Time Series Analysis

1. Stationary ARMA models

Andrew Lesniewski

Baruch College
New York

Spring 2017

Outline

Time series

- A *time series* is a sequence of data points X_t indexed a discrete set of (ordered) dates t, where $-\infty < t < \infty$.
- Each X_t can be a simple number or a complex multi-dimensional object (vector, matrix, higher dimensional array, or more general structure).
- We will be assuming that the times t are equally spaced throughout, and denote the time increment by h (e.g. second, day, month). Unless specified otherwise, we will be choosing the units of time so that $h = 1$.
- Typically, time series exhibit significant irregularities, which may have their origin either in the nature of the underlying quantity or imprecision in observation (or both).
- Examples of time series commonly encountered in finance include:
  - (i) prices,
  - (ii) returns,
  - (iii) index levels,
  - (iv) trading volums,
  - (v) open interests,
  - (vi) macroeconomic data (inflation, new payrolls, unemployment, GDP, housing prices, . . .)

Time series

- For modeling purposes, we assume that the elements of a time series are random variables on some underlying probability space.
- *Time series analysis* is a set of mathematical methodologies for analyzing observed time series, whose purpose is to extract useful characteristics of the data.
- These methodologies fall into two broad categories:
  - (i) *non-parametric*, where the stochastic law of the time series is not explicitly specified;
  - (ii) *parametric*, where the stochastic law of the time series is assumed to be given by a model with a finite (and preferably tractable) number of parameters.
- The results of time series analysis are used for various purposes such as
  - (i) data interpretation,
  - (ii) forecasting,
  - (iii) smoothing,
  - (iv) back filling, ...
- We begin with stationary time series.

Stationarity and ergodicity

- A time series (model) is *stationary*, if for any times $t_1 < \ldots < t_k$ and any τ the joint probability distribution of $(X_{t_1+\tau}, \ldots, X_{t_k+\tau})$ is identical with the joint probability distribution of $(X_{t_1}, \ldots, X_{t_k})$.
- A stationary time series model is *ergodic* if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{1\le k\le T} X_{t+k} = \mu, \tag{1}$$

  i.e. if the time average of X_t is equal to the (ensemble) average.
- Ergodicity is a desired property of a financial time series, as we are always faced with a single realization of a process rather than an ensemble of alternative outcomes.
- The limit in (1) is usually understood in the sense of squared mean convergence.
- The notions of stationarity and ergodicity are hard to verify in practice. Luckily, there is a more practical concept.

Autocovariance and stationarity

- A time series is *covariance-stationary* (a.k.a. *weakly stationary*), if:
  - (i) $\mathrm{E}(X_t) = \mu$ is a constant,
  - (ii) For any τ, the *autocovariance* $\mathrm{Cov}(X_s, X_t)$ is time translation invariant,

$$\mathrm{Cov}(X_{s+\tau}, X_{t+\tau}) = \mathrm{Cov}(X_s, X_t), \tag{2}$$

    i.e. $\mathrm{Cov}(X_s, X_t)$ depends only on the difference $t - s$. We will write it as Γ_{t-s}.
- For covariance stationary series, $\Gamma_{-t} = \Gamma_t$ (show it!).
- Notice that $\Gamma_0 = \mathrm{Var}(X_t)$.
- The autocorrelation function of a time series is defined as

$$R_{s,t} = \frac{\mathrm{Cov}(X_s, X_t)}{\sqrt{\mathrm{Var}(X_s)}\sqrt{\mathrm{Var}(X_t)}}\,. \tag{3}$$

- For covariance-stationary time series, $R_{s,t} = R_{t-s}$, and

$$R_t = \frac{\Gamma_t}{\Gamma_0}\,. \tag{4}$$

Autocovariance and stationarity

- Note that μ, Γ, and R are usually unknown, and are estimated from sample data. The estimated sample mean $\widehat{\mu}$, autocovariance $\widehat{\Gamma}$, and autocorrelation $\widehat{R}$ are calculated as follows.
- Consider a finite sample $X_0, X_1, \ldots, X_T$. Then

$$
\begin{aligned}
\widehat{\mu} &= \frac{1}{T}\sum_{t=1}^{T} X_t,\\
\widehat{\Gamma}_t &= \begin{cases} \frac{1}{T}\sum_{j=t+1}^{T}(X_j-\widehat{\mu})(X_{j-t}-\widehat{\mu}), & \text{for } t=0,1,\ldots,T-1,\\ \widehat{\Gamma}_{-t}, & \text{for } t=-1,\ldots,-(T-1). \end{cases}\\
\widehat{R}_t &= \frac{\widehat{\Gamma}_t}{\widehat{\Gamma}_0}.
\end{aligned}
\tag{5}
$$

- Notice that this method allows us to compute up to $T-1$ estimated sample autocorrelations.

Models of time series

- For practical applications, it is convenient to model a time series as a discrete-time stochastic process with a small number of parameters.
- Time series models have typically the following structure:

$$X_t = p_t + m_t + \varepsilon_t, \tag{6}$$

  where the three components on the RHS have the following meaning:
  - p_t is a periodic function called the *seasonality*,
  - m_t is a slowly varying process called the *trend*,
  - ε_t is a stochastic component called the *error* or *disturbance*.
- Classic linear time series models fall into three broad categories:
  - *autoregressive*,
  - *moving average*,
  - *integrated*,

  and their combinations.

White noise

- The source of randomness in the models discussed in these lectures is *white noise*. It is a process specified as follows:

$$X_t = \varepsilon_t, \tag{7}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ are i.i.d. (= independent, identically distributed) normal random variables.

- Note that

$$\begin{aligned} \mathrm{E}(\varepsilon_t) &= 0, \\ \mathrm{Cov}(\varepsilon_s, \varepsilon_t) &= \begin{cases} \sigma^2, & \text{if } s = t, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{8}$$

- The white noise process is stationary and ergodic (show it!).
- The white noise process with *linear drift*

$$X_t = at + b + \varepsilon_t, \quad a \neq 0, \tag{9}$$

is not stationary, as $\mathrm{E}(X_t) = at + b$.

Autoregressive model $AR(1)$

- The first class of models that we consider are the *autoregressive models* $AR(p)$. Their key characteristic is that the current observation is directly correlated with the lagged p observations.
- The simplest among them is $AR(1)$, the autoregressive model with a single lag.
- The model is specified as follows:

$$X_t = \alpha + \beta X_{t-1} + \varepsilon_t. \tag{10}$$

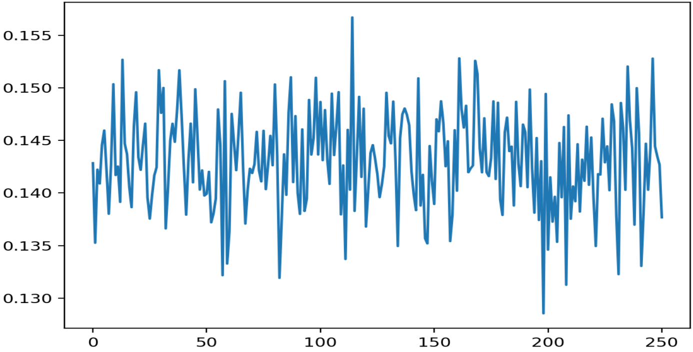
- Here, $\alpha, \beta \in \mathbb{R}$, and $\varepsilon_t \sim N(0, \sigma^2)$ is a white noise.
- A particular case of the $AR(1)$ model is the *random walk model*, namely

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

in which the current value of X is the previous value plus a “white noise” disturbance.

Autoregressive model $AR(1)$

- The graph below shows a simulated $AR(1)$ time series with the following choice of parameters: $\alpha = 0.1$, $\beta = 0.3$, $\sigma = 0.005$.



Autoregressive model $AR(1)$

- Here is the code snippet used to generate this graph in Python:

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_model import ARMA

alpha=0.1
beta=0.3
sigma=0.005

#Simulate AR(1)
T=250
x0=alpha/(1-beta)
x=np.zeros(T+1)
x[0]=x0
eps=np.random.normal(0.0,sigma,T)
for i in range(1,T+1):
   x[i]=alpha+beta*x[i-i]+eps[i-1]

#Take a look at the simulated time series
plt.plot(x)
plt.show()
```

Autoregressive model $AR(1)$

- Let us investigate the circumstances under which an $AR(1)$ process is covariance-stationary.
- For $\mu = \mathrm{E}(X_t)$ to be independent of t we must have from (10):

$$\mu = \alpha + \beta\mu.$$

This equation has a solution iff $\beta \neq 1$ (except for the random walk case corresponding to $\alpha = 0$, $\beta = 1$). In this case,

$$\mu = \frac{\alpha}{1-\beta}\,. \tag{11}$$

- Let us now compute the autocovariance. To this end, we rewrite (10) as

$$X_t - \mu = \beta(X_{t-1} - \mu) + \varepsilon_t. \tag{12}$$

Notice that the two terms on the RHS of this equation are independent of each other.

Autoregressive model $AR(1)$

- For $\Gamma_0 = \operatorname{Var}(X_t)$ to be independent of t, this implies that

$$\Gamma_0 = \beta^2 \Gamma_0 + \sigma^2,$$

and so

$$\Gamma_0 = \frac{\sigma^2}{1 - \beta^2}\,. \tag{13}$$

- Since $\Gamma_0 > 0$, this equation implies that $|\beta| < 1$.
- Multiplying (12) by $X_{t-1} - \mu$, we find that $\Gamma_1 = \beta \Gamma_0$. Iterating, we find that

$$\Gamma_k = \beta^k \Gamma_0, \tag{14}$$

with Γ_0 given by (14). The autocorrelation function is decaying exponentially fast as a function of lag between two observations.

- In conclusion, the condition for a $AR(1)$ process to be covariance-stationary is that $|\beta| < 1$.

Autoregressive model $AR(1)$

- The $AR(1)$ with $|\beta| < 1$ has a natural interpretation that can be gleaned from the following "explicit" representation of X_t. Namely, iterating (10) we find that:

$$
\begin{aligned}
X_t &= \alpha + \beta X_{t-1} + \varepsilon_t \\
&= \alpha(1+\beta) + \beta^2 X_{t-2} + \varepsilon_t + \beta\varepsilon_{t-1} \\
&= \dots \\
&= \alpha(1+\beta+\ldots+\beta^{L-1}) + \beta^L X_{t-L} + \varepsilon_t + \beta\varepsilon_{t-1} + \ldots + \beta^{L-1}\varepsilon_{t-L+1} \\
&= \mu(1-\beta^L) + \beta^L X_{t-L} + \sqrt{\Gamma_0(1-\beta^{2L-1})}\,\xi_t
\end{aligned}
\tag{15}
$$

  where $\xi_t \sim N(0,1)$.

- This implies that

$$
\begin{aligned}
\mathrm{E}(X_t|X_{t-L}) &= \mu(1-\beta^L) + \beta^L X_{t-L}, \\
\mathrm{Var}(X_t|X_{t-L}) &= \Gamma_0(1-\beta^{2L-1}).
\end{aligned}
\tag{16}
$$

Autoregressive model $AR(1)$

- Since $\beta^L \to 0$ exponentially fast, for large L we have

$$X_t \approx \mu + \sqrt{\Gamma_0}\, \xi_t. \tag{17}$$

- In other words, the $AR(1)$ model describes a *mean reverting* time series. After a large number of observations, X_t takes the form (17), i.e. it is equal to its mean value plus a Gaussian noise.
- The rate of convergence to this limit is given by $|\beta|$: the smaller this value, the faster X_t reaches its limit behavior.
- The next question is: given a set of observations, how do we determine the values of the parameters α, β, and σ in (10)?

Maximum likelihood estimation

- *Maximum likelihood estimation* (MLE) is a commonly used method of estimating the parameters of a statistical model given a set of observations.
- It is based on the premise that the best choice of the parameter values should maximize the likelihood of making the observations given these parameters.
- Given a statistical model with parameters $\theta = (\theta_1, \ldots, \theta_d)$, and a set of data $y = (y_1, \ldots, y_N)$, we construct the *likelihood function* $\mathcal{L}(\theta|y)$, which links the model with the data in such a way as if the data were drawn from the assumed model.
- In practice, $\mathcal{L}(\theta|y)$ is the joint probability density function (PDF) $p(y|\theta)$ under the model, evaluated at the observed values.
- In particular, if the observations y_i are independent, then

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{N} p(y_i|\theta), \tag{18}$$

where $p(y_i|\theta)$ denotes the PDF of a single observation.

Maximum likelihood estimation

- The value θ^* that maximizes $\mathcal{L}(\theta|y)$ serves as the best fit between the model specification and the data.
- It is usualy more convenient to consider the *log liklihood function* (LLF) $-\log\mathcal{L}(\theta|y)$. Then, θ^* is the value at which the LLF attains its minimum.
- As an illustration, consider a sample $y = (y_1, \ldots, y_N)$ drawn from the normal distribution $N(\mu, \sigma^2)$. Its likelihood function is given by

$$\mathcal{L}(\theta|y) = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{N} \exp\Big(-\frac{(y_i-\mu)^2}{2\sigma^2}\Big), \tag{19}$$

and the LLF is

$$-\log\mathcal{L}(\theta|y) = \frac{1}{2} N \log\sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i-\mu)^2 + const. \tag{20}$$

Maximum likelihood estimation

- Taking the μ and σ derivatives and setting them to 0, we readily find that that the MLE estimates of μ and σ are

$$\begin{aligned} \mu^* &= \frac{1}{N} \sum_{i=1}^{N} y_i \,, \\ \sigma^* &= \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu^*)^2 \,. \end{aligned} \tag{21}$$

  respectively.
- Note that, while μ^* is *unbiased*, the estimator σ^* is *biased* (N in the denominator above, rather than the usual $N - 1$).
- The fact that the MLE estimator of a parameter is biased is a common occurance. One can show, however, that MLE estimators are *consistent*, i.e. in the limit $N \to \infty$ they converge to the appropriate value.
- Going forward, we will use the notation $\widehat{\theta}$ rather than θ^* for the MLE estimators.

MLE for $AR(1)$

- Consider now the $AR(1)$ model and a time series of data $x_0, \ldots, x_T$, believed to be drawn from this model. The easiest way to construct the likelihood function is to focus on the conditional PDF $p(x_1, \ldots, x_T | x_0, \theta)$. This leads to the *conditional* MLE method.
- Let
$$\widehat{\varepsilon}_t = x_t - \alpha - \beta x_{t-1}, \tag{22}$$
for $t = 1, \ldots, T$, be the disturbances implied from the data. According to the model specification, each $\widehat{\varepsilon}_t$ is independently drawn from $N(0, \sigma^2)$, and thus
$$\begin{aligned} p(x_1, \ldots, x_T | x_0, \theta) &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\Big(- \frac{1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2 \Big) \\ &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\Big(- \frac{1}{2\sigma^2} \sum_{t=1}^{T} (x_t - \alpha - \beta x_{t-1})^2 \Big) \end{aligned} \tag{23}$$
- Hence the LLF is given by
$$-\log \mathcal{L}(\theta | y) = \frac{1}{2} T \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{t=0}^{T-1} (x_{t+1} - \alpha - \beta x_t)^2 + const. \tag{24}$$

MLE for $AR(1)$

- Minimizing this function yields:

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} T & \sum_{t=0}^{T-1} x_t \\ \sum_{t=0}^{T-1} x_t & \sum_{t=0}^{T-1} x_t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=0}^{T-1} x_{t+1} \\ \sum_{t=0}^{T-1} x_t x_{t+1} \end{pmatrix}, \tag{25}$$

$$\widehat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} (x_t - \widehat{\alpha} - \widehat{\beta} x_{t-1})^2 .$$

- This can also be explicitly rewritten as

$$\widehat{\beta} = \frac{\sum_{t=0}^{T-1} (x_t - \widehat{x})(x_{t+1} - \widehat{x}_+)}{\sum_{t=0}^{T-1} (x_t - \widehat{x})^2}, \tag{26}$$

$$\widehat{\alpha} = \widehat{x}_+ - \widehat{\beta}\widehat{x},$$

where

$$\widehat{x} = \sum_{t=0}^{T-1} x_t, \qquad \widehat{x}_+ = \sum_{t=0}^{T-1} x_{t+1} . \tag{27}$$

MLE for $AR(1)$

- The *exact* MLE method attempts to infer the likelihood of x_0 from the probability distribution. Since $x_0 \sim N(\mu, \Gamma_0)$,

$$p(x_0|\theta) = \sqrt{\frac{1-\beta^2}{2\pi\sigma^2}} \exp\Big(-\frac{(x_0-\alpha/(1-\beta))^2}{2\sigma^2/(1-\beta^2)}\Big). \tag{28}$$

- On the other hand, for $t = 1, \ldots, T$,

$$p(x_t|x_{t-1}, \ldots, x_1, \theta) = \frac{1}{2\pi\sigma^2} \exp\Big(-\frac{\big(x_t-\alpha-\beta x_{t-1}\big)^2}{2\sigma^2}\Big). \tag{29}$$

- From the definition of conditional probability we have the following identity:

$$p(x_0, x_1, \ldots, x_T|\theta) = p(x_0|\theta) \prod_{t=1}^{T} p(x_t|x_{t-1}, \ldots, x_1, \theta). \tag{30}$$

MLE for $AR(1)$

- Therefore, the LLF is given by

$$
\begin{aligned}
-\log \mathcal{L}(\theta|x) = {} & \frac{1}{2}\log\frac{\sigma^2}{1-\beta^2} + \frac{1}{2}T\log\sigma^2 \\
& + \frac{(x_0-\alpha/(1-\beta))^2}{2\sigma^2/(1-\beta^2)} + \frac{1}{2\sigma^2}\sum_{t=1}^{T}(x_t-\alpha-\beta x_{t-1})^2 + const.
\end{aligned} \tag{31}
$$

- Unlike the conditional case, the minimum of the exact LLF cannot be calculated in closed form, and the calculation has to be done by means of a numerical search.

MLE for $AR(1)$

- Here is the Python code snippet implementing the MLE for $AR(1)$:

```
#Conditional MLE estimate
y=x[0:T]
yp=x[1:(T+1)]
m=np.sum(y)/T
mp=np.sum(yp)/T
betaCMLE=np.inner(y-m,yp-mp)/np.inner(y-m,y-m)
alphaCMLE=mp-betaCMLE*m
sigmaCMLE=np.sqrt(np.inner(yp-betaCMLE*y-alphaCMLE,
                          yp-betaCMLE*y-alphaCMLE)/T)
```

- Alternatively, one can use statsmodels functions:

```
#MLE estimate with statsmodels
model=ARMA(x,order=(1,0)).fit(method='mle')
alphaMLE=model.params[0]
betaMLE=model.params[1]
sigmaMLE=np.std(model.resid)
```

Second order autoregressive model $AR(2)$

- A *second order autoregressive model* $AR(2)$ model is specified as follows:

$$X_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t, \tag{32}$$

where $\alpha, \beta_1, \beta_2 \in \mathbb{R}$, and $\varepsilon_t \sim N(0, \sigma^2)$ is a white noise.

- Under this specification, the state variable depends on its two lags (rather than one lag as in $AR(1)$.
- Let us determine the conditions under which the model is covariance-stationary.
- From the requirement that $\mathsf{E}(X_t) = \mu$,

$$\mu = \frac{\alpha}{1 - \beta_1 - \beta_2}, \tag{33}$$

and so we can can rewrite (32) in the following form:

$$X_t - \mu = \beta_1(X_{t-1} - \mu) + \beta_2(X_{t-2} - \mu) + \varepsilon_t. \tag{34}$$

Second order autoregressive model $AR(2)$

- Multiplying (34) by $X_{t-j} - \mu$, for $j = 0, 1, 2$, and calculating expectations, we find that

$$\Gamma_k = \begin{cases} \beta_1 \Gamma_1 + \beta_2 \Gamma_2 + \sigma^2, & \text{if } k = 0, \\ \beta_1 \Gamma_{k-1} + \beta_2 \Gamma_{k-2}, & \text{if } k = 1, 2. \end{cases} \tag{35}$$

  This identity is called the *Yule-Walker equation* for the autocovariance.
- Dividing (57) by Γ_0 yields the Yule-Walker equation for the autocorrelation:

$$R_k = \beta_1 R_{k-1} + \beta_2 R_{k-2}, \tag{36}$$

  for $k = 1, 2$.
- This equation allows us calculate explicitly the ACF for $AR(2)$.
- Namely, plugging in $k = 1$ and remembering that $R_{-1} = R_1$ yields $R_1 = \beta_1 + \beta_2 R_1$, or

$$R_1 = \frac{\beta_1}{1 - \beta_2} . \tag{37}$$

Second order autoregressive model $AR(2)$

- Plugging in $k = 2$ yields $R_2 = \beta_1 R_1 + \beta_2$, or

$$R_2 = \beta_2 + \frac{\beta_1^2}{1 - \beta_2}\,. \tag{38}$$

- Finally, substituting $k = 0$ in (34) yields

$$\Gamma_0 = (\beta_1 R_1 + \beta_2 R_2)\Gamma_0 + \sigma^2. \tag{39}$$

Solving this, we obtain

$$\Gamma_0 = \frac{(1 - \beta_2)\sigma^2}{(1 + \beta_2)((1 - \beta_2)^2 - \beta_1^2)}\,. \tag{40}$$

Lag operators and characteristic roots

- We have not yet addressed the question under what condition is an $AR(2)$ time series covariance-stationary. We will now introduce the concepts that will settle this issue and will allow us to formulate criteria for stationarity for more general models,
- Let us define the *lag operator* L as a (linear) mapping:

$$LX_t = X_{t-1}. \tag{41}$$

  In other words, the lag operator shifts the time index back by one unit.
- Applying the lag operator k times shifts the time index by k units:

$$L^k X_t = X_{t-k}. \tag{42}$$

  We refer to L^k as the k-th power of L.
- Finally, if $\psi(z) = \psi_0 + \psi_1 z + \ldots + \psi_n z^n$ is a polynomial in z, we associate with it an operator $\psi(L)$ defined by

$$\psi(L) = \psi_0 + \psi_1 L + \ldots + \psi_n L^n. \tag{43}$$

Lag operators and characteristic roots

- Notice that equation (32) can be stated as

$$\psi(L)X_t = \alpha + \varepsilon_t, \tag{44}$$

where $\psi(z) = 1 - \beta_1 z - \beta_2 z^2$.

- Solving this equation amounts to finding the inverse $\psi(L)^{-1}$ of $\psi(L)$:

$$X_t = \frac{\alpha}{\psi(1)} + \psi(L)^{-1}\varepsilon_t. \tag{45}$$

- Suppose that we can write $\psi(L)^{-1}$ as an infinite series

$$\psi(L)^{-1} = \sum_{j=0}^{\infty} \gamma_j L^j, \tag{46}$$

with

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \tag{47}$$

Lag operators and characteristic roots

- Then

$$X_t = \frac{\alpha}{\psi(1)} + \sum_{j=0}^{\infty} \gamma_j \varepsilon_{t-j}, \tag{48}$$

with

$$\mathrm{E}(X_t) = \frac{\alpha}{\psi(1)}, \tag{49}$$

and

$$\mathrm{Cov}(X_t, X_{t+k}) = \sum_{j=0}^{\infty} \gamma_j \gamma_{j+k}, \text{ for } k \geq 0, \tag{50}$$

independently of t. The series is thus covariance-stationary.

- In the case of $AR(1)$, $\psi(L) = 1 - \beta L$, it is clear that the geometric series does the job:

$$(1 - \beta L)^{-1} = \sum_{j=0}^{\infty} \beta^j L^j, \tag{51}$$

- Condition (47) holds as long as $|\beta| < 1$. Another way of saying this is that the root $z_1 = 1/\beta$ of $1 - \beta z$ lies outside of the unit circle.

Lag operators and characteristic roots

- Now, if $\psi(z)$ is a polynomial with non-zero roots $z_1, \ldots, z_n$. Then

$$\psi(L) = (-1)^n \Big(\prod_{j=1}^{n} z_j\Big) \prod_{j=1}^{n} (1 - z_j^{-1} L). \tag{52}$$

- If each of the roots z_j (they may be complex) lies outside of the unit circle, i.e. $|z_j^{-1}| < 1$, then we can invert $\psi(L)$ by applying (51) to each factor in the product above.
- It is not hard to verify that the convergence criterion (47), and thus the time series is stationary.
- We can summarize these arguments by stating that *a time series model given by the lag form equation* (44) *is covariance stationary if the roots of the polynomial* $\psi(z)$ *lie outside of the unit circle.*

General autoregressive model $AR(p)$

- The *p-th order autoregressive model* $AR(p)$ model is specified as follows:

$$X_t = \alpha + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon_t, \tag{53}$$

where $\alpha, \beta_j \in \mathbb{R}$, and $\varepsilon_t \sim N(0, \sigma^2)$ is a white noise.

- For the covariance-stationarity, the requirement that $\mathrm{E}(X_t) = \mu$ yields

$$\mu = \frac{\alpha}{1 - \beta_1 - \ldots - \beta_p}. \tag{54}$$

- Furthermore, we require that the roots of the characteristic polynomial $\psi(z) = 1 - \alpha - \beta_1 z - \ldots - \beta_p z^p$ lie outside of the unit circle.
- We can rewrite (53) in the following form:

$$X_t - \mu = \beta_1 (X_{t-1} - \mu) + \ldots + \beta_p (X_{t-p} - \mu) + \varepsilon_t. \tag{55}$$

General autoregressive model $AR(p)$

- Multiplying this equation by $X_{t-j} - \mu$, for $j = 0, \ldots, p$, and calculating expectations yields the Yule-Walker equation for the autocovariance:

$$\Gamma_k = \begin{cases} \beta_1 \Gamma_1 + \cdots + \beta_p \Gamma_p + \sigma^2, & \text{if } k = 0, \\ \beta_1 \Gamma_{k-1} + \ldots + \beta_p \Gamma_{k-p}, & \text{if } k = 1, \ldots, p. \end{cases} \tag{56}$$

- Dividing (56) by Γ_0 yields the Yule-Walker equation for the autocorrelation:

$$R_k = \beta_1 R_{k-1} + \ldots + \beta_p R_{k-p}, \tag{57}$$

for $k = 1, \ldots, p$.

- Note that the autocorrelations satisfy essentially the same equation as the process defining X_t.
- The ACF R_k can be found as the solution to the Yule-Walker equation and are expressed in terms of the roots of the characteristic polynomial.

Choosing the number of lags in $AR(p)$

- In practice, the number of lags p is unknown, and has to be determined empirically.
- This can be done by regressing the variable on its lagged values with $p = 1, 2, \ldots$, and assessing the impact of each added lag on the fit.
- It is important not to overfit the model ("torture it until it confesses") by adding too many lags.
- Useful quantitative guides for model selection are various information criteria.
- The *Akaike information criterion* defined as follows:

$$\mathrm{AIC} = 2k - 2\log\mathcal{L}(\widehat{\theta}|x). \tag{58}$$

  Here $k = \#\theta$ is the number of model parameters, $-\log\mathcal{L}(\widehat{\theta}|x)$ denotes the optimized value of the LLF.
- According to this criterion, among the candidate models the model with the lowest value of AIC is the preferred one.

Choosing the number of lags in $AR(p)$

- This is in contrast with picking the model whose optimized LLF is the lowest: this may be the result of overfitting. The AIC criterion penalizes the number of parameters, and thus discourages overfitting.
- Another popular information criteria is the *Bayesian information criterion* (a.k.a the *Schwarz criterion*), which is defined as follows:

$$\mathrm{BIC} = \log(N)k - 2\log\mathcal{L}(\widehat{\theta}|x), \tag{59}$$

  where $N = \#x$ is the number of data points.
- According to this criterion, the model with the smallest value of BIC is the preferred model.

Moving average model $MA(1)$

- The *moving average* model $MA(1)$ is specified as follows:

$$X_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}., \tag{60}$$

  where μ and θ are constants, and ε_t is white noise.
- The key feature of the $MA(1)$ model is that its are autocorrelated with lag 1.
- The expected value of X_t is

$$\mathsf{E}(X_t) = \mu, \tag{61}$$

  as $\mathsf{E}(\varepsilon_t) = \mu$, for all t.
- Its variance is

$$\begin{aligned}\mathsf{E}((X_t - \mu)^2) &= \mathsf{E}((\varepsilon_t + \theta\varepsilon_{t-1})^2)\\ &= \mathsf{E}(\varepsilon_t^2) + 2\theta\mathsf{E}(\varepsilon_t\varepsilon_{t-1}) + \theta^2\mathsf{E}(\varepsilon_{t-1}^2)\\ &= (1+\theta^2)\sigma^2.\end{aligned}$$

Moving average model $MA(1)$

- For the first autocovariance, we have

$$\begin{aligned} \mathsf{E}((X_t-\mu)(X_{t-1}-\mu)) &= \mathsf{E}((\varepsilon_t+\theta\varepsilon_{t-1})(\varepsilon_{t-1}+\theta\varepsilon_{t-2})) \\ &= \theta\sigma^2 . \end{aligned}$$

- All autocovariances with lag ≥ 2 are zero (show it!).
- As a result, $MA(1)$ is (unlike $AR(1)$) always covariance-stationary with

$$\Gamma_t = \begin{cases} (1+\theta^2)\sigma^2, & \text{if } t=0, \\ \theta\sigma^2, & \text{if } |t|=1, \\ 0, & \text{if } |t|\geq 2. \end{cases} \tag{62}$$

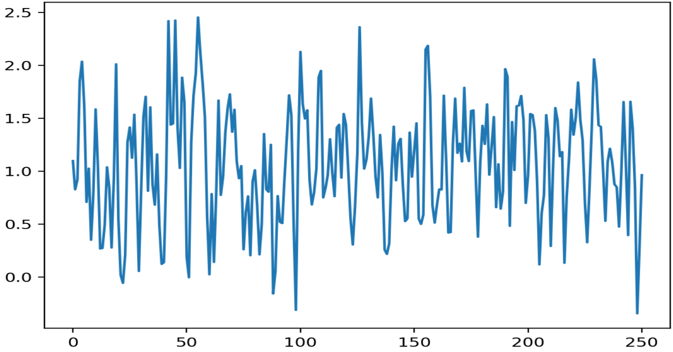
- As a result, the first autocorrelation $R_1 = \Gamma_1/\Gamma_0$ is given by

$$R_1 = \frac{\theta}{1+\theta^2}, \tag{63}$$

with all higher order autocorrelations equal zero.

Moving average model $MA(1)$

- The graph below shows a simulated $MA(1)$ time series with the following choice of parameters: $\mu = 1.1$, $\beta = 0.6$, $\sigma = 0.5$.



Moving average model $MA(1)$

- Here is the code snippet used to generate this graph in Python:

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_model import ARMA

mu=1.1
theta=0.6
sigma=0.5

#Simulate MA(1)
T=250
x0=mu
x=np.zeros(T+1)
x[0]=x0
eps=np.random.normal(0.0,sigma,T+1)
for i in range(1,T+1):
   x[i]=mu+eps[i]+theta*eps[i-1]

#Take a look at the simulated time series
plt.plot(x)
plt.show()
```

MLE for $MA(1)$

- As in the case of $AR(1)$, there are two natural approaches to MLE of an $MA(1)$ model: conditional on the initial value of ε and exact.
- We begin with the *conditional* MLE method, which is somewhat easier.
- Since the value of ε_0 cannot be calculated from the observed data, we are free to set it arbitrarily; we choose $\varepsilon_0 = 0$. All the probabilities calculated below are conditional on this choice.
- We then have, for $t = 1, \ldots, T$,

$$\varepsilon_t = x_t - \mu - \theta\varepsilon_{t-1}, \tag{64}$$

and so the conditional PDF of x_t is

$$p(x_t|x_{t-1}, \ldots, x_1, \varepsilon_0 = 0, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{\varepsilon_t^2}{2\sigma^2}\Big). \tag{65}$$

- This expression is deceivingly simply: in reality ε_t is a nested function of all x_s with $s \leq t$.
- The liklihood function of the sample $x_1, \ldots_T$ is given by the product of the probabilities above, and so

$$\mathcal{L}(\theta|x, \varepsilon_0 = 0) = \prod_{t=1}^{T} p(x_t|x_{t-1}, \ldots, x_1, \varepsilon_0 = 0, \theta), \tag{66}$$

MLE for $MA(1)$

- The log liklihood has thus the following form:

$$-\log \mathcal{L}(\theta|x, \varepsilon_0 = 0) = \frac{1}{2}\, T \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2 + const. \tag{67}$$

- This is a quadratic function of the x_t's. It is cumbersome to write it down explicitly, but easy to code it in a programming language. Its minimum is easiest to find by means of a numerical search.
- In case of $|\theta| < 1$, the impact of the choice $\varepsilon_0 = 0$ phases out as we iterate through time steps. For $|\theta| > 1$ the impact of this choice accumulates, and the method cannot be used.
- For the *exact* MLE method, we notice that the joint PDF of x is given by

$$p(x|\theta) = \frac{1}{(2\pi)^{T/2} \det(\Omega)^{1/2}} \exp\Big(-\frac{1}{2}\,(x-\mu)^{\mathrm{T}} \Omega^{-1} (x-\mu) \Big), \tag{68}$$

and thus

$$-\log \mathcal{L}(\theta|x) = \frac{1}{2} \log\det(\Omega) + \frac{1}{2}\,(x-\mu)^{\mathrm{T}} \Omega^{-1} (x-\mu). \tag{69}$$

MLE for $MA(1)$

- Here, Ω is a band diagonal matrix:

$$\Omega = \sigma^2 \begin{pmatrix} 1+\theta^2 & \theta & 0 & \dots & 0 \\ \theta & 1+\theta^2 & \theta & \dots & 0 \\ 0 & \theta & 1+\theta^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 1+\theta^2. \end{pmatrix} \tag{70}$$

- The numerics of minimizing (69) can be handled either by (i) a clever triangular factorization of Ω, or by the Kalman filter method (we will discuss Kalman filters later in this course).
- Unlike the conditional MLE method, the exact method does not suffer from instabilities if $|\theta| \geq 1$.

MLE for $MA(1)$

- Here is the Python code snippet implementing the MLE for $MA(1)$ using statsmodels:

```
#MLE estimate with statsmodels
model=ARMA(x,order=(0,1)).fit(method='mle')
muMLE=model.params[0]
thetaMLE=model.params[1]
sigmaMLE=np.std(model.resid)
```

General moving average model $MA(q)$

- A *q-th order moving average* model $MA(q)$ is specified as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}., \tag{71}$$

  where μ and θ_j are constants, and ε_t is white noise.
- In other words, the $MAq1)$ model fluctuates around μ with disturbances which are autocorrelated with lag q.
- The expected value of X_t is

$$\mathrm{E}(X_t) = \mu, \tag{72}$$

  while its autocovariance is

$$\Gamma_j = \begin{cases} (1 + \theta_1^2 + \ldots + \theta_q^2)\sigma^2, & \text{if } j = 0, \\ (\theta_j + \theta_{j+1}\theta_1 + \ldots + \theta_q \theta_{q-j})\sigma^2, & \text{if } j = 1, \ldots, q, \\ 0, & \text{if } j > q. \end{cases} \tag{73}$$

$ARMA(p, q)$ model

- A *mixed autoregressive moving average* model $ARMA(p, q)$ is specified as follows:

$$X_t = \alpha + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}., \tag{74}$$

  where α and β_j, θ_k are constants, and ε_t is white noise.

- The equation above has the following lag operator representation:

$$\psi(L) X_t = \alpha + \varphi(L) \varepsilon_t, \tag{75}$$

  where

$$\begin{aligned} \psi(z) &= 1 - \beta_1 z - \ldots - \beta_p z^p, \\ \varphi(z) &= 1 + \theta_1 z + \ldots + \theta_q z^q. \end{aligned} \tag{76}$$

- The process (45) is covariance stationary if the roots of ψ lie outside of the unit circle.

Parameter estimation and model selection for $ARMA(p, q)$

- The parameters ARMA models are estimated by means of the MLE method. The complexity of computation required to minimize the LLF increases with the number of parameters.
- Information criteria, such as AIC or BIC, remain useful quantitative guides for model selection.

Forecasting time series with $ARMA(p, q)$

- An important function of time series analysis is making predictions about future values of the observed data, i.e. *forecasting*.
- Data based forecasting problem can be formulated as follows: given the observations $X_{1:t} = X_1, \ldots, X_t$, what is the best forecast $X^*_{t+1|1:t}$ of X_{t+1}?
- In mathematical terms, the problem requires minimizing a suitable loss function. We choose to minimize the *mean squared error* (MSE) given by

$$\mathrm{E}_t\big((X_{t+1} - X^*_{t+1|1:t})^2\big). \tag{77}$$

- We claim that $X^*_{t+1|1:t}$ is, indeed, given given by the conditional expected value:

$$X^*_{t+1|1:t} = E_t(X_{t+1}). \tag{78}$$

Here E_t denotes expectation, conditional on the information up to time t,

$$\mathrm{E}_t(\cdot) = \mathrm{E}(\,\cdot\,|X_{1:t}). \tag{79}$$

Forecasting time series with $ARMA(p, q)$

- Indeed, if Z is any random variable measurable with respect to the information set generated by $X_{1:t}$, then

$$\begin{aligned}
\mathrm{E}\big((X_{t+1} - Z)^2\big) &= \mathrm{E}\big((X_{t+1} - E_t(X_{t+1}) + E_t(X_{t+1}) - Z)^2\big) \\
&= \mathrm{E}\big((X_{t+1} - E_t(X_{t+1}))^2\big) + \mathrm{E}\big((E_t(X_{t+1}) - Z)^2\big) \\
&\quad + 2\mathrm{E}\big((X_{t+1} - E_t(X_{t+1}))(E_t(X_{t+1}) - Z)\big).
\end{aligned}$$

- We argue that the cross term above is zero. Indeed

$$\begin{aligned}
\mathrm{E}_t\big((X_{t+1} - E_t(X_{t+1}))(E_t(X_{t+1}) - Z)\big) &= \mathrm{E}_t\big(X_{t+1} - E_t(X_{t+1})\big)\big(E_t(X_{t+1}) - Z\big) \\
&= \big(\mathrm{E}_t(X_{t+1}) - E_t(X_{t+1})\big)\big(E_t(X_{t+1}) - Z\big) \\
&= 0.
\end{aligned}$$

Since $\mathrm{E}(\cdot) = \mathrm{E}(\mathrm{E}_t(\cdot)|X_t)$, the claim follows.

Forecasting time series with $ARMA(p, q)$

- As a result

$$\mathsf{E}\big((X_{t+1} - Z)^2\big) = \mathsf{E}\big((X_{t+1} - E_t(X_{t+1}))^2\big) + \mathsf{E}\big((E_t(X_{t+1}) - Z)^2\big),$$

  which has its minimum at $Z = E_t(X_{t+1})$.
- For example, a single period forecast in an $AR(1)$ model is

$$\begin{aligned} X^*_{t+1|1:t} &= \mathsf{E}_t(X_{t+1}) \\ &= \mathsf{E}_t(\alpha + \beta X_t + \varepsilon_t) \\ &= \alpha + \beta X_t. \end{aligned} \tag{80}$$

- The forecast error is ε_{t+1}, and so the variance of the forecast error is σ^2.
- Likewise, a single period forecast in an $AR(p)$ model is

$$X^*_{t+1|1:t} = \alpha + \beta_1 X_t + \ldots + \beta_p X_{t-p+1}. \tag{81}$$

  with forecast error is ε_{t+1}, and the variance of the forecast error is σ^2.

Forecasting time series with $ARMA(p, q)$

- A two period forecast in an $AR(1)$ model is given by

$$\begin{aligned} X^*_{t+2|1:t} &= \mathrm{E}_t(X_{t+2}) \\ &= \mathrm{E}_t(\alpha + \beta X_{t+1} + \varepsilon_{t+1}) \\ &= (1+\beta)\alpha + \beta^2 X_t. \end{aligned} \tag{82}$$

- The error of the two period forecast is $\varepsilon_{t+2} + \beta\varepsilon_{t+1}$, and its variance is $(1+\beta)\sigma^2$.
- A one period forecast in an $MA(1)$ model is

$$\begin{aligned} X^*_{t+1|1:t} &= \mathrm{E}_t(X_{t+1}) \\ &= \mathrm{E}_t(\mu + \varepsilon_{t+1} + \theta\varepsilon_t) \\ &= \mu + \theta\varepsilon_t. \end{aligned} \tag{83}$$

- The forecast error is ε_{t+1}, and its variance is σ^2.
- These calculations can be generalized to produce a general formula for a multi-period forecast in an $ARMA(p, q)$ model. This result is known as the *Wiener-Kolmogorov prediction formula* and its discussion can be found in [1].

References

Hamilton, J. D.: *Time Series Analysis*, Princeton University Press (1994).

Tsay, R. S.: *Analysis of Financial Time Series*, Wiley (2010).