Transformation of variables: Transform $X_t \to Y_t = h(X_t)$ s. th. there is an additive decomposition

$$Y_t = T_t + S_t + I_t$$

into trend, seasonal component and irregular component. This means in particular that the amplitude of the seasonal cycle remains constant in time. Ideally, I_t is (approximately) stationary. The most common transformations h are from the Box-Cox family

$$h(x) = \frac{x^\lambda - 1}{\lambda} \ (\lambda \neq 0), \quad = \log(x) \ (\lambda = 0).$$

Parametric models for trend and seasonal:

$$Y_t = \sum_{k=0}^{K} \alpha_k t^k + \sum_{j=1}^{J} (\beta_j \cos(2\pi \frac{jt}{M}) + \gamma_j \sin(2\pi \frac{jt}{M})) + I_t.$$

where M is the number of time points in one cycle and $J \leq M/2$. Estimate the parameters by least squares. Usually this is not flexible enough (e.g. the seasonal component is assumed to be constant over time).

Nonparametric decompositions: We can estimate the trend by the moving averages:

$$\widehat{T_t} = \frac{1}{M}(\frac{1}{2}Y_{t-M/2} + Y_{t-M/2+1} + \cdots + \frac{1}{2}Y_{t+M/2}).$$

The seasonal component can then be estimated from $R_t = X_t - \widehat{T}_t$:

$$\widehat{S}_t = \frac{1}{2k+1}(R_{t-kM} + R_{t-(k-1)M} + \cdots + R_{t+kM})$$

The function stl in R provides a more sophisticated version of this basic idea.

Differencing: Since T_t changes slowly, $T_t \approx T_{t-1}$. Hence the differenced series $\Delta Y_t = Y_t - Y_{t-1}$ is approximately trend free. Similarly, $S_{t-M} \approx S_t$, i.e. seasonal differences $\Delta_M Y_t = Y_t - Y_{t-M}$ is approximately free of the seasonal component. In order to eliminate both, we can consider

$$\Delta_M(\Delta Y)_t = \Delta(\Delta Y_M)Y_t = Y_t - Y_{t-1} - (Y_{t-M} - Y_{t-M-1}).$$

2 The autocovariance function

If (X_t) is a stationary process whose first and second moments exist, then $E(X_t)$ is independent of t, $E(X_t) = \mu$ and $Cov(X_t, X_s)$ depends only on $t - s$:

$$\mathrm{Cov}(X_t, X_s) = C(t - s).$$

C is called the autocovariance function of the process. $C(0)$ is the variance of X_t and thus

$$\rho(u) = \frac{C(u)}{C(0)}$$

is the autocorrelation function. Because

$$C(-h) = \mathrm{Cov}(X_{t-h}, X_t) = \mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(X_{t+h}, X_t) = C(h),$$

the autocovariance and the autocorrelation are symmetric.

Example: White noise If (X_t) is i.i.d., $C(h) = 0$ for all $h \neq 0$.

Example: Moving averages. If

$$X_t = \sum_{k=0}^{K} \alpha_k \varepsilon_{t-k}$$

with (ε_t) i.i.d., then

$$E(X_t) = E(\varepsilon_t) \sum_{k=1}^{K} \alpha_k$$

and

$$C(h) = \mathrm{Var}(\varepsilon_t) \sum_{k=h}^{K} \alpha_k \alpha_{k-h} \ (0 \leq h \leq K), \quad C(h) = 0 \ (h > K).$$

This result can be extended to two-sided infinite moving averages

$$X_t = \sum_{k=-\infty}^{\infty} \alpha_k \varepsilon_{t-k}$$

with $\sum_k |\alpha_k| < \infty$. It still holds

$$E(X_t) = E(\varepsilon_t) \sum_{k=-\infty}^{\infty} \alpha_k, \quad C(h) = \mathrm{Var}(\varepsilon_t) \sum_{k=-\infty}^{\infty} \alpha_k \alpha_{k-h}.$$

(One has to address the issues of convergence and the exchange of expectation and limits, this can be done).

A process with $E(X_t) = \mu$ and $\mathrm{Cov}(X_t, X_s) = C(t - s)$ for all t, s is called weakly stationary. For Gaussian processes, stationarity and weak stationarity are equivalent, but for other processes the two concepts differ.

Example: Autoregression Iterating the equation $X_t = \phi X_{t-1} + \varepsilon_t$ gives

$$X_t = \phi^t X_0 + \sum_{j=0}^{t-1} \phi^j \varepsilon_{t-j}.$$

Taking expectations gives

$$E(X_t) = \phi^t E(X_0) + \sum_{j=0}^{t-1} \phi^j E(\varepsilon_t).$$

Hence, if $|\phi| < 1$, $E(X_t) \to E(\varepsilon_t)/(1-\phi)$. If moreover $X_0, \varepsilon_1, \varepsilon_2, \ldots$ are all independent, we obtain for $h \geq 0$ and $t \to \infty$

$$\mathrm{Cov}(X_{t+h}, X_t) \to \mathrm{Var}\,(\varepsilon_1) \sum_{j=0}^{\infty} \phi^j \phi^{h+j} = \frac{\mathrm{Var}\,(\varepsilon_1)}{1-\phi^2} \phi^h.$$

In particular, we have shown that the process is asymptotically weakly stationary if $|\phi| < 1$.

ARCH processes The strategy to iterate the recursion does not generalize beyond linear autoregressions. Because of this, there are in general no simple formulae for nonlinear autoregressions. But for ARCH processes $X_t = F(X_{t-1})\varepsilon_t$ with $E(\varepsilon_t) = 0$, we can show that the autocorrelations are zero. Note that by definition, X_t depends only on X_0 and $\varepsilon_1, \ldots, \varepsilon_t$. Because we assume that the random variables ε_t are i.i.d. and independent of X_0, it follows that ε_t is independent of X_s for $s < t$. Therefore

$$E(X_t) = E(F(X_{t-1})\varepsilon_t) = E(F(X_{t-1}))E(\varepsilon_t) = 0$$

and for $h > 0$

$$\mathrm{Cov}(X_{t+h}, X_t) = E(X_{t+h}X_t) = E(\varepsilon_{t+h}F(X_{t+h-1})X_t) = E(\varepsilon_{t+h})E(F(X_{t+h-1})X_t) = 0.$$

Although X_{t+h} and X_t are uncorrelated, they are dependent: One can verify for instance that $|X_{t+h}|$ and $|X_t|$ are correlated, and so are X_{t+h}^2 and X_t^2.

The standard estimators of mean, autocovariance and autocorrelation are

$$\widehat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t$$

$$\widehat{C}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n)$$

and

$$\widehat{\rho}(h) = \frac{\widehat{C}(h)}{\widehat{C}(0)}.$$

The reason why the denominator in $\widehat{C}(h)$ is n and not $n - |h|$ will be discussed below. If the time series plot does not show clear evidence against stationarity, one usually looks next at a plot of $\widehat{\rho}(h)$ against h in order to obtain information about linear dependence in the series.

2.1 Linear prediction and partial autocorrelations

The best linear prediction of X_t based on $(X_r, X_{r+1}, \dots X_s)$ for $r \le s < t$ or $t < r \le s$is the linear combination

$$\widehat{X}_{t|r:s} = \alpha + \sum_{k=0}^{s-r} \beta_k X_{s-k}$$

which minimzes the mean square error of prediction:

$$E((X_t - \widehat{X}_{t|r:s})^2).$$

$\widehat{X}_{t|r:s}$ is determined by a system of linear equations which involve the mean and autocovariance of X_t only:

$$\begin{aligned} E(X_t - \widehat{X}_{t|r:s}) &= 0 \\ E((X_t - \widehat{X}_{t|r:s})X_u) &= 0 \quad (r \le u \le s) \end{aligned}$$

Example: Let $r = s = t - 1$. One easily verifies that

$$\widehat{X}_{t|t-1} = \mu + \rho(1)(X_{t-1} - \mu), \quad E((X_t - \widehat{X}_{t|t-1})^2) = C(0)(1 - \rho(1)^2).$$

The Durbin-Levinson algorithm allows to compute the coefficients of the linear predictions

$$\widehat{X}_{p|0:p-1} = \alpha^{(p)} + \sum_{k=1}^{p} \beta_k^{(p)} X_{p-k}$$

and the mean square errors $\sigma_p^2 = E((X_p - \widehat{X}_{p|0:p-1})^2)$ recursively. We start with $\sigma_0^2 = C(0)$ and $\alpha_0 = \mu$. Then we have

$$\begin{aligned} \beta_k^{(p)} &= \beta_k^{(p-1)} + \tau(p)\beta_{p-k}^{(p-1)} \quad (1 \le k < p), \\ \beta_p^{(p)} &= \tau(p), \\ \alpha^{(p)} &= \mu(1 - \sum_{k=1}^{p} \beta_k^{(p)}), \\ \sigma_p^2 &= \sigma_{p-1}^2(1 - \tau(p)^2) \end{aligned}$$

where

$$\tau(p) = \frac{C(p) - \sum_{k=1}^{p-1} \beta_k^{(p-1)} C(p-k)}{\sigma_{p-1}^2}$$

where $\tau(p)$ is the so-called partial autocorrelation of lag p. According to the above formula, τ_p is the coefficient of X_0 in $\widehat{X}_{p|0:p-1}$, and $1 - \tau_p^2$ gives the reduction in mean square error if one more observation from the past becomes available. For a derivation of these formulae, see for instance Brockwell and Davies, Chapter 5.2.

Example: For an autoregression we have $C(h) = C(0)\phi^{|h|}$ and therefore

$$\tau(2) = \frac{C(2) - \phi \cdot C(1)}{\sigma_1^2} = 0.$$

This means that if we know X_{t-1}, then there is no additional information in X_{t-2} that can be used for predicting X_t. This holds because $X_t = \phi X_{t-1} + \varepsilon_t$ and ε_t is independent of all past values.

If we allow arbitrary (non-linear) functions of $(X_r, X_{r+1}, \dots X_s)$, we obtain the conditional mean as the best prediction. To compute it, we need the joint distribution of $(X_r, X_{r+1}, \dots X_s, X_t)$, not only the first and second moment. For Gaussian processes, the best linear and the best prediction coincide.

2.2 Regression with correlated errors

The autocovariance function is needed for the variance of the arithmetic mean, or more generally for the variance of least squares estimators with time series errors.

Variance of the arithmetic mean:

Lemma 1. *Let (X_t) be stationary with autocovariance C. Then it holds*

a)

$$\operatorname{Var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) = \frac{1}{n}\sum_{k=-n+1}^{n-1}(1 - \frac{|k|}{n})C(k).$$

b) If $\sum_{k=1}^{\infty} |C(k)| < \infty$, then as $n \to \infty$

$$n \operatorname{Var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) \to \sigma_\infty^2 = \sum_{k=-\infty}^{\infty} C(k) = \operatorname{Var}(X_t)\,(1 + 2\sum_{k=1}^{\infty}\rho(k)).$$

Proof. The first claim follows from

$$\begin{aligned}\operatorname{Var}\left(\sum_{t=1}^{n} X_t\right) &= \sum_{t=1}^{n}\sum_{s=1}^{n} C(t-s) \\ &= \sum_{k=-n+1}^{n-1} C(k) \cdot \underbrace{(\text{Number of pairs with } t-s=k)}_{(=n-|k|)}\end{aligned}$$

For the second claim, we write

$$\sum_{k=-n+1}^{n-1}(1 - \frac{|k|}{N})C(k) = \sum_{k=-\infty}^{\infty} \underbrace{\max(0, 1 - \frac{|k|}{n})C(k)}_{\to C(k) \text{ for } n\to\infty}$$

The claim thus follows by dominated convergence (Lebesgue's Theorem). □

Hence dependence typically inflates the standard deviation of the arithmetic mean by a constant. In order to estimate this standard deviation, we have to estimate the sum of the autocovariances. It will become clear later how this can be done.

In models of long range dependence, the autocorrelations decay like a constant times $|h|^{2d-1}$ with $0 < d < \frac{1}{2}$, and thus the condition $\sum_{k=1}^{\infty} |C(k)| < \infty$ does not hold. One can show that in such a situation, the variance of the arithmetic mean decays like n^{2d-1} times another constant, that is more slowly than under independence.

Regression model:

$$Y_t = \sum_{k=1}^{p} \beta_k x_{t,k} + \varepsilon_t \quad (t = 1, \ldots, n)$$

where ε_t is stationary with autocovariance C. The covariance of the ordinary least squares estimator is in matrix form

$$(X^T X)^{-1} (X^T \Sigma_n X)(X^T X)^{-1}$$

where Σ_n is the $n \times n$ matrix with elements $C(i-j)$. Like in the case of the mean, the correlation of the errors changes the covariance of the estimated coefficients – often substantially. In order to detect correlations of the errors, the Durbin-Watson test can be used. It is based on the test statistic

$$T = \frac{\sum_{i=1}^{n-1} (r_{i+1} - r_i)^2}{\sum_{i=1}^{n} r_i^2} \approx 2 \left(1 - \frac{\sum_{i=1}^{n-1} r_i r_{i+1}}{\sum_{i=1}^{n} r_i^2} \right).$$

In order to take the correlations of the errors into account for the estimated standard errors, the standard procedure is to assume a parametric model (e.g. an ARMA-model, to be introduced below) for the autocovariance function of the errors ε_t and to estimate these parameters from the residuals.

The generalized least squares estimator is the best linear unbiased estimator of the coefficients β_k. It is the maximum likeleihood estimator of β and it is given by

$$\widehat{\beta} = (X^T \Sigma_n^{-1} X)^{-1} X^T \Sigma_n^{-1} y,$$

and has covariance $(X^T \Sigma_n^{-1} X)^{-1}$. Again, one can use a parametric model for Σ_n and then estimate β and the parameters in Σ_n by (joint) maximum likelihood.

2.3 Properties of estimated ACF

The variance of the estimated autocovariances depends on the fourth moments of the process and are very complicated. It turns out that in the case of a moving average, the asymptotic variance of the estimated autocorrelations depends only on the autocorrelations

Theorem 1. *If* $X_t = \sum_k a_k \varepsilon_{t-k}$ *where the* ε_t *are i.i.d. with* $E(\varepsilon_t^2) < \infty$, $\sum_k |a_k| < \infty$ *and* $\sum_k a_k^2 |k| < \infty$, *then for* $n \to \infty$ *the standardized sequence* $\sqrt{n}(\widehat{\rho}(h) - \rho(h))$ *is asymptotically normal with mean zero and covariances*

$$\sum_{j=1}^{\infty} (\rho(j+h) + \rho(j-h) - 2\rho(j)\rho(h))(\rho(j+k) + \rho(j-k) - 2\rho(j)\rho(k)).$$

We omit the proof (see Brockwell and Davies, Chapter 7.3)

In general, the asymptotic variance of $\widehat{\rho}(h)$ is thus complicated and depends on the unknown autocorrelations for all lags. Some simplification occurs in special cases: If X_t is white noise, then the estimated autocorrelations are asymptotically independent and $\mathcal{N}(0, 1/n)$-distributed. For a moving average process $X_t = \sum_{k=0}^{K} \alpha_k \varepsilon_{t-k}$, the asymptotic variance of $\widehat{\rho}(h)$ is for $|h| > K$ equal to $(1 + 2\rho(1)^2 + \ldots + 2\rho(K)^2)/n$. Because the estimates are dependent for different lags, the interpretation of the sample autocorrelation function is not straightforward.

2.4 Herglotz's theorem: The spectrum

The autocovariance function of any stationary process has the property that the matrix Σ_n with elements $C(i-j)$ for $1 \leq i, j \leq n$ is positive definite for any n:

$$\sum_{t=1}^{n} \sum_{s=1}^{n} C(t-s) a_t a_s = \mathrm{Var}\left(\sum_{t=1}^{n} a_t X_t \right) \geq 0$$

for any $a_1, \ldots, a_n$. Such a function is called positive definite. The converse is also true: Any positive definite function is the autocovariance of some stationary process.

The reason for the denominator n in the definition of the estimated autocovariance $\widehat{C}(h)$ is that with this choice $\widehat{C}$ is guaranteed to be positive definite: If we set $X_u = \bar{X}$ for $u > n$, then for $1 \leq t, s \leq n$

$$\widehat{C}(t-s) = \frac{1}{n} \sum_{u=0}^{n-1} (X_{u+t} - \bar{X})(X_{u+s} - \bar{X}).$$

This implies (by exchanging the order of summation)

$$\sum_{t=1}^{n} \sum_{s=1}^{n} \widehat{C}(t-s) a_t a_s = \frac{1}{n} \sum_{u=0}^{n-1} \left(\sum_{t=1}^{n} a_t (X_{u+t} - \bar{X}) \right)^2 \geq 0.$$

With the denominator $n - |h|$, this is not true in general: Consider for instance the case of three observations $X_1 = 1$, $X_2 = 0$, $X_3 = -1$.

It is in general difficult to decide whether a function is positive definite. The Theorem of Herglotz gives a characterization in terms of the Fourier transform:

Theorem 2. *A function $C : \mathbb{Z} \to \mathbb{R}$ with $C(h) = C(-h)$ is positive definite iff there exists a finite, symmetric measure S on $[-\frac{1}{2}, \frac{1}{2}]$ such that*

$$C(h) = \int_{-1/2}^{1/2} \exp(2\pi i h\lambda) S(d\lambda) = \int_{-1/2}^{1/2} \cos(2\pi h\lambda) S(d\lambda).$$

If $\sum |C(k)| < \infty$, then S has the density

$$s(\lambda) = \sum_{h=-\infty}^{\infty} C(h) \exp(-2\pi i h\lambda) = \sum_{h=-\infty}^{\infty} C(h) \cos(2\pi h\lambda),$$

that is

$$C(h) = \int_{-1/2}^{1/2} \exp(2\pi i h\lambda) s(\lambda) d\lambda = \int_{-1/2}^{1/2} \cos(2\pi h\lambda) s(\lambda) d\lambda.$$

S is called the spectral measure, and s the spectral density. Note that $s(0)$ is the asymptotic variance of the mean (for summable covariances).

Proof. The "if" part is easy:

$$\sum_{t=1}^{n}\sum_{s=1}^{n} C(t-s)a_t a_s = \int \sum_{t=1}^{n}\sum_{s=1}^{n} \exp(2\pi i(t-s)\lambda) a_t a_s S(d\lambda) = \int \left|\sum_{t=1}^{n} \exp(2\pi i t\lambda) a_t\right|^2 S(d\lambda) \geq 0.$$

For the "only if" part, we start with

$$0 \leq \sum_{t=1}^{n}\sum_{s=1}^{n} C(t-s) \exp(-2\pi i(t-s)\lambda) = \sum_{h=-n+1}^{n-1} (n - |h|) C(h) \exp(-2\pi i h\lambda)$$

If $\sum |C(h)| < \infty$, we can divide by n and let n go to infinity to obtain

$$s(\lambda) = \sum_{h=-\infty}^{\infty} C(h) \exp(-2\pi i h\lambda) \geq 0.$$

Finally, multiplying both sides by $\exp(2\pi i k\lambda)$ and integrating over λ gives

$$C(k) = \int_{-1/2}^{1/2} \exp(2\pi i k\lambda) s(\lambda) d\lambda.$$

For the general case, we have to look at the spectral distribution function:

$$S_n(\lambda) = \int_{-1/2}^{\lambda} \sum_{h=-n+1}^{n-1} (1 - |h|/n) C(h) \exp(-2\pi i h\nu) d\nu$$

is monotonically increasing and $S_n(\frac{1}{2}) = C(0)$. By compactness of the space of distributions on $[-1/2, 1/2]$ (Prohorov's Theorem), one then obtains convergence of S_n. □

Examples:

- White noise: We find easily

$$s(\lambda) \equiv \mathrm{Var}(X_t).$$

  Since white light has a flat spectrum, this explains the name "white noise".

- Moving average: We have seen before that $C(h) = \sum_k \alpha_k \alpha_{k-h}$. Hence we have

$$C(h)\exp(-2\pi i h\lambda) = \sum_k \alpha_k \exp(-2\pi i k\lambda)\alpha_{k-h}\exp(2\pi i(k-h)\lambda).$$

  By a change of summation we obtain therefore

$$s(\lambda) = \sum_{h=-\infty}^{\infty} C(h)\exp(-2\pi i h\lambda) = \left|\sum_k \alpha_k \exp(-2\pi i k\lambda)\right|^2$$

- Autoregressive process. Because $C(h) = \phi^{|h|}C(0)$, we obtain

$$\begin{aligned} s(\lambda) &= C(0)\left(\sum_{h=0}^{\infty}(\phi^h\exp(-2\pi i h\lambda) + \phi^h\exp(2\pi i h\lambda)) - 1\right) \\ &= C(0)\left(\frac{1}{1-\phi\exp(-2\pi i\lambda)} + \frac{1}{1-\phi\exp(2\pi i\lambda)} - 1\right) \\ &= \frac{C(0)(1-\phi^2)}{|1-\phi\exp(2\pi i\lambda)|^2}. \end{aligned}$$

- Random harmonic oscillations. Consider the process

$$X_t(\omega) = \sum_{j=1}^{J}(A_j(\omega)\cos(2\pi\lambda_j t) + B_j(\omega)\sin(2\pi\lambda_j t))$$

  where the A_j and B_j are independent with means zero and variances σ_j^2 and the λ_j are deterministic. This process has mean zero and

$$\begin{aligned} \mathrm{Cov}(X_{t+h}, X_t) &= \sum_j \sigma_j^2(\cos(2\pi\lambda_j(t+h))\cos(2\pi\lambda_j t) + \sin(2\pi\lambda_j(t+h))\sin(2\pi\lambda_j t)) \\ &= \sum_j \sigma_j^2\cos(2\pi\lambda_j h) \end{aligned}$$

  by the well known formula for $\cos(x-y)$. Hence it is weakly stationary, and the spectrum is the sum of point masses at $\pm\lambda_j$ with weights $\sigma_j^2/2$.

In the last example, the process was a superposition of harmonics, and the spectrum encodes the information about frequencies present and variances of the amplitudes. The example is however not very useful for applications because

for each realization of $X_t(\omega)$ the amplitudes A_j and B_j are fixed and thus we cannot obtain information about the average size σ_j of the amplitude. Later, we will see that all stationary processes can be decomposed in (infinitely many) harmonics whose amplitudes are determined by the spectrum. Moreover, we typically can recover the spectrum from a single long realization.

The argument above for the moving average can be generalized.

Lemma 2. *If (X_t) is a weakly stationary process with autocovariance C_X and spectrum S_X and if (α_k) are coefficients with $\sum_k |\alpha_k| < \infty$, then the process*

$$Y_t = \sum_k \alpha_k X_{t-k}$$

is also weakly stationary with

$$C_Y(h) = \sum_j C_X(j) \sum_k \alpha_k \alpha_{j+k-h}$$

and

$$S_Y(d\lambda) = \left| \sum_k \alpha_k \exp(-2\pi i k \lambda) \right|^2 S_X(d\lambda).$$

Proof.

$$C_Y(h) = \mathrm{Cov}(\sum_k \alpha_k X_{t+h-k}, \sum_\ell \alpha_\ell X_{t-\ell}) = \sum_{k,\ell} \alpha_k \alpha_\ell C_X(h+\ell-k) = \sum_{k,j} \alpha_k \alpha_{k+j-h} C_X(j).$$

Because $|C_X(j)| < C_X(0)$ for all j, the double sum on the right hand side converges. This also shows that

$$\mathrm{Var}(\sum_{|k|>n} \alpha_k X_{t-k}) \to 0 \quad (n \to \infty),$$

that is $\sum_k \alpha_k X_{t+h-k}$ converges in L_2 and thus we can indeed exchange covariance and sum above.

Finally, by the spectral representation of $C_X(j)$

$$C_Y(h) = \int \sum_{k,\ell} \alpha_k \alpha_\ell \exp(2\pi i (h+\ell-k)\lambda) S_X(d\lambda) = \int \exp(2\pi i h \lambda) \left| \sum_k \alpha_k \exp(-2\pi i k \lambda) \right|^2 S_X(d\lambda).$$

□

This allows a different derivation of the spectrum of an autoregression: Because $X_t - \phi X_{t-1} = \varepsilon_t$ and because ε_t is white noise, we conclude

$$S_\varepsilon(d\lambda) = |1 - \phi \exp(-2\pi i \lambda)|^2 S_X(d\lambda) = C_\varepsilon(0) d\lambda.$$

3 ARMA models

3.1 Definition of ARMA models: Causality, stationarity, invertibility

3.1.1 Linear difference equations

We collect here some results about the solutions of homogeneous difference equations that will be useful in the following. Hence for given (real) coefficients $\phi_1, \dots \phi_p$ with $\phi_p \neq 0$ we consider complex valued sequences $(u_t)_{t\in\mathbb{Z}}$ which satisfy

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \ldots + \phi_p u_{t-p} \quad (t \in \mathbb{Z}).$$

Theorem 3. *The set of sequences (u_t) that satisfy the above difference equation is a vector space of dimension p. A basis is given by the sequences of the form*

$$u_t = t^j \lambda^t$$

where λ^{-1} is a root of the polynomial

$$\Phi(z) = 1 - \phi_1 z - \dots \phi_p z^p$$

with multiplicity M and $0 \leq j < m$.

Proof. It is clear that a linear combination of two solutions is again a solution. Moreover, if p consecutive values $u_{k+1}, \dots, u_{k+p}$ of a solution (u_t) are given, then the solution is unique: Values u_t for $t > k+p$ follow by forward iteration, those for $t \leq k$ follow by backward iteration

$$u_{t-p} = \frac{u_t - \phi_1 u_{t-1} - \ldots - \phi_{p-1} u_{t-p+1}}{\phi_p}.$$

Therefore the dimension of the vector space is p.

Next, we show that the above sequences are indeed solutions. First we take $j = 0$:

$$\lambda^t - \phi_1 \lambda^{t-1} - \dots \phi_p \lambda^{t-p} = \lambda^t \Phi(\lambda^{-1}) \equiv 0.$$

Similarly, for $j = 1$ we have

$$t\lambda^t - \phi_1 (t-1)\lambda^{t-1} - \dots \phi_p (t-p)\lambda^{t-p} = \lambda^t t\Phi(\lambda^{-1}) - \lambda^{t-1}\Phi'(\lambda^{-1}) \equiv 0.$$

The general case follows because

$$\Phi^{(j)}(\lambda^{-1}) = -\lambda^j \sum_{k=j}^{p} \phi_k k(k-1)\cdots(k-j+1)\lambda^{-k} = 0 \quad (j < m)$$

implies that also

$$\sum_{k=1}^{p} \phi_k k^j \lambda^{-k} = 0 \quad (j < m).$$

The proof will be completed if we can show that the above solutions are linearly independent since the number of zeroes of a polynomial of degree p counted with their multiplicity is equal to p. For a proof of the linear independence, we refer to Brockwell and Davies, Theorem 3.6.2. □

Often we are interested in real valued solutions. These can be obtained easily because complex valued solution of real polynomials occur in conjugate pairs: Hence if $r\exp(i\nu)$ is a zero of $\Phi(z)$, then so is $r\exp(-i\nu)$. By the vector space property

$$t^j r^{-t}\frac{\exp(i\nu t)+\exp(-i\nu t)}{2} = t^j r^{-t}\cos(\nu t)$$

and

$$t^j r^{-t}\frac{\exp(i\nu t)-\exp(-i\nu t)}{2i} = t^j r^{-t}\sin(\nu t)$$

are also solutions, and one can easily show that they are linearly independent.

The above theorem shows that all solutions have the property $u_t \to 0$ as $t \to \infty$ iff $|z| > 1$ for all zeroes of $\Phi(z)$. In this case, all solutions decay exponentially to zero. If we require only that the solutions remain bounded as $t \to \infty$, we can allow zeroes with multiplicity 1 on $|z| = 1$.

All the results here remain valid if we consider sequences $(u_t)_{t\geq t_0}$ which satisfy the recursion for all $t \geq t_0 + p$.

3.1.2 Causal and stationary autoregressions

A stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called a Markovian autoregressive process of order p if

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

where ε_t is independent of all X_s, $s < t$. The variable ε_t is called the innovation at time t.

For a Markovian autoregression, $\phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + E(\varepsilon_t)$ is the best prediction of X_t from the past. Furthermore, the innovations at different times are independent: For $t > s$ ε_t is independent of $X_s - \phi_1 X_{s-1} - \ldots - \phi_p X_{s-p} = \varepsilon_s$.

When is a Markovian autoregression stationary ? First it is clear that under stationarity, the innovations are not only independent, but also identically distributed. For an AR(1)-process, we obtain by iteration

$$X_t = \sum_{j=0}^{t-1}\phi^j\varepsilon_{t-j} + \phi^t X_0.$$

Hence if second moments exist and if (X_t) is stationary, then

$$C(0) = \mathrm{Var}(\varepsilon)\sum_{j=0}^{t-1}\phi^{2j} + \phi^{2t}C(0)$$

since by assumption all terms on the right are independent. Clearly this implies that $|\phi| < 1$. The general case is covered by the next Theorem.

Theorem 4. *A stationary Markovian autoregression with finite second moments exists iff all zeroes of the polynomial*

$$\Phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$$

are outside the unit circle $\{z; |z| \leq 1\}$. *In that case the process has the representation*

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where the coefficients ψ_j *are the solution of the recursion*

$$\psi_j = \phi_1 \psi_{j-1} + \ldots + \phi_p \psi_{j-p} \quad (j \geq 1)$$

with initial conditions $\psi_0 = 1$, $\psi_{-1} = \ldots \psi_{1-p} = 0$ *and thus converge to zero exponentially fast. Moreover, if we define for* $t > 0$

$$X_t^* = \phi_1 X_{t-1}^* + \ldots \phi_p X_{t-p}^* + \varepsilon_t$$

with arbitrary initial conditions $X_0^*, X_{-1}^*, \ldots, X_{1-p}^*$, $X_t - X_t^* \to 0$ *almost surely and in* L_1.

Proof. We write the autoregression of order p as a vector autoregression of order 1: If we set $Z_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})^T$, $\eta_t = (\varepsilon_t, 0, \ldots, 0)^T$ and

$$\Phi = \begin{pmatrix} \phi_1 & \ldots & \phi_{p-1} & \phi_p \\ & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \end{pmatrix}$$

(with I_{p-1} the identity matrix in dimension $p-1$), then

$$Z_t = \Phi Z_{t-1} + \eta_t.$$

Iterating this autoregression, we obtain

$$Z_t = \sum_{j=0}^{t-1} \Phi^j \eta_{t-j} + \Phi^t Z_0.$$

If Z_t is stationary with finite variance, then Φ^t must converge to zero, and this is known to be equivalent to the condition that all eigenvalues of Φ are smaller than one in absolute value. The characteristic polynomial of Φ is however nothing else than the polynomial $z^p \Phi(1/z) = z^p - \phi_1 z^{p-1} - \ldots - \phi_p$.

Taking the limit in the above recursion, we obtain

$$Z_t = \sum_{j=0}^{\infty} \Phi^j \eta_{t-j}$$

Because of the way Φ is defined, the first column of Φ^t, say $c^{(t)}$, satisfies the recursion

$$c^{(t)} = ((\phi_1, \ldots, \phi_p)^T c^{(t-1)}, c_1^{(t-1)}, \ldots, c_{p-1}^{(t-1)}).$$

This implies the recursion for ψ_t. □

This Theorem shows that a stationary Markovian autoregression can be written as a linear combination of past innovations. A process with such a representation is called causal. It is clear that a causal autoregression with independent ε_t is Markovian.

Without the Markovian (or the causal) assumption, the Theorem is false. To see why, take any $|\phi| > 1$ and set

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} \varepsilon_{t+j}.$$

Clearly this is stationary if the ε_t are i.i.d. Moreover,

$$\phi X_{t-1} = -\varepsilon_t + X_t,$$

so the recursion is satisfied. However, ε_t contributes to the sum defining X_s for $s < t$, and thus the two variables are dependent.

Example: AR(2). The roots of $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2$ are

$$z_{1,2} = -\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2\phi_2}$$

One can verify that z_1 and z_2 are both outside the unit circle iff

$$-1 < \phi_2 < 1, \quad \phi_2 < 1 - |\phi_1|$$

Hence the set of parameters which correspond to a stationary Markovian autoregression is a triangle. The roots are complex for $\phi_2 < -\frac{1}{4}\phi_1^2$.

3.1.3 Invertible moving averages

A linear moving average of order q

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots \theta_q \varepsilon_{t-q}$$

with ε_t i.i.d. is always stationary. Moreover ε_t is always independent of X_s for $s < t$. However we cannot call ε_t the innovation of the process unless the other terms $\theta_1 \varepsilon_{t-1} + \ldots \theta_q \varepsilon_{t-q}$ on the right hand side can be expressed with the values X_s for $s < t$.

We therefore call a moving average invertible if there are coefficients (π_j) with $\sum |\pi_j| < \infty$ such that

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

(An autoregression is always invertible, just set $\pi_0 = 1$, $\pi_j = -\phi_j$ for $1 \le j \le p$ and $\pi_j = 0$ for $j > p$).

Theorem 5. *A moving average is invertible iff all zeroes of the polynomial*

$$\Theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$$

are outside the unit circle $\{z; |z| \leq 1\}$. *In that case the coefficients* π_j *are the solution of the recursion*

$$\pi_j = -\theta_1 \pi_{j-1} - \ldots - \theta_q \pi_{j-q}$$

with initial conditions $\pi_0 = 1$, $\pi_{-1} = \ldots \pi_{1-q} = 0$ *and thus converge to zero exponentially fast.*

Proof. For $q = 1$, we simply iterate the equation $X_t = \varepsilon_t + \theta \varepsilon_{t-1}$. For $q > 1$, we write the process as a vector moving average of order 1. □

3.1.4 ARMA-Processes

An autoregressive moving average process of order (p, q) (ARMA(p, q)) combines the properties of the two previous models. The recursion is

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t.$$

For a reasonable model ε_t should again be independent of X_s for $s < t$ and ε_t should depend only on past values $X_s, s \leq t$, i.e. the model should be invertible

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

Again it then follows that the variables ε_t are independent for different times t, and if (X_t) is stationary, the ε_t are even i.i.d. Causality is defined as in the autoregressive case: There are summable coefficients ψ_j such that

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

For a causal ARMA model with (ε_t) i.i.d., ε_t is obviously independent of X_s for $s < t$.

If one wants to generalize the arguments for the autoregressive case one sees that a problem occurs: For instance for any ϕ

$$X_t = \phi X_{t-1} + \varepsilon_t - \phi \varepsilon_{t-1}$$

has the stationay solution $X_t = \varepsilon_t$ which is also invertible and ε_t is independent of X_s for $s < t$. The reason for this problem is that Φ and Θ have common zeroes.

If we assume that Φ and Θ have no common zeroes, then the conditions that all zeroes of Φ and Θ are outside of the unit circle are again necessary and

sufficient for the existence of a stationary ARMA model which is invertible and causal.

For a more compact notation, I introduce now the backshift operator B which acts on infinite sequences $(BX)_t = X_{t-1}$. The recursion of the ARMA process can then be written as

$$\Phi(B)X_t = \Theta(B)\varepsilon_t.$$

Formally, we thus can write

$$X_t = \Phi(B)^{-1}\Theta(B)\varepsilon_t, \quad \varepsilon_t = \Theta(B)^{-1}\Phi(B)X_t.$$

If $\Phi(z)$ has no zeroes in $\{z; |z| \leq 1\}$, the Taylor series

$$\frac{\Theta(z)}{\Phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j$$

converges on $\{z; |z| \leq 1\}$ and thus we can define

$$\Phi(B)^{-1}\Theta(B) = \sum_{j=1}^{\infty} \psi_j B^j.$$

From the equality

$$\Theta(z) = \Phi(z) \cdot \sum_{j=0}^{\infty} \psi_j z^j,$$

we obtain by comparing the coefficient of z^j on both sides the equations

$$\psi_j - \sum_{k=1}^{\min(p,j)} \phi_k \psi_{j-k} = \theta_j \ (0 \leq j \leq q), \quad = 0 \ (j > q)$$

(we set $\theta_0 = 1$). This is the most convenient way to compute ψ_j numerically. In particular, ψ_j again satisfies a difference equation except for an initial part of length q.

A similar argument applies for the coefficients in the invertibility representation

$$\varepsilon_t = \Theta(B)^{-1}\Phi(B)X_t.$$

3.2 Properties

Let (X_t) be a stationary, causal and invertible ARMA(p, q) process. Then by the linearity of the expectation we obtain

$$E(X_t) = E(\varepsilon_t)\frac{\Theta(1)}{\Phi(1)}.$$

Hence the mean centered processes also satisfy the ARMA recursion.

We next compute the autocovariance $C(h)$ of the process. Because the covariance is linear in both arguments, we obtain for $h \geq 0$

$$\begin{aligned} C(h) &= \operatorname{Cov}(X_h, X_0) = \sum_{j=1}^{p} \phi_j \operatorname{Cov}(X_{h-j}, X_0) + \sum_{k=0}^{q} \theta_k \operatorname{Cov}(\varepsilon_{h-k}, X_0) \\ &= \sum_{j=1}^{p} \phi_j C(h-j) + \sum_{k=h}^{q} \theta_k \operatorname{Cov}(\varepsilon_{h-k}, X_0). \end{aligned}$$

In the last equality, we have used the property that ε_t is independent and thus uncorrelated with X_0 for $t > 0$. For $h > q$, the second sum on the right runs over an empty set and is thus zero. Therefore, we have shown that for $h \geq \max(p, q+1)$ the autocovariance function satisfies the difference equation

$$C(h) = \sum_{j=1}^{p} \phi_j C(h-j).$$

In particular, it decays to zero exponentially fast. Moreover, the properties are closely linked to properties of the zeroes of the polynomial Φ. If Φ has two zeroes $r\exp(\pm i\nu)$ with r close to one, then the covariance is (approximately) a damped harmonic with period $2\pi/\nu$.

In order to compute the values $C(h)$ for $h < \max(p, q+1)$, we need $\operatorname{Cov}(\varepsilon_s, X_0)$ for $s \leq 0$. These covariances can be computed from the causal representation:

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \Rightarrow \operatorname{Cov}(\varepsilon_s, X_0) = \operatorname{Var}(\varepsilon)\psi_{-s} \quad (s \leq 0).$$

Example: Autoregressions. For autoregressions, we only need $\operatorname{Cov}(\varepsilon_0, X_0)$ which is equal to $\operatorname{Var}(\varepsilon)$. The autocovariances $C(h)$ for $0 \leq h \leq p$ are then obtained from the equations

$$\begin{aligned} C(0) - \sum_{j=1}^{p} \phi_j C(j) &= \operatorname{Var}(\varepsilon) \\ C(h) - \sum_{j=1}^{h} \phi_j C(h-j) - \sum_{j=h+1}^{p} \phi_j C(j-h) &= 0 \quad (1 \leq h \leq p). \end{aligned}$$

These equations are called Yule-Walker equations. In Section 2.1 we computed the best linear prediction $\widehat{X}_{p|0:p-1}$: There we started with the covariances and computed the coefficients $\beta_k^{(p)}$. We end up with the same equations and $\beta_k^{(p)} = \phi_k$.

Example: ARMA(1,1). From

$$X_t = \phi X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$$

we obtain

$$X_t = \phi^2 X_{t-2} + \varepsilon_t + (\phi + \theta)\varepsilon_{t-1} + \phi\theta\varepsilon_{t-2}$$

and therefore $\psi_0 = 1$, $\psi_1 = (\phi + \theta)$. Hence the autocovariances $C(0)$ and $C(1)$ can be found by solving the equations

$$\begin{aligned} C(0) &= \phi C(1) + \mathrm{Var}(\varepsilon)(1 + \theta(\phi + \theta)) \\ C(1) &= \phi C(0) + \mathrm{Var}(\varepsilon)\theta. \end{aligned}$$

This gives the variance

$$C(0) = \mathrm{Var}(\varepsilon)\frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}$$

and the autocorrelations

$$\rho(1) = \phi + \frac{\theta(1 - \phi^2)}{1 + 2\theta\phi + \theta^2}, \quad \rho(h) = \phi^{h-1}\rho(1) \ (h > 1).$$

The spectrum of an ARMA model follows easily from Lemma 2:

$$s(\lambda) = \mathrm{Var}(\varepsilon)\frac{|\Theta(\exp(-2\pi i\lambda))|^2}{|\Phi(\exp(-2\pi i\lambda))|^2}.$$

If Φ has two zeroes $r\exp(\pm i\nu)$ with r close to 1, then the spectral density will have a peak near $\lambda = \nu/(2\pi)$ and the process will have an approximately periodic behavior with period $2\pi/\nu$.

Prediction from the infinite past: We assume that both X_t and ε_t have mean zero (i.e. we have subtracted the mean). For a causal and invertible ARMA model

$$\widehat{X}_{t|-\infty:t-1} = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

is then the best prediction of X_t based on the infinite past $(X_s, s < t)$, and ε_t is the prediction error. In order to compute it, one can either express ε_{t-j} with past observations by computing the coefficients π_j according to the formula given above, or one can set $\varepsilon_{s-1} = \cdots = \varepsilon_{s-q} = 0$ for a time point $s << t$ and then iterate the relation

$$\varepsilon_u = X_u - \sum_{j=1}^{p} \phi_p X_{u-p} - \sum_{j=1}^{q} \theta_k \varepsilon_{u-k}$$

for $u = s, s+1, \ldots t$. The error due to assuming $\varepsilon_{s-1} = \cdots = \varepsilon_{s-q} = 0$ decays exponentially as $t - s \to \infty$.

Predictions from the infinite past for more than one time step ahead can be made as follows: Because the best prediction is linear, we obtain for $k > 0$

$$\widehat{X}_{t+k|-\infty:t-1} = \sum_{j=1}^{\min(p,k-1)} \phi_j \widehat{X}_{t+k-j|-\infty:t-1} + \sum_{j=k}^{p} \phi_j X_{t+k-j} + \sum_{j=k}^{q} \theta_j \varepsilon_{t+k-j}.$$

Hence we see that the predictions for different lead times satisfy the difference equation associated with the AR part, except for finitely many lead times at the beginning. In particular, as the lead time increases, the predictions tend to zero, the mean of X_t.

3.3 Statistical inference for ARMA models

3.3.1 Estimation of coefficients

Estimation of the unknown parameters ϕ_j, θ_k and $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_t)$ is usually done with exact or approximate Gaussian maximum likelihood (MLE). An unknown mean is usually estimated first by the arithmetic mean of the data and then subtracted.

We have the following general formula for the density of $X_1, \ldots, X_n$

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) \cdots f(x_n|x_1, \ldots, , x_{n-1}).$$

In the Gaussian case, the conditional densities $f(x_t|x_1, \ldots, , x_{t-1})$ are again Gaussian with mean equal to the best linear prediction $\widehat{X}_{t|1:t-1}$ and variance equal to $\mathrm{Var}(X_t - \widehat{X}_{t|1:t-1})$. For the exact MLE, one computes these means and variances exactly as a function of the unknown parameters. An approximate likelihood uses $\widehat{X}_{t|-\infty:t-1}$ and $\mathrm{Var}(\varepsilon_t)$ instead where $\widehat{X}_{t|-\infty:t-1}$ is computed recursively starting with $\varepsilon_0 = \cdots = \varepsilon_{1-q} = 0$. In order to reduce the effect of these artificial starting values, one typically omits the first $r = \max(p, q+1)$ factors in the likelihood, that is one takes

$$f(x_{r+1}, \ldots, x_n|x_1, \ldots, x_r) = \prod_{t=r+1}^{n} f(x_t|x_1, \ldots x_{t-1}).$$

For the AR(p) model, this reduces to the least squares estimator

$$\arg\min \sum_{t=p+1}^{n} (x_t - \sum_{j=1}^{p} \phi_j x_{t-j})^2$$

which is particularly simple to compute. In the autoregressive case, there are other estimators: The Yule-Walker estimator determines the unknown ϕ_j and σ_ε^2 from the Yule-Walker equations with estimated covariances $\widehat{C}(h)$ ($0 \le h \le p$). . The Burg estimator proceeds recursively with respect to p, that is, it estimates the partial autocorrelations, and it does this by minimizing forward and backward prediction errors.

For long series, all the different versions give similar estimates, but for shorter series and parameters close to the boundary of the causality and invertibility region, the choice of the estimator can matter. Usually one prefers the exact MLE or the Burg estimator.

3.3.2 Asymptotic properties of estimators

One can show that all estimators introduced in the previous section are consistent, and the vector

$$\sqrt{n}((\widehat{\phi} - \phi)^T, (\widehat{\theta} - \theta)^T)^T$$

is asymptotically normal with mean zero and covariance matrix $\Gamma(\phi, \theta)$. Here the elements of Γ^{-1} are given by the covariances of the two autoregressive processes

$$U_t = \sum_{j=1}^{p} \phi_j U_{t-j} + Z_t, \quad V_t = \sum_{k=1}^{q} \theta_k V_{t-j} + Z_t$$

where Z_t is i.i.d. with mean zero and variance one. Because the same innovations are used, these two processes are correlated. More precisely,

$$\begin{aligned} \left(\Gamma^{-1}\right)_{jk} &= \mathrm{Cov}(U_j, U_k) \quad (j \le p, k \le p) \\ &= \mathrm{Cov}(V_{j-p}, V_{k-p}) \quad (j > p, k > p), \\ &= \mathrm{Cov}(U_j, V_{k-p}) \quad (j \le p, k > p) \end{aligned}$$

This holds if (X_t) is a causal and invertible ARMA model with no common zeroes in $\Phi(z)$ and $\Theta(z)$ and the innovations are i.i.d. with mean zero and variance σ^2. It is not required that the innovations are normal although we use the Gaussian MLE (the same is true in regression).

Example: For an AR(1) process $\sqrt{n}(\widehat{\phi} - \phi)$ is asymptotically normal with mean zero and variance $(1 - \phi^2)$ because $\mathrm{Cov}(U_1, U_1) = 1/(1 - \phi^2)$.

3.3.3 Order selection

A simple technique is to identify the orders p and q from the plot of the autocorrelations and partial autocorrelations. For an MA(q) process, all autocrrelations $\rho(h) = 0$ for $h > q$ whereas the partial autocorrelations $\tau(h)$ decay exponentially or like a damped harmonic as $h \to \infty$. For an AR(p) process, the partial autocorrelations $\tau(h)$ are zero for $h > p$ and the autocorrelations decay exponentially or like a damped harmonic as $h \to \infty$. For an ARMA(p,q) process with $p > 0$ and $q > 0$ both $\tau(h)$ and $\rho(h)$ decay exponentially or like a damped harmonic.

Nowadays it is also possible to fit ARMA(p,q) models for all $p \le p_0$ and $q \le q_0$ and to choose the one with the best fit afterwards. The most popular methods to choose the order are then the selection critera AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). They are defined as follows:

$$-2 \sup \ell(\phi, \theta, \sigma_\varepsilon^2) + C(p + q)$$

where ℓ is the log likelihood function and $C = 2$ in case of the AIC and $C = \log(n)$ in case of the BIC. If the order increases, the first term always increases because the supremeum is taken over a larger set. The second term is a penalty for the complexity of the model. If the estimates $\widehat{\phi}$ and $\widehat{\theta}$ are based on an approximate likelihood, then one uses this approximate likelihood in the AIC or BIC instead of the exact likelihood ℓ.

I do not discuss here the justification of these criteria, but just mention two results: 1) The AIC is an unbiased estimate of a distance between the fitted and the true model. 2) The AIC favors complex models and does not provide a consistent estimate of the true order if the true order is finite.

3.3.4 Goodness of fit

Once a model has been fitted (that is both the orders and the parameters have been estimated), one should check whether the fit is adequate. As a minimum, one should look at the time series plot and the acf of the residuals $\widehat{\varepsilon}_t$ which approximate the innovations ε_t and thus should be approximately i.i.d.. It is also a good idea to simulate from the fitted model and compare the plot of a simulated series with the plot of the original series. Ideally, the two plots should be visually indistinguishable. One can also look for nonlinear dependence among the residuals, by using for instance lag plots ($\widehat{\varepsilon}_{t+h}$ versus $\widehat{\varepsilon}_t$) or the acf of the squared residuals, or for non-Gaussianity with a normal plot of the residuals.

3.4 ARIMA-Models

So far all ARMA models were stationary. One way to analyze nonstationary data is to take differences, see 1.2. This can be included in the ARMA model:

$$\Phi(B)(1-B)^d X_t = \Phi^*(B)X_t = \Theta(B)\varepsilon_t.$$

The polynomial $\Phi^*(z)$ has degree $d+p$ and it has a root at $z=1$ of multiplicity d and p roots outside of the unit circle. Such a model is called an ARIMA(p,d,q) model (autoregressive integrated moving average). Note that an ARIMA model is not unique: If (X_t, ε_t) satisfies the above recursion, then so does $(X_t + A_0 + \ldots A_{d-1}t^{d-1}, \varepsilon_t)$ for arbitrary coefficients $A_0, \ldots, A_{d-1}$. In other words, the ARIMA model only specifies the conditional distribution of $X_1, X_2, \ldots$ given the initial values $X_0, X_{-1}, \ldots, X_{d-1}$, and not the distribution of these initial values. Also note that if $E(\varepsilon_t) \neq 0$, then $E(X_t)$ contains a term ct^d with $c \neq 0$. Because of this, one usually assumes that $E(\varepsilon_t) = 0$ if $d > 0$.

Whether we should choose $d > 0$ usually becomes clear from the inspection of the time series plot (slowly changing level or slowly changing slope of the series) and of the acf (behaviour of $\widehat{\rho}(h) \sim 1 - \text{const.}h$ with a small value of const.). Identifying p and q and estimating the coefficients is then done based on the differenced series $Y_t = (1-B)^d X_t$.

For forecasting, one usually assumes that the initial values $X_0, X_{-1}, \ldots, X_{d-1}$ are independent of the differenced series $Y_t = (1-B)^d X_t$. Then the same formula can be used for recursive computation of the forecast k steps ahead as in the stationary case.

If a series contains a seasonal component, then we often need to take also seasonal differences to achieve stationarity. This means that we use a model of the form

$$\Phi(B)(1-B)^d(1-B^M)^D X_t = \Theta(B)\varepsilon_t$$

where M is the number of observations in one seasonal cycle. Moreover, empirically the seasonal behavior also shows up in the structure of the polynomials Φ and Θ. For instance in the autoregressive case, X_t depends usually

on X_{t-1}, X_{t-M} and maybe X_{t-M-1}. This leads to the so-called seasonal ARIMA(p, d, q, P, D, Q) model:

$$\Phi(B)\Phi_M(B^M)(1-B)^d(1-B^M)^D X_t = \Theta(B)\Theta_M(B^M)\varepsilon_t$$

An example is given by the so-called airline model (because it fits the data on airline passengers, one of the standard data sets in R, well):

$$(1-B)(1-B^M)X_t = (1-\theta_1 B)(1-\theta_{M,1}B^M)\varepsilon_t.$$

4 Spectral methods

4.1 The spectral representation

4.1.1 Some results from deterministic spectral analysis

Fourier theory is concerned with the representation of signals g as a superposition of harmonics with different frequencies and amplitudes. If g is a signal in continuous time t with finite energy

$$\int_{-\infty}^{\infty} g(t)^2 dt < \infty,$$

then it can be represented as

$$g(t) = \int_{-\infty}^{\infty} G(\nu)\exp(i2\pi\nu t)d\nu \tag{1}$$

where

$$G(\nu) = \int_{-\infty}^{\infty} g(t)\exp(-i2\pi\nu t)dt. \tag{2}$$

Hence g is a superposition of harmonics with continuous frequencies ν. If we write $G(\nu)$ in polar coordinates $G(\nu) = |G(\nu)|\exp(i\phi(\nu))$, we see that the harmonic $G(\nu)\exp(i2\pi\nu t)$ has amplitude $|G(\nu)|$ and phase $\phi(\nu)$. Since g is real, $G(-\nu) = \overline{G(\nu)}$ and we also have a representation in terms of sine and cosine functions with frequencies $\nu > 0$. Moreover, Parseval's theorem says that

$$\int_{-\infty}^{\infty} g(t)^2 dt = \int_{-\infty}^{\infty} |G(\nu)|^2 d\nu,$$

i.e. the energy is the integral of squared amplitudes.

Next, we consider a signal (g_t) observed at time points $t\Delta$ with $t = 0, \pm 1, \ldots$ with finite energy $\sum_{t=-\infty}^{\infty} g_t^2 < \infty$. If we replace the integrand in (2) by a function which is constant on intervals of length Δ, then we obtain

$$G_p(\nu) = \Delta \sum_{t=-\infty}^{\infty} g_t \exp(-i2\pi t\Delta\nu), \tag{3}$$

and we can represent the signal with G_p:

$$g_t = \int_{-1/(2\Delta)}^{1/(2\Delta)} G_p(\nu) \exp(i2\pi t\Delta\nu) d\nu. \tag{4}$$

Hence again g is a superposition of harmonics, but the continuous frequencies ν are now restricted to $|\nu| \leq 1/(2\Delta)$. The reason for this is that in discrete time we cannot distinguish between harmonics at frequencies ν, $\nu \pm 1/\Delta$, $\nu \pm 2/\Delta$ etc. . This is called aliasing, and $1/(2\Delta)$ is called the Nyquist frequency.

If we consider G_p as a function of arbitrary ν, then it is periodic with period $1/\Delta$ (this is the reason for the subscript p). Note that $G_p(\nu) \neq G(\nu)$ for $|\nu| \leq 1/\Delta$, but rather

$$G_p(\nu) = \sum_{k=-\infty}^{\infty} G(\nu + k/\Delta) = G(\nu) + \sum_{k=1}^{\infty} (G(\nu + k/\Delta) + \overline{G(-\nu + k/\Delta)}).$$

This means that we add up the amplitudes at all frequencies we cannot distinguish. Finally, for a discrete time signal, Parseval's theorem says that

$$\Delta \sum_{t=-\infty}^{\infty} g_t^2 = \int_{-1/(2\Delta)}^{1/(2\Delta)} |G_p(\nu)|^2 d\nu.$$

In the last step, we consider a signal g observed at finitely many discrete time points $t\Delta$ with $t = 0, 1, \ldots n-1$. By replacing the integrand in (4) by a function which is constant on intervals of length $1/(n\Delta)$, we obtain the representation

$$g_t = \frac{1}{n\Delta} \sum_{k=0}^{n-1} G_k \exp(i2\pi t\Delta \frac{k}{n\Delta}) = \frac{1}{n\Delta} \sum_{k=0}^{n-1} G_k \exp(i2\pi tk/n) \tag{5}$$

whose inversion is

$$G_k = \Delta \sum_{t=0}^{n-1} g_t \exp(-i2\pi tk/n). \tag{6}$$

Hence the signal is now a superposition of harmonics with a finite number of frequencies $\nu_k = k/(\Delta n)$, the so-called Fourier frequencies. Again Parseval's theorem holds

$$\Delta \sum_{t=0}^{n-1} g_t^2 = \frac{1}{n\Delta} \sum_{k=0}^{n-1} |G_k|^2.$$

If we use (5) or (6) to define g_t for any $t \in \mathbb{Z}$ or G_k for any $k \in \mathbb{Z}$, we obtain periodic sequences. If we restrict an infinite sequence with Fourier representation

$$g_t = \int_{-1/(2\Delta)}^{1/(2\Delta)} G_p(\nu) \exp(i2\pi t\Delta\nu) d\nu,$$

to $0 \leq t < n$, then the relation between $G_p(\nu)$ and the discrete amplitudes G_k is

$$G_k = n\Delta \int_{-1/(2\Delta)}^{1/(2\Delta)} G_p(\nu) \exp(-i\pi(n-1)(\nu_k - \nu)\Delta) D_n(|\nu - \nu_k|\Delta) d\nu$$

where D_n is the so-called Dirichlet kernel

$$D_n(\nu) = \frac{\sin(n\pi\nu)}{n\sin(\pi\nu)}.$$

This means that the amplitude G_k in the discrete representation is a weighted average of the amplitudes $G_p(\nu)$ for ν around ν_k. The phase shift in the above formula occurs because the time points are not symmetric around the origin. The proof of this formula uses the the summation formula of a geometric series

$$\sum_{t=0}^{n-1} e^{i\lambda t} = \frac{e^{i\lambda n}-1}{e^{i\lambda}-1} = e^{i(n-1)\lambda/2}\frac{e^{in\lambda/2}-e^{-in\lambda/2}}{e^{i\lambda/2}-e^{-i\lambda/2}} = e^{i(n-1)\lambda/2}\frac{\sin(n\lambda/2)}{\sin(\lambda/2)}.$$

The discrete Fourier transform $(g_t) \to (G_k)$ can be computed by the Fast Fourier Transform (FFT) with $O(n\log_2(n))$ operations instead of $O(n^2)$ operations in a naive implementation. This algorithm is crucial for the widespread use of Fourier methods in many applications.

4.1.2 The spectral representation of stationary stochastic processes

For a stationary stochastic process $(X_t; t \in \mathbb{Z})$, the energy $\sum_{t=-\infty}^{\infty} X_t^2$ is infinite, but if second moments exist, the power (energy per time unit) converges to a finite value

$$\frac{1}{2T+1}\sum_{t=-T}^{T} X_t^2 \to E(X_t^2).$$

Hence we cannot expect to have a representation of the form

$$X_t(\omega) = \int_{-1/2}^{1/2} \exp(i2\pi\nu t) Z(\nu,\omega) d\nu.$$

However, a deep result says that we have the representation

$$X_t(\omega) = E(X_t) + \int_{-1/2}^{1/2} \exp(i2\pi\nu t) Z(d\nu,\omega)$$

where Z is a (complex) stochastic process with uncorrelated increments:

1. $Z(-\nu) - Z(-\nu-h) = \overline{Z(\nu+h)-Z(\nu)}$ for all ν, h.
2. $E(Z(\nu+h)-Z(\nu)) = 0$ for all ν, h.
3. $E|Z(\nu+h)-Z(\nu)|^2 = S(\nu+h)-S(\nu)$ where S is the spectral distribution function $S(\nu) = S([-1/2,\nu])$.
4. For $\nu < \nu+h < \nu' < \nu'+h'$, $E((Z(\nu+h)-Z(\nu))\overline{(Z(\nu'+h')-Z(\nu'))}) = 0$.

Here the integral is defined as the limit of

$$\sum_j \exp(i2\pi\nu_j t)(Z(\nu_j,\omega) - Z(\nu_{j-1},\omega))$$

as the partition $\nu_0 = -1/2 < \nu_1 < \ldots < \nu_J = 1/2$ becomes finer. Hence intuitively, the process is a superposition of harmonics with uncorrelated mean zero amplitudes, and the variance of the amplitudes are given by the increments of the spectrum. In other words, the spectrum spectrum says how strongly the different frequencies are represented in the process. If the spectral density exists, $E(|Z(\nu_j,\omega) - Z(\nu_{j-1},\omega)|^2)$ is of the order $\nu_j - \nu_{j-1}$ and therefore $|Z(\nu_j,\omega) - Z(\nu_{j-1},\omega)|$ is typically of the order $\sqrt{\nu_j - \nu_{j-1}} > \nu_j - \nu_{j-1}$. This is the crucial difference between the spectral representation here and the representations in the previous subsection.

Formally, we can write the properties of Z as

$$E(Z(d\nu)\overline{Z(d\nu')}) = \delta_{\nu,\nu'} S(d\nu)$$

where $\delta_{\nu,\nu'} = 0$ for $\nu \neq \nu'$ and $\delta_{\nu,\nu} = 1$ (the Kronecker delta). We then obtain the spectral representation of the autocovariances (Herglotz's Theorem)

$$\begin{aligned} C(k) &= \mathrm{Cov}(X_{t+k}(\omega), X_t(\omega)) = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \exp(i2\pi\nu(t+k))\exp(-i2\pi\nu' t)E(Z(d\nu,\omega)\overline{Z(d\nu',\omega)}) \\ &= \int_{-1/2}^{1/2} \exp(i2\pi\nu k)S(d\nu). \end{aligned}$$

In particular,

$$E((X_t - E(X_t)^2) = \int_{-1/2}^{1/2} S(d\nu)$$

which is the analogue of Parseval's theorem.

4.1.3 Linear filters

A (time invariant) linear filter is a transformation of an input time series (X_t) into an output time series (Y_t) of the following form

$$Y_t = \sum_k a_k X_{t-k}$$

The input or output can be either deterministic or stochastic. Usually one assumes that $\sum_k |a_k| < \infty$ or some other condition in order that the right hand side is well defined.

If the input is an impulse at time zero, $X_t = \delta_{t0}$, then the output is equal to $Y_t = a_t$. Because of this, the coefficients a_k are called the impulse response coefficients. If the input is a harmonic with frequency ν, $X_t = G\exp(i2\pi\nu t)$ then the output is again a harmonic with the same frequency

$$Y_t = GA(\nu)\exp(i2\pi\nu t), \quad A(\nu) = \sum_k a_k \exp(-i2\pi\nu k).$$

There is however a change in amplitude by $|A(\nu)|$ and also a phase shift unless the coefficients are symmetric ($a_{-k} = a_k$). $A(\nu)$ is called the transfer function.

By linearity of the linear filter, a superposition of harmonic oscillations is transformed into another superposition of harmonics where the amplitudes and phases are changed by the transfer function. Stationary stochastic processes are superpositions of harmonic oscillations:

$$X_t = E(X_t) + \int_{-1/2}^{1/2} \exp(i2\pi\nu t) Z(d\nu).$$

If the coefficients (a_k) are summable,

$$Y_t = A(0)E(X_t) + \int_{-1/2}^{1/2} \exp(i2\pi\nu t) A(\nu) Z(d\nu).$$

Therefore the spectral increment process of (Y_t) is $A(\nu)Z(d\nu)$ and we have the following relation between the spectral measures of (X_t) and (Y_t):

$$S_Y(d\nu) = |A(\nu)|^2 S_X(d\nu)$$

(compare Lemma 2).

4.2 The periodogram

The periodogram of a time series of length n with sampling interval Δ is defined as

$$I_n(\nu) = \frac{\Delta}{n} \left| \sum_{t=1}^{n} (X_t - \overline{X}) \exp(-i2\pi\nu t\Delta) \right|^2 .$$

In words, we compute the absolute value squared of the Fourier transform of the sample, that is we consider the squared amplitude and ignore the phase.

Note that I_n is periodic with period $1/\Delta$ and that $I_n(0) = 0$ because we have centered the observations at the mean. The centering has no effect for Fourier frequencies $\nu = k/(n\Delta)$, $k \neq 0$.

By mutliplying out the absolute value squared on the right, we obtain

$$I_n(\nu) = \frac{\Delta}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} (X_t - \overline{X})(X_s - \overline{X}) \exp(-i2\pi\nu(t-s)\Delta) = \Delta \sum_{h=-n+1}^{n-1} \widehat{C}(h) \exp(-i2\pi\nu h).$$

Hence the periodogram is nothing else than the Fourier transform of the estimated acf. In the following, we assume that $\Delta = 1$ in order to simplify the formula (although for applications the value of Δ in the original time scale matters for the interpretation of frequencies).

By the above result, the periodogram seems to be the natural estimator of the spectral density

$$s(\nu) = \sum_{h=-\infty}^{\infty} C(h) \exp(-i2\pi\nu h).$$

However, a closer inspection shows that the periodogram has two serious shortcomings: It has large random fluctuations, and also a bias which can be large.

We first consider the bias. Using the spectral representation, we see that up to a term which involves $E(X_t) - \overline{X}$

$$I_n(\nu) = \frac{1}{n}\left|\int \sum_{t=1}^{n} e^{-i2\pi(\nu-\nu')t} Z(d\nu')\right|^2 = n\left|\int e^{-i\pi(n+1)(\nu-\nu')} D_n(\nu-\nu')Z(d\nu')\right|^2.$$

Taking the expectation on both sides and using the properties of Z, we obtain

$$E(I_n(\nu)) = n\int D_n(\nu-\nu')^2 s(\nu)d\nu'.$$

In order to gain insight from this formula, we need to understand the behavior of the Dirchlet kernel D_n and the so-called Fejér kernel

$$F_n(\nu) = nD_n(\nu)^2.$$

It can be checked that $F_n(0) = n$, $F_n(\nu) \to 0$ as $n \to \infty$ for all $0 < |\nu| \leq 1/2$ and $\int_{-1/2}^{1/2} F_n(\nu) = 1$ for all n. Hence F_n approximates the Dirac delta function and for a continuous density we obtain $E(I_n(\nu)) \to s(\nu)$ for any $\nu \neq 0$.

Still, for some applications, the bias of the periodogram can be substantial. In such cases the bias is reduced if we use a so-called taper. This is a set of weights $h_1, h_2, \ldots, h_n$ which are one for t close to $n/2$ and decay smoothly to zero for t near 1 and n. With these weights, we compute the tapered periodogram as follows

$$I_n^h(\nu) = \frac{1}{\sum_{t=1}^n h_t^2}\left|\sum_{t=1}^{n} h_t(X_t - \overline{X})\exp(-i2\pi\nu t)\right|^2.$$

If we use a taper, then we obtain

$$E(I_n^h(\nu)) = \int H_n(\nu-\nu')s(\nu')d\nu'$$

where

$$H_n(\nu) = \frac{1}{\sum_{t=1}^n h_t^2}\left|\sum_{t=1}^{n} h_t \exp(-2\pi i\nu t)\right|^2.$$

If h_t is as described above, $H_n(\nu)$ has smaller sidelobes than the Fejér kernel.

The variances and covariances of the periodogram depend in principle on the fourth moments of the process. However, for many processes a Central Limit Theorem applies for the Fourier transform and thus the real and imaginary part of $\sum h_t(X_t - \overline{X})\exp(2\pi i\nu t)$ have asymptotically a normal distribution with mean zero and variance $s(\nu)/2$ for $\nu \neq 0, 1/2$. Because of this

$$\frac{I_n^h(\nu)}{s(\nu)} \text{ approximately } \sim \text{Exp}(1).$$

In particular, the periodogram is an asymptotically unbiased, but not consistent estimator for the spectral density, and

$$\left[\frac{I_n^h(\nu)}{-\log(0.025)}, \frac{I_n^h(\nu)}{-\log(0.975)}\right]$$

is an approximate 95% confidence interval for $s(\nu)$. On the logarithmic scale, this interval has constant width.

For two different frequencies $\nu \neq \nu'$, the periodogram values are asymptotically independent, in particular the covariance tends to zero. This explains the irregular behavior of the periodogram as a function of frequency. Because of this and because of the inconsistency, the periodogram is of limited value.

For two frequencies close together, we have the following approximation

$$\mathrm{Cov}(I_n^h(\nu), I_n^h(\nu')) \approx \frac{s(\nu)s(\nu')}{\sum_{t=1}^n h_t^2} \left|\sum_{t=1}^n h_t^2 \exp(-2\pi i(\nu-\nu')t)\right|^2 .$$

Without a taper, i.e. for $h_t \equiv 1$, the periodogram values at two Fourier frequencies j/n and j'/n are thus approximately uncorrelated. This does not hold if we use a taper.

I refer to the literature for exact statements and proofs of these results.

4.3 Smoothing the periodogram

The reason why the periodogram is not consistent is that as the length n of the time series increases , we obtain independent estimates of the spectral density at an increasingly dense set of Fourier frequencies $\nu_k = k/n$. If the spectral density is smooth, we can therefore pool the information from nearby frequencies.

The tapered and smoothed spectral estimate is

$$\hat{s}^{(ts)}(k/n) = \sum_{j=-J}^{J} w_j I_n^h((k-j)/n),$$

where the w_j's are weights with the following properties

$$w_j > 0,\ w_j = w_{-j}\ (-J \leq j \leq J),\ \sum_{j=-J}^{J} w_j = 1.$$

If $k \leq J$, the smoothing includes the periodogram at the origin which is equal or very close to zero if the mean μ is estimated. In this case, we exclude $j = k$ from the sum and renormalize the weights.

The properties of this estimator can be derived by the same arguments that are used for kernel smoothers in nonparametric regression. If we neglect the bias of the tapered periodogram, the bias of $\hat{s}^{(ts)}$ is approximately

$$\frac{s''(k/n)}{2}\frac{1}{n^2}\sum_{j=-J}^{J} j^2 w_j.$$

The variance of $\hat{s}^{(ts)}(k/n)$ depends on whether or not a taper is used. Without a taper the summands are approximately uncorrelated, and we obtain for $k \neq 0, n/2$

$$\mathrm{Var}(\hat{s}^{(ts)}(k/n)) \approx s(k/n)^2 \sum_{j=-J}^{J} w_j^2.$$

With a taper, we have to take the correlation of the summands into account. We skip the details and just state that in this case the variance is increased by the factor

$$M(h) = \frac{\frac{1}{n}\sum_{t=1}^{n} h_t^4}{(\frac{1}{n}\sum_{t=1}^{n} h_t^2)^2}.$$

By Cauchy-Schwarz, $M(h)$ is strictly greater than one unless h_t is constant, and thus asymptotically tapering entails some loss of precision. However, this is often more than compensated by a reduction in bias.

The choice of J, that is the number of frequencies involved in the smoothed estimate, is difficult. Small values of J give a small bias, but a large variance, and vice versa. Asymptotically, the optimal choice is $J = O(n^{4/5})$, but the constants involve both s and s'' which are unknown. In practice, one often looks at the estimate for different values of J and then makes a subjective choice.

The above results imply that to a first approximation

$$E\left(\frac{\hat{s}^{(ts)}(k/n)}{s(k/n)}\right) = 1, \quad \mathrm{Var}\left(\frac{\hat{s}^{(ts)}(k/n)}{s(k/n)}\right) = \sum_{j=-J}^{J} w_j^2 \; M(h).$$

Because the periodogram values have asymptotically an exponential distribution and the sum of m independent exponential random variables is distributed as $1/2$ times a chisquared random variable with $2m$ degrees of freedom, one approximates the distribution of $\hat{s}^{(ts)}(k/n)/s(k/n)$ by Z_d/d where $Z_d \sim \chi_d^2$ and the degrees of freedom d are chosen to match the variance given above. This then leads to the following confidence interval for $s(k/n)$

$$\left[\frac{\hat{s}^{(ts)}(k/n)\; d}{\chi^2_{d,1-\alpha/2}}, \frac{\hat{s}^{(ts)}(k/n)\; d}{\chi^2_{d,\alpha/2}}\right] \quad \text{where } d = \frac{2}{\sum_{j=-J}^{J} w_j^2 \; M(h)}.$$

4.4 Alternative estimators of the spectrum

So far, we have averaged over the values of the periodogram at the Fourier frequencies k/n because they are approximately independent in the case of no taper and because the fast Fourier transform can be used for computation. We can also use a different grid k/n' with $n' > n$ (we then have to set $X_t = \bar{X}$ for $n < t \leq n'$ in order to use the fast Fourier transform). In the limit we then have a continuous average

$$\widehat{s}^{(lw)}(\nu) = \int W(\nu - \nu') I_n^h(\nu') d\nu'.$$

This can be shown to be equal to

$$\sum_{k=-n+1}^{n-1} w_k \widehat{C}^h(k) \exp(-2\pi i \nu k)$$

where

$$w_k = \int W(\nu) \exp(2\pi i \nu k) d\nu$$

and

$$\widehat{C}^h(k) = \frac{1}{\sum_{t=1}^n h_t^2} \sum_{t=1}^{n-|k|} h_t (X_t - \bar{X}) h_{t+|k|} (X_{t+|k|} - \bar{X})$$

are the autocovariances of the tapered series. In other words, smoothing of the periodogram is equivalent to downweighting the estimated autocovariances in the inversion formula

$$s(\nu) = \sum_{k=-\infty}^{\infty} C(k) \exp(-2\pi i k \nu).$$

This estimator is therefore called a lag weight estimator (which explains the superscript lw). For computational reasons, $\widehat{s}^{(ts)}$ is usually prefered.

A different approach consists in averaging the periodograms for segments of $m < n$ consecutive observations:

$$\widehat{s}^{(os)}(\nu) = \frac{1}{J \sum_{t=1}^m h_t^2} \sum_{j=0}^{J-1} \left| \sum_{t=1}^m h_t (X_{t+jd} - \bar{X}) e^{-2\pi i \nu t} \right|^2$$

where J is the integer part of $(n-m)/d$. The parameter d regulates how much the segments overlap: For $d = 1$ we have maximal overlap whereas for $d = m$ there is no overlap (os stands for overlapping segments). It can be shown that in case of maximal overlap, this is essentially a lag weight estimator. It has however the advantage that it gives also information about changes in the periodogram over time. It is thus the first step towards a time-frequency analysis where one wants to analyze how strongly different frequencies are present at different times. This is however an ill-posed question since by Heisenberg's uncertainty principle a high resolution in time entails a low resolution in frequency and vice versa.

An entirely different approach to spectral estimation consists in using the spectral density of a fitted autoregressive model. Usually, one chooses the order of the autoregression by AIC. This usually gives very smooth estimates, but sometimes details are lost that can be detected by $\widehat{s}^{(ts)}$. A combination of both methods fits an autoregression, usually of low order without assuming that the innovations

$$\varepsilon_t = X_t - \sum_{k=1}^p \phi_k X_{t-k}$$

are exactly white noise. In any case, the general formula

$$s_X(\nu) = \frac{s_\varepsilon(\nu)}{|1 - \sum \phi_k \exp(-2\pi i \nu k)|^2}.$$

holds, and one estimates $s_\varepsilon(\nu)$ by smoothing the periodogram of the residuals. Even when $s_\varepsilon(\nu)$ is not exactly constant, it is at least much flatter than $s_X(\nu)$ and thus the problems with the bias are less serious. This approach is called prewhitening.

4.5 Wavelets in time series analysis

Wavelets are a rather recent invention which is suitable both for smoothing time series and for a time-frequency analysis. We can only give a very brief introduction. The discrete wavelet transform decomposes an equispaced time series of length n as follows:

$$X_t = \sum_{j=1}^{J} \sum_{k=0}^{2^{-j}n-1} d_{j,k} 2^{-j/2} \psi(2^{-j}t-k) + \sum_{k=0}^{2^{-J}n-1} a_{J,k} 2^{-J/2} \phi(2^{-J}t-k) \quad (t = 0, 1, \ldots, n-1)$$

where ψ is the so-called mother wavelet – a small wave located near zero – and ϕ is the so-called father wavelet or scaling function which represents a smooth part. Hence we have a decomposition into oscillations with frequencies 2^{-j} located at times $k2^j$ for $j = 1, 2, \ldots, J$ and a part which contains the lower frequencies. The simplest example is the Haar wavelet where

$$\psi(t) = 1_{[0,1/2)}(t) - 1_{[1/2,1)}(t), \quad \phi(t) = 1_{[0,1)}(t).$$

For other cases, ψ and ϕ are defined through a limiting operation and thus have to be calculated numerically.

The amplitudes $d_{j,k}$ and $a_{J,k}$ are computed from the original series by iterative application of an orthogonal transformation. We start with $a_{0,t} = X_t$ and set for $j = 1, 2, \ldots, J \le \log_2(n)$

$$a_{j,t} = \sum_{\ell=0}^{L-1} g_\ell \; a_{j-1,2t+1-\ell}, \quad d_{j,t} = \sum_{\ell=0}^{L-1} h_\ell \; a_{j-1,2t+1-\ell} \quad (t = 0, 1, \ldots, 2^{-j}n - 1).$$

(all indices are extended periodically). In words, we take the coefficients $a_{j-1,t}$ for odd times and apply to them two linear filters with impulse response coefficients g_ℓ and $h_\ell = (-1)^\ell g_{L-\ell-1}$, respectively. The coefficients g_ℓ are defined through the father wavelet (details omitted). They can be chosen arbitrarily subject to the constraints that L must be even and

$$\sum_{\ell=0}^{L-1} g_\ell = \sqrt{2}, \qquad \sum_{\ell=0}^{L-1-2n} g_\ell g_{\ell+2n} = \delta_{n,0} \quad (n = 0, 1, \ldots, L/2 - 1).$$

For $L = 2, 4$ there is essentially only one solution, e.g. for $L = 2$ we have $g_0 = g_1 = 1/\sqrt{2}$. For $L \ge 6$, there are several solutions.

Because the discrete wavelet transform is a product of orthogonal linear transformations and thus is again linear and orthogonal, the computation of the inverse is easy. For smoothing, one typically sets $d_{j,k}$ and $a_{J,k}$ equal to zero if their absolute value is small and then applies the inverse transform. This retains features in the data which are not smooth in a conventional sense.

In the maximal overlap discrete wavelet transform, one uses the above recursions without omitting coefficients $a_{j-1,t}$ for even t:

$$\tilde{a}_{j,t} = 2^{-j/2} \sum_{\ell=0}^{L-1} g_\ell \, \tilde{a}_{j-1,t-2^{j-1}\ell}, \quad \tilde{d}_{j,t} = 2^{-j/2} \sum_{\ell=0}^{L-1} h_\ell \, \tilde{a}_{j-1,t-2^{j-1}\ell} \ (t = 0, 1, \ldots, n-1).$$

This creates redundancies, but is sometimes easier for a time-frequency interpretation.

If X_t is a stochastic process, the amplitudes $a_{j,t}$ and $d_{j,t}$ are random variables, and one can study their distributions. Because the wavelet transform is orthogonal, these amplitudes are again i.i.d. for Gaussian white noise. It turns out that also under dependence they become apprxoimately independent like the periodogram values. In addition, the average of the $a_{j,t}^2$ for fixed j is essentially an estimate of the spectrum integrated over the frequency interval $[2^{-j-1}, 2^{-j}]$. A key difference is however that this holds also for integrated processes: We only need that $(1-B)^d X_t$ is stationary for some $d < L/2$.

5 Further topics

5.1 Multivariate time series

In a multivariate time series we have at each observation time a random vector $\mathbf{X}_t = (X_{ti})$. The definition of stationarity is unchanged. The mean $E(\mathbf{X}_t)$ is then also a vector which is independent of t under the assumption of stationarity. The covariance $\mathrm{Cov}(\mathbf{X}_{t+h}, \mathbf{X}_t)$ is now a matrix which depends only on h in the stationary case:

$$C(h)_{ij} = \mathrm{Cov}(X_{t+h,i}, X_{tj}).$$

It is called the cross covariance function. Because the covariance is symmetric, we obtain

$$C(-h)_{ij} = \mathrm{Cov}(X_{t-h,i}, X_{tj}) = \mathrm{Cov}(X_{ti}, X_{t+h,j}) = C(h)_{ji},$$

that is $C(-h) = C(h)^T$. The cross correlations are defined as

$$\rho(h)_{i,j} = \frac{C(h)_{ij}}{\sqrt{C(0)_{ii} C(0)_{jj}}}.$$

The estimation of the mean, the cross covariances and the cross correlations is done as in the univariate case. The judgement of estimated cross correlations is however delicate because the variance depends on cross- and autocorrelations

of all lags. If either (X_{t1}) or X_{t2} is white noise and if the true value $\rho_{12}(h) = 0$, then $\widehat{\rho}_{12}(h)$ is asymptotically normal with mean zero and variance $1/n$.

Multivariate ARMA-models are defined similarly to the univariate case:

$$\mathbf{X}_t = \sum_{j=1}^{p} \phi_j \mathbf{X}_{t-j} + \boldsymbol{\varepsilon}_t + \sum_{j=1}^{q} \theta_j \boldsymbol{\varepsilon}_{t-j}$$

where the ϕ_j and θ_j are now matrices and $\boldsymbol{\varepsilon}_t$ is so-called multivariate white noise, that is $C_{\boldsymbol{\varepsilon}}(h) = 0$ for $h \neq 0$ whereas $C_{\boldsymbol{\varepsilon}}(0)$ is an aribtrary positive definite matrix (the innovations for different components of the time series can be correlated). Causality and invertibility of multivariate ARMA models are defined as in the univariate case. The condition for causality becomes

$$\det(\Phi(z)) \neq 0 \text{ for } |z| \leq 1, \text{ where } \Phi(z) = I - \sum_{j=1}^{p} \phi_j z^j.$$

and for invertibility

$$\det(\Theta(z)) \neq 0 \text{ for } |z| \leq 1, \text{ where } \Theta(z) = I + \sum_{j=1}^{p} \theta_j z^j.$$

Also the computation of cross covariances and of linear predictions for known parameters is similar to the univariate case. The estimation of the unknown parameters is however much more difficult in the multivariate case than in the univariate case: First, the number of parameters grows quickly with the orders p and q, and the likelihood surface can have easily multiple maxima.

In a multivariate AR model, X_{ti} is influenced by all other components of past observations. In some applications, one assumes that the influence goes only in one direction:

$$X_{t2} = \sum_{j=0}^{\infty} \beta_j X_{t-j,1} + U_t$$

where (U_t) is a stationary process uncorrelated with (X_{t1}). This is called a transfer function model. In order to be able to estimate the infinitely many coefficients β_j, we assume that they are obtained through a Taylor series of a rational function

$$\sum_{j=0}^{\infty} \beta_j z_j = \frac{1 + \sum_{j=1}^{q} \alpha_j z^j}{1 - \sum_{j=1}^{p} \gamma_j z^j}.$$

Moreover, one also assumes that the error process U_t is an ARMA process. Then for given orders, we have a parametric model and can estimate the parameters. The choice of the orders can however be difficult.

Cointegration is a rather recent method (for which Clive Granger got the Nobel Prize in Economics in 2003). This is a special method for multivariate integrated series. In order to explain the main idea, We consider the bivariate case and assume that both (X_{t1}) and X_{t2} are nonstationary, but $(1 - B)X_{t1}$

and $(1-B)X_{t2}$ are stationary. Then it can happen that some linear combination $Y_t = \beta_1 X_{t1} + \beta_2 X_{t2}$ is stationary. For instance, a macroeconomic series of two countries can both be nonstationary, but the difference can be stationary because of strong economic relations between the two countries. If this is the case, then (β_1, β_2) is called a cointegrating vector. If such a cointegrating vector exists, then fitting a model to the differenced series $(1-B)\mathbf{X}_t$ is not a good procedure.

As in the univariate case, the cross covariance cannot be an aribtrary function of the lag h. There is a condition of positive definiteness which is satisfied if and only if we have the representation

$$C(h)_{jk} = \int_{-1/2}^{1/2} e^{2\pi i h \nu} S(d\nu)_{jk}$$

where S is a non-negative definite matrix distribution. If $\sum_h |C(h)_{jk}| < \infty$, then the S_{jk} have densities $(s(\nu)_{jk})$ where

$$s(\nu)_{jk} = \sum_{h=-\infty}^{\infty} C(h)_{jk} e^{-2\pi i \nu h}.$$

However, the densities $s_{jk}(\nu)$ are typically not real for $j \neq k$, but $s_{jk}(-\nu) = \overline{s_{jk}(\nu)} = s_{kj}(\nu)$ holds and the matrix $(s_{jk}(\nu))$ is nonnegative definite. The real part of s_{jk} is called the cospectrum and minus the imaginary part the quadspectrum.

Spectra of multivariate ARMA processes. The following formula holds

$$s(\nu) = \Phi(e^{2\pi i \nu})^{-1} \Theta(e^{2\pi i \nu}) C_{\boldsymbol{\varepsilon}}(0) \Theta(e^{-2\pi i \nu})^T \Phi(e^{-2\pi i \nu})^{-T}.$$

In the multivariate case, we have again the spectral representation

$$X_{tj} = \int_{-1/2}^{1/2} e^{2\pi i \nu t} Z_j(d\nu)$$

where

$$E(Z_j(d\nu)\overline{Z_k(d\nu')}) = \delta_{\nu,\nu'} S(d\nu)_{jk}.$$

This says that the amplitudes in the spectral representation are uncorrelated for different frequencies, and the spectrum contains the information about the absolute values and the linear relations between amplitudes at the same frequencies.

As an example of the use of the spectrum, we mention the problem of approximating X_{t2} linearly by the values of the series (X_{s1}). Assuming that both series have mean zero,

$$\widehat{X}_{t2} = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j,1}.$$

We want to find the coefficients (ψ_j) such that the mean square error $E((X_{t2} - \widehat{X}_{t2})^2)$ is minimal. It turns out that the solution is most easily found in the frequency domain:

$$\psi_j = \int_{-1/2}^{1/2} \frac{s_{21}(\nu)}{s_{11}(\nu)} e^{2\pi i \nu j} d\nu$$

and

$$E((X_{t2} - \widehat{X}_{t2})^2) = \int_{-1/2}^{1/2} \left(1 - \frac{|s_{2,1}(\nu)|^2}{s_{11}(\nu) s_{22}(\nu)}\right) s_{22}(\nu) d\nu.$$

Without the first series, the best prediction is zero and the mean squared error is $\int_{-1/2}^{1/2} s_{22}(\nu) d\nu$.

Estimation of the spectrum begins with the matrix periodogram

$$I_n(\nu)_{jk} = \frac{1}{n} (\sum_{t=1}^{n} (X_{tj} - \bar{X}_{.j}) e^{-2\pi i t \nu}) (\sum_{t=1}^{n} (X_{tk} - \bar{X}_{.k}) e^{2\pi i t \nu})$$

and smoothes it by averaging over neighboring frequencies.

5.2 Long range dependence

A stationary process (X_t) is called long range dependent (or long memory) with parameter $d \in (0, 0.5)$, if

$$C(h) = \mathrm{Cov}(X_{t+h}, X_t) \sim c\ h^{2d-1} \quad (h \to \infty).$$

Note that for such a process $\sum_h |C(h)| = \infty$. Under additional technical conditions, this is equivalent to

$$\mathrm{Var}(\sum_{t=1}^{n} X_t) \sim \frac{c}{d(1+2d)} n^{1+2d}$$

and

$$S(d\nu) = s(\nu) d\nu, \text{ with } s(\nu) \sim \frac{c\Gamma(1-d)\Gamma(d)}{(2\pi)^{2d}\Gamma(1-2d)} |\nu|^{-2d} \ (\nu \to 0).$$

The first result implies that the variance of the arithmetic mean decays to zero like n^{-1+2d}, i.e. at a lower rate than in the case where the autocovariances are summable.

There are two explicit models which show this behavior. The first one are the increments of fractional Brownian motion $(B_d(t); t \geq 0)$ with parameter d: This is a Gaussian (nonstationary) process with $E(B_d(t)) = 0$, $\mathrm{Var}(B_d(t)) = t^{2d+1}$ and $\mathrm{Var}(B_d(t) - B_d(s)) = \mathrm{Var}(B_d(t-s))$. Hence, the increments $X_t = B_d(t) - B_d(t-1)$ form a stationary Gaussian process with autocovariance

$$C(h) = \frac{1}{2}(|h+1|^{2d+1} - 2|h|^{2d+1} + |h-1|^{2d+1})$$

This process is central because for any Gaussian long range dependent process, the block sums

$$Z_t = \frac{1}{n^{1/2+d}} \sum_{s=(t-1)n+1}^{tn} (X_s - E(X_s))$$

converge in distribution to it. The formula for the spectral density is however rather complicated.

The other important example of a long range dependent process is given by fractional differences:

$$(1-B)^d X_t = \varepsilon_t \Leftrightarrow X_t = (1-B)^{-d}\varepsilon_t$$

where B is the backshift operator that we used for ARMA models and (ε_t) is white noise. We define fractional differences through the Taylor expansion of $(1-z)^{-d}$, that is

$$X_t = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} \varepsilon_{t-j}.$$

The coefficients on the right are not summable, but their squares are and one can show that the series converges and defines a stationary process. Moreover, the theory of linear filters gives the spectral density

$$s(\nu) = \frac{\sigma_\varepsilon^2}{|1-\exp(2\pi i\nu)|^{2d}} = \frac{\sigma_\varepsilon^2}{(2\sin(\pi\nu))^{2d}} \sim \frac{\sigma_\varepsilon^2}{(2\pi)^{2d}} |\nu|^{-2d}.$$

For fractional differences, one can also compute the acf. One obtains

$$C(h) = \frac{\sigma_\varepsilon^2 \Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(h+d)}{\Gamma(h-d+1)} \sim \frac{\sigma_\varepsilon^2 \Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} h^{2d-1}.$$

Fractional differences exist also for $d < 0$: In this case the spectrum has a zero at $\nu = 0$. Combined with integer differences we can therefore define the fractionally differenced process for any d: It is stationary for $d < 1/2$, whereas for $d \geq 1/2$ the $[d+1/2]$-th difference is stationary. Finally, instead of assuming that $(1-B)^d X_t$ is white noise, we can assume that this is a stationary, causal and invertible ARMA process:

$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t.$$

This is the fractional ARIMA(p,d,q) model.

The best way to decide whether long range dependence is present, is to look at the periodogram in log-log scale, that is we plot $\log(I_n(k/n))$ versus $\log(k/n)$ for $k = 1, 2, \ldots n/2$. If these points scatter around a line with negative slope, this indicates long range dependence. Moreover, we can estimate d as $-\frac{1}{2}$ times the slope. The approximate independence of periodogram values still holds. In fact, we can obtain estimates of the parameters of a fractional ARIMA model

by treating the $I_n(k/n)$ as independent exponential$(1/s(k/n))$-random variables and using maximum likelihood. We thus maximise

$$-\sum_{k=1}^{n/2}\left(\log s(k/n) + \frac{I_n(k/n)}{s(k/n)}\right)$$

with respect to the unknown parameters which appear in the spectrum $s(\nu)$ of the model. It is asymptotically equivalent to the exact Gaussian MLE, but much easier to compute. It also turns out that the estimator of d converges with rate $n^{-1/2}$, despite the long rnage dependence.

5.3 State space models

Due to lack of time, this topic is not covered in this course.

5.4 Nonlinear parametric models

I intended to discuss threshold autoregressions, ARCH and GARCH models. Due to lack of time, I could not cover it.