Learning material for *Time series analysis seminar.*
2020. május 5.

Fundamentals of time series theory

Definition. **Stochastic process:** $(X_t)_{t\in T}$, where T is the parameter space and for each t, X_t is a random variable.
On this course *usually* $T=\mathbb{Z}$ or $T=\mathbb{Z}_+$, in stochastic analysis theory $T=\mathbb{R}$ or $T=\mathbb{R}_+$.

Definition. **Gaussian process:** a stochastic process for which every finite marginal distribution is Gaussian, i.e. for every $n\in\mathbb{Z}_+$, $t_1\in T,\dots,t_n\in T$, $(X_{t_1},\dots,X_{t_n})$ is jointly Gaussian distributed.

Time series: A stochastic process, when we consider the parameter space T as '*time*'.

Definition. **Strict stationarity.** $(X_t)_{t\in T}$ is strictly stationary, if for each $n\in\mathbb{Z}_+$, $t_1\in T,\dots,t_n\in T$ and $h\in T$, $(X_{t_1},\dots,X_{t_n})$ has the same joint distribution as its h-shifted version, i.e. $(X_{t_1+h},\dots,X_{t_n+h})$.

Definition. **Autocovariance function:** $R(t,s)=\operatorname{cov}(X_t,X_s)$

Definition. **Weak stationarity.** $(X_t)_{t\in T}$ is weakly stationary, if EX_t does not depend on t (i.e. it is constant), and the autocovariance function $R(t,s)$ depends only on the $t-s$ differences.
Therefore the autocovariance function of a weakly stationary time series is only the function of 1 variable, this function will be denoted also by R: $R(t,s)=R(t-s)$. So the autocovariance function of a weakly stationary function can be computed as $R(h)=\operatorname{cov}(X_{t+h},X_t)$.

Comment. Strict stationarity does not follow from weak stationarity and weak does not follow from strict, either (it can happen that the process does not have finite moments).

Comment. The word *stationarity* means stability, invariance/permanence in time.

Comment. Stationarity is a popular assumption on empirical time series, however, it is really hard to check whether an empirical time series is stationary. If there is apparent trend or seasonality in a time series, then it is not stationary.

In the following we assume that the time series is weakly stationary and its parameter space is discrete.

Proposition. $R(0)=D^2X_t$ for all t.

Definition. **Autocorrelation function (ACF):** $r(h)=\operatorname{cor}(X_t,X_{t+h})$, $h\in\mathbb{Z}$.

Proposition. $r(h)=\frac{R(h)}{R(0)}$

Definition. **Partial autocorrelation function (PACF):**
$\rho(h)=\operatorname{cor}(X_t,X_{t+h}\,|\,X_{t+1},X_{t+2},\dots,X_{t+h-1})$, $h\in\mathbb{Z}$.

Notation.

- $r_h:=r(h)$
- $R_h:=\begin{bmatrix}1 & r_1 & r_2 & \dots & r_{h-1}\\ r_1 & 1 & r_1 & \dots & r_{h-2}\\ r_2 & r_1 & 1 & \dots & r_{h-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ r_{h-1} & r_{h-2} & r_{h-3} & \dots & 1\end{bmatrix}$
- $R_h^{\#}$: we change the last row of the R_h matrix above to $r_1,\dots,r_h$.

Proposition. $\rho(h)=\begin{cases}r_1 & \text{if } |h|=1\\ \frac{\det R_h^{\#}}{\det R_h} & \text{if } |h|>1\end{cases}$

Definition. **White independent noise process:** $\varepsilon_t\sim i.i.d.(0,\sigma^2)$, if $E\varepsilon_t=0$, $D^2\varepsilon_t=\sigma^2$, furthermore ε_t and ε_s are independent for each $t\neq s$.

Definition. **White noise process:**
$\varepsilon_t\sim WN(0,\sigma^2)$, ha $E\varepsilon_t=0$, $D^2\varepsilon_t=\sigma^2$ and $\operatorname{cov}(\varepsilon_t,\varepsilon_s)=0$ for each $t\neq s$.

Comment. It is often convenient to assume about a white noise to be a Gaussian process as well, in this case we call it a Gaussian white noise (GWN).

Definition. **Linear process:**
$X_t=\mu+\sum_{j=-\infty}^{\infty}\beta_j\varepsilon_{t-j}$, where $\mu\in\mathbb{R}$, $\varepsilon_t\sim WN(0,\sigma^2)$ and $\sum_{j=-\infty}^{\infty}|\beta_j|<\infty$.

Comment. In the definition of the linear process, ε_t noise process is often called *innovation*. This naming is used in case of other processes as well.

Proposition. The linear process is stationary.

Definition. **MA(∞) process:** a linear process that does not depend on the future values of the noise, i.e. $\beta_{-1}=\beta_{-2}=\ldots=0$.

Comment. MA=moving average.

Comment. If there is an $x_1,\dots,x_n$ empirical sample on hand, then taking moving average may be a good smoothing method (not the only one) to remove the non-stationary components from a time series. For instance, if we have quarterly data, then it is worth calculating a 4-point moving average: $\widetilde{x}_t=\frac{x_t+x_{t+1}+x_{t+2}+x_{t+3}}{4}$, $t=1,2,\dots,n-3$. With taking the moving average, the time series will be shorter, hence we lose data.

Results from the theory of the autocorrelation's estimation

Let $X_1,\dots,X_n$ be a sample from a stationary process with expected value μ, autocovariance function $R(h)$ and autocorrelation function $r(h)$. Let us consider the following estimates:

- estimate of μ: $\overline{X}_n = \frac{X_1+\ldots+X_n}{n}$;
- estimate of $R(h)$: $\widehat{R}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}\left(X_{t+|h|}-\overline{X}_n\right)\left(X_t-\overline{X}_n\right)$, where $|h|<n$ integer;
- estimate of $r(h)$: $\widehat{r}(h) = \frac{\widehat{R}(h)}{\widehat{R}(0)}$, where $h=-(n-1),\ldots,-1,0,\ldots,n-1$.

Comment. We can choose between several statistics to estimate the autocovariance function, see the 4th part of László Márkus's notes, which also details its statistical properties.

Let us use the following notations:
- $\mathbf{r}(h) = (r(1),\ldots,r(h))^T$
- $\widehat{\mathbf{r}}(h) = (\hat{r}(1),\ldots,\widehat{r}(h))^T$
- $\mathbf{0_h}$: vector of length h containing 0 values
- $\mathbf{I}_h$: $h\times h$ sized unit matrix

Theorem. **Bartlett theorem.** Let X_t be a linear process, $\sum_{j=-\infty}^{\infty}\beta_j^2|j|<\infty$. Then $\left(\frac{1}{n}\mathbf{W}\right)^{-\frac{1}{2}}\left(\hat{\mathbf{r}}(h)-\mathbf{r}(h)\right)\xrightarrow[n\to\infty]{d} N_h\left(\mathbf{0_h},\mathbf{I}_h\right)$, where $\mathbf{W}$ is the so-called Bartlett matrix whose elements are $[\mathbf{W}]_{i,j} = \sum_{k=-\infty}^{\infty}[r(k+i)r(k+j)+r(k-i)r(k+j)+2r(i)r(j)r^2(k)-2r(i)r(k)r(k+j)-2r(j)r(k)r(k+i)]$.

Comment. Writing the result of the previous theorem much simpler but much less precisely: $\hat{\mathbf{r}}(h)\approx N_h\left(\mathbf{r}(h),\frac{1}{n}\mathbf{W}\right)$.

Corollary. If $X_1,\ldots,X_n$ is a sample from a white noise, then $\hat{\mathbf{r}}(h)\approx N_h\left(\mathbf{0}_h,\frac{1}{n}\right)$, therefore we can construct common confidence intervals with confidence level $1-\alpha$ for each autocorrelation: $0\pm\frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}$.

Comment. If in the previous corollary $\alpha=0.05$, then we get approximately the interval estimation $\left[-\frac{2}{\sqrt{n}};\frac{2}{\sqrt{n}}\right]$. When we calculate the autocorrelation function of a time series with the help of the **R** function `acf`, then the two blue dashed lines correspond to the values $\pm\frac{2}{\sqrt{n}}$. Now we can describe the statistical test that we needed all the preparation this far for – let α be the fixed value of the type I. error's probability:

H_0: the sample stems from an unautocorrelated process
H_1: at least 1 autocorrelation value is not 0

Test statistic: $\mathbf{T}(\mathbf{X}) = \hat{\mathbf{r}}(n-1)$, which is a vector of $n-1$ length

Acceptance region: $\mathcal{X}_e = \left\{\mathbf{X}:\ \text{minden } 1\le i\le n-1\text{-re } |\hat{r}_i|\le\frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}\right\}$

So if we find that any of the empirical autocorrelations are bigger than $\frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}$ in absolute value, then with $100\cdot(1-\alpha)\%$ confidence, the sample does not stem from an unautocorrelated process. Usually we can say even more, the autocorrelation values outside the inner "strip" refer to the type of the non-stationarity and to the time series model we should try to explain the data with.

Goodness-of-fit (GoF) testing of time series

The final goal of the modeller is to choose the best fitting model to the empirical sample (now: empirical time series) on hand. This is often a pretty challenging task. The vast majority of time series models originates from a white noise process (that is also called innovation process and usually we think about it as the series of observational, measurement errors or the sum of miscellaneous random factors). We examine the goodness-of-fit of a given time series model by first estimating the unknown parameters (fitting), then we test the innovations, whether they stem from a white noise process or not. If we *cannot* reject that the estimated innovations, also called residuals originate from a white noise, then we do *not* have statistical evidence against the usage of the given time series model.

What features can ruin a time series to be a white noise:
- autocorrelation – the residuals are autocorrelated
- heteroscedasticity – the variance of the residuals changes with time ($\longleftrightarrow$ homoskedasticity: the variance is constant with time)

Tests that check whether a given $X_1,\ldots,X_n$ sample stems from an i.i.d. white noise:

a.) The empirical autocorrelation function is equal to 0 *in each point*. For the description of this test, see the previous section.
When we calculate the autocorrelation function of a time series with the help of the **R** function `acf`, then the two blue dashed lines correspond to the values $\pm\frac{2}{\sqrt{n}}$.
Our decision is that the sample does NOT stem from a white noise in the following cases:
- one of the autocorrelation values is very outlying, it is "strongly" outside of the blue strip. What does "strongly" mean: the absolute value of the given outstanding autocorrelation value is at least as much as the length of the strip.
- although there is no strongly outlying autocorrelation value, but there are relatively too many values outside of the blue strip. Rule of thumb what is considered to be relatively too many: if we have m pieces of autocorrelation values, then more than $0,05\cdot m$ autocorrelation values can be found outside of the blue strip. The basic option in **R** is $m=40$, given this, if at least 3 values are a little bit outside of the blue strip, then we reject the null hypothesis and say that the sample does not originate from a white noise.

Usually we can say even more, the autocorrelation values outside the inner "strip" refer to the type of the non-stationarity and to the time series model we should try to explain the data with.

b.) Portmanteau tests. Nullhypothesis: the first couple (usually 20/30) of autocorrelation values are *all* equal to 0. Hence if H_0 is rejected, then we can assert that the sample does NOT originate from a white noise. The most basic among them is the *Ljung-Box test*. Test statistic: $Q_{LB} = n(n+2)\sum_{i=1}^{h}\frac{\hat{r}(i)}{n-i}$, which tends to χ^2_h in distribution in case of H_0.

c.) *Difference-sign test.* Let Z_n be the number of time point when the process is

increasing, i.e. $Z_n = \sum_{i=2}^{n} Y_i$, where $Y_i = I(X_i > X_{i-1})$, $i = 2, 3, \ldots, n$. test statistic: $\frac{Z_n - \frac{n-1}{2}}{\sqrt{\frac{n-1}{12}}}$, that converges in distribution to the standard Gaussian distribution in case of the null hypothesis.

d.) *Turning point test.* Let Z_n be the number of turning points in the sample. Sample point x_i is a turning point, if either ($x_i > x_{i-1}$ and $x_i > x_{i+1}$), or ($x_i < x_{i-1}$ and $x_i < x_{i+1}$). It can be seen that the test statistic $\frac{Z_n - \frac{2}{3}(n-2)}{\sqrt{(n-2)\frac{8}{45}}}$ tends to the standard Gaussian distribution in case of the null hypothesis.

Heteroskedasticity test for time series: *Mcleod Li test.*

Definition. m**-dependence.** The series of random variables $(X_n)_{n\in\mathbb{Z}}$ is m-dependent, if $\sigma(\ldots, X_{k-1}, X_k)$ and $\sigma(X_{k+m+1}, X_{k+m+2}, \ldots)$ σ-algebras are independent for each $k \in \mathbb{Z}$.

Comment. The MA(p) process without a unit root is p-dependent.

Theorem. **Central limit theorem for m-dependent variables.**
Let $X_1, X_2, \ldots,$ be m-dependent, weakly stationary random variables,

$\sigma_m^2 := D^2 X_1 + 2\sum_{k=1}^{m} \operatorname{cov}(X_1, X_{1+k})$. Then $\frac{\sum_{i=1}^{n}(X_i - EX_i)}{\sqrt{n}\sigma_m} \xrightarrow[n\to\infty]{d} N(0;1)$.

ARMA processes

Definition. **Backshift operator.**
$B : L_2(\Omega) \longrightarrow L_2(\Omega), \quad BX_t = X_{t-1}$

Proposition. The operator B above is

- iterable: $B^k X_t = X_{t-k}$, $k \in \mathbb{Z}_+$;
- invertible: $B^{-1} X_t = B_{t+1}$;
- unitary.

Definition. **ARMA process.** $X_t = \underbrace{\sum_{i=1}^{p} \alpha_i X_{t-i}}_{\text{AR}(p)\text{ part}} + \underbrace{\sum_{j=0}^{q} \beta_j \varepsilon_{t-j}}_{\text{MA}(q)\text{ part}}$, where p, q positive integer (*orders*); $\alpha_i, \beta_j \geq 0$ real numbers; $\varepsilon_t \sim WN(0, \sigma^2)$.

Comment. ARMA: autregressive moving average

Comment. Another form or the ARMA process: $\sum_{i=0}^{p} \tilde{\alpha}_i X_{t-i} = \sum_{j=0}^{q} \beta_j \varepsilon_{t-j}$, where $\tilde{\alpha}_0 = 1$ és $\tilde{\alpha}_i = -\alpha_i$ $(i = 1, \ldots, p)$.

Definition. **Causality.** The ARMA(p, q) process is causal, if there exist ψ_i, $i \in \mathbb{Z}$ constants, for which $\sum_{i=1}^{\infty} |\psi_i| < \infty$ and $X_t = \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i}$ $\forall t$.

Comment. If the process is causal then it means that:

- the ARMA "equation" has a "solution";
- the ARMA process has an MA(∞) form.

Definition. **Invertability.** The ARMA(p, q) process is invertible, if there exist π_i, $i \in \mathbb{Z}$ constants, for which $\sum_{i=1}^{\infty} |\pi_i| < \infty$ and $\varepsilon_t = \sum_{i=1}^{\infty} \pi_i X_{t-i}$ $\forall t$.

Comment. If the ARMA process is invertible, we usually say that is has an AR(∞) form.

Definition. **Characteristic polynomials of the ARMA process:**
$P(x) = \sum_{i=0}^{p} \tilde{\alpha}_i x^{p-i}, \quad \widetilde{P}(x) = \sum_{i=0}^{p} \tilde{\alpha}_i x^i, \quad Q(x) = \sum_{i=0}^{q} \beta_i x^i.$

Proposition. $P(x) = x^p \cdot \widetilde{P}\left(\frac{1}{x}\right)$

Proposition. The roots of $P(x)$ lie inside the unit circle $\iff$ the roots of $\widetilde{P}(x)$ lie outside the unit circle

Theorem. **The "solution" and the "inverse" of the ARMA process.**

- If the roots of $P(x)$ lie inside the unit circle, then the process has a strictly stationary solution;
- If the roots of $Q(x)$ lie outside the unit circle, then the process is invertible.

Corollary. The AR(1) process $X_t = \alpha X_{t-1} + \varepsilon_t$ is stationary $\iff$ $|\alpha| < 1$

Theorem. **Spectral density of the ARMA process:** $f(x) = \frac{\sigma^2}{2\pi}\left|\frac{Q(e^{ix})}{\widetilde{P}(e^{ix})}\right|^2$

Using our notations introduced this far, the ARMA process can be written in the following operator form: $\widetilde{P}(B)X_t = Q(B)\varepsilon_t$. Now we can express the process and the white noise with the other one:

- $X_t = \left[\widetilde{P}(B)\right]^{-1} Q(B)\varepsilon_t$
- $\varepsilon_t = [Q(B)]^{-1} \widetilde{P}(B) X_t$

It is absolutely not trivial how to express the inverse of these operators above and how to create the multiplication of the operators. On this point, we call the help of the theory of the homogenous linear difference equations. Let us give the explicit solution of X_t above, the other would go similarly. Let us denote by $S(x) = \sum_{i=0}^{\infty} s_i x^i$ the polynomial, for which $S(B) = \left[\widetilde{P}(B)\right]^{-1} Q(B)$. Here the inverse can be calculated by looking for the $T(x) = \sum_{i=0}^{\infty} t_i x^i$ polynomial, for which $T(x)\widetilde{P}(x) = 1$. On this point we face the problem that we have a recursion (difference equation) for the s_i coefficients that we should be solved. Let us consider the following homogeneous difference equation:

$$s_t + \gamma_1 s_{t-1} + \ldots + \gamma_k s_{t-k} = 0,$$

where

- $k \le t \in \mathbb{Z}$;
- $\gamma_1, \dots, \gamma_k \in \mathbb{Z}$; $\gamma_i \neq 0 \ \forall i$;
- $s_0, s_1, \dots, s_{k-1}$ initial values are given.

On the seminar we will only solve 2nd order difference equations. The following propositions describe the solution of these.

Proposition. The general solution of the first order homogeneous linear difference equation $s_t + \gamma_1 s_{t-1} = (1 - \xi^{-1}B)s_t = 0$ is $s_t = s_0 \xi^{-t}$.

Proposition. The general solution of the second order homogeneous linear difference equation $s_t + \gamma_1 s_{t-1} + \gamma_2 s_{t-2} = (1 - \xi_1^{-1}B)(1 - \xi_2^{-1}B)s_t = 0$ is:

case 1: $\xi_1 \neq \xi_2$ are real $\rightsquigarrow$ $s_t = c_1\xi_1^{-t} + c_2\xi_2^{-t}$, where c_1 és c_2 are real constants that can be derived using the initial values, solving the following system of equations:
$\begin{cases} c_1 + c_2 & = s_0 \\ c_1\xi_1^{-1} + c_2\xi_2^{-1} & = s_1 \end{cases}$ (similarly for the two other cases).

case 2: $\xi_1 = \xi_2$ are real $\rightsquigarrow$ $s_t = (c_1 + c_2 t)\xi_1^{-t}$, where c_1 and c_2 are real constants that can be derived using the initial values.

case 3: $\xi_1 \neq \xi_2$ are complex $\rightsquigarrow$ $s_t = c_1\xi_1^{-t} + c_2\xi_2^{-t}$, where c_1 and c_2 are complex constants that can be derived using the initial values. The general solution can also be written in a different form. We know from algebra that $\xi_2 = \overline{\xi}_1$. Using the exponential form $\xi_1 = de^{\mathrm{i}\theta}$, we get $s_t = ad^{-t}\cos(\theta t + b)$, where a and b are real constants that can be derived with the help of the initial values.

The basic properties of ARMA processes (useful to choose the best ARMA model):

Model	$R(h)$ autocorrelation function	$r(h)$ partial autocorr. function
AR(p)	tends to 0, if $\lvert h\rvert \to \infty$	not 0, if $\lvert h\rvert \le p$; otherwise 0
MA(q)	not 0, if $\lvert h\rvert \le q$; otherwise 0	tends to 0, if $\lvert h\rvert \to \infty$
ARMA(p,q)	tends to 0, if $\lvert h\rvert \to \infty$, convergence starts after the first q values	tends to 0, if $\lvert h\rvert \to \infty$, convergence starts after the first p values

Model choice: choosing the "best" ARMA(p,q) process. Steps:

1. Choosing orders p and q.
   The following information criteria can help us:
   - **Akaike information criterion**: $AIC = 2k - 2\log\widehat{L}$, where
     - k: number of parameters to estimate
     - $\widehat{L}$ value of the likelihood function if the ML estimate is used (in case of Gaussian errors it concurs with the least square estimate)

     The smaller, the better.
   - **Bayes information criterion**: $BIC = \log n \cdot k - 2\log\widehat{L} \rightsquigarrow$ the smaller the better

   Looking at the empirical autocovariance function, the table above can help us in finding the proper orders. Decisions based on the information criteria are generally better.
2. Estimation of the unknown parameters.
   We can apply the classical methods: momemtum method (Yule-Walker equations derived from the autocovariances), ML method, least squares. The ML method is usually the best, however, as it looks for a numerical maximum, it is the slowest as well.
   After estimation, we have to check whether the coefficients are significant. Every statistical software gives the parameter's point estimation $\hat{\vartheta}$ and its standard error $\widehat{D(\vartheta)}$. Rule of thumb rule for our parameter ϑ of interest to be significant, i.e. to be different from zero: is 0 included in the interval $\hat{\vartheta} \pm 2 \cdot \widehat{D(\vartheta)}$. If 0 is not included in the interval, then the parameter is significant.

Model diagnostic: after having fitted a model we have to check whether the assumptions in the definition of the time series hold. In case of ARMA models this means we have to check whether the residuals may stem from a white noise process. Usually the following tests are carried out:

- Investigating the values outside the strip in the empirical autocorrelation function
- Ljung-Box test or a generalisation of it
- Heteroskedasticity test – if we get that the time series is not homoscedastic, we fit a GARCH model to the residuals

Comment. The term 'model diagnostic' can also include the checking of the coefficents' significance.

If we have come to the conclusion that the goodness-of-fit of the chosen ARMA process is poor, we can do the following:

- Is there further trend/seasonality in the time series (we have to start the analysis of time series with eliminating these, for further detail see section 'Non-stationary time series' later)
- We try to fit another time series family

At last we state one of the most important theorems of time series theory, namely why is the MA(∞) model family so important. The notions 'predictable process' and 'deterministic' will not be defined here because they would need more preparation, see for example section 2.6 of the Brockwell–Davis book and the sections before for further details.

Theorem. **Wold decomposition.** Every X_t stationary process can be written in the following form: $X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} + V_t$, where

- $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$;
- $\varepsilon_t \sim WN(0, \sigma^2)$;
- V_t 'deterministic' process;
- $\mathrm{cov}(\varepsilon_t, V_s) = 0 \ \ \forall s, t$-re;
- ε_t és V_t 'predictable processes' .

GARCH processes

Definition. **GARCH process.** $(X_t)_{t\in\mathbb{Z}}$ GARCH(p,q) process, if $X_t = \sigma_t\varepsilon_t$, where $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2$ is the time-dependent variance of the process, $\varepsilon_t \sim i.i.d.(0;1)$, $E\varepsilon_t^4 < \infty$, furthermore $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for each i and j.

Specifically, if $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2$, then X_t is called an ARCH(p) process.

Comment. A process can be written in the $X_t = \sqrt{\alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2}\cdot\varepsilon_t$ form.

Comment. The name of the process: GARCH = generalised autoregressive conditional heteroskedastic.

Proposition. The GARCH(p,q) process is a (NOT independent) white noise.

Theorem. **The weak stationarity of the GARCH process.**

- If $\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$, then the GARCH($p$,$q$) process has a weak stationary solution ("solution": MA(∞) form of the X_t process).
- If the GARCH(p,q) process has a weakly stationary solution and $\alpha_0 > 0$, then $\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$.

Characteristics of an empirical sample from a GARCH process:

- The data are not correlated, the standard deviance is not constant in time;
- The distribution of the data is heavy/fat tailed;
- The squares and the absolute values of the sample are strongly autocorrelated;
- The outlying values appear in clusters.

In case we experience the features above for an empirical sample, it is worth considering modelling with a GARCH process. These characteristics are really common in financial time series (i.e. at log-returns of stock prices, exchange rates).

Comment. If we intend to model our observations with some heteroscedastic process, then the relatively easy GARCH(1,1) process is often a really good choice.

Forecasting of time series

Forecasting: let $(X_n)_{n\in\mathbb{Z}}$ a stationary process with μ expected value and $R(h)$ autocovariance function. Our goal is the following: using the first n values $X_n,\dots,X_1$ of the process, let us give the estimation for the following h values $X_{n+h},\dots,X_{n+1}$ so that the mean square error is minimised.

Let us denote the best linear predictor by $\mathbb{P}_nX_{n+h} := a_0 + a_1X_n + \ldots + a_nX_1$.

Theorem. **Properties of the best linear predictor operator.**

a.) The coefficients of the best linear predictor satisfies the following equations:
$E(X_{n+h} - \mathbb{P}_nX_{n+h}) = 0$
$E\left[(X_{n+h} - \mathbb{P}_nX_{n+h})X_{n+1-j}\right] = 0 \quad j = 1,\dots,n$

b.) The coefficients of the best linear predictor satisfies the following system of equations: $\mathbf{\Gamma}_n\mathbf{a}_n = \mathbf{R}_n(h)$, where
- $\mathbf{\Gamma}_n = [R(i-j)]_{i,j=1}^n$;
- $\mathbf{a}_n = (a_1,\dots,a_n)^T$;
- $\mathbf{R}_n(h) = (R(h), R(h+1),\dots,R(h+n-1))^T$.

c.) The forecast can be written in the form $\mathbb{P}_nX_{n+h} = \mu + \sum_{i=1}^{n}a_i(X_{n+1-i}-\mu)$, therefore at the beginning of the exercises we can assume that $\mu = 0$, then using this formula, we can get the forecast;

d.) the mean square error of the forecast is $R(0) - \mathbf{a}_n^T\mathbf{R}_n(h)$.

Proposition. If $R(0) > 0$ and $R(h) \overset{h\to\infty}{\longrightarrow} 0$, then the matrix $\mathbf{\Gamma}_n$ above is invertible, so the a_i coefficients can be calculated the following way: $\mathbf{a}_n = \mathbf{\Gamma}_n^{-1}\mathbf{R}_n(h)$.

Generalised prediction operator.

Let $\mathbf{W} = (W_n,\dots,W_1)^T$ be a random vector, Y a random variable, both with finite standard deviation/covariance matrix. Let us denote by $\mathbf{\Gamma} = \mathrm{cov}(\mathbf{W}) = \Sigma(\mathbf{W})$ the covariance matrix. Our goal is to get the best least square linear estimate of Y with the help of $\mathbf{W}$.

Let us denote the forecast operator by $\mathbb{P}(\bullet|\mathbf{W}) : L_2(\Omega) \longrightarrow L_2(\Omega)$, we are looking for a linear estimate: $\mathbb{P}(Y|\mathbf{W}) = a_0 + \mathbf{a}_n^T\mathbf{W}$, where a_0 and $\mathbf{a}_n = (a_1,\dots,a_n)^T$ are the unknown values to estimate.

Theorem. **Properties of the generalised best linear predictor operator.**

Let $\alpha_1,\dots,\alpha_n,\beta$ arbitrary real numbers and Z an arbitrary random variable with finite variance. Then

a.) $\mathbb{P}(Y|\mathbf{W}) = EY + \mathbf{a}_n^T(\mathbf{W} - E\mathbf{W})$, where $\mathbf{\Gamma}\mathbf{a}_n = \mathrm{cov}(Y,\mathbf{W})$;

b.) $E(Y - \mathbb{P}(Y|\mathbf{W}) = 0$
$E[(Y - \mathbb{P}(Y|\mathbf{W}))\mathbf{W}] = 0$;

c.) $MSE = E[(Y - \mathbb{P}(Y|\mathbf{W}))^2] = D^2Y - \mathbf{a}_n^T\mathrm{cov}(Y,\mathbf{W})$;

d.) $\mathbb{P}(\alpha_1Y + \alpha_2Z + \beta|\mathbf{W}) = \alpha_1\mathbb{P}(Y|\mathbf{W}) + \alpha_2\mathbb{P}(Z|\mathbf{W}) + \beta$;

e.) $\mathbb{P}\left(\sum_{i=1}^{n}\alpha_iW_i + \beta\middle|\mathbf{W}\right) = \sum_{i=1}^{n}\alpha_iW_i + \beta$;

f.) ha $\mathrm{cov}(Y,\mathbf{W}) = \mathbf{0}$, akkor $\mathbb{P}(Y|\mathbf{W}) = EY$.

Comment. The connection between $\mathbb{P}$ generalised prediction operator and $\mathbb{P}_n$ operator: $\mathbb{P}_n(X_{n+h}) = \mathbb{P}(X_{n+h}|\mathbf{X}_n)$, where $\mathbf{X}_n = (X_n,\dots,X_1)^T$.

Comment. From part a.) of the theorem above we can conclude that knowing $\mathbf{a}_n^T$, we can determine the constant a_0: $a_0 = EY - \mathbf{a}_n^TEW$. In the special case if we are working with operator $\mathbb{P}_n$, then $a_0 = \mu\cdot\left(1 - \sum_{i=1}^{n}a_i\right)$.

Comment. From part b.) of the theorem above we can conclude that the best least square linear estimate is unbiased.

Comment. With the help of the generalised forecast operator we are able to estimate the values of missing of potentially wrong data.

Modelling non-stationary time series

Definition. **Lag-1 difference operator.** $\nabla = 1 - B$, where is B the lag operator.
Definition. **Lag-d difference operator.** $\nabla_d = 1 - B^d$.

Using the above, $\nabla X_t = X_t - X_{t-1}$ and $\nabla_d X_t = X_t - X_{t-d}$. We can take the arbitrary positive integer exponent of the difference operator, then $\nabla^m X_t = \nabla^{m-1}(X_t - X_{t-1})$, which can be iterated further.

Definition. **ARIMA model.** Process X_t is an ARIMA(p, d, q) process, if $Y_t = (1 - B)^d X_t$ process is an ARMA(p, q) model.

Comment. Character I in the acronym ARIMA is the initial of the word 'Integrated'.

In case of ARIMA models the so-called **unit root tests** are the tools of a statistician to find the best value for parameter d. If the characteristic polynomial $P(x)$ of the model has a unit root (there exists a root with absolute value 1), then the process is not stationary. The unit root tests are looking for such non-stationarity that can be eliminated by taking differences.
The most often applied unit root test is the **augmented Dickey–Fuller test**. To understand how it works, first let us describe its simplest version.
Let $X_t = \alpha X_{t-1} + \varepsilon_t$, $\varepsilon_t \sim GWN(0, \sigma^2)$ be an AR(1) process. It is known that if $\alpha = 1$, then this the well-known random walk that is not stationary and $X_t = \sum_{i=1}^{t} \varepsilon_i$; in case of $|\alpha| < 1$ it is stationary. The AR(1) equation can be written in the $\nabla X_t = \phi X_{t-1} + \varepsilon_t$ form, where $\phi = \alpha - 1$. The hypothesis of the test: $H_0 : \alpha = 1$ and $H_1 : |\alpha| < 1$, so the alternative hypothesis asserts that the process is stationary; while the zero hypothesis states that although the process is NOT stationary, but by taking the difference, it can be made stationary. To conduct the test, we regress ∇X_t on X_{t-1}, the test statistic is the standardised regression coefficient of X_{t-1}: $\frac{\hat{\alpha}-1}{D(\hat{\alpha}-1)}$. Using Donsker's theorem it can be shown that in case of the null hypothesis this test statistic converges in distribution to a random variable which quantiles can only be computed via simulation or numerical approximation.

The augmented version of the test is valid for AR(p) processes and admits the presence of drift μ and trend ηt in the process: $X_t = \mu + \eta t + \sum_{j=1}^{p} \alpha_j X_{t-j} + \varepsilon_t$. Here the form applied for regression is $\nabla X_t = \mu + \eta t + \phi X_{t-1} + \sum_{j=1}^{p} \psi_j \nabla X_{t-j} + \varepsilon_t$, where $\phi = \sum_{j=1}^{p} \alpha_j - 1$ and $\psi_j = -\sum_{k=j}^{p} \alpha_k$. The non-stationarity is tested again with $H_0 : \phi = 0$, the test statistic is again the standardised estimate of ϕ, using the estimate of the regression coefficients.

When conducting a unit root test on a process, it is the responsibility of the modeler to determine the AR order p and whether we count on the presence of drift and/or trend.

Classical time series decomposition model: $X_t = m_t + s_t + Y_t$, where

- m_t: trend component – long term changes in a time series; often described by a linear or quadratic function of t.
- s_t: seasonal component – short term changes, deviations in the time series that are repeating regularly, have the same period and have a regular amplitude. If $d \in \mathbb{Z}$ denotes the length of the periods of seasonality, then $s_t = s_{t+d}$ for each t. We assume that $\sum_{i=1}^{d} s_i = 0$. In economic applications the value of d is 4 for quarterly data and 12 for monthly data.
- Y_t: random noise component, that is stationary and can be modeled by a time series model. We assume that $EY_t = 0$, otherwise the constant could be put into the trend.

To estimate and eliminate the trend, there are two approaches:

1. Estimation of the trend using a function
   - *Parametric*: We fit a proper function, usually as polynomial, i.e. the trend has the form $m_t = \sum_{i=0}^{p} c_k t^k$, where coefficients c_k can be estimated by the least squares method.
     The main disadvantage of this approach is that we assume the chosen function will describe the dynamic of the time series in the future, although a crisis or another sudden event can easily change the trend dramatically.
   - *Non-parametric*: the most common method is using a linear filter, less common is applying exponential smoothing. The usage of the LOESS smoothing, which is a local regression fitting, is becoming more and more popular.
     Linear filter: $F = \sum_{i=-\infty}^{\infty} a_i B^i$ operator (a_i values are real coefficients), which "smooths" the process – we take the averages of some consecutive values of the process to decrease the magnitudes and amounts of spikes and to make the trend visible. The linear filter is often used to estimate the trend: $m_t \approx FX_t = \sum_{i=-\infty}^{\infty} a_i X_{t-i}$. In practice, we choose most of the a_i coefficients equal to 0.
     A special type of linear filters are the so-called finite symmetric **moving averages**: $a_i = \frac{1}{1+2q}$, if $-q \leq i \leq q$ and $a_i = 0$, if $|i| > q$. Then $\hat{m}_t = \frac{1}{1+2q} \sum_{i=-q}^{q} X_{t-i}$.
     The main disadvantage of non-parametric approaches is that they smooth only locally, therefore it is not possible to use them *directly* for forecasting. In case we still need make forecasts, we have to try to find a function that fits reasonably well to the smoothed process – but it is absolutely not guaranteed

that we will succeed in finding such a function.

2. Elimination by differencing – we use the ∇ difference operator as many times as needed to eliminate the trend. For example having a linear trend at hand means using differencing only once is sufficient to eliminate the trend.

Definition. A linear filter F passes an arbitrary polynomial of degree k without distortion, if $m_t = \sum_{l=0}^{k} c_l t^l \quad \Longrightarrow \quad m_t = F m_t$ for each t.

Now we will look through the methods to detect and eliminate trend AND seasonal components as well:

1. Calculating season indices
   Let $x_1, x_2, \ldots, x_n$ be an empirical sample. First we eliminate the trend with a parametric or non-parametric method mentioned above, then we estimate the individual season indices from the deviations $x_t - \hat{m}_t$ via averaging. At last, we make a correction so that the sum of the season indices be 0. So the steps of the calculations:
   I. Estimating uncorrected seasonal indices (We create d sets where we put every dth deviation, then calculate the average of values in each set. For instance the first set contains the 1st,(d+1)th,(2d+1)th etc. deviations, the second set the 2nd,(d+2)th,(2d+2)th etc. deviations):
   $\tilde{s}_k = \frac{\sum_{i:\ 1\le k+id\le n} (x_{k+id} - \hat{m}_{k+id})}{\sum_{i:\ 1\le k+id\le n} 1}, \quad k = 1, 2, \ldots, d$
   II. Correction of individual seasonal indices: $\hat{s}_k = \tilde{s}_k - \frac{1}{d}\sum_{i=1}^{d} \tilde{s}_i, \quad k = 1, 2, \ldots, d$
2. Differencing – we apply the difference operator ∇ on the process as many times as needed to eliminate the trend. After that, we check whether there has still remained any seasonality in the residuals, and if yes, then using a properly chosen d, we apply the ∇_d difference operator. The graph of the remainder process and the ACF/PACF functions can help us in chosing the best d value.
3. Filtering – applying a well chosen linear filter, we can eliminate trend and seasonal components at once. There exist such a filter in case of polynomial trend.

Let us summarize our knowledge this far! If we intend to model our empirical time series featuring strong temporal dependence in the time domain (in opposition with the frequency domain), it is advised to follow the following steps.

Main steps of time series modeling (Box-Jenkins modelling):

1. Depict the empirical time series with a line chart
   - after first glance, partition the sample to homogenous parts that seem to be looking the same (and the sample size is not too small)
   - handle suspicious/missing values: delete/correct/leave them alone
2. Transformation of the data if needed, e.g. if there is clear exponential trend, it is worth taking the logarithm.
3. Eliminating the trend component
4. Eliminating the seasonal component
5. Model fitting – steps:
   (i) Finding the proper model family (on this practice we only learned about ARMA/ARIMA processes)
   (ii) finding the best model (e.g. ARMA(1,1)) based on an information criterion
   (iii) parameter estimation via a properly chosen method (moment / ML / least squares estimation)
6. Model diagnostics – steps:
   (i) are the coefficients significant?
   (ii) may the estimated residuals stem from a white noise process?
   (iii) are the residuals heteroscedastic? If yes, then we fit a GARCH model on them.
   If the model fits well, we can go on, otherwise let us go to step 5 or step 2.
7. Forecasting: usually this is the ultimate goal, we intend to give good forecasts for the future behaviour of our time series with the help of the past values.

Several financial time series can be modeled well with a properly chosen element of the model family defined below.

Definition. **SARIMA (seasonal ARIMA) model.** X_t process follows a SARIMA$(p,d,q)\times(P,D,Q)_s$ process with s period, if $Y_t = (1-B)^d(1-B^s)^D X_t$ process is a seasonal ARMA(p,q) process, i.e. it can be written in the following form: $\widetilde{P}(B)\widetilde{R}(B^s)Y_t = Q(B)S(B^s)\varepsilon_t$, where

- $\widetilde{P}(x) = 1 - \alpha_1 x - \ldots - \alpha_p x^p$
- $\widetilde{R}(x) = 1 - \gamma_1 x - \ldots - \gamma_P x^P$
- $Q(x) = 1 + \beta_1 x + \ldots + \beta_q x^q$
- $S(x) = 1 + \delta_1 x + \ldots + \delta_Q x^Q$

Spectral theory of time series

Definition. Let X_t a stationary process, $EX_t = 0$, $\sum_{h=-\infty}^{\infty} |R(h)| < \infty$. Then the **spectral density** of X_t is $f(x) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ihx} R(h), \quad -\infty < x < \infty$.

Comment. It is enough to look at it at interval $[-\pi;\pi]$, because it is periodic.

Proposition. Properties of the spectral density:

- f is even, i.e. $f(-x) = f(x) \quad \forall x \in \mathbb{R}$
- $f(x) \ge 0 \quad \forall x \in \mathbb{R}$
- $R(h) = \int_{-\pi}^{\pi} e^{ihx} f(x)\, dx$

Comment. The spectral density function is (with 1 probability) unique.

Comment. The Fourier coefficients of f are the autocovariances $R(h)$.

Proposition. Function $f : [-\pi;\pi] \longrightarrow \mathbb{R}$ is the spectral density of a stationary process $\iff$

- f is even;
- $f \ge 0$;
- $\int_{-\pi}^{\pi} f(x)\, dx < \infty$.

Corollary.

The absolutely summable function $R(h)$ is the autocovariance function of a weakly stationary process $\iff$

- R is even;
- $f(x) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ihx} R(h) \geq 0$ for all $x \in [-\pi; \pi]$.

Definition. **Positive semidefinit function.** $g : \mathbb{Z} \to \mathbb{C}$ function is positive semidefinite, if matrix $[g(|t_i - t_j|)]_{i,j=1,\ldots,n}$ is positive semidefinite for each $n \in \mathbb{Z}_+$ and for each $t_1, \ldots, t_n \in \mathbb{Z}$.

Theorem. **The spectral representation of the autocovariance function.**

$R(h)$ function is the autocovariance function of a weadkly stationary process $\iff$ $\exists F : [-\pi; \pi] \longrightarrow \mathbb{R}$ right-continuous, non-decreasing, bounded, $F(-\pi) = 0$ function for which $R(h) = \int_{-\pi}^{\pi} e^{ihx}\, dF(x)$.

The name of F in the theorem above is **spectral distribution function**. If

- $\exists f : \; F(x) = \int_{-\pi}^{x} f(y)\, dy$, then we say that the process has *continuous spectrum* with spectral density function f;
- F is a jump function, then we say that the process has *discrete spectrum.*

Corollary. $F(\pi) = R(0) = D^2 X_t$, therefore F is NOT a probability distribution function, as it can occur that $F(\pi) \neq 1$!

In the following we will claim two general theorems.

Theorem. **Herglotz theorem.** Let $R : \mathbb{Z} \to \mathbb{C}$ be a positive semidefinite function. Then there exists a Q finite measure on interval $[-\pi; \pi]$ for which $R(h) = \int_{[-\pi;\pi]} e^{ihx}\, dQ(x)$.

In time series analysis we apply Herglotz theorem on the autocovariance function denoted by R, which is positive semidefinite. The spectrum Q in Herglotz theorem is called **spectral measure**. The spectral distribution function from this: $F(x) = Q([-\pi, x))$, furthermore in case Q is absolutely continuous with respect to the λ Lebesgue measure, then the $f = \frac{dQ}{d\lambda}$ Radon-Nikodym derivative is the spectral density function.

Comment. This theorem contains a Lebesgue integral, while in the theorem before there was a Riemann–Stieltjes integral.

A little outlook: not only the autocovariance function of a stochastic process does have a spectral representation, but the process itself as well. This is described in the following theorem.

Theorem. **Spectral representation of a stochastic process.**
Let $T = \mathbb{Z}$, X_k be a stationary process with 0 expectation. Then there exists a ϕ ortogonal stochastic measure for which $X_k = \int_{[-\pi;\pi]} e^{ik\mu}\, d\phi(\mu)$, $k \in \mathbb{Z}$.

In this theorem two things are not properly defined:

- what is a stochastic measure (the essence: assigns a random variable to a set) and when is it ortogonal;
- how shall we define an integral with respect to such a measure.
