Time Series Analysis

1. Stationary ARMA models

Andrew Lesniewski

Baruch College
New York

Fall 2023

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 $h = 1$.
- *Time series analysis* is a branch of data science, which develops mathematical methodologies for analysis of observed time series.

Time series

- Examples of time series commonly encountered in finance include:
  - (i) asset prices,
  - (ii) portfolio returns,
  - (iii) index levels,
  - (iv) trading volumes,
  - (v) futures open interests,
  - (vi) implied volatilities,
  - (vii) interest rates,
  - (viii) macroeconomic data (inflation, new payrolls, unemployment, GDP, housing prices, . . .).

Time series

- Typically, time series exhibit significant irregularities, which may have their origin either in (i) the nature of the underlying quantity or (ii) imprecision in its observation (or both).
- The objective of time series analysis is to design analytic methodologies for making inferences from these irregularities.
- These methodologies fall into two broad categories: parametric and non-parametric.
- In the *parametric* approach, we assume a probability model: the elements of a time series are random variables on an underlying probability space, and the stochastic law of the time series is characterized by a probability distribution with a finite number of parameters.
- In the *non-parametric* approach the observed data is analyzed without assuming a probability model, and no stochastic law of the time series is explicitly specified.

Time series

- The results of time series analysis are used for various purposes such as
  - (i) data interpretation,
  - (ii) forecasting,
  - (iii) smoothing,
  - (iv) back filling, ...
- Also, parametric models allow for hypothesis testing; hypothesis testing is not possible within the non-parametric approach.

Stationarity and ergodicity

- We begin with the concept of a stationary time series, which falls into the category of parametric models.
- 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 $(X_{t_1}, \ldots, X_{t_k})$.
- In other words, the joint probability distribution of $(X_{t_1}, \ldots, X_{t_k})$ remains the same, if each observation time t_i is shifted by the same amount (time translation invariance).
- For a stationary time series, the expected value $\mathrm{E}(X_t)$ is independent of t and is called the *mean* of X_t; we will denote its value by μ.

Stationarity and ergodicity

- A stationary time series model is *ergodic* if

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

  i.e. if the time average of X_t is equal to its mean.
- The limit in (2) is usually understood in the sense of squared mean convergence, i.e.

$$\lim_{T\to\infty} \mathrm{E}\Big(\big(\frac{1}{T} \sum_{k=1}^{T} X_{t+k} - \mu \big)^2 \Big) = 0. \tag{2}$$

- In case of financial time series, we are always faced with a single realization of a process, rather than an ensemble of alternative outcomes, and we hope that the time average represents the “true” expected value of the state variable.
- Therefore, ergodicity is a desired if illusive property of a financial time series.

Stationarity and ergodicity

- The notions of stationarity and ergodicity are hard to verify in practice.
- In particular, there is no practical statistical test for ergodicity.
- For this reason, a weaker but more practical concept of stationarity has been introduced.

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,

$$\mathrm{Cov}(X_{s+\tau}, X_{t+\tau}) = \mathrm{Cov}(X_s, X_t), \tag{3}$$

  i.e. $\mathrm{Cov}(X_s, X_t)$ depends only on the difference $t - s$.
- We will use the notation $\Gamma_{t-s} = \mathrm{Cov}(X_s, X_t)$.
- Weak stationarity implies that $\Gamma_{-t} = \Gamma_t$ (show it!).
- Notice that $\Gamma_0 = \mathrm{Var}(X_t)$.

Autocorrelation function

- The *autocorrelation function* (ACF) of a time series is defined as

$$R_{s,t} = \frac{\text{Cov}(X_s, X_t)}{\sqrt{\text{Var}(X_s)}\sqrt{\text{Var}(X_t)}} . \tag{4}$$

- The ACF is the (Pearson) correlation coefficient between the values of the time series at two time instances, $R_{s,t} = \text{Corr}(X_s, X_t)$.
- For covariance-stationary time series, $R_{s,t} = R_{s-t,0}$, i.e. the ACF is a function of the difference $s - t$ only.
- For simplicity, we will write $R_t = R_{t,0}$, and note that

$$R_t = \frac{\Gamma_t}{\Gamma_0} . \tag{5}$$

- The parameters μ, Γ, and R are usually unknown, and have to estimated from observed data.

Autocorrelation function

- The estimated sample mean $\widehat{\mu}$, autocovariance $\widehat{\Gamma}$, and autocorrelation $\widehat{R}$ are calculated as follows.
- Consider a finite sample $x_1, \ldots, x_T$; then

$$\widehat{\mu} = \frac{1}{T} \sum_{t=1}^{T} x_t,$$

$$\widehat{\Gamma}_t = \begin{cases} \frac{1}{T} \sum_{j=t+1}^{T} (x_j - \widehat{\mu})(x_{j-t} - \widehat{\mu}), & \text{for } t = 0, 1, \ldots, T-1, \\ \widehat{\Gamma}_{-t}, & \text{for } t = -1, \ldots, -(T-1). \end{cases} \tag{6}$$

$$\widehat{R}_t = \frac{\widehat{\Gamma}_t}{\widehat{\Gamma}_0}.$$

- These quantities are called the *sample mean*, *sample autocovariance*, and *sample ACF*, respectively.

Autocorrelation function

- Notice that this method allows us to compute up to $T-1$ estimated sample autocorrelations.
- The estimator $\widehat{\mu}$ of the mean is *unbiased*, i.e. $\mathrm{E}(\widehat{\mu}) = \mu$.
- Usually, $\widehat{R}_t$ is a *biased* estimator of R_t, i.e. $\mathrm{E}(\widehat{R}_t) \neq R_t$.
- The *bias* $\mathrm{E}(\widehat{R}_t) - R_t$ goes to zero as $1/T$, as the sample size T increases, $T \to \infty$, i.e. the estimator $\widehat{R}_t$ is *consistent*[1].

[1] under some technical assumptions

Partial autocorrelation function

- Another useful metric of dependence between the values of a stationary time series at two time instances is the partial autocorrelation function (PACF).
- In order to define it, we let $P_k(X_{k+1})$ denote the OLS regression of X_{k+1} on the variables $1, X_2, \ldots, X_k$, i.e.

$$P_k(X_{k+1}) = \widehat{\pi}_0 + \sum_{j=2}^{k} \widehat{\pi}_j X_j. \tag{7}$$

- The coefficients $\widehat{\pi}_0, \widehat{\pi}_2, \ldots, \widehat{\pi}_k$ are determined by the requirement

$$\widehat{\pi}_0, \widehat{\pi}_2, \ldots, \widehat{\pi}_k = \underset{\pi_0, \pi_2, \ldots, \pi_k}{\arg\min} \, \mathrm{E}\Big(\big(X_{k+1} - \pi_0 - \sum_{j=2}^{k} \pi_j X_j \big)^2 \Big). \tag{8}$$

- Likewise, we let $P_k(X_1)$ denote the OLS regression of X_1 on the variables $1, X_2, \ldots, X_k$.

Partial autocorrelation function

- The PACF at lag k, denoted by α_k, is defined as follows [2], [1]:
  - (i) $\alpha_1 = \text{Corr}(X_2, X_1)$
  - (ii) $\alpha_k = \text{Corr}(X_{k+1} - P_k(X_{k+1}), X_1 - P_k(X_1))$, for $k \geq 2$
- In other words, α_k measures the correlation between the residuals of the regressions of X_{k+1} and X_1 on the observations $X_1, \ldots, X_k$.
- There is an algorithmic way of computing the PACF in terms of the ACF (Durbin-Levinson algorithm, see [1]).
- Similar expressions hold for observed data samples, with ensemble values replaced by the corresponding sample estimates.

Autocovariance and stationarity

- One can use the above estimators to test the hypothesis $H_0: \ R_t = 0$ versus $H_a: \ R_t \neq 0$.
- The relevant t-stat is

$$r = \frac{\widehat{R}_t}{\sqrt{\frac{1}{T}\left(1 + 2\sum_{i=1}^{t-1} \widehat{R}_i^2\right)}} \, .$$

- If X_t is a stationary Gaussian time series with $R_s = 0$ for $s > t$, this t-stat is normally distributed, asymptotically as $T \to \infty$.
- We thus reject H_0 with confidence $1 - \alpha$, if $|r| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution.

Autocovariance and stationarity

- Another test, the *Portmanteau test*, allows us to test jointly for the presence of several autocorrelations, i.e. $H_0: \ R_1 = \ldots = R_k = 0$, versus $H_a: \ R_i \neq 0$, for some $1 \leq i \leq k$.
- The relevant t-stat is defined as

$$Q^*(k) = T \sum_{i=1}^{k} \widehat{R}_i^2.$$

- Under the assumption that X_t is i.i.d., $Q^*(k)$ is asymptotically distributed according to $\chi^2(k)$.
- The power of the test is increased if we replace the statistics above with the *Ljung-Box stat*:

$$Q(k) = T(T+2) \sum_{i=1}^{k} \frac{\widehat{R}_i^2}{T-i}.$$

- H_0 is rejected if $Q(k)$ is greater than the $1-\alpha$ percentile of the $\chi^2(k)$ distribution.

Models of time series

- For practical applications, it is convenient to model a time series as a discrete-time stochastic process X_t.
- The number of parameters of the probability distribution of X_t should be manageable, so that they can be estimated from the training data.
- *Classic linear time series models* fall into three broad categories:
  - *autoregressive*,
  - *moving average*,
  - *integrated*,

  and their combinations.

White noise

- The source of randomness frequently assumed in time series models is *white noise*.
- It is a process specified as follows:

$$X_t = \varepsilon_t, \tag{9}$$

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

- *Colored noise* refers to more complicated sources of randomness. We will discuss examples later.

White noise

- Note that

$$\begin{aligned} \mathrm{E}(\varepsilon_t) &= 0, \\ \mathrm{Cov}(\varepsilon_s, \varepsilon_t) &= \begin{cases} \sigma^2, & \text{if } s = t, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

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

$$X_t = at + b + \varepsilon_t, \quad a \neq 0, \tag{11}$$

is not stationary, as $\mathrm{E}(X_t) = at + b$ is a linear function of time.

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:

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

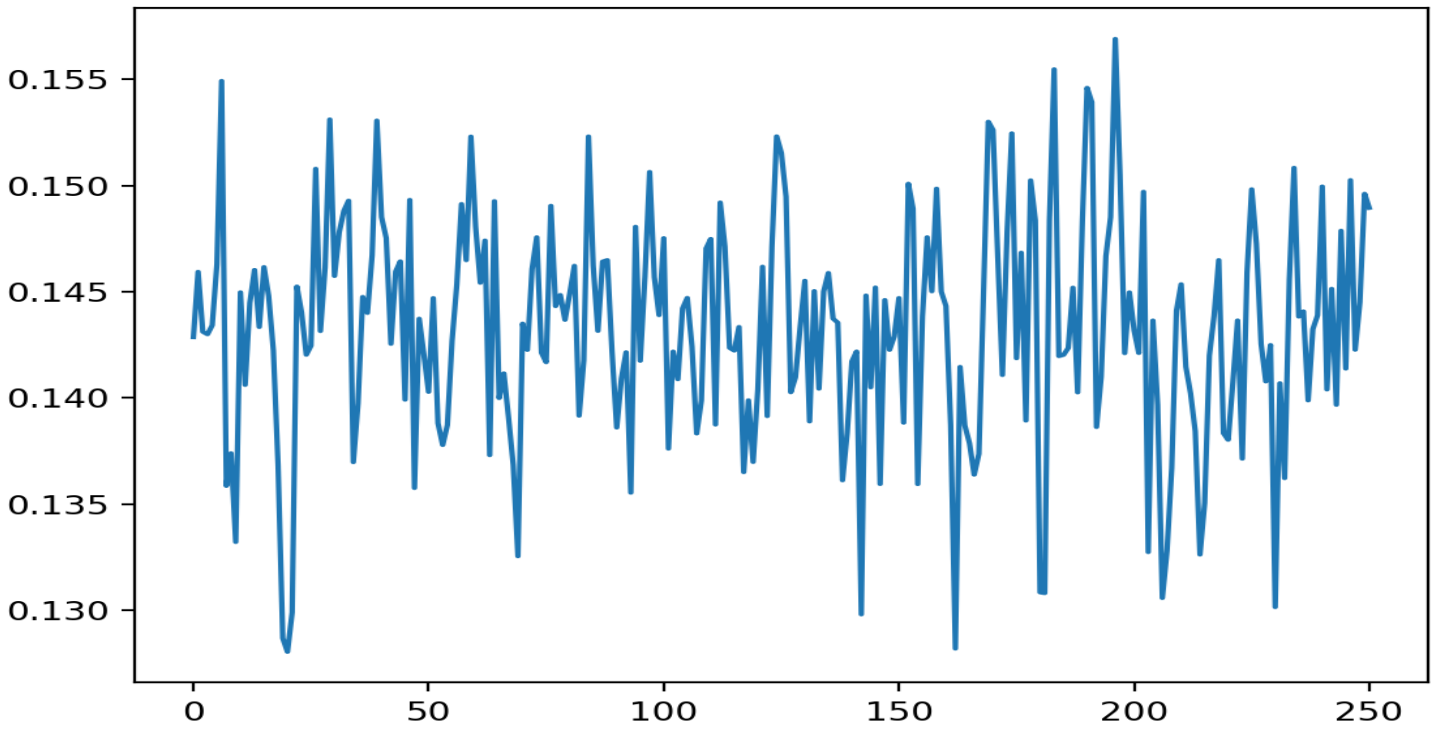
- Here, $\alpha, \beta \in \mathbb{R}$, and $\varepsilon_t \sim N(0, \sigma^2)$ is a white noise.
- A particular case of the $AR(1)$ model is the *random walk model*, namely

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

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 $AR(1)$

- 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 (12):

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

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

$$\mu = \frac{\alpha}{1-\beta}\,. \tag{13}$$

Autoregressive model $AR(1)$

- Let us now compute the autocovariance.
- To this end, we rewrite (12) as

$$X_t - \mu = \beta(X_{t-1} - \mu) + \varepsilon_t. \tag{14}$$

- Notice that the two terms on the RHS of this equation are independent of each other.
- Indeed, X_{t-1} depends only on $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$, which are all independent of ε_t.
- For $\Gamma_0 = \mathrm{Var}(X_t)$ to be independent of t, this implies that

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

and so

$$\Gamma_0 = \frac{\sigma^2}{1 - \beta^2}\,. \tag{15}$$

Autoregressive model $AR(1)$

- Since $\Gamma_0 > 0$, this is possible if and only if $|\beta| < 1$.
- Multiplying (14) by $X_{t-1} - \mu$, we find that $\Gamma_1 = \beta\Gamma_0$.
- Iterating, we find that

$$\Gamma_k = \beta^k \Gamma_0, \tag{16}$$

  with Γ_0 given by (24).
- As a consequence,

$$R_k = \beta^k. \tag{17}$$

- 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 $AR(1)$

- We can calculate the PACF α_k of the $AR(1)$ process.
- Clearly, $\alpha_1 = \beta$.
- In order to find α_2, notice that

$$\mathrm{E}\left((X_3 - \pi_0 - \pi_2 X_2)^2\right) = \pi_0^2 + (\Gamma_0 + \mu^2)\pi_2^2 + 2\mu\pi_0\pi_2 - 2\mu\pi_0 - 2(\Gamma_0\beta + \mu^2)\pi_2 + const,$$

  and so $\widehat{\pi}_0 = \mu(1-\beta)$, $\widehat{\pi}_2 = \beta$.
- Therefore,

$$P_2(X_3) = \mu(1-\beta) + \beta X_2.$$

- Similarly,

$$P_2(X_1) = \mu(1-\beta) + \beta X_2.$$

Autoregressive model $AR(1)$

- As a consequence,

$$\begin{aligned}\alpha_2 &= \mathrm{Corr}\big((X_3 - \mu(1-\beta) - \beta X_2)(X_1 - \mu(1-\beta) - \beta X_2)\big) \\ &= \frac{\mathrm{Cov}\big(((X_3 - \mu) - \beta(X_2 - \mu))((X_1 - \mu) - \beta(X_2 - \mu))\big)}{\text{denominator}}.\end{aligned}$$

- But

$$\begin{aligned}\mathrm{Cov}\big(((X_3 - \mu) - \beta(X_2 - \mu))((X_1 - \mu) - \beta(X_2 - \mu))\big) &= \Gamma_2 - 2\beta\Gamma_1 + \beta^2\Gamma_0 \\ &= 0,\end{aligned}$$

  and so $\alpha_2 = 0$.
- One can show that, in fact, in the $AR(1)$ model,

$$\alpha_k = 0, \text{ for all } k \geq 2. \tag{18}$$

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 (12) we find that:

$$\begin{aligned} X_t &= \alpha + \beta X_{t-1} + \varepsilon_t \\ &= \alpha(1+\beta) + \beta^2 X_{t-2} + \varepsilon_t + \beta\varepsilon_{t-1} \\ &= \dots \\ &= \alpha(1+\beta+\ldots+\beta^{L-1}) + \beta^L X_{t-L} + \varepsilon_t + \beta\varepsilon_{t-1} + \ldots + \beta^{L-1}\varepsilon_{t-L+1} \\ &= \mu(1-\beta^L) + \beta^L X_{t-L} + \sqrt{\Gamma_0(1-\beta^{2L})}\,\xi_t \end{aligned} \tag{19}$$

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

- This implies that

$$\begin{aligned} \mathrm{E}(X_t|X_{t-L}) &= \mu(1-\beta^L) + \beta^L X_{t-L}, \\ \mathrm{Var}(X_t|X_{t-L}) &= \Gamma_0(1-\beta^{2L}). \end{aligned} \tag{20}$$

Autoregressive model *AR*(1)

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

$$X_t \approx \mu + \sqrt{\Gamma_0}\,\xi_t. \tag{21}$$

- In other words, the *AR*(1) model describes a *mean reverting* time series.
- After a large number of observations, X_t takes the form (21), 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 (12)?

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* $\mathcal{L}(\theta|y)$.
- This function 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.

Maximum likelihood estimation

- In particular, if the observations y_i are independent, then

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{N} p(y_i|\theta), \tag{22}$$

  where $p(y_i|\theta)$ denotes the PDF of a single observation.
- The value θ^* that maximizes $\mathcal{L}(\theta|y)$ serves as the best fit between the model specification and the data.

Maximum likelihood estimation

- It is usualy more convenient to consider the (negative) *log likelihood 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

$$\mathcal{L}(\theta|y) = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{N} \exp\Big(- \frac{(y_i - \mu)^2}{2\sigma^2} \Big), \tag{23}$$

and the LLF is

$$-\log \mathcal{L}(\theta|y) = \frac{1}{2} N \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 + const. \tag{24}$$

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

$$\begin{aligned} \mu^* &= \frac{1}{N}\sum_{i=1}^{N} y_i \,, \\ (\sigma^*)^2 &= \frac{1}{N}\sum_{i=1}^{N} (y_i - \mu^*)^2 \,. \end{aligned} \tag{25}$$

  respectively.
- Note that, while μ^* is *unbiased*, the estimator σ^* is *biased* (N in the denominator above, rather than the usual $N-1$).

Maximum likelihood estimation

- The fact that the MLE estimator of a parameter is biased is a common occurrence: typically MLE estimators are biased.
- One can show[2], 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.

[2] under suitable assumptions

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 $p(x_1, \ldots, x_T | x_0, \theta)$ (the *conditional* MLE method).
- Let

$$\widehat{\varepsilon}_t = x_t - \alpha - \beta x_{t-1}, \tag{26}$$

for $t = 1, \ldots, T$, be the residuals implied from the data.

- According to the model specification, each $\widehat{\varepsilon}_t$ is independently drawn from $N(0, \sigma^2)$, and thus

$$\begin{aligned} p(x_1, \ldots, x_T | x_0, \theta) &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2 \right) \\ &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} (x_t - \alpha - \beta x_{t-1})^2 \right). \end{aligned} \tag{27}$$

MLE for *AR*(1)

- Hence the LLF is given by

$$-\log \mathcal{L}(\theta|y) = \frac{1}{2} T \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{t=0}^{T-1} (x_{t+1} - \alpha - \beta x_t)^2 + const. \tag{28}$$

- Minimizing this function yields:

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} T & \sum_{t=0}^{T-1} x_t \\ \sum_{t=0}^{T-1} x_t & \sum_{t=0}^{T-1} x_t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=0}^{T-1} x_{t+1} \\ \sum_{t=0}^{T-1} x_t x_{t+1} \end{pmatrix}, \tag{29}$$

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

MLE for $AR(1)$

- This can also be explicitly rewritten as

$$\widehat{\beta} = \frac{\sum_{t=0}^{T-1}(x_t - \widehat{x})(x_{t+1} - \widehat{x}_+)}{\sum_{t=0}^{T-1}(x_t - \widehat{x})^2}, \tag{30}$$
$$\widehat{\alpha} = \widehat{x}_+ - \widehat{\beta}\widehat{x},$$

where

$$\widehat{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t, \qquad \widehat{x}_+ = \frac{1}{T}\sum_{t=0}^{T-1} x_{t+1}. \tag{31}$$

MLE for *AR*(1)

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

$$p(x_0|\theta) = \sqrt{\frac{1-\beta^2}{2\pi\sigma^2}} \exp\Big(-\frac{(x_0 - \alpha/(1-\beta))^2}{2\sigma^2/(1-\beta^2)}\Big). \tag{32}$$

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

$$p(x_t|x_{t-1}, \ldots, x_1, \theta) = \frac{1}{2\pi\sigma^2} \exp\Big(-\frac{(x_t - \alpha - \beta x_{t-1})^2}{2\sigma^2}\Big). \tag{33}$$

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

$$p(x_0, x_1, \ldots, x_T|\theta) = p(x_0|\theta) \prod_{t=1}^{T} p(x_t|x_{t-1}, \ldots, x_1, \theta). \tag{34}$$

MLE for $AR(1)$

- Therefore, the LLF is given by

$$-\log \mathcal{L}(\theta|x) = \frac{1}{2}\log\frac{\sigma^2}{1-\beta^2} + \frac{1}{2}T\log\sigma^2 + \frac{(x_0-\alpha/(1-\beta))^2}{2\sigma^2/(1-\beta^2)} + \frac{1}{2\sigma^2}\sum_{t=1}^{T}(x_t-\alpha-\beta x_{t-1})^2 + const. \tag{35}$$

- 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.

Second order autoregressive model *AR*(2)

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

$$X_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t, \tag{36}$$

  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.

Second order autoregressive model $AR(2)$

- From the requirement that $\mathrm{E}(X_t) = \mu$,

$$\mu = \frac{\alpha}{1 - \beta_1 - \beta_2}, \tag{37}$$

and so we can can rewrite (36) in the following form:

$$X_t - \mu = \beta_1(X_{t-1} - \mu) + \beta_2(X_{t-2} - \mu) + \varepsilon_t. \tag{38}$$

- We note (as in the case of $AR(1)$) that X_{t-1} and X_{t-2} are independent of ε_t.
- Multiplying (38) by $X_{t-j} - \mu$, for $j = 0, 1, 2$, and calculating expectations, we find that

$$\Gamma_k = \begin{cases} \beta_1\Gamma_1 + \beta_2\Gamma_2 + \sigma^2, & \text{if } k = 0, \\ \beta_1\Gamma_{k-1} + \beta_2\Gamma_{k-2}, & \text{if } k = 1, 2. \end{cases} \tag{39}$$

Second order autoregressive model $AR(2)$

- This identity is called the *Yule-Walker equation* for the autocovariance.
- Dividing (39) by Γ_0 yields the Yule-Walker equation for the autocorrelation:

$$R_k = \beta_1 R_{k-1} + \beta_2 R_{k-2}, \tag{40}$$

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

$$R_1 = \frac{\beta_1}{1 - \beta_2}. \tag{41}$$

Second order autoregressive model $AR(2)$

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

$$R_2 = \beta_2 + \frac{\beta_1^2}{1 - \beta_2}. \tag{42}$$

- Finally, substituting $k = 0$ in (38) yields

$$\Gamma_0 = (\beta_1 R_1 + \beta_2 R_2)\Gamma_0 + \sigma^2. \tag{43}$$

- Solving this, we obtain

$$\Gamma_0 = \frac{(1 - \beta_2)\sigma^2}{(1 + \beta_2)((1 - \beta_2)^2 - \beta_1^2)}. \tag{44}$$

- One can show that the PACF of the $AR(2)$ model satisfies

$$\alpha_k = 0, \text{ for all } k \geq 3. \tag{45}$$

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:

$$LX_t = X_{t-1}. \tag{46}$$

- 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:

$$L^k X_t = X_{t-k}. \tag{47}$$

- We refer to L^k as the k-th power of L.

Lag operators and characteristic roots

- 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

$$\psi(L) = \psi_0 + \psi_1 L + \ldots + \psi_n L^n. \tag{48}$$

- For example, if $\psi(z) = a + bz$, then $\psi(L)X_t = aX_t + bX_{t-1}$.
- Notice that equation (36) can be stated as

$$\psi(L)X_t = \alpha + \varepsilon_t, \tag{49}$$

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

Lag operators and characteristic roots

- Solving this equation amounts to finding the *inverse* $\psi(L)^{-1}$ of $\psi(L)$, i.e. an operator which satisfies $\psi(L)\psi(L)^{-1} = I$ and $\psi(L)^{-1}\psi(L) = I$.
- Then

$$X_t = \frac{\alpha}{\psi(1)} + \psi(L)^{-1}\varepsilon_t. \tag{50}$$

- In order to determine $\psi(L)^{-1}$, we assume that it can be written as an infinite power series

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

with

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \tag{52}$$

Lag operators and characteristic roots

- Then

$$X_t = \frac{\alpha}{\psi(1)} + \sum_{j=0}^{\infty} \gamma_j \varepsilon_{t-j}, \tag{53}$$

with

$$\mathrm{E}(X_t) = \frac{\alpha}{\psi(1)}, \tag{54}$$

and

$$\mathrm{Cov}(X_t, X_{t+k}) = \sum_{j=0}^{\infty} \gamma_j \gamma_{j+k}, \text{ for } k \geq 0, \tag{55}$$

independently of t.

- Consequently, if the inverse $\psi(L)^{-1}$ exists, the time series X_t is covariance-stationary.

Lag operators and characteristic roots

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

$$(1 - \beta L)^{-1} = \sum_{j=0}^{\infty} \beta^j L^j, \tag{56}$$

- In other words,

$$(1 - \beta L)^{-1} X_t = \sum_{j=0}^{\infty} \beta^j X_{t-j}, \tag{57}$$

- Condition (52) 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, consider a polynomial $\psi(z)$ of degree n, whose non-zero roots are $z_1, \ldots, z_n$, i.e. $\psi(z) = \psi_n \prod_{j=1}^{n}(z - z_j)$.
- Then

$$\psi(L) = c \prod_{j=1}^{n}(1 - z_j^{-1} L), \tag{58}$$

  where c is the constant $c = (-1)^n \psi_n \prod_{j=1}^{n} z_j$.
- If each of the roots z_j (they may be complex) lies outside of the unit circle, i.e. $|z_j| > 1$, then we can invert $\psi(L)$ by applying (56) to each factor in (58).
- It is possible to verify that the convergence criterion (52), 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* (49) *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:

$$X_t = \alpha + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon_t, \tag{59}$$

  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

$$\mu = \frac{\alpha}{1 - \beta_1 - \ldots - \beta_p}. \tag{60}$$

- Furthermore, we require that the roots of the characteristic polynomial $\psi(z) = 1 - \beta_1 z - \ldots - \beta_p z^p$ lie outside of the unit circle.
- We can rewrite (59) in the following form:

$$X_t - \mu = \beta_1 (X_{t-1} - \mu) + \ldots + \beta_p (X_{t-p} - \mu) + \varepsilon_t. \tag{61}$$

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:

$$\Gamma_k = \begin{cases} \beta_1 \Gamma_1 + \cdots + \beta_p \Gamma_p + \sigma^2, & \text{if } k = 0, \\ \beta_1 \Gamma_{k-1} + \ldots + \beta_p \Gamma_{k-p}, & \text{if } k = 1, \ldots, p. \end{cases} \tag{62}$$

- Dividing (62) by Γ_0 yields the Yule-Walker equation for the autocorrelation:

$$R_k = \beta_1 R_{k-1} + \ldots + \beta_p R_{k-p}, \tag{63}$$

  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.
- The PACF α_k has the property that

$$\alpha_k = 0, \text{ for all } k \geq p + 1. \tag{64}$$

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

- In practice, we face the issue of identifying the number of lags p (assuming that the model $AR(p)$ is appropriate).
- This can be done by first calculating the PACFs α_k, and assessing whether it rapidly drops to zero beyond a certain threshold $k = p$.
- That threshold may be the appropriate number of lags and warrants further investigation.
- Various information criteria may be helpful in the decision making process, and prevent model overfitting ("torture it until it confesses") by adding too many lags.

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

- The *Akaike information criterion* defined as follows:

$$\mathrm{AIC} = 2k - 2\log \mathcal{L}(\widehat{\theta}|x). \tag{65}$$

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

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

- Another popular information criteria is the *Bayesian information criterion* (a.k.a the *Schwarz criterion*), which is defined as follows:

$$\mathrm{BIC} = \log(N)k - 2\log\mathcal{L}(\widehat{\theta}|x), \tag{66}$$

  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:

$$X_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}., \tag{67}$$

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

$$\mathrm{E}(X_t) = \mu, \tag{68}$$

  as $\mathrm{E}(\varepsilon_t) = 0$, for all t.

Moving average model *MA*(1)

- Its variance is

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

- For the first autocovariance, we have

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

Moving average model *MA*(1)

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

$$\Gamma_t = \begin{cases} (1+\theta^2)\sigma^2, & \text{if } t = 0, \\ \theta\sigma^2, & \text{if } |t| = 1, \\ 0, & \text{if } |t| \geq 2. \end{cases} \tag{69}$$

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

$$R_1 = \frac{\theta}{1+\theta^2}\,, \tag{70}$$

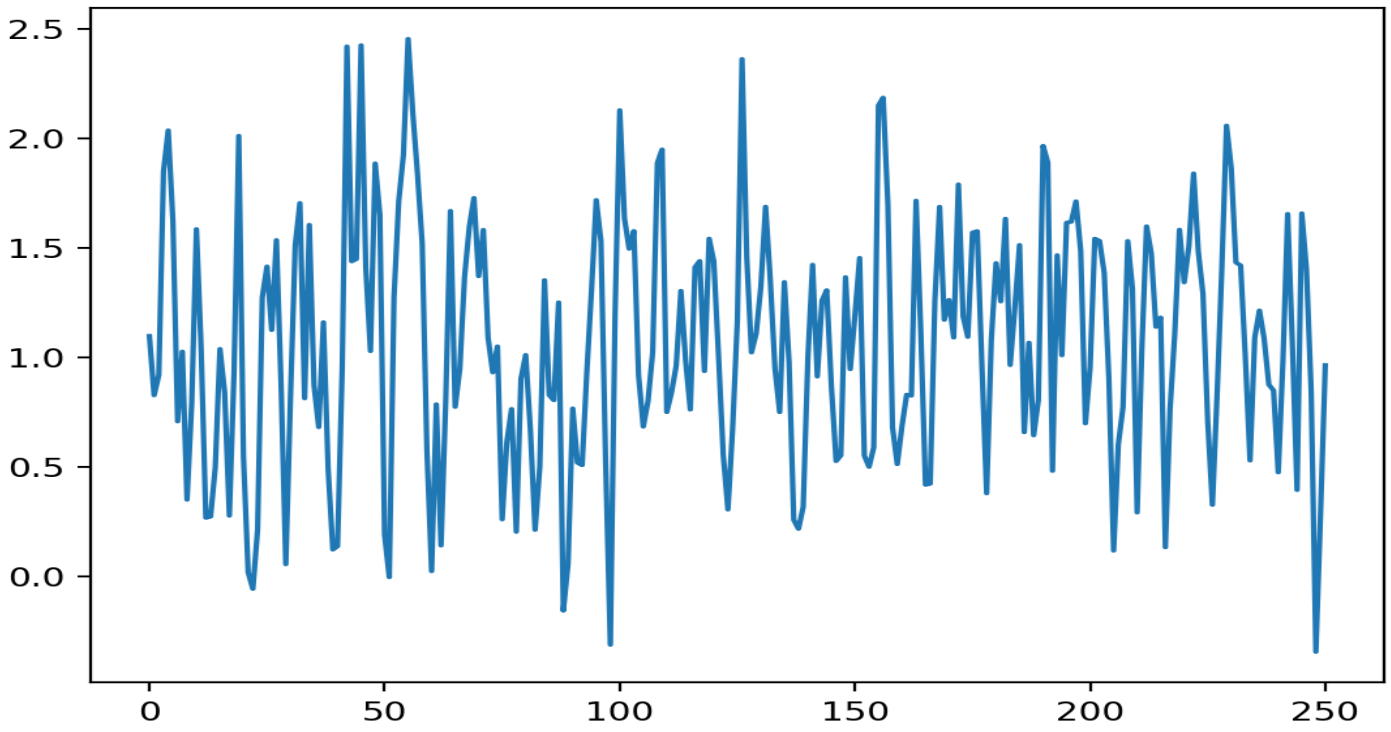
with all higher order autocorrelations equal zero:

$$R_k = 0, \text{ for all } k \geq 2. \tag{71}$$

- The PACF α_k goes to zero as $k \to \infty$, albeit at a slower rate.

Moving average model *MA*(1)

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



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 $t = 1, \ldots, T$,

$$\varepsilon_t = x_t - \mu - \theta\varepsilon_{t-1}, \tag{72}$$

and so the conditional PDF of x_t is

$$p(x_t|x_{t-1}, \ldots, x_1, \varepsilon_0 = 0, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right). \tag{73}$$

- This expression is deceivingly simply: in reality ε_t is a nested function of all x_s with $s \leq t$.

MLE for *MA*(1)

- The likelihood function of the sample $x_1, \ldots_T$ is given by the product of the probabilities above, and so

$$\mathcal{L}(\theta|x, \varepsilon_0 = 0) = \prod_{t=1}^{T} p(x_t|x_{t-1}, \ldots, x_1, \varepsilon_0 = 0, \theta), \tag{74}$$

- The log likelihood has thus the following form:

$$-\log \mathcal{L}(\theta|x, \varepsilon_0 = 0) = \frac{1}{2} T \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2 + const. \tag{75}$$

- 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.

MLE for *MA*(1)

- 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

$$p(x|\theta) = \frac{1}{(2\pi)^{T/2}\det(\Omega)^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)\Big), \tag{76}$$

and thus

$$-\log\mathcal{L}(\theta|x) = \frac{1}{2}\log\det(\Omega) + \frac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu). \tag{77}$$

MLE for *MA*(1)

- Here, Ω is a band diagonal matrix:

$$\Omega = \sigma^2 \begin{pmatrix} 1+\theta^2 & \theta & 0 & \dots & 0 \\ \theta & 1+\theta^2 & \theta & \dots & 0 \\ 0 & \theta & 1+\theta^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 1+\theta^2. \end{pmatrix} \tag{78}$$

- The numerics of minimizing (77) 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 $|\theta| \geq 1$.

General moving average model *MA*(q)

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

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}., \tag{79}$$

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

$$\mathrm{E}(X_t) = \mu, \tag{80}$$

  while its autocovariance is

$$\Gamma_j = \begin{cases} (1 + \theta_1^2 + \ldots + \theta_q^2)\sigma^2, & \text{if } j = 0, \\ (\theta_j + \theta_{j+1}\theta_1 + \ldots + \theta_q\theta_{q-j})\sigma^2, & \text{if } j = 1, \ldots, q, \\ 0, & \text{if } j > q. \end{cases} \tag{81}$$

ARMA(p, q) model

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

$$X_t = \alpha + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}, \tag{82}$$

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

- The equation above has the following lag operator representation:

$$\psi(L) X_t = \alpha + \varphi(L) \varepsilon_t, \tag{83}$$

where

$$\begin{aligned} \psi(z) &= 1 - \beta_1 z - \ldots - \beta_p z^p, \\ \varphi(z) &= 1 + \theta_1 z + \ldots + \theta_q z^q. \end{aligned} \tag{84}$$

- The process (64) is covariance stationary if the roots of ψ lie outside of the unit circle.

ARMA(p, q) model

- In this case, we can write the model in the form

$$X_t = \mu + \gamma(L)\varepsilon_t, \tag{85}$$

where $\mu = \alpha/\psi(1)$, and $\gamma(L) = \psi(L)^{-1}\varphi(L)$.

- Explicitly, $\gamma(L)$ is an infinite series:

$$\gamma(L) = \sum_{j=0}^{\infty} \gamma_j L^j, \tag{86}$$

with

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \tag{87}$$

- This form of the model specification is called the *moving average form*.
- Notice that the noise term in (85) is an example of colored noise.

ARMA(p, q) model

- There are various methods for estimating the parameters of an *ARMA*(p, q) model.
- The MLE method follows the steps similar to those described for the *AR*(p) and *MA*(q) models.
- A simpler method is the *two step regression estimation*, which proceeds as follows.
- Suppose we are given a set of data $\{x_0, x_1, \ldots, x_T\}$.

ARMA(p, q) model

- *Step 1*. Run the least square regression:

$$x_t = \pi_0 + \pi_1 x_{t-1} + \ldots + \pi_p x_{t-p} + \varepsilon_t, \text{ for } t = p, \ldots, T, \tag{88}$$

where ε_t denote the residuals.

- Let $\widehat{\pi}_i$ denote the estimated parameters, and compute the residuals of this regression:

$$\widehat{\varepsilon}_t = x_t - \widehat{\pi}_0 - \widehat{\pi}_1 x_{t-1} - \ldots - \widehat{\pi}_p x_{t-p}. \tag{89}$$

- *Step 2*. Run the second regression:

$$x_t = \beta_0 + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \theta_1 \widehat{\varepsilon}_{t-1} \ldots + \theta_q \widehat{\varepsilon}_{t-q}. \tag{90}$$

- The estimated parameters $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q$ are consistent estimators of the parameters in (64).
- Analysis of the ACF and PACF, as well as the information criteria, such as AIC or BIC, remain useful quantitative guides for model selection in the case of *ARMA*(p, q) models.

Mean squared error forecasting

- 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

$$\mathrm{E}\big((X_{t+1} - X^*_{t+1|1:t})^2\big). \tag{91}$$

Mean squared error forecasting

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

$$X^*_{t+1|1:t} = E_t(X_{t+1}). \tag{92}$$

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

$$\mathrm{E}_t(\cdot) = \mathrm{E}(\,\cdot\,|X_{1:t}). \tag{93}$$

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

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

Mean squared error forecasting

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

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

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

Mean squared error forecasting

- As a result

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

  which has its minimum at $Z = E_t(X_{t+1})$.
- This proves (92).
- The argument above is, in fact, quite general, and it easily extends to general k-period forecasts $X^*_{t+k|1:t}$.
- Minimizing the corresponding MSE yields:

$$X^*_{t+k|1:t} = E_t(X_{t+k}). \tag{94}$$

- Later we will generalize this method to time series models with more complex structure.

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

- As an example, a single period forecast in an $AR(1)$ model is

$$\begin{aligned} X^*_{t+1|1:t} &= \mathrm{E}_t(X_{t+1}) \\ &= \mathrm{E}_t(\alpha + \beta X_t + \varepsilon_{t+1}) \\ &= \alpha + \beta X_t. \end{aligned} \tag{95}$$

- 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

$$X^*_{t+1|1:t} = \alpha + \beta_1 X_t + \ldots + \beta_p X_{t-p+1}. \tag{96}$$

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

$$\begin{aligned} X^*_{t+2|1:t} &= \mathrm{E}_t(X_{t+2}) \\ &= \mathrm{E}_t(\alpha + \beta X_{t+1} + \varepsilon_{t+2}) \\ &= (1+\beta)\alpha + \beta^2 X_t. \end{aligned} \tag{97}$$

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

$$\begin{aligned} X^*_{t+1|1:t} &= \mathrm{E}_t(X_{t+1}) \\ &= \mathrm{E}_t(\mu + \varepsilon_{t+1} + \theta\varepsilon_t) \\ &= \mu + \theta\varepsilon_t. \end{aligned} \tag{98}$$

- The forecast error is ε_{t+1}, and its variance is σ^2.

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

- These calculations can be generalized to produce a general expression 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 [2].

References


[1] Brockwell, P. J. and Davis, R. A.: *Introduction to Time Series and Forecasting*, Springer (2016).


[2] Brockwell, P. J. and Davis, R. A.: *Time Series: Theory and Methods*, Springer (2009).


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