21

TIME SERIES ANALYSIS

21.1 Overview of time series analysis

21.1.1 Introduction to time series

A **time series** is a set of observations, x_t, each made at a specific time or period. A time series usually exhibits variation and fluctuation across time. The source of these dynamical changes can be classified as[1, p. 10]:

- **seasonal effect** that patterns repeating over **known, fixed periods** of time.
- **trend** that are long term change in the mean.
- **irregular fluctuations** due to disturbance.

In Figure 21.1.1, we show example time series include a white noise process, a seasonable time series with periodicity 20, and social science data.

We are in large part concerned with time series with stochastic disturbance and thus adopt a probabilistic approach to modeling time series. On a high level, we let random variables $X_1, ..., X_t$ denote the observation and models various quantities of interest, including the joint distributions of $(X_1, ..., X_k)$, conditional distributions like $P(X_t|X_{t-1}, ..., X_{t-K})$, and different moments of the joint distribution, such as $E[X_t]$, $E[X_t X_{t+h}]$, etc.

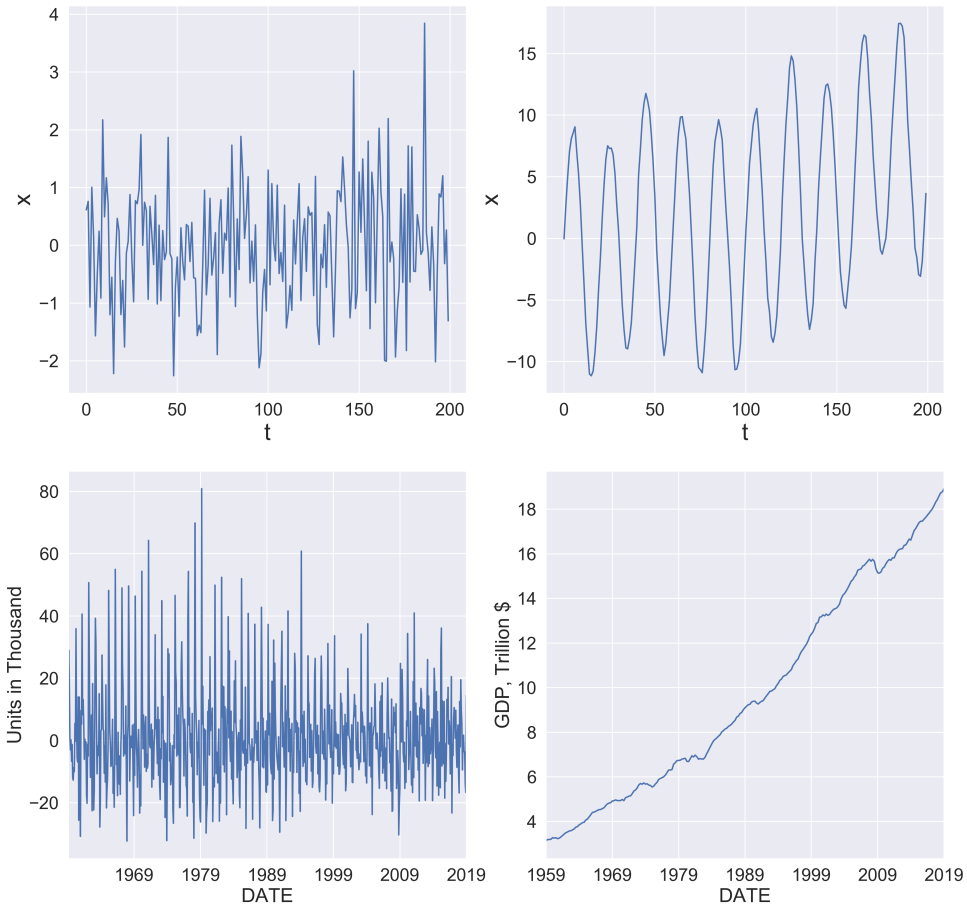

Figure 21.1.1: Example time series including a white noise process (upper left), a seasonable time series with periodicity 20, the US new privately owned housing [source], and the US GDP time series[source].

21.1.2 Stationarity

21.1.2.1 *Stationarity concept*

We modeling efforts are mainly directed toward stationary time series. Although in practical application most time series are not stationary, there are a number of tools [subsection 21.1.3] to convert them to stationary time series.

Definition 21.1.1 (mean and covariance function). *Let $\{X_t\}$ be a discrete time series with $E[X_t^2] < \infty$. The mean function of $\{X_t\}$ is*

$$\mu_X(t) = E[X_t]$$

The covariance function of $\{X_t\}$ is

$$\gamma_X(r,s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for all integer r, s.

Definition 21.1.2 (weakly and strictly stationary). *[2, p. 15] Let $\{X_t\}$ be a discrete time series with $E[X_t^2] < \infty$. $\{X_t\}$ is **weakly stationary** if*

- *its mean function $\mu_X(t)$ is independent of t,*
- *its covariance function $\gamma_X(t+h, t)$ is independent of t for each h.*

Further, $\{X_t\}$ is a strictly stationary time series if the joint distributions of $(X_1, ..., X_n)$ and joint distribution of $(X_{1+h}, ..., X_{n+h})$ are equal for all integers h and $n \geq 1$. Or we write

$$P(X_1, ..., X_n) = P(X_{1+h}, ..., X_{n+h}).$$

Lemma 21.1.1 (properties of strictly stationary process). *[2, p. 49] Let $\{X_t\}$ be a strictly stationary time series, then we have*

1. *The random variable X_t are identically distributed;*
2. *$P(X_t, X_{t+h}) = P(X_1, X_{1+h}$ for all integers t and h;*
3. *$\{X_t\}$ is weakly stationary if $E[X_t^2] < \infty$ for all t;*
4. *Weak stationarity does not imply strict stationarity*
5. *All iid sequence is strictly stationary.*

Proof. (1)(2)(3) Directly from definition, joint distribution equal implies marginal distribution equal. □

Remark 21.1.1. A stationary process will not contain trends and periodic trends.A stationary process will not contain periodic/seasonal change and trends.

Finally, we define autocovariance and autocorrelation of a stationary time series and give their properties.

Definition 21.1.3 (autocovariance and autocorrelation of a stationary time series). *Let $\{X_t\}$ be a stationary time series. The autocovariance function of $\{X_t\}$ at lag h is*

$$\gamma_X(h) = Cov(X_{t+h}, X_t)$$

The autocorrelation function of $\{X_t\}$ at lag h is

$$\rho_X(h) = \frac{Cov(X(h), X(0))}{\sqrt{Cov(X(h), X(h)}\sqrt{Cov(X(0), X(0))}} = \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t)$$

Lemma 21.1.2 (basic property of $\gamma(h)$). *For a stationary $\{X_t\}$, its autocovariance function*

- $\gamma(0) \geq 0$
- $\left|\gamma(h)\right| \leq \gamma(0), \forall h$
- $\gamma(h)$ *is even, i.e.,* $\gamma(h) = \gamma(-h), \forall h$.

Proof. (1) It is variance. (2) From Cauchy inequality [Corollary 11.9.4.1]. (3) $\gamma(h) = Cov(X_t, X_{t+h}) = Cov(X_{t+h}, X_t) = \gamma(-h)$. □

Example 21.1.1.

- A white noise process $Z_t \sim N(0,1)$ is a strictly stationary (therefore weakly stationary) time series. It has zero mean and $\gamma_X(h) = 1$ if $h = 0$ and $\gamma_X(h) = 0$ if $h > 0$.
- A process $X_t = \sin(t) + Z_t, Z_t \sim N(0,1)$ is not a stationary process because it does not have constant mean. In general, a stationary process will not contain trends and periodic/seasonal trends.
- A process $X_t = 0.9X_t + Z_t, Z_t \sim N(0,1)$ is a weakly stationary process but not a strictly stationary process. It has zero mean and $\gamma_X(h) = 0.9^h, h \geq 0$ [Lemma 21.2.5].

21.1.2.2 *Rolling analysis*

The most salient characteristics of a weakly stationary time series is the constancy of mean and variance. Consider the analysis of a univariate time series y_t over a sample from $t = 1, ..., T$. We can evaluate the mean, variance, and volatility constancy over the entire sample using the following rolling analysis.

Definition 21.1.4 (rolling analysis). *Consider the analysis of a univariate time series y_t over a sample from $t = 1, 2, ..., T$. Let n denote the width of a sub-sample or window and defien the **rolling** sample means, variances and standard deviation*

$$\hat{\mu}_t(n) = \frac{1}{n}\sum_{i=0}^{n-1} y_{t-i}$$
$$\hat{\sigma}_t^2(n) = \frac{1}{n}\sum_{i=0}^{n-1}(y_{t-i} - \hat{\mu}_t(n))^2$$
$$\hat{\sigma}_t(n) = \sqrt{\hat{\sigma}_t^2(n)}$$

21.1.3 Remove trend and seasonality

In face of a time series with trend and seasonality, we can remove trend and seasonality component using following decomposition model[2, p. 31]:

$$X_t = Y_t + s_t + m_t,$$

where m_t is the trend process, Y_t is the residual that we require to have zero mean $E[Y_t] = 0$, and s_t is the seasonal component required to satisfy $s_{t+d} = s_t$ and $\sum_{j=1}^{d} s_j = 0$.

The most commonly way estimate the trend $\hat{m}_t$ in a time series is to apply a **moving average linear filter** with window size $2h + 1$ defined by [1]

$$\hat{m}_t(n) = \frac{1}{2h+1}\sum_{i=n-h}^{n+1} X_{t+i}.$$

The filtering method can estimate the trend of a non-seasonal trending time series given by $X_t = m_t + Y_t, E[Y_t] = 0$, if m_t is linear within the window $[t - h, t + h]$. To show this, we have

$$\hat{m}_t(n) = \frac{1}{2h+1}\sum_{i=n-h}^{n+1} X_{t+i} = \frac{1}{2h+1}\sum_{i=n-h}^{n+1} m_t.$$

If X_t contains a seasonal component of period of d, then we should choose the window size to be d to remove both the noise and the seasonality effect.

1 for moving window of even length, the weight for left and right most observations are $1/2$.

Based on the assumption on the time series generation model, we can also apply one-sided linear filter, such as the **exponentially weighted moving averages(EWMA)** filter. An n period EWMA of a time series X_t is defined as

$$\hat{\mu}_t(n) = \sum_{i=0}^{n-1} w_i X_{t-i}, w_i = \frac{\lambda^{i-1}}{\sum_{i=0}^{n-1} \lambda^{i-1}},$$

where $0 < \lambda < 1$ is the decay parameter. Note that in EWMA, we put more weight on the most recent observations.

By **taking difference between consecutive observation** is alternative method to remove trend. For example,

- If $X_t = \beta_0 + \beta_1 t + E_t$, then

$$Y_t = X_t - X_{t-1} = \nabla X_t = \beta_1 + \nabla E_t$$

- If $X_t = \sum_{i=0} \beta_i t^i + E_t$, then

$$\nabla^k X_t = k!\beta_k + \nabla^k E_t$$

After estimating trend component $\hat{m}_t$ using moving average filter, we compute the average w_k of the deviations $\{(x_{k+jd} - \hat{m}_{k+jd}), q < k + jd \leq n - q\}$. Because the $\{w_k\}$ do not necessarily sum to zero, we can estimate the seasonal component s_k as

$$\hat{s}_k = w_k - d^{-1} \sum_{i=1}^{d} w_i, k = 1, ..., d.$$

Then the final *de-trend* and *deseasonalized* data is given by

$$\hat{Y}_t = X_t - \hat{m}_t - \hat{s}_t.$$

More robust and well-established algorithms to remove trend and seasonality include the STL[3] algorithm, which decompose a time series into three components: trend, season(al) and residual. STL uses LOESS (locally estimated scatter plot smoothing) to achieve smooths estimates of these three components[Figure 21.1.2].

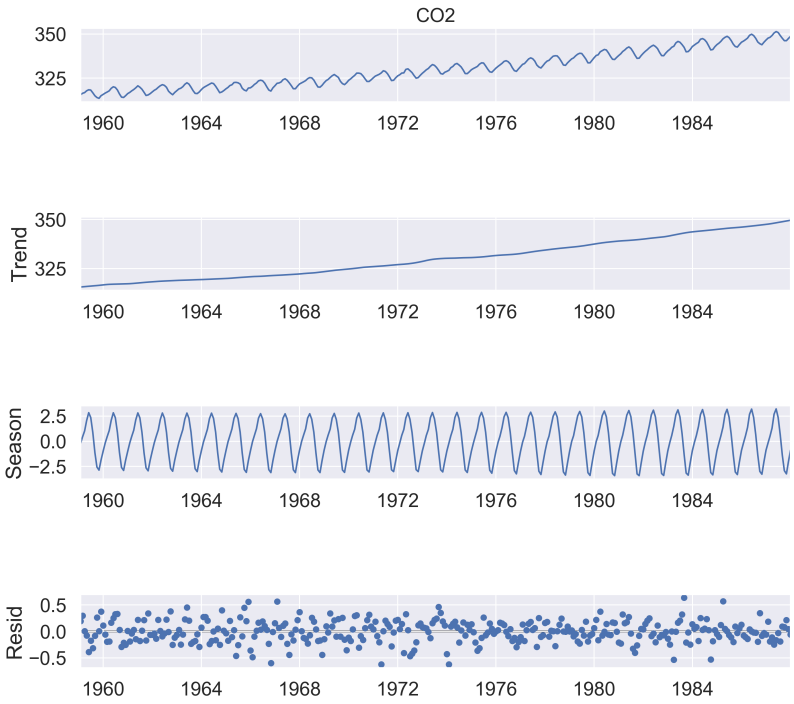

Figure 21.1.2: Demonstration on using STL to decompose an example CO_2 concentration time series.

21.2 Linear stationary process theory

21.2.1 Preliminaries: the lag operator and polynomial

Definition 21.2.1 (lag operator and lag polynomial). *A lag operator is defined as* $Lx_t = x_{t-1}$. *A lag polynomial is defined as*

$$P(L) = a_0 + a_1 L + a_2 L^2 + ... + a_p L^p.$$

An inverse of a lag polynomial is defined as as

$$P^{-1}(L)P(L) = P(L)P^{-1}(L) = 1.$$

Remark 21.2.1.

- The lag operator can operate(addition,subtraction,multiplication, division) like integers.

Lemma 21.2.1 (inverse of factor). *The inverse of* $P(L) = 1 - aL, a \in \mathbb{C}$ *exists when* $|a| < 1$*and the inverse is given as*

$$P^{-1}(L) = 1 + aL + a^2L^2 + a^3L^3 + ...$$

Proof. directly use the definition to verify $P(L)P^{-1}(L) = 1$ based on the fact of $a^\infty = 0$. □

Lemma 21.2.2 (factoring lag polynomial). *Any lag polynomial of degree q can be written as*

$$P(L) = (1 - \beta_1 L)(1 - \beta_2 L)...(1 - \beta_q L)$$

where $\beta_1, \beta_2, ..., \beta_q \in \mathbb{C}$ *are the q roots of* $P(x) = 0$.

Proof. Directly from fundamental theorem of algebra Theorem 4.18.3 □

Lemma 21.2.3 (inverse of general lag polynomial). *A lag polynomial* $P(L)$ *of degree q has its inverse if all its roots* $\beta_i \in \mathbb{C}$ *satisfying*

$$|\beta_i| < 1, i = 1, ..., q$$

and the inverse is given as:

$$P^{-1}(L) = (1 - \beta_1 L)^{-1} ... (1 - \beta_q L)^{-1}$$

where $\beta_1, \beta_2, ..., \beta_q$ are the q roots of $P(x) = 0$.

Proof. Based on the assumption, each factor exists and can be inverted(Lemma 21.2.1]). Then the finite product of these factor will exist. □

Definition 21.2.2 (difference operator). *The difference operator Δ is $1 - L$, i.e., $\Delta x_t = x_t - x_{t-1}$. $\Delta^2 = (1 - L)^2 = 1 - 2L + L^2$,i.e., $\Delta^2 x_t = x_t - 2x_{t-1} + x_{t-2}$*

Example 21.2.1. Let $y_t = \nabla x_t = (1 - L)x_t$, we can solve $x_t = (1 - L)^{-1} y_t = \sum_{-\infty}^{t} y_i$, or $x_t - x_0 = \sum_{i=1}^{t} y_i$

21.2.2 Linear process

Definition 21.2.3 (linear process). *[2, p. 51] The time series $\{X_t\}$ is a linear process if it has the representation*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \phi_j Z_{t-j}$$

for all t, where $\{Z_t\}$ $WN(0, \sigma^2)$ and $\{\phi_j\}$ is a sequence of constants with absolute summability $\sum_{j=-\infty}^{\infty} \left|\phi_j\right| < \infty$.

Lemma 21.2.4 (basic property of linear process, stationarity). *A linear process*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \phi_j Z_{t-j}$$

with $\sum_{j=-\infty}^{\infty} \left|\phi_j\right| < \infty$ has the following properties:

- $E[X_t] = \mu$.
- *stationary covariance*

$$Cov[X_t, X_{t+\tau}] = \sigma^2 \sum_{i=-\infty}^{\infty} \phi_i \phi_{i-\tau}.$$

That is, a linear process is weakly stationary.

Proof. (1) We are able to exchange the summation and the integral due to Theorem 3.9.8. To apply this theorem, note that $E[\![Z_t|]\!] = \sigma\sqrt{2/\pi}$ and $\sum_{j=-\infty}^{\infty} \sigma\sqrt{2/\pi}\left|\phi_j\right| < \infty$. (2)The convergence of $\sum_{i=-\infty}^{\infty} \phi_i\phi_{i-\tau}$ is discussed in Theorem 1.7.1 □

Definition 21.2.4 (causality). *A linear process $\{X_t\}$ is causal(strictly, a causal function of $\{W_t\}$) if there exists a*

$$\phi(B) = \phi_0 + \phi_1 B + \phi_2 B^2 + ...$$

with $\sum_{i=1}^{\infty}\left|\phi_i\right| < \infty$ and $X_t = \phi(B)W_t$, where B is the lag operator.

Definition 21.2.5 (invertibility). *A linear process $\{X_t\}$ is invertible(strictly, an invertible function of $\{W_t\}$) if there exists a*

$$\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + ...$$

with $\sum_{i=1}^{\infty}\left|\phi_i\right| < \infty$ and $W_t = \phi(B)X_t$, where B is the lag operator.

Theorem 21.2.1 (linear combination of stationary process is stationary). *[2, p. 52] Let Y_t be a stationary time series with mean 0 and covariance function γ_Y. If $\sum_{j=-\infty}^{\infty}\left|\psi_j\right| < \infty$, then the time series*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} = \psi(L)Y_t$$

is stationary with mean 0 and autocovariance function

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \psi_j\psi_k\gamma_Y(h+k-j)$$

In particular, if $Y(t)$ is the white noise process, then

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j\psi_{j+h}\sigma^2$$

Proof. (1) We are able to exchange the summation and the integral due to Theorem 3.9.8. To apply this theorem, note that $E[\![Y_t|]\!] = 0$ and $\sum_{j=-\infty}^{\infty}\left|\psi_j\right| < \infty$. (2) (3) Lemma 21.2.4. □

21.2.3 Autoregressive (AR) process

21.2.3.1 *Basics*

Autoregressive (AR) processes are processes where the observation is generated by the summation of its own history and shocks.We first introduce lag operator for retrogressive processes.

Definition 21.2.6 (autoregressive operator of order q). *[4, p. 11]The autoregressive operator of order q is defined as*

$$\phi_q(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_q B^q$$

where ϕ_i are constants.

The AR process with different order q is given as follows.

Definition 21.2.7 (AR process of order p).

- *A process $\{X_t\}$ is called an autoregressive(AR) process of order p if*

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + Z_t, Z_t \sim WN(0, \sigma^2)$$

  or

$$\theta_p(B) X_t = Z_t.$$

- *Specifically, $AR(1)$ process has the following form*

$$X_t = a_1 X_{t-1} + Z_t.$$

To gain some intuition for the AR processes, we demonstrate some representative realizations of AR(1) model with different choices of a in Figure 21.2.1. When $0 < a < 1$, the trajectory will have some memory effect. When $-1 < a < 0$, the trajectory will revert to the mean zero or oscillates around mean zero, similar to the OU process. Note that a stationary AR(1) process can be viewed as a discrete-time version of OU processes [section 19.4]. When $|a| > 1$, the trajectory will blow up in the long run.

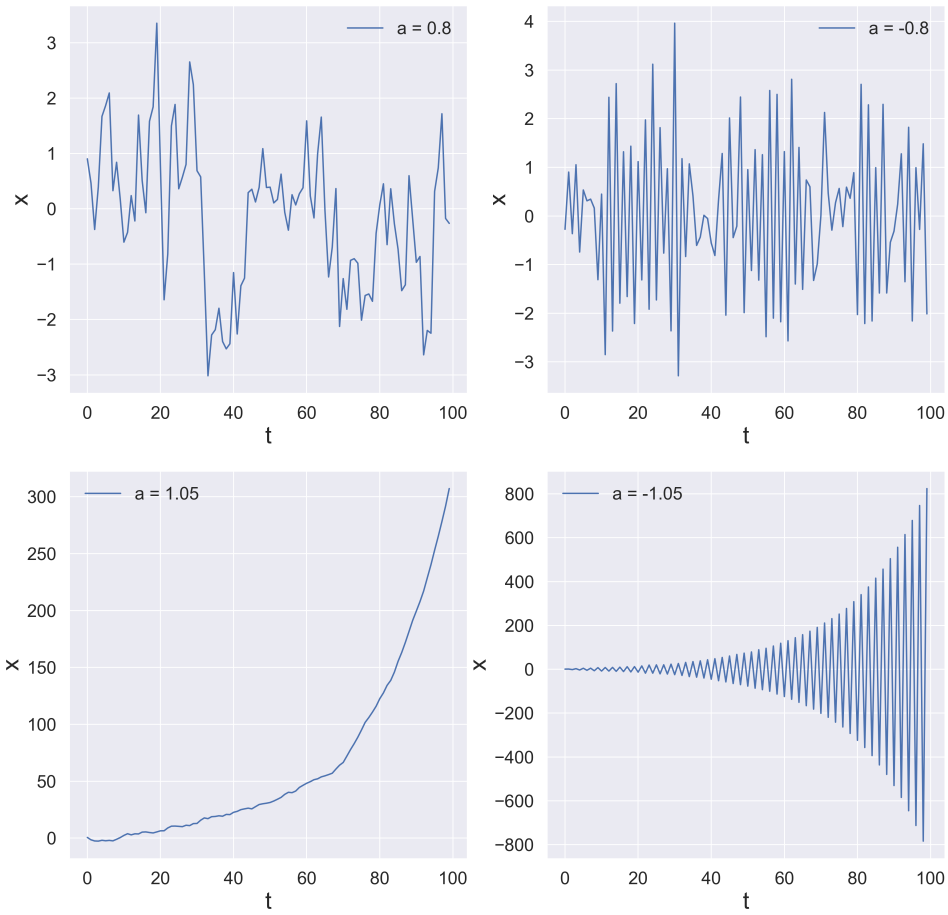

Figure 21.2.1: Example trajectories of AR(1) models ($X_t = aX_{t-1} + Z_t$.) with different choices of a.

Now we list basic properties of an AR(1) process. For other $AR(q)$ process, similar approaches can be used derived their properties.

Lemma 21.2.5 (AR(1) properties). *The first order autoregressive process $AR(1)$, given by,*

$$X_t = a_1 X_{t-1} + Z_t, Z_t \sim WN(0, \sigma^2), X_0 = x_0, |a| < 1,$$

has the following property:

- *It has solution*

$$X_t = a^t X_0 + \sum_{i=1}^{t} a^{i-1} Z_{t-i+1}$$

- *mean and variance*

$$\begin{aligned} E[X_t] &= 0 \\ Var[X_t] &= \sigma^2(1 + a^2 + a^4 + \cdots + a^{2(t-1)}) = \sigma^2 \frac{1 - a^{2t}}{1 - a^2} \\ Var[X_t] &\to \sigma^2/(1 - a^2), as\ t \to \infty \end{aligned}$$

- *covariance*

$$Cov[X_t X_s] = a^{t-s} Var[X_s], t \geq s$$

.

- $$\gamma(k) \triangleq \begin{cases} a^k Var(X_t), k \geq 0 \\ \gamma(-k), k < 0 \end{cases} .$$

- $$\rho(k) = \begin{cases} a^k, k \geq 0 \\ \rho(-k), k < 0 \end{cases} .$$

Proof. (1)

$$\begin{aligned} X_t &= aX_{t-1} + Z_t \\ &= a(aX_{t-2} + Z_{t-1}) + Z_t \\ &= a^2 X_{t-2} + aZ_{t-1} + Z_t \\ &= a^3 X_{t-3} + a^2 Z_{t-2} + aZ_{t-1} + Z_t \\ &= a^t X_0 + \sum_{i=1}^{t} a^{i-1} Z_{t-i+1} \end{aligned}$$

(2) Directly from (1). Use the result of variance of sum of normal random variables; (3)(4)(5)

$$\begin{aligned} & Cov[X_t X_s] \\ =& Cov[(a^{t-s}X_s + \sum_{j=1}^{t-s} a^{t-s-j} Z_{s+j}), X_s] \\ =& Cov[a^{t-s}X_s, X_s] \\ =& a^{t-s} Var[X_s] \end{aligned}$$

□

21.2.3.2 *Stationarity and invertibility condition*

Lemma 21.2.6 (stationarity condition for AR(1)). *[4, p. 54]The AR(1) process $X_t = \phi_1 X_{t-1} + Z_t$ has the necessary condition of stationarity of the roots of $1 - \phi_1 t = 0$ has roots lying outside the unit circle, i.e., $|t^*| > 1$; or equivalently, $|\phi_1| < 1$.*

Proof. In lag polynomial form, we have

$$(1 - \phi_1 B)X_t = Z_t \Rightarrow X_t = (1 - \phi_1 B)^{-1} Z_t = \sum_{i=0}^{\infty} \phi_1^i B^i Z_t$$

where we know that $|\phi_1| < 1$ is necessary for the variance to be finite, which is equivalent to the condition of $|t| = \left|\phi_1^{-1}\right| > 1$ from $|t\phi_1| = 1$, where t is the root of $1 - \phi_1 t = 0$. □

Remark 21.2.2. One example for $AR(1)$ to be nonstationary is the random walking process $X_t = X_{t-1} + W_t$; Even though the mean is 0, but its variance,i.e., $\gamma(0)$ is changing with time.

Lemma 21.2.7 (condition of stationarity). *[4, p. 54]*

- *The $AR(q)$ process $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_q X_{t-q} + Z_t$ is stationary of the roots of polynomial $\phi(t) = 1 - \phi_1 t - \phi_2 t^2 - \phi_3 t^3 ... = 0$ must all lying outside the unit circle,i.e., $|t^*| > 1$.*
- *Another necessary condition for stationary is*

$$|\phi_i| \leq 1, \forall i = 1, 2, ..., q.$$

Proof. (1) From Lemma 21.2.3. The condition ensures that AR(q) can be represented as a MA process, and further MA process are stationary [Lemma 21.2.13]. (2) □

Lemma 21.2.8 (AR process is always invertible). *[4, p. 57]An AR(q) process* $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_q X_{t-q} + Z_t$ *is alway intertible.*

Proof. From the definition of invertible process [Definition 21.2.5]. □

21.2.3.3 *Forecasting*

Lemma 21.2.9 (best least square forecasting for AR(1)). *[5, p. 55] Consider an AR(1) process given by*

$$X_t = a_1 X_{t-1} + Z_t + c_0, Z_t \sim WN(0, \sigma^2).$$

Let $\mathcal{F}_t$ *denote the information available upto* t*. It follows that*

- *the 1-step ahead best least-square forecast is*

$$\hat{X}_{t+1} \triangleq E[X_{t+1}|\mathcal{F}_t] = c_0 + a_1 X_t,$$

  and the error is $E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = \sigma^2$.
- *the 2-step ahead best least-square forecast is*

$$\hat{X}_{t+2} \triangleq E[X_{t+2}|\mathcal{F}_t] = c_0 + a_1 \hat{X}_{t+1},$$

  and the error is $E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = (1 + a_1^2)\sigma^2$.
- *the m-step*$(m > 2)$ *ahead best least-square forecast is*

$$\hat{X}_{t+m} \triangleq E[X_{t+m}|\mathcal{F}_t] = c_0 + a_1 \hat{X}_{t+m-1},$$

  and the error is

$$E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] = \sigma^2 + a_1^2 E[(X_{t+m-1} - \hat{X}_{t+m-1})^2|\mathcal{F}_t] = \sigma^2 \frac{a_1^m - 1}{a_1 - 1}.$$

Proof. We use the least-square minimizing property of conditional expectation [Lemma 11.7.6]. We have (1)

$$\begin{aligned}
\hat{X}_{t+1} &= E[Z_{t+1} + a_1 X_t + c_0|\mathcal{F}_t] \\
&= E[Z_{t+1}|\mathcal{F}_t] + E[a_1 X_t|\mathcal{F}_t] + E[c_0|\mathcal{F}_t] \\
&= 0 + a_1 X_t + c_0 \\
&= a_1 X_t + c_0
\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[Z_{t+1}^2|\mathcal{F}_t] = \sigma^2.$$

(2)

$$\begin{aligned}\hat{X}_{t+2} &= E[Z_{t+2} + a_1 X_{t+1} + c_0 | \mathcal{F}_t] \\ &= E[Z_{t+2}|\mathcal{F}_t] + E[a_1 X_{t+1}|\mathcal{F}_t] + E[c_0|\mathcal{F}_t] \\ &= 0 + a_1 \hat{X}_{t+1} + E[c_0|\mathcal{F}_t] \\ &= 0 + a_1(a_1 X_t + c_0) + c_0 \\ &= a_1^2 X_t + a_1 c_0 + c_0\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = E[(Z_{t+2} + a_1(X_{t+1} - \hat{X}_{t+1}))^2|\mathcal{F}_t] = (1 + a^2)\sigma^2.$$

(3) same as (1)(2). To get the error, we have

$$\begin{aligned}&E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] \\ &= E[(Z_{t+m} + a_1(X_{t+m-1} - E[X_{t+m-1}|\mathcal{F}_t])^2|\mathcal{F}_t] \\ &= \sigma^2 + a_1^2 E[(X_{t+m-1} - \hat{X}_{t+m-1})^2|\mathcal{F}_t]\end{aligned}$$

Continue the recursive relation and we will get the result. □

Lemma 21.2.10 (best least square forecasting for AR(q)). *[5, p. 55] Consider an AR(q) process given by*

$$X_t = a_1 X_{t-1} + ... + a_q X_{t-q} + Z_t + c_0, Z_t \sim WN(0, \sigma^2).$$

Let $\mathcal{F}_t$ denote the information available upto t. It follows that

- *the 1-step ahead best least-square forecast is*

$$\hat{X}_{t+1} \triangleq E[X_{t+1}|\mathcal{F}_t] = c_0 + a_1 X_t + ... + a_q X_{t-q+1},$$

  and the error is

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[Z_{t+1}^2|\mathcal{F}_t] = \sigma^2.$$

- *the 2-step ahead best least-square forecast is*

$$\hat{X}_{t+2} \triangleq E[X_{t+2}|\mathcal{F}_t] = c_0 + a_1 \hat{X}_{t+1} + a_2 X_t + ... + a_q X_{t-q+2},$$

  and the error is $E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = (1 + a_1^2)\sigma^2$.

- *the 3-step ahead best least-square forecast is*

$$\hat{X}_{t+3} \triangleq E[X_{t+3}|\mathcal{F}_t] = c_0 + a_1\hat{X}_{t+2} + a_2\hat{X}_{t+1} + a_3X_t + ... + a_qX_{t-q+3},$$

  and the error is $E[(X_{t+3} - \hat{X}_{t+3})^2|\mathcal{F}_t] = (1 + a_1^2 + (a_1^2 + a_2)^2)\sigma^2$.
- *the m-step($m > q$) ahead best least-square forecast is*

$$\hat{X}_{t+m} \triangleq E[X_{t+m}|\mathcal{F}_t] = c_0 + a_1\hat{X}_{t+m-1} + ... + a_q\hat{X}_{t+m-q}.$$

Proof. We use the least-square minimizing property of conditional expectation [Lemma 11.7.6]. We have (1)

$$\begin{aligned}\hat{X}_{t+1} &= E[Z_{t+1} + a_1X_t + c_0|\mathcal{F}_t] \\ &= E[Z_{t+1}|\mathcal{F}_t] + E[a_1X_t|\mathcal{F}_t] + E[c_0|\mathcal{F}_t] \\ &= 0 + a_1X_t + c_0 \\ &= a_1X_t + c_0\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[Z_{t+1}^2|\mathcal{F}_t] = \sigma^2.$$

(2)

$$\begin{aligned}\hat{X}_{t+2} &= E[Z_{t+2} + a_1X_{t+1} + c_0|\mathcal{F}_t] \\ &= E[Z_{t+2}|\mathcal{F}_t] + E[a_1X_{t+1}|\mathcal{F}_t] + E[c_0|\mathcal{F}_t] \\ &= 0 + a_1\hat{X}_{t+1} + E[c_0|\mathcal{F}_t] \\ &= 0 + a_1(a_1X_t + c_0) + c_0 \\ &= a_1^2X_t + a_1c_0 + c_0\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = E[(Z_{t+2} + a_1(X_{t+1} - \hat{X}_{t+1}))^2|\mathcal{F}_t] = (1 + a_1^2)\sigma^2.$$

(3)(4) same as (1)(2). To get the error, we have

$$\begin{aligned}E[(X_{t+3} - \hat{X}_{t+3})^2|\mathcal{F}_t] &= E[(Z_{t+3} + a_1(X_{t+2} - \hat{X}_{t+2}) + a_2(X_{t+1} - \hat{X}_{t+1}))^2|\mathcal{F}_t] \\ &= E[(Z_{t+3} + a_1Z_{t+2} + a_1(a_1(X_{t+1} - \hat{X}_{t+1})) + a_2(X_{t+1} - \hat{X}_{t+1}))^2|\mathcal{F}_t] \\ &= (1 + a_1^2 + (a_1^2 + a_2)^2)\sigma^2\end{aligned}$$

□

21.2.4 Moving average (MA) process

21.2.4.1 *Basics*

Moving average (MA) processes are processes where the observation is generated by moving average of shocks. We first introduce lag operator for moving average processes.

Definition 21.2.8 (moving average operator of order q). *[4, p. 10] A moving average operator of order q can be defined as*

$$\theta_q(B) = 1 + \theta_1 B + \theta_2 B + ... + \theta_q B^q.$$

The MA process with different order q is given as follows.

Definition 21.2.9 (MA(q) process). *[2, p. 50] A time series $\{X_t\}$ is a moving-average process of order q if*

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$, and $\theta_1, \theta_2, ..., \theta_q$ are constants.

To gain some intuition for the MA processes, we demonstrate some representative realizations of MA(1) and MA(2) model with different choices of θ in Figure 21.2.2. There are a couple of observations: First, trajectories will fluctuates around mean (i.e., zero value here). Second, when there are negative coefficients, trajectories tend to revert to its mean more often.

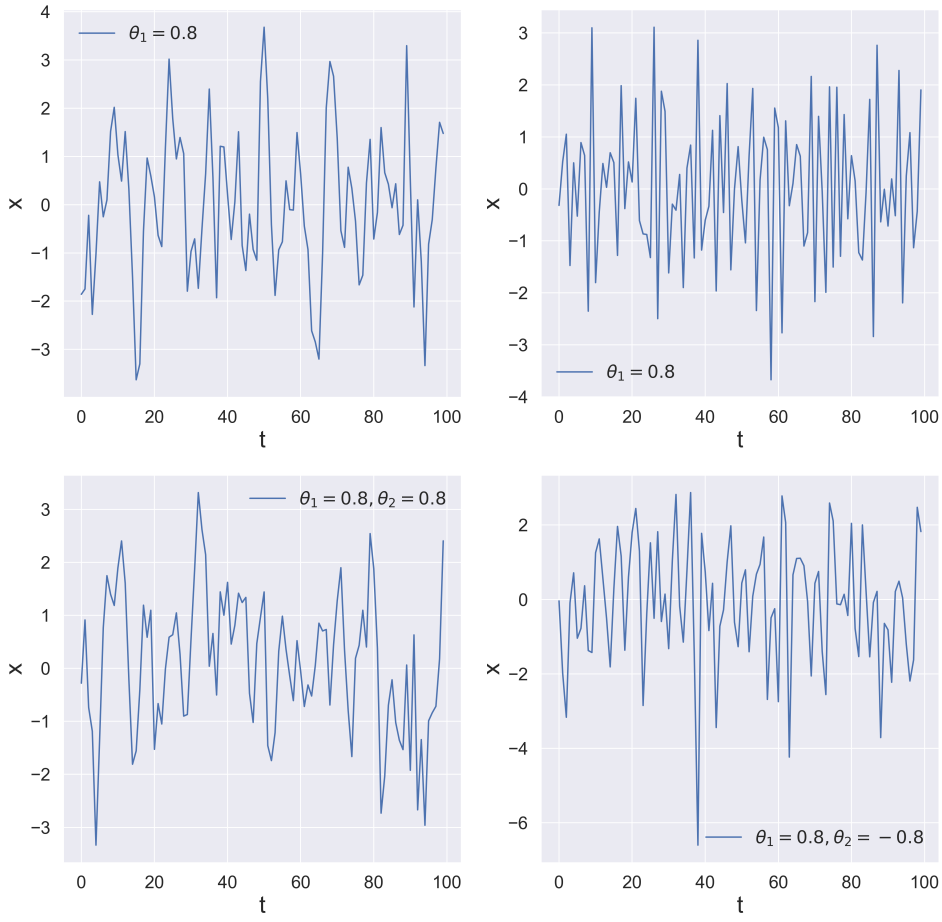

Figure 21.2.2: Example trajectories of MA(1) models (upper) and MA(2) models (lower).

Now we list basic properties of an MA(q) process.

Theorem 21.2.2 (basic statistics of MA(q)). *[1, p. 33] Given a MA process of order q,*

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q},$$

we have

$$E[X_t] = 0$$

$$Var(X_t) = \sigma^2 \sum_{i=0}^{q} \theta_i^2$$

and

$$\gamma(k) = \begin{cases} 0, k > q \\ \sigma^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k}, k = 0, 1, ..., q \\ \gamma(-k), k < 0 \end{cases},$$

and

$$\rho(k) = \begin{cases} 1, k = 0 \\ 0, k > q \\ \sigma^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k} / \sum_{i=0}^{q} \theta_i^2, k = 1, ..., q \\ \rho(-k), k < 0 \end{cases},$$

where $\theta_0 = 1$ is used.

Proof. (1)

$$E[X_t] = E[Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}] = E[Z_t] + \theta_1 E[Z_{t-1}] + ... + \theta_q E[Z_{t-q}] = 0.$$

(2)

$$Var[X_t] = Var[Z_t] + Var[\theta_1 Z_{t-1}] + ... + Var[\theta_q Z_{t-q}] + .cross.terms = \sigma^2 + \theta_1 \sigma^2 + ... + \theta_q^2 \sigma^2.$$

(3)

$$Cov(X_t, X_{t+k}) = Cov(Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}, Z_t + \theta_1 Z_{t-1+k} + ... + \theta_q Z_{t-q+k})$$

□

Corollary 21.2.2.1 (basic statistics of MA(1)). *Given a MA process of order* 1*,*

$$X_t = Z_t + \theta_1 Z_{t-1},$$

we have

$$E[X_t] = 0$$
$$Var(X_t) = \sigma^2(1 + \theta_1)$$
$$\gamma(1) = \sigma^2 \theta_1$$
$$\rho(1) = \theta_1 / (1 + \theta_1^2)$$

Theorem 21.2.3 (MA Representation Theorem). *[2, p. 50] If $\{X_t\}$ is a stationary q correlated time series with mean 0, then it can be represented as $MA(q)$ process.*

Proof. First, An $MA(q)$ process has exactly $q+1$ unknown parameters $(\sigma, \theta_1, ..., \theta_q)$, which can be determined from the autocovariance function of the given q correlated time series. Note that we can first uniquely determine θ_q by matching $\gamma(q)$, and then θ_{q-1}, and so on. □

Lemma 21.2.11 (sufficient condition for equivalent AR representation). *An $MA(1)$ process $X_t = W_t + \theta W_{t-1}$ can be represented as an AR process of infinite terms:*

$$X_t = \sum_{j=1}^{\infty} -(-\theta)^j X_{t-j} + W_t$$

if $|\theta| < 1$.

Proof.

$$X_t = W_t + \theta W_{t-1} = (1 + \theta B) W_t$$

then if $|\theta| < 1$, we have

$$W_t = (1 + \theta B)^{-1} X_t = \sum_{j=0}^{\infty} (-\theta)^j B^j X_t$$

□

21.2.4.2 *Stationarity and invertibility*

Lemma 21.2.12 (invertibility of MA). *An $MA(q)$ process given as*

$$X_t = \theta_q(B) W_t$$

is invertible, if $\theta_q(x) \neq 0, \forall |x| \leq 1$. That is, the polynomial $\theta_q(x)$ having all root lying outside unit circile.

Remark 21.2.3 (invertibility and uniqueness). Consider $X_t = W_t + 0.5W_{t-1}, W_t \sim WN(0, 25)$ and $Y_t = V_t + 5V_{t-1}, V_t \sim WN(0, 1)$ are the same, but one is invertible, but one is not.

Lemma 21.2.13 (MA process is always weakly stationary). *[5, p. 59] The MA process is ch:time-series-analysis:fig:ma1and2simulatedtraj for any value of $\theta_1, \theta_2, ..., \theta_q$.*

Proof. See linear combination of white noise process is still stationary process Theorem 21.2.1. Note that any MA process can be viewed as a linear combination of stationary process $Z_1, Z_2, ...$ □

Remark 21.2.4. For an MA process, we are not concerned with its stationarity property but concerned with its invertibility.

21.2.4.3 *Forecasting*

Definition 21.2.10 (best least square forecast). *[5, p. 54] Consider a time series $\{X_t\}$. Suppose we are currently at time index h. Let $\hat{X}_{h+l}$ be the forecast of X_{t+h}, where the positive integer l is the forecast horizon. Let $\mathcal{F}_t$ be a σ algebra representing all the information upto t. We say $\hat{X}_{h+l}$ is the **best least square forecast** of X_{h+l} if*

$$E[(X_{h+l} - \hat{X}_{h+l})^2 | \mathcal{F}_h] = \min_g E[(X_{h+l} - g)^2 | \mathcal{F}_h],$$

where g is all the random variables measurable with respect to $\mathcal{F}_h$.

Remark 21.2.5 (interpretation). Note that the forecast $\hat{X}_{h+l}$ is random variable instead of a single number.

Lemma 21.2.14 (best least square forecasting for MA(1) with observed shocks). *[5, p. 62] Consider an MA(1) process given by*

$$X_t = Z_t + \theta_1 Z_{t-1} + c_0, Z_t \sim WN(0, \sigma^2).$$

Let $\mathcal{F}_t$ denote the information available upto t. It follows that

- *the 1-step ahead best least-square forecast is*

$$\hat{X}_{t+1} \triangleq E[X_{t+1} | \mathcal{F}_t] = c_0 + \theta_1 Z_t,$$

  and the error is $E[(X_{t+1} - \hat{X}_{t+1})^2 | \mathcal{F}_t] = \sigma^2$.
- *the 2-step ahead best least-square forecast is*

$$\hat{X}_{t+2} \triangleq E[X_{t+2} | \mathcal{F}_t] = c_0,$$

and the error is $E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = (1+\theta_1^2)\sigma^2$.

- *the m-step($m > 2$) ahead best least-square forecast is*

$$\hat{X}_{t+m} \triangleq E[X_{t+m}|\mathcal{F}_t] = c_0,$$

and the error is $E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] = (1+\theta_1^2)\sigma^2$.

Proof. We use the least-square minimizing property of conditional expectation [Lemma 11.7.6]. We have (1)

$$\begin{aligned}\hat{X}_{t+1} &= E[Z_{t+1} + \theta_1 Z_t + c_0|\mathcal{F}_t] \\ &= E[Z_{t+1}|\mathcal{F}_t] + E[\theta_1 Z_t|\mathcal{F}_t] + E[c_0|\mathcal{F}_t] \\ &= 0 + \theta_1 Z_t + c_0 \\ &= \theta_1 Z_t + c_0\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[Z_{t+1}^2|\mathcal{F}_t] = \sigma^2.$$

(2)

$$\begin{aligned}\hat{X}_{t+2} &= E[Z_{t+2} + \theta_1 Z_{t+1} + c_0|\mathcal{F}_t] \\ &= E[Z_{t+2}|\mathcal{F}_t] + E[\theta_1 Z_{t+1}|\mathcal{F}_t] + E[c_0|\mathcal{F}_t] \\ &= 0 + 0 + c_0 \\ &= c_0\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = E[(Z_{t+2} + \theta_1 Z_{t+1})^2|\mathcal{F}_t] = (1+\theta_1^2)\sigma^2.$$

(3) same as (1)(2). To get the error, we have

$$E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] = E[Z_{t+m}|\mathcal{F}_t] = E[(Z_{t+m} + \theta_1 Z_{t+m-1})^2|\mathcal{F}_t] = (1+\theta_1^2)\sigma^2.$$

□

Lemma 21.2.15 (best least square forecasting for MA(q) with observed shocks). *[5, p. 62] Consider an MA(q) process given by*

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q} + c_0, Z_t \sim WN(0, \sigma^2).$$

Let $\mathcal{F}_t$ denote the information available upto t. It follows that

- *the 1-step ahead best least-square forecast is*

$$\hat{X}_{t+1} \triangleq E[X_{t+1}|\mathcal{F}_t] = c_0 + \theta_1 Z_t + ... + \theta_q Z_{t-q+1},$$

  and the error is $E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = \sigma^2$.
- *the 2-step ahead best least-square forecast is*

$$\hat{X}_{t+2} \triangleq E[X_{t+2}|\mathcal{F}_t] = \theta_2 Z_t + ... + \theta_q Z_{t-q+2} + c_0,$$

  and the error is $E[(X_{t+2} - \hat{X}_{t+2})^2|\mathcal{F}_t] = (1 + \theta_1^2)\sigma^2$.
- *the m-step*$(3 \leq m \leq q)$ *ahead best least-square forecast is*

$$\hat{X}_{t+m} \triangleq E[X_{t+m}|\mathcal{F}_t] = c_0 + \sum_{i=m}^{q} \theta_i Z_{t-i+m},$$

  and the error is $E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] = \sigma^2(\sum_{i=1}^{m-1} \theta_i^2), \theta_0 = 1$.
- *the m-step*$(m > q)$ *ahead best least-square forecast is*

$$\hat{X}_{t+m} \triangleq E[X_{t+m}|\mathcal{F}_t] = c_0,$$

  and the error is $E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] = \sigma^2(\sum_{i=1}^{q} \theta_i^2), \theta_0 = 1$.

Proof. We use the least-square minimizing property of conditional expectation [Lemma 11.7.6]. We have (1)

$$\begin{aligned}
\hat{X}_{t+1} &= E[Z_{t+1} + \theta_1 Z_t + ... + \theta_q Z_{t-q+1} + c_0|\mathcal{F}_t] \\
&= E[Z_{t+1}|\mathcal{F}_t] + E[\theta_1 Z_t|\mathcal{F}_t] + E[\theta_2 Z_{t-1} + ... + \theta_q Z_{t-q+1} + c_0|\mathcal{F}_t] \\
&= 0 + \theta_1 Z_t + ... + \theta_q Z_{t-q+1} + c_0 \\
&= \theta_1 Z_t + ... + \theta_q Z_{t-q+1} + c_0
\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[Z_{t+1}^2|\mathcal{F}_t] = \sigma^2.$$

(2)

$$\begin{aligned}
\hat{X}_{t+2} &= E[Z_{t+2} + \theta_1 Z_{t+1} + ... + \theta_q Z_{t-q+2} + c_0|\mathcal{F}_t] \\
&= E[Z_{t+2} + \theta_1 Z_{t+1}|\mathcal{F}_t] + E[\theta_2 Z_t|\mathcal{F}_t] + E[(\theta_3 Z_{t-1} + ... + \theta_q Z_{t-q+2} + c_0)^2|\mathcal{F}_t] \\
&= 0 + \theta_2 Z_t + ... + \theta_q Z_{t-q+2} + c_0 \\
&= \theta_2 Z_t + ... + \theta_q Z_{t-q+2} + c_0
\end{aligned}$$

where we use the property of taking-out-known of conditional expectation [subsection 11.7.4]. To get the error, we have

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[Z_{t+2} + \theta_1 Z_{t+1}|\mathcal{F}_t] = (1 + \theta_1^2)\sigma^2.$$

(3) (4) same as (1)(2). □

Lemma 21.2.16 (best least square forecasting for MA(q) without observed shocks). *[5, p. 62] Consider an MA(q) process given by*

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q} + c_0, Z_t \sim WN(0, \sigma^2).$$

Let $\mathcal{F}_t$ denote the information available upto t but not including the information about the shocks It follows that

- *the 1-step ahead best least-square forecast is*

$$\hat{X}_{t+1} \triangleq E[X_{t+1}|\mathcal{F}_t] = c_0,$$

  and the error is $E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = \sigma^2$.
- *the m-step($m \geq 1$) ahead best least-square forecast is*

$$\hat{X}_{t+m} \triangleq E[X_{t+m}|\mathcal{F}_t] = c_0,$$

  and the error is $E[(X_{t+m} - \hat{X}_{t+m})^2|\mathcal{F}_t] = \sigma^2(\sum_{i=1}^m \theta_i^2), \theta_0 = 1$.

Proof. We use the least-square minimizing property of conditional expectation [Lemma 11.7.6]. We have (1)

$$\begin{aligned}\hat{X}_{t+1} &= E[Z_{t+1} + \theta_1 Z_t + ... + \theta_q Z_{t-q+1} + c_0|\mathcal{F}_t] \\ &= E[Z_{t+1}|\mathcal{F}_t] + E[\theta_1 Z_t|\mathcal{F}_t] + E[\theta_2 Z_{t-1} + ... + \theta_q Z_{t-q+1} + c_0|\mathcal{F}_t] \\ &= 0 + 0 + ... + 0 + c_0\end{aligned}$$

To get the error, we have

$$E[(X_{t+1} - \hat{X}_{t+1})^2|\mathcal{F}_t] = E[(Z_{t+1} + \theta_1 Z_t + ... + \theta_q Z_{t-q+1})^2|\mathcal{F}_t] = \sigma^2(\sum_{i=1}^m \theta_i^2), \theta_0 = 1.$$

(2) same as (1). □

Remark 21.2.6 (practical consideration of observability of shocks). In reality, we cannot observe shocks; we can only observe X_t.

21.2.5 ARMA process

21.2.5.1 *Basic properties*

Definition 21.2.11 (ARMA process). *An $ARMA(p,q)$ model can be represented as*

$$\theta_p(L)X_t = \phi_q(L)W_t$$

where θ_p is the AR operator of order p,ϕ_q is the MA operator of order q.

Remark 21.2.7 (Parameter redundancy in ARMA). When θ and ϕ share a common factor, the model can be simplified. For example, $(1-0.5L)(1-1/3L)X_t = (1-0.5L)W_t \Rightarrow (1-1/3B)X_t = W_t$.

Lemma 21.2.17 (stationarity of ARMA process). *[6, p. 95] The $ARMA(p,q)$ process is stationary if the roots of $\theta_q(z) = 0$ satisfies $|z| > 1$.*

Proof. See Lemma 21.2.7. □

Lemma 21.2.18 (invertibility of ARMA process). *[6, p. 95] An ARMA model is **invertible**, i.e., $W_t = \phi_q(L)^{-1}\theta_p(L)X_t = \pi(L)W_t = \sum_{j=0}^{\infty}\pi_j X_{t-j}$ if and only if $\phi_p(z) \neq 0$ for $|z| < 1$.*

Proof. When ϕ_q is invertible, then we have $W_t = \phi_q^{-1}(L)\theta_p(L)X_t$. Therefore, it is invertible. □

Lemma 21.2.19 (causal form, moving average representation of AR process). *[6, p. 95] An ARMA model can be written as a **causal form**, i.e. $X_t = \theta_p(L)^{-1}\phi_q(L)W_t = \psi(L)W_t = \sum_{j=0}^{\infty}\psi_j W_{t-j}$ if and only if $\theta_p(z) \neq 0$ for $|z| < 1$.*

Proof. See Lemma 21.2.7 and Lemma 21.2.3. □

Methodology 21.2.1 (Hannan-Rissanen). *Two step regression for $ARMA(p,q)$*

$$X_t - \phi_1 X_{t-1} - \cdots + \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $Z_t \sim WN(0,\sigma^2)$

In the first step, we regress X_t on the past $X_{t-1}, ..., X_{t-p}$ using OLS to get the coefficient estimatoes $\hat{\phi}_1, ..., \hat{\phi}_p$. Secondly, Estimate the residuals given by

$$\hat{Z}_t = X_t - \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}.$$

Thirdly, we use the following ARMA form with estimated residuals,

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \hat{Z}_t + \theta_1 \hat{Z}_{t-1} + \cdots + \theta_q \hat{Z}_{t-q}$$

to estimate the coefficients $(\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ by OLS. Finally, update the estimate of residuals and refit model until convergence.

21.2.6 Unit root AR process

21.2.6.1 *Unit root process*

Definition 21.2.12 (AR(p) process as a unit root process).

- *An AR(p) process*

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \epsilon_t, \epsilon_t \sim WN(0,1)$$

is a unit root process, if the characteristic polynomial

$$p(z) = 1 - a_1 z - a_2 z^2 - \cdots - a_p z^p$$

has at least one unit root(a root equal to 1) and all other roots are outside the complex unit circle. That is, if $p(z)$ has a unit root, then

$$p(1) = 1 - a_1 - a_2 - \cdots - a_p = 0.$$

- *An AR(1) process*

$$y_t = y_{t-1} + \epsilon_t, \epsilon_t \sim WN(0,1)$$

is a unit root process

Example 21.2.2. In Figure 21.2.3, we show representative simulated trajectories for AR(1) process with coefficient 1, which forms a unit-root process, and with coefficient -1.

Note that

$$y_t = y_{t-1} + \epsilon_t, \epsilon_t \sim WN(0,1)$$

is a random walk and is also an $AR(1)$ process with unit root to the characteristic polynomial $1 - z = 0$.

Also note that $y_t = -y_{t-1} + \epsilon_t$, with solution $y_n = y_0 + \sum_{i=0}^{n}(-1)^i \epsilon_i$ will exhibit oscillation patter and is not considered a unit root process, although its variance will similarly increase with time.

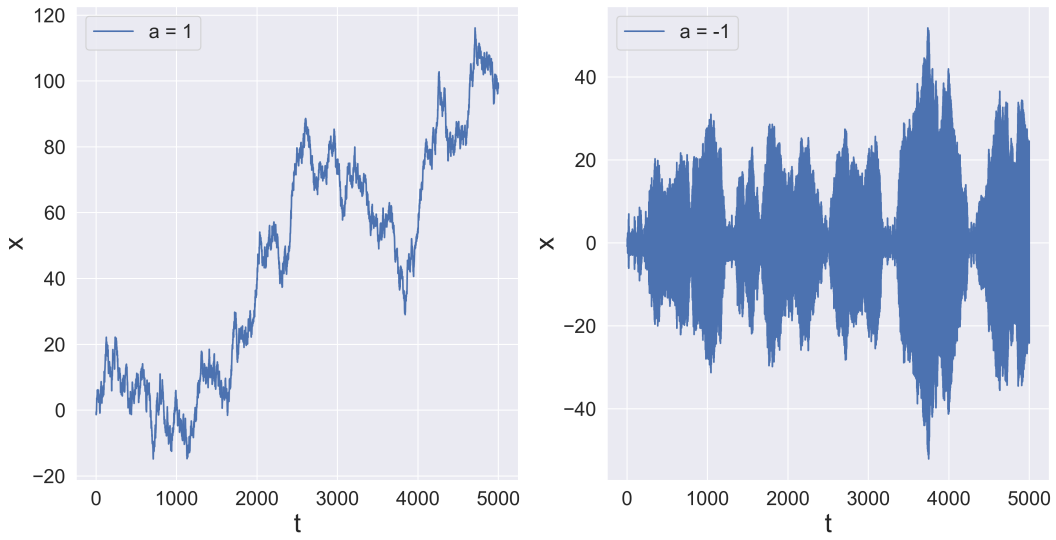

Figure 21.2.3: Representative simulated trajectories for AR(1) process with coefficient 1, which forms a unit-root process, and with coefficient -1.

Lemma 21.2.20 (basic properties of AR(1) unit root process). *Consider an AR(1) process given by*

$$y_t = y_{t-1} + \epsilon_t, \epsilon_t \sim WN(0, \sigma^2),$$

where initial condition y_0. It follows that

- *The solution is*

$$y_t = y_0 + \sum_{j=1}^{t} \epsilon_j.$$

- *The mean and variance is*

$$E[y_t] = y_0, Var[y_t] = \sigma^2 t$$

- *The long-run correlation is given by*

$$corr(y_t, y_{t+h}) = \frac{E[(\sum_{i=1}^{t} u_i)(\sum_{i=1}^{t+h} u_i)]}{(t\sigma^2(t+h)\sigma^2)^{0.5}} = \frac{t}{\sqrt{t(t+h)}} \to 1$$

as $t \to \infty$.

Proof. (1) Just repeat substitution. (2)(3) Use (1). □

21.2.6.2 *Trend stationarity vs. unit root process*

Note 21.2.1 (deterministic trend vs. stochastic trend). Many time series are trending. It is important to distinguish between two important cases:

- A stationary process with a deterministic trend: shocks have **transitory** effects.
- A process with a stochastic trend or a unit root: shocks have **permanent** effects.

Example 21.2.3 (deterministic trend vs. stochastic trend).

- Consider a stationary AR(1) model with a deterministic linear trend term

$$Y_t = \theta Y_{t-1} + a_0 + a_1 t + \epsilon_t, \epsilon_t \sim WN(0, \sigma^2),$$

where $|\theta| < 1$, and Y_0 is an initial value.
  - Note that the mean and variance is given by

$$E[Y_t] = \theta^t Y_0 + \mu + \mu_1 t \to \mu + \mu_1 t, as\ t \to \infty,$$

and

$$Var[Y_t] = \frac{\sigma^2}{1-\theta},$$

where $\mu = a_0/(1-\theta), \mu_1 = a_1/(1-\theta)$.
  - Y_t is not a stationary process since its mean is time dependent.However, the process $Y_t - E[Y_t]$ is a stationary process, called trend-stationary.
  - Also note that stochastic part of Y_t is stationary and shocks have transitory effects. We say that the process is mean reverting to its long-run $\mu + \mu_1 t$.

21.2.6.3 *Unit root test*

Now we introduce the most widely unit root test - Dickey-Fuller test. Other tests include augmented Dickey-Fuller test.

Definition 21.2.13 (Dickey-Fuller ρ test). *[7, p. 574] Consider the AR(1) model*

$$y_t = \theta y_{t-1} + \epsilon_t$$

with $T+1$ observations $\{y_0, y_1, ..., y_T\}$ and the hypothesis

$$H_0 : \theta = 1; H_1 : \theta < 1.$$

The test statistic,known as Dickey-Fuller ρ statistic, is given as

$$DF = T(\hat{\theta} - 1)$$

where $\hat{\theta}$ is the least square estimation of θ from linear regression.

Definition 21.2.14 (Dickey-Fuller t test). *[7, p. 574] Consider the AR(1) model*

$$y_t = \theta y_{t-1} + \epsilon_t$$

with $T+1$ observations $\{y_0, y_1, ..., y_T\}$ and the hypothesis

$$H_0 : \theta = 1; H_1 : \theta < 1.$$

The test statistic,known as Dickey-Fuller t statistic, is given as

$$DF_t = \frac{\hat{\theta} - 1}{SSE[\hat{\theta}/(n-2)]}$$

where $\hat{\theta}$ is the least square estimation of θ from linear regression, such that

$$\hat{\theta} = \frac{\sum_{i=1}^{T} y_{i-1} y_i}{\sum_{i=1}^{T} y_i^2}$$

21.2.6.4 *Forecasting*

21.2.7 Correlation analysis

21.2.7.1 *Autocorrelation statistical analysis*

Definition 21.2.15. *[2, p. 19] Let $x_1, ..., x_n$ be observations of a time series. The **sample mean** is*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

*The **sample autovariance function** is*

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_{i+h} - \bar{x})$$

*The **sample autocorrelation function** is*

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$$

Remark 21.2.8. Note that the divisor in the covariance is $1/n$.

Lemma 21.2.21 (bounded variance of sample autocorrelation function).

- *The sample autocorrelation function of leg h defined by $\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$ is random variable with support $[-1, 1]$ and bounded variance of $Var[\hat{\rho}] \leq 1$.*
- *If n is the number of samples used in the calculation of ACF $\hat{\rho}$, then $Var[\hat{\rho}] \leq 1/n$.*

Proof. (1) $|\hat{\rho}| \leq 1$ is showed in Cauchy inequality [Corollary 11.9.4.1]. (2) The boundedness the variance is at Lemma 11.9.2. □

Definition 21.2.16 (Ljung and Box serial correlation test). *[5, p. 32] Let ρ_i be the autocorrelation function of a time series $\{X_t\}$. Consider the null hypothesis and alternative hypothesis*

- *null hypothesis $H_0 : \rho_1 = \cdots = \rho_m = 0$.*
- *alternative hypothesis $H_1 : \rho_i \neq 0, for\ some\ i \in \{1, 2, ..., m\}$.*

The test statistic for the hypothesis is given by

$$Q(m) = T(T+2)\sum_{i=1}^{m}\frac{\hat{\rho}_l^2}{T-l}.$$

The decision rule is to reject H_0 if $Q(m) > \chi^2_\alpha$(where χ^2_α denote the α percentile of a chi-squared distribution).

Remark 21.2.9 (interpretation). If there exists correlation, $Q(m)$ tends to be large.

21.2.7.2 *Partial autocorrelation function theory*

Definition 21.2.17 (partial autocorrelation function, PACF). *[8, p. 100] Let $\{X_t\}$ be a zero mean stationary process. The partial autocorrelation at lag h for $h \geq 1$, denoted by $\pi_X(h)$, is defined as the direct correlation between X_t and X_{t+h} with the **linear dependence** between the intermediate variables $X_s, t < s < t+h$ removed. Specifically, we have*

$$PACF(X_t, X_{t+h}) = Cov(X_t - Var[Y]^{-1}E[YX_i]Y, X_{t+h} - Var[Y]^{-1}E[YX_{t+h}]]Y),$$

where $Y = (X_{t+1}, X_{t+2}, ..., X_{t+h-1})$.

Theorem 21.2.4 (computation of PACF). *Let $\{X_t\}$ be a zero-mean stationary process.*

- *The partial autocorrelation at lag h for $h \geq 2$, denoted by $\pi_X(h)$, is equal to the coefficient a_h from the optimal linear prediction of X_{t+h} on the observations $X_t, X_{t+1}, ..., X_{t+h-1}$ given by*

  $$X_{t+h} = a_h X_t + \beta_1 X_{t+1} + ... + \beta_{h-1} X_{t+h-1}.$$

  For $h = 0$, $\pi_X(0) = 1$;$h = 1$, $\pi_X(1) = \rho_X(1)$.
- *The partial autocorrelation at lag h for $h \geq 2$, denoted by $\pi_X(h)$, is given by the partial correlation conditioned on $(X_{t+1}, X_{t+2}, ..., X_{t+h-1})$,i.e.,*

  $$\begin{aligned}\pi_X(h) &= corr(X_{t+h} - P_Y(X_{t+h}), X_t - P_Y(X_t)) \\ &= \frac{Cov(X_{t+h} - P_Y(X_{t+h}), X_t - P_Y(X_t))}{\sqrt{Var[(X_{t+h} - P_Y(X_{t+h}))]}\sqrt{Var[X_t - P_Y(X_t)]}}\end{aligned}$$

  where $Y = (X_{t+1}, X_{t+2}, ..., X_{t+h-1})$, and $P_Y(X_{t+h})$ is the projection of X_{t+h} onto the subspace spanned by $(X_{t+1}, X_{t+2}, ..., X_{t+h-1})$.

Proof. From the Hilbert space approximation theory [Theorem 5.4.4], we know that the coefficient a_h is given by

$$\pi_X(h) = \frac{Cov(X_{t+h} - P_Y(X_{t+h}), X_t - P_Y(X_t))}{\sqrt{Var[(X_t - P_Y(X_t))]}\sqrt{Var[X_t - P_Y(X_t)]}}.$$

From the definition partial correlation coefficient, we have

$$\pi_X(h) = \frac{Cov(X_{t+h} - P_Y(X_{t+h}), X_t - P_Y(X_t))}{\sqrt{Var[(X_{t+h} - P_Y(X_{t+h}))]}\sqrt{Var[X_t - P_Y(X_t)]}}.$$

To show these two are equivalent, our goal is to show $Var[(X_t - P_Y(X_t))]$ and $Var[(X_{t+h} - P_Y(X_{t+h}))]$ are equivalent. We note that $Var[(X_t - P_Y(X_t))]$ is fully determined by $Var[X(t)]$, the covariance structure $Cov[Y]$, and the covariance structure of $Cov[Y, X(t)]$. Due to the weak stationarity, $Var[X(t)] = Var[X(t+h)]$, the $Cov[Y, X(t)], Cov[Y, X(t+h)]$ only differ at the labeling. □

Note 21.2.2 (caution on calculating PACF and its application).

- Suppose we have stationary time series given by

  $$X_{t+1} = \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_q X_{t-q} Z_t,$$

  when we use linear regression model to compute

  $$X_{t+h} = a_h X_t + a_1 X_{t+1} + ... + a_{h-1} X_{t+h-1}.$$

  where $1 < h \leq q$, only when $h = q$, the coefficient $\pi(h) = a_h = \beta_q$; when $h > q$, $a_h = 0$, when $h < q$, a_h is not equivalent to β_h.
- Therefore, PACF $\pi(h)$ is the method to estimate the order of AR process; it cannot be used to estimate the coefficient in the AR model; The estimation of the coefficient usually rely on maximum likelihood or least square.

Lemma 21.2.22 (bounded variance of sample partial autocorrelation function). *[4, p. 66]*

- *The partial autocorrelation function of leg h, denoted by $\hat{\pi}(h)$, with for a single random sample, can be viewed as a random variable with support $[-1, 1]$ and bounded variance of $Var[\hat{\pi}] \leq 1$.*
- *If n is the number of samples used in the calculation of PACF $\hat{\pi}$, then $Var[\hat{\pi}] \leq 1/n$.*

Proof. (1) $|\hat{\rho}| \leq 1$ is showed in Cauchy inequality [Corollary 11.9.4.1]. (2) The boundedness the variance is at Lemma 11.9.2. □

Remark 21.2.10 (interpretation of PACF as conditional correlation).

- Consider a regression context in which y is the response variable and x_1, x_2, x_3 as predictor variables. The partial correlation between y and x_3 is given by

  $$\frac{Cov(y, x_3 | x_1, x_2)}{\sqrt{Var(y|x_1, x_2)}\sqrt{Var(x_3|x_1, x_2)}}.$$

- We can find out the partial correlation between y and x_3 by constructing an optimal linear regression
$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$
and take β_3 as the PCAF.

Lemma 21.2.23 (PACF for AR and MA process).

- *Consider a AR(p) process given by*
$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + Z_t, Z_t \sim WN(0, \sigma^2),$$
then
$$\pi_X(h) = \begin{cases} a_h, h = 1, 2, .., p \\ 0, otherwise \end{cases}.$$
- *Consider a MA(1) process given by*
$$X_t = Z_t + \theta_1 Z_{t-1}, Z_t \sim WN(0, \sigma^2),$$
then
$$\pi_X(h) = \begin{cases} a_h, h = 1, 2, .., p \\ 0, otherwise \end{cases}.$$

Proof. (1)From Lemma 21.2.11 directly. (2)Lemma 21.2.11 □

Note 21.2.3 (ACF vs. PACF for application in order identification)**.**

- PACF is useful in identifying the order of AR process; however, ACF has nonzero coefficient for all lags.
- ACF is useful in identifying the order of MA process; however, PACF has nonzero coefficient for all lags.

21.2.7.3 *Correlogram analysis example*

In time series, we usually call these autocorrelation vs. lag plots **correlograms**.

To start with, we first examine the correlogram for a while noise process [Figure 21.2.4]. Both ACF and PACF are near zeros, matching the expectation. For an AR(1) process with coefficient 0.8 [Figure 21.2.5], PACF in recovers the coefficient in the first lag and ACF shows peaks $\sim 0.8 = \rho, \sim 0.64 = \rho^2, ...$ For an MA(1) process with coefficient 0.8 [Figure 21.2.6], PACF shows multiple alternating peaks and ACF shows a single peak at first lag with value around $\frac{0.8}{1+0.8^2} = 0.4878$.

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off beyond lag p	Cuts off after lag p	Tails off beyond lag p
PACF	Cuts off after lag p	Tails off beyond lag p (with possible oscillations)	Tails off beyond lag p

Table 21.2.1: Summary of PACF and ACF for AR, MA, and ARMA processes.

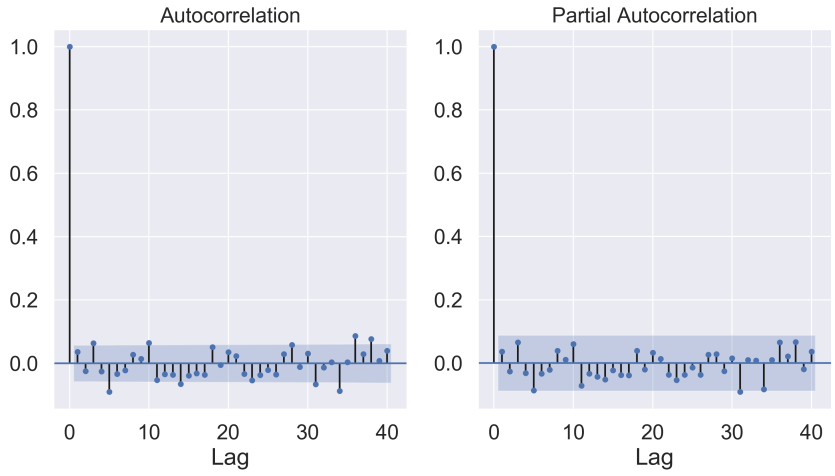

Figure 21.2.4: The ACF and PACF corrlogram for a white noise process.

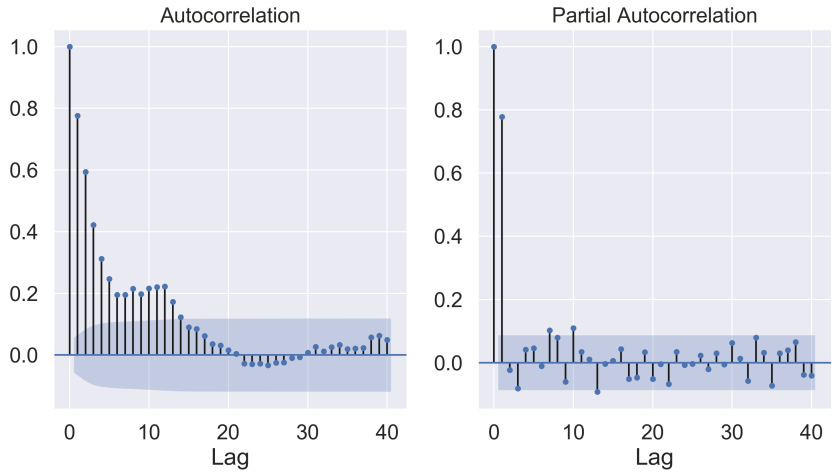

Figure 21.2.5: The ACF and PACF corrlogram for an AR(1) process with coefficient 0.8.

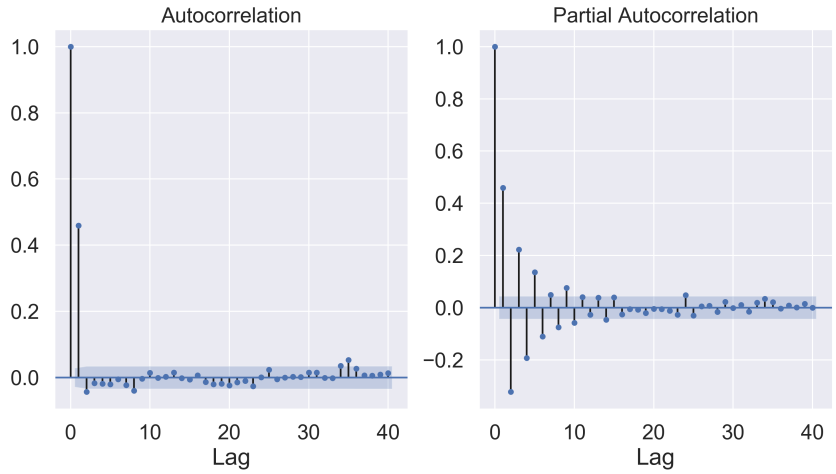

Figure 21.2.6: The ACF and PACF corrlogram for an MA(1) process.

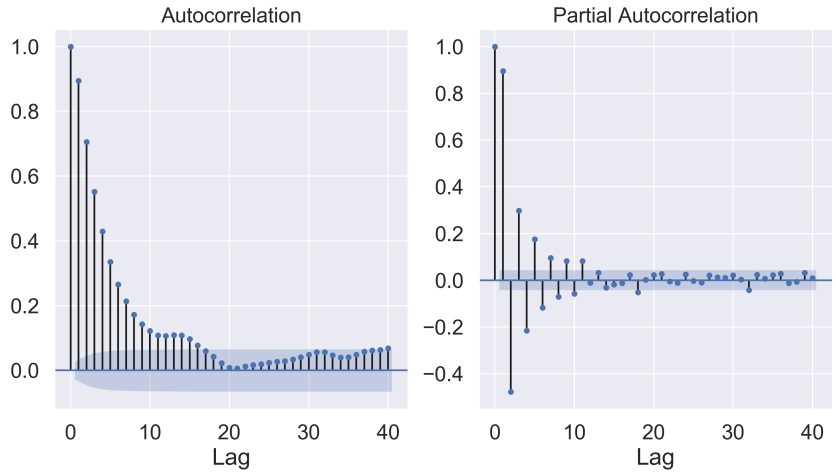

Figure 21.2.7: The ACF and PACF corrlogram for an ARMA(1,1) process.

21.2.8 Model analysis and calibration

21.2.8.1 *Order selection*

Lemma 21.2.24 (order identification for MA processes). *Suppose we have random sample from the MA time series* $\{X_t\}, t = 1, 2, ..., n$. *The ACF of order* h *is zero with* 95% *confindence level if*

$$|\hat{\rho}(h)| \leq \frac{2}{\sqrt{n}}.$$

Proof. Consider the null hypothesis that $\rho(h) = 0$. Note that the calculation $\hat{\rho}$ is given by

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$$

The **sample autovariance function** is

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_{i+h} - \bar{x}).$$

The null hypothesis ensures that each term $(x_i - \bar{x})(x_{i+h} - \bar{x})$ is independent and the Cauchy inequality ensures that it is random variable with support $[-1, 1]$, mean o, and variance σ^2 bounded by 1.

From central limit theorem [Theorem 11.11.3], $\hat{\rho} \sim N(0, \frac{\sigma^2}{n})$. Take 95% confidence level, when

$$|\hat{\rho}| \geq \frac{2}{\sqrt{n}} \geq \frac{2\sigma}{\sqrt{n}},$$

we reject the null hypothesis. □

Remark 21.2.11 (order identification). We cannot use the PACF to identify the order of MA process, because PACF has non-zero value extending to infinityLemma 21.2.23.

Remark 21.2.12 (model checking). Let x_t be the time series observation and $\hat{x}_t$ be the linear optimal prediction. Let $r_t = x_t - \hat{x}_t$ be the residual. We can use the serial correlation [Definition 21.2.16] to check whether the residual is correlated or not. If the model we use is proper, then there should be no correlation among the residuals.

21.2.8.2 *Yule-Walker equations and related methods*

The Yule-Walker equations establish the connection between coefficients in AR(q) equation and time series autocorrelation. By calculating sample autocorrelation, we can use Yule-Walker equations to reversely calculate the model coefficients.

To start with, we discuss the Yule-Walker equations for AR(q) process in the following.

Theorem 21.2.5 (Yule-Walker equations for AR(q) process). *Given an zero-mean $AR(p)$ model,*

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + Z_t, Z_t \sim WN(0, \sigma^2).$$

The autocovariance and the coefficients $a_1, a_2, ..., a_q$ are related in the following ways:

- $$\gamma(h) = a_1\gamma(h-1) + a_2\gamma(h-2) + \cdots + a_q\gamma(h-q), \ for\ h = 1, ..., q;$$

  or in matrix form

$$\begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(q) \end{bmatrix} = \underbrace{\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(q-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(q-2) \\ \vdots & \vdots & \ddots & \cdots \\ \gamma(q-1) & \gamma(q-2) & \cdots & \gamma(0) \end{bmatrix}}_{\Gamma} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{bmatrix}.$$

  In particular, the matrix Γ is symmetric positive definite when X_ts are not perfectly correlated.
- *If both sides divided by $\gamma(0)$, we have*

$$\begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(q) \end{bmatrix} = \begin{bmatrix} 1 & \rho(1) & \cdots & \rho(q-1) \\ \rho(1) & 1 & \cdots & \rho(q-2) \\ \vdots & \vdots & \ddots & \cdots \\ \rho(q-1) & \rho(q-2) & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{bmatrix}.$$

- $$\gamma(0) = a_1\gamma(1) + a_2\gamma(2) + \cdots + a_p\gamma(q) + \sigma^2.$$
-

Proof. (1)(2)Given the zero mean $AR(q)$ model,

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + Z_t, (*)$$

we can multiply both sides by X_{t-h} for $h = 1, ..., p$, and get

$$X_t X_{t-h} = a_1 X_{t-1} X_{t-h} + a_2 X_{t-2} X_{t-h} + ... + a_p X_{t-p} X_{t-h} + Z_t X_{t-h}.$$

Take expectation and we get

$$\gamma(h) = a_1\gamma(h-1) + a_2\gamma(h-2) + \cdots + a_q\gamma(h-q).$$

The positive definiteness of Γ can be seen from the fact that

$$\Gamma = E[XX^T], X = (X_1, X_2, ..., X_q),$$

and for any nonzero $p \in \mathbb{R}^q$, $q^T \Gamma q = Var[q^T X] > 0$. (3)If we take $h = 0$, we will get

$$X_t X_t = a_1 X_{t-1} X_t + a_2 X_{t-2} X_t + ... + a_p X_{t-p} X_t + Z_t X_t.$$

Take expectation, we get

$$\gamma(0) = a_1 \gamma(1) + a_2 \gamma(2) + \cdots + a_p \gamma(q) + \sigma^2.$$

□

Example 21.2.4. Consider the AR(2) process given by

$$X_n = Z_n + a_1 X_{n-1} + a_2 X_{n-2}.$$

The Yule-Walker equations are

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}.$$

Solving the Yule-Walker equation, we obtain

$$\rho_1 = \frac{a_1}{1 - a_2}, \rho_2 = \frac{a_1^2}{1 - a_2} + a_2.$$

Methodology 21.2.2 (method of moments parameter estimation via Yule-Walker equation). *Given an zero-mean $AR(p)$ model,*

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + Z_t, Z_t \sim WN(0, \sigma^2).$$

The autocovariance and the coefficients $a_1, a_2, ..., a_q$ are related in the following ways:

- *Let $\hat{\rho} = (\hat{\rho}(1), \hat{\rho}(2), ..., \hat{\rho}(q))^T$, $a = (a_1, a_2, ..., a_q)$, and $\hat{\Gamma}$ be the matrix of*

$$\begin{bmatrix} 1 & \hat{\rho}(1) & \cdots & \hat{\rho}(q-1) \\ \hat{\rho}(1) & 1 & \cdots & \hat{\rho}(q-2) \\ \vdots & \vdots & \ddots & \cdots \\ \hat{\rho}(q-1) & \hat{\rho}(q-2) & \cdots & 1 \end{bmatrix},$$

then $\hat{a} = (\hat{\Gamma})^{-1}\hat{\rho}$.

- $$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{a}_1\hat{\gamma}(1) - \hat{a}_2\hat{\gamma}(2) - \cdots \hat{a}_q\hat{\gamma}(q)$$

21.2.8.3 *Linear regression approach*

The AR(q) model is given by

$$X_t = a_1X_{t-1} + a_2X_{t-2} + ... + a_pX_{t-p} + Z_t, Z_t \sim WN(0, \sigma^2).$$

The formalism also fits into the linear regression framework where the outcome variable is X_t, the predictor variables are $X_{t-1}, ..., X_{t-p}$, and the model parameters to be determined are $a_1, ..., a_p$.

To accommodate the linear regression framework, we can construct corresponding observation matrices from time series and solve the model parameter using least square method.

Methodology 21.2.3 (least square estimation for AR(q)). *Consider the AR(q) process given by*

- *The correlation parameter* $\beta = (\beta_1, \beta_2, .., \beta_q)^T$

$$Y = \begin{bmatrix} x_{q+1} \\ x_{q+2} \\ \vdots \\ x_N \end{bmatrix}, X = \begin{bmatrix} x_q & x_{q-1} & \cdots & x_1 \\ x_{q+1} & x_q & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1} & x_{N-2} & \cdots & x_{N-q} \end{bmatrix},$$

  and the $Y = X\beta$ *model gives the least square estimator*

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

- *The estimation of* σ *is given by*

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1}.$$

Proof. Use the least square solution in linear regression analysis [Theorem 15.1.1] and the corresponding variance estimation [Theorem 15.1.5]. □

Remark 21.2.13 (connection of least square approach to conditional likelihood approach)**.**

- Note that we can decompose a full distribution as the product of conditional distributions given by

$$\begin{aligned} &f(X_N, X_{N-1}, ..., X_1) \\ =&f(X_N|X_{N-1}, ..., X_{N-q})f(X_{N-1}|X_{N-2}, ..., X_{N-q-1}) \cdots f(X_{q+1}|X_q, ..., X_1). \end{aligned}$$

  If the conditional distribution is Gaussian, then the maximum likelihood will give the same formula as the least square solution.
- In contrast, full distribution likelihood approach requires the auto-covariance structure of an $AR(q)$ process, which is difficult to obtain.

Methodology 21.2.4 (least square estimation of the correlation for AR(1)). *[1, p. 53][5, p. 46] Consider the AR(1) process given by*

- *The correlation parameter ρ as a solution to*

$$\min_{\rho} \sum_{i=2}^{N} (x_i - \rho x_{i-1})^2$$

  is given by

$$\hat{\rho} = \frac{\sum_{t=1}^{N-1} x_t x_{t+1}}{\sum_{t=1}^{N-1} x_t^2}.$$

  Alternatively,

$$Y = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}, X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{bmatrix},$$

  and the $Y = X\rho$ model gives the least square estimator

$$\hat{\rho} = (X^T X)^{-1} X^T Y.$$

- *The estimation of σ is given by*

Example 21.2.5 (least square estimation for AR(2)). Consider the AR(2) process given by

- The correlation parameter $\beta = (\beta_1, \beta_2)^T$

$$Y = \begin{bmatrix} x_3 \\ x_4 \\ \vdots \\ x_N \end{bmatrix}, X = \begin{bmatrix} x_2 & x_1 \\ x_3 & x_2 \\ \vdots & \vdots \\ x_{N-1} & x_{N-2} \end{bmatrix},$$

  and the $Y = X\beta$ model gives the least square estimator

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

- The estimation of σ is given by

21.2.8.4 *Maximum likelihood estimation*

An MA(q) process given by

$$X_t = \mu + Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \cdots + \beta_q Z_{t-q},$$

does not fit into the Yule-Walker and linear regression approach framework.

Still, we can use maximum likelihood estimation as long as we can express the likelihood function in terms of model parameters, as we show below.

Methodology 21.2.5 (maximum likelihood estimation of the coefficients in invertible MA(q)). *Consider an invertible* $MA(q)$

$$X_t = \mu + Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \cdots + \beta_q Z_{t-q},$$

where $Z_t \sim WN(0, \sigma^2)$.

Let $(X_1, X_2, ..., X_n)$ *be the observations. Then the MLE is given by*

$$L(\mu, \beta_1, ...\beta_q, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp(-\frac{1}{2}(X - \mu)^T \Sigma^{-1} (X - \mu)),$$

where

$$\Sigma_{ij} = Var[X_t]\rho(i-j), \rho(k) = \begin{cases} 1, k = 0 \\ 0, k > q \\ \sigma^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k} / \sum_{i=0}^{q} \beta_i^2, k = 1, ..., q \\ \gamma(-k), k < 0 \end{cases},$$

where $\beta_0 = 1$ *is used.*

Example 21.2.6 (MLE for invertible MA(1)). Consider an invertible $MA(1)$

$$X_t = \mu + Z_t + \beta_1 Z_{t-1},$$

where $Z_t \sim WN(0, \sigma^2)$ and $|\beta_1| \leq 1$.

Let $(X_1, X_2, ..., X_n)$ be the observations. Then the MLE is given by

$$L(\mu, \beta_1, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)),$$

where

$$\Sigma = \sigma^2(1+\beta_1^2) \begin{bmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ 0 & \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \rho = \frac{\beta_1}{\sigma^2(1+\beta_1^2)}.$$

21.2.8.5 *Example: a toy example*

Now we consider a toy example where we fit an AR(2) model (ground truth model is $X_t = 0.4X_{t-1} + 0.4X_{t-2} + Z_t$) using linear regression method or conditional MLE [Methodology 21.2.3]. The fitting results are summarized below.

Dep. Variable:	y	**No. Observations:**	500
Model:	AutoReg(2)	**Log Likelihood**	-723.520
Method:	Conditional MLE	**S.D. of innovations**	1.034
Date:	Mon, 10 Feb 2020	**AIC**	0.084
Time:	01:00:06	**BIC**	0.118
Sample:	2	**HQIC**	0.097

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0089	0.046	0.193	0.847	-0.082	0.100
y.L1	0.4467	0.041	10.912	0.000	0.366	0.527
y.L2	0.4069	0.041	9.932	0.000	0.327	0.487

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	1.1121	+0.0000j	1.1121	0.0000
AR.2	-2.2100	+0.0000j	2.2100	0.5000

Further support on the fitting results are from visual diagnosis plots in terms of residuals, residual sample histograms, residual normal QQ plot and correlogram for the residuals, as showed in Figure 21.2.8.

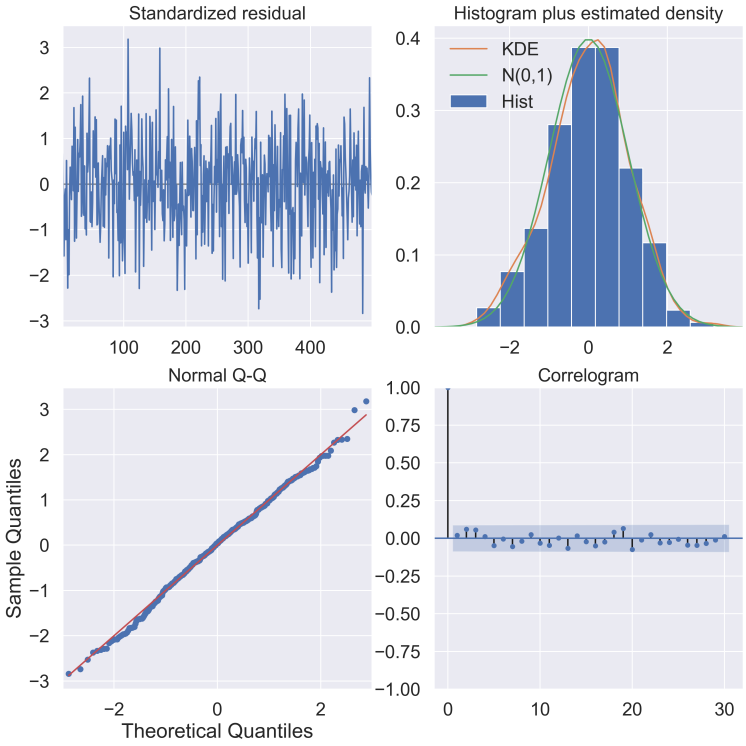

Figure 21.2.8: Diagnosis plot of residuals for AR(2) model estimation.

What happen if we mis-specify the model? Consider the case where we use AR(1) model to fit the data generated by AR(2), visual diagnosis plots in Figure 21.2.9 show that the major difference is correlogram where residuals have significant non-zero auto-correlation at lags greater than 0.

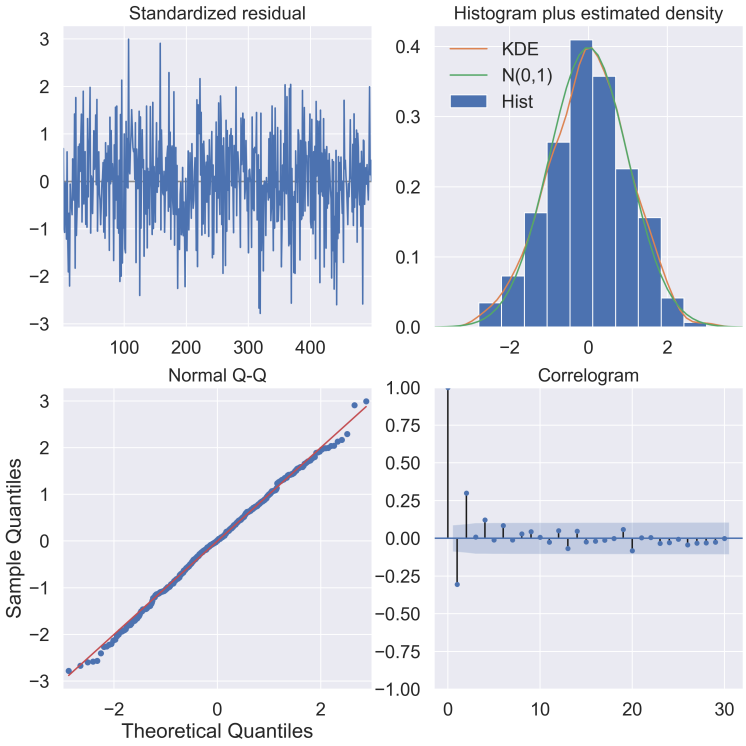

Figure 21.2.9: Diagnosis plot of residuals for AR(1) model fitted to time series generated by AR(2) ground truth model.

21.2.9 Wold Representation theorem

Definition 21.2.18 (orthogonal projection). *Let $X_t, t \in \mathbb{Z}$ be a covariance-stationary process. The random variable*

$$T[X_{t+h}|X_{t-1}, ..., X_{t-N}] = a_0 + a_1 X_{t-1} + ... + a_N x_{t-N}$$

where coefficient $a_1, ..., a_N$ are such that

$$E[(X_{t+h} - T[X_{t+h}|X_{t-1}, ..., X_{t-N}])^2]$$

*is minimun, is called **orthogonal projection** of X_{t+h} on $X_{t-1}, ..., X_{t-N}$.*

$$T[X_{t+h}|X_{t-1}, X_{t-2}, ...] = a_0 + a_1 X_{t-1} + a_N x_{t-2} + ...$$

Definition 21.2.19 (linearly deterministic process). *A covariance-stationary process, X_t, is called (linearly) deterministic if*

$$T[X_t|X_{t-1}, X_{t-2}, ...] = X_t.$$

Remark 21.2.14 (implications).

- A stationary process X_t is deterministic if X_t can be predicted correctly (with zero error) using the entire past $X_{t-1}, X_{t-2}, ...$
- For a deterministic process, the one-step linear prediction error is zero.
- A **nonlinear dynamical system might not be linearly deterministic**.

Example 21.2.7. The process $X_t = A\cos(t) + Bcos(t)$ is deterministic process because A, B can be determined from past history and then the future can be predicted with zero error.

Theorem 21.2.6 (Wold Representation theorem, Wold decomposition theorem). *Any zero-mean nondeterministic covariance stationary time series X_t can be **uniquely** decomposed as*

$$X_t = V_t + S_t$$

where

- *V_t is a linearly determinstic process*
- *S_t is an infinitely moving average process given as*

$$S_t = \phi_\infty(L)W_t, \phi_\infty(L) = I + \phi_1 L + ...$$

$$\sum_{i=1}^{\infty} \phi_i^2 < \infty$$

- $W_t \sim WN(0, \sigma^2)$
- $E[V_t W_s] = 0, \forall t, s$

Definition 21.2.20 (purely deterministic process). *zero-mean nondeterministic covariance stationary time series X_t is called **purely nondeterministic process** if in this Wold decomposition, the deterministic process component is zero, i.e. $V_t = 0$.*

Theorem 21.2.7 (multivariate Wold decomposition). *Any m dimensional convariance stationary time series X_t can be decomposed as*

$$\begin{aligned} X_t &= V_t + \eta_t + \psi_1 \eta_{t-1} + \psi_2 \eta_{t-2} + ... \\ &= V_t + \sum_{k=0}^{\infty} \psi_k \eta_{t-k} \end{aligned}$$

where

- $V_t \in \mathbb{R}^m$ *an m dimensional linearly deterministic process.*
- $\eta_t \in \mathbb{R}^m$ *is multivariate white noise process, i.e.* $E[\eta_t] = 0, E[\eta_t \eta_s^T] = \Sigma \delta_{st}$
- $cov(\eta_t, V_s) = 0, \forall t, s$
- $\psi_k \in \mathbb{R}^{m \times m}$ *and* $\psi_0 = I_m$, $\sum_{i=0}^{\infty} \psi_i \psi_i^T$ *converges.*

21.3 Extensions to multivariate time series

21.3.1 Introduction

Definition 21.3.1 (multivariate time series). *A K dimensional **multivariate time series** $\{X_t\}_{t\in T}$ is a collection of K dimensional random vectors X_t taking values in $\mathbb{R}^K$ whose the indexing set T is the set of integers.*

Definition 21.3.2 (mean vector, covariance matrix, cross-correlation matrix). *Given a K dimensional multivariate time series, the mean vector is defined as*

$$\mu(t) = E[X_t], \mu(t) \in \mathbb{R}^K.$$

The covariance matrix is defined as

$$\Gamma_0(t) = E[(X_t - \mu)(X_t - \mu)^T], \Gamma_0(t) \in \mathbb{R}^{K\times K}.$$

And the cross-correlation matrix

$$\Gamma_k(t) = Cov(X_t, X_{t-k}) = E[(X_t - \mu)(X_{t-k} - \mu)^T], \Gamma_k(t) \in \mathbb{R}^{K\times K}.$$

Definition 21.3.3 (weakly stationary). *The multivariate time series X_t is weakly stationary if $\mu(t)$ is constant and $\Gamma_k(t)$ does not depend on t.*

Definition 21.3.4 (Matrix polynomial of lag operator). *A matrix polynomial of lag operator of order pis defined as*

$$\Theta_p(B) = I_k + \Theta_1 B + \Theta_2 B^2 + ... + \Theta_p B^p$$

where B is the lag operator(we can view as a special scalar), $I_k, \Theta_i \in \mathbb{R}^{k\times k}$

Definition 21.3.5 (invertibility). *[9, p. 7] A multivariate(m dimension) time series z_t is said to be **invertible** if it can be written as*

$$z_t = c + a_t + \sum_{j=1}^{\infty} \pi_j z_{t-j},$$

where

- $c \in \mathbb{R}^m$ *is a constant vector;*
- $\pi_i \in \mathbb{R}^{m \times m}, i = 1, 2, \ldots$ *are matrices;*
- $\{a_t\}$ *is a white noise process with* $Var[a_t] = \Sigma_a \in \mathbb{R}^{m \times m}$.

To ensure convergence, we also require $\pi_i \to 0$ *as* $i \to \infty$.

21.3.2 Vector autoregressive models

21.3.2.1 *VAR(1) model*

Definition 21.3.6 (VAR(1)). *An m dimensional multivariate time series* X_t *is called* $VAR(1)$ *process if*

$$X_t = C + \Phi_1 X_{t-1} + \eta_t$$

where $X_t \in \mathbb{R}^m, \Phi_i \in \mathbb{R}^{m \times m}, \eta_t \sim MN(0, \Sigma), \Sigma \in \mathbb{R}^{m \times m}, \eta_t, C \in \mathbb{R}^m$.

Example 21.3.1. A bivariate VAR(1) model is given explicitly by

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \begin{bmatrix} \Phi_{1,11} & \Phi_{1,12} \\ \Phi_{1,21} & \Phi_{1,22} \end{bmatrix} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix}$$

Lemma 21.3.1 (mean-adjusted of VAR(1)). *A stable VAR(1) process* $X_t = C + AX_{t-1} + u_t$ *can be written as*

$$X_t - \mu = A(X_{t-1} - \mu) + u_t$$

where $\mu = (I_K - A)^{-1} C$.

Proof.

$$X_t - \mu = A(X_{t-1} - \mu) + u_t \implies X_t - (I - A)\mu = AX_{t-1} + u_t.$$

Therefore, we can set $(I - A)\mu = C$. □

Lemma 21.3.2 (stationary condition for VAR(1) process). *[9, p. 32] A VAR(1) process*

$$X_t = C + AX_{t-1} + u_t$$

is stationary if all eigenvalues of A satisfy $|\lambda| < 1$. *Or equivalent*

$$det(I_K - Az) \neq 0, \forall z \in \mathbb{C}, |z| \leq 1.$$

*Such VAR(1) process is said to be **stationary or stable**.*

Proof. (1)

$$\begin{aligned} X_t &= AX_{t-1} + u_t \\ &= A(AX_{t-2} + u_{t-1}) + u_t \\ &= \vdots \\ &= A^{t-v}X_v + \sum_{i=0}^{t-1} A^i u_{t-i} \end{aligned}$$

As $v \to -\infty$, A^∞ will have a limit if the spectral radius of A is smaller than 1 [Theorem 4.13.4]. (2) To show that equivalence, suppose there exists a z_0 such that $|z_0| \leq 1$ and $det(I_K - Az_0) = 0$. Then

$$det(I_K - Az_0) = z_0^K det(I_k/z_0 - A) = 0;$$

that is, there exists $1/z_0 = \lambda_0, |\lambda_0| \geq 1$ as the eigenvalue of A. □

Note 21.3.1 (implications of stability). From Theorem 4.13.5, a stable VAR(1) process will ensure $(I_K - A)^{-1}$ exist; and

$$(I_K - A)^{-1} = I_K + A + A^2 + ...$$

Then

$$X_t = (I_K - AB)^{-1}(C + u_t),$$

where B is the lag operator.

Lemma 21.3.3 (Vector moving averaging representation of stable VAR(1)). *[9, p. 36] A stable K dimensional VAR(1) process*

$$X_t = C + AX_{t-1} + u_t$$

has the moving averaging representation as

$$X_t = (I_K + A + A^2 + ...)C + (I_K u_t + Au_{t-1} + A^2 u_{t-2} + ...),$$

or equivalently,

$$X_t = (I_K - A)^{-1}C + (I_K - AB)^{-1}u_t,$$

where B is the lag operator.

Proof. Directly recursively expand the original equation. □

Lemma 21.3.4 (basic properties of a stable VAR(1)). *[10, p. 16] A **stable** K dimensional VAR(1) process*

$$X_t = C + AX_{t-1} + u_t$$

has

- *constant mean* $\mu = (I_K - A)^{-1}C$.
- *constant covariance variance:* $Cov(X_t, X_t) = E[(X_t - \mu)^2] = (I_K - A)^{-1}Cov(\mu_t, \mu_t)(I_K - A)^{-T} = \sum_{i=0}^{\infty} A^{h+i}Cov(u_t, u_t)(A^i)^T$.
- *shift-invariant cross-covariance matrix:*

$$\Gamma(h) = \sum_{i=0}^{\infty} A^{h+i}\Sigma_u(A^i)^T.$$

- *it is a weakly stationary process; moreover, since* u_t *is multivariate Gaussian, it is also a strongly stationary process.*

Proof. (1)(2) use moving average representation [Lemma 21.3.3] to prove.

$$X_t = (I_K + A + A^2 + ...)C + (I_K u_t + Au_{t-1} + A^2 u_{t-2} + ...)$$
$$X_t = (I_K - A)^{-1}C + (I_K - A)^{-1}u_t \implies E[X_t] = (I_K - A)^{-1}C$$

(3) Note that

$$X_t - \mu = (I_K - AB)^{-1}u_t.$$

Then,

$$Cov(X_{t+h}, X_t) = E[(X_{t+h} - \mu)(X_t - \mu)^T] = A^h E[(X_t - \mu)^(X_t - \mu)^T].$$

(4) directly from (1)(2)(3). □

21.3.2.2 *VAR(2) model*

Definition 21.3.7 (VAR(2)). *An m dimensional multivariate time series* X_t *is called* $VAR(2)$ *process if*

$$X_t = C + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \eta_t$$

where $X_t \in \mathbb{R}^m, \Phi_1, \Phi_2 \in \mathbb{R}^{m\times m}, \eta_t \sim MN(0, \Sigma), \Sigma \in \mathbb{R}^{m\times m}, \eta_t, C \in \mathbb{R}^m$.

Lemma 21.3.5 (VAR(1) representation of VAR(2) process). *Every m dimensional $VAR(2)$ process X_t is equivalent to a $2m$ dimensional $VAR(1)$ process by using the following transformation procedures: Define*

$$Z_t = (X_t^T, X_{t-1}^T)^T$$
$$Z_{t-1} = (X_{t-1}^T, X_{t-2}^T)^T$$

where $Z_t \in \mathbb{R}^{2m}$. Then

$$Z_t = D + AZ_{t-1} + F$$

where $D \in \mathbb{R}^{2m}, A \in \mathbb{R}^{2m \times 2m}, F \in \mathbb{R}^{2m}$ are given as

$$D = \begin{bmatrix} C \\ 0_m \end{bmatrix}, A = \begin{pmatrix} \Phi_1 & \Phi_2 \\ I_m & 0 \end{pmatrix}, F = \begin{bmatrix} \eta_t \\ 0_m \end{bmatrix}$$

Lemma 21.3.6 (stability condition for VAR(2) process). *[9, p. 38] The stability condition for a VAR(2) process is*

$$det(I_{2m} - \phi_1 z - \phi_2 z^2) \neq 0, \forall |z| \leq 1.$$

Proof. Use the equivalent representation VAR(1) of VAR(2) and stability condition for VAR(1) Lemma 21.3.2. We have

$$\begin{aligned} det(I_{2m} - Az) &= \begin{vmatrix} I_m - \phi_1 z & -\phi_2 z \\ -I_m z & I_m \end{vmatrix} \\ &= \begin{vmatrix} I_m - \phi_1 z - \phi_2 z^2 & -\phi_2 z \\ 0 & I_m \end{vmatrix} \\ &= \left| I_m - \phi_1 z - \phi_2 z^2 \right| \end{aligned}$$

where we multiply the second column block matrix by z and add to the first column(which will not change the determinant value [Lemma A.8.12]). □

21.3.2.3 *VAR(p) model*

Definition 21.3.8 (VAR operator). *A VAR operator of order q is a matrix polynomial of lag operator given as*

$$\Phi_p(B) = I_k - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p$$

where B is the lag operator, $\Theta_i \in \mathbb{R}^{k\times k}$

Definition 21.3.9 (VAR(p) process). *[5, p. 27] An m dimensional multivariate time series X_t is called $VAR(p)$ process if*

$$X_t = C + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \eta_t$$

where $X_t \in \mathbb{R}^m, \Phi_i \in \mathbb{R}^{m\times m}, \eta_t \sim MN(0,\Sigma), \Sigma \in \mathbb{R}^{m\times m}, \eta_t, C \in \mathbb{R}^m$. Using VAR operator, we have

$$\Phi_p(B)X_t = C + \eta_t$$

Lemma 21.3.7 (VAR(1) representation of VAR(p) process). *Every m dimensional $VAR(p)$ process X_t is equilvalent to a mp dimensional $VAR(1)$ process by using the following transformation procedures: Define*

$$Z_t = (X_t^T, X_{t-1}^T, ..., X_{t-p+1}^T)^T$$
$$Z_{t-1} = (X_{t-1}^T, X_{t-2}^T, ..., X_{t-p}^T)^T$$

where $Z_t \in \mathbb{R}^{m\times p}$. Then

$$Z_t = D + AZ_{t-1} + F$$

where $D \in \mathbb{R}^{mp}, A \in \mathbb{R}^{mp\times mp}, F \in \mathbb{R}^{mp}$ are given as

$$D = \begin{bmatrix} C \\ 0_m \\ 0_m \\ \vdots \\ 0_m \\ 0_m \end{bmatrix}, A = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \dots & \dots & \phi_p \\ I_m & 0 & 0 & \dots & \dots & 0 \\ 0 & I_m & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & & & 0 \\ 0 & 0 & \ddots & I_m & 0 & 0 \\ 0 & 0 & \dots & 0 & I_m & 0 \end{pmatrix}, F = \begin{bmatrix} \eta_t \\ 0_m \\ 0_m \\ \vdots \\ 0_m \\ 0_m \end{bmatrix}$$

Lemma 21.3.8 (stability condition for VAR(p) process). *The stability condition for a VAR(p) process is*

$$det(I_m - \phi_1 z - ... - \phi_p z^p) \neq 0, \forall |z| \leq 1.$$

or equivalently,

$$det(I_{mp} - \boldsymbol{A}z) \neq 0, \forall |z| \leq 1.$$

Proof. Use the VAR(1) representation and stability condition for VAR(1) process. Note that

$$\begin{aligned}
det(I_{mp} - Az) &= \begin{vmatrix} I_m - \phi_1 z & -\phi_2 z & -\phi_3 z & \dots & \dots & -\phi_p z \\ -I_m z & I_m & 0 & \dots & \dots & 0 \\ 0 & -I_m z & I_m & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & & & 0 \\ 0 & 0 & \ddots & -I_m z & I_m & 0 \\ 0 & 0 & \dots & 0 & -I_m z & I_m \end{vmatrix} \\
&= \begin{vmatrix} I_m - \phi_1 z - \phi_2 z^2 - \cdots \phi_p z^p & -\phi_2 z - \phi_3 z^2 - \cdots & -\phi_3 z & \dots & \dots & -\phi_p z \\ 0 & I_m & 0 & \dots & \dots & 0 \\ 0 & & I_m & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & & & 0 \\ 0 & 0 & \ddots & 0 & I_m & 0 \\ 0 & 0 & \dots & 0 & 0 & I_m \end{vmatrix} \\
&= \left| I_m - \phi_1 z - \phi_2 z^2 - \cdots \phi_p z^p \right|
\end{aligned}$$

Note that we iterative eliminate the off-diagonal entries in the lower triangle matrix(starting from the p-1 column, then p-2 column, until the first column). □

Lemma 21.3.9 (invertibility of VAR(p) operator and MA representation). *[10, p. 23] Consider the VAR(p) process given as*

$$y_t = C + (A_1 B + ... + A_p B^p) y_t + u_t$$

and define

$$A(B) = I_K - A_1 B - ... - A_p B^p.$$

We have

- *If* $det(I_K - A_1 z - ... - A_p z^p) \neq 0, \forall |z| \leq 1$, *then* $A(B)$ *is invertible.*
- *If* $A(B)$ *is invertible, let* $\Phi(B)$ *denote its inverse such that* $\Phi(B)A(B) = I_K$, *then*

$$\Phi(B) = \sum_{i=1}^{\infty} \Phi_i B^i$$

where

$$\begin{aligned} I_K &= \Phi_0 \\ 0 &= \Phi_1 - \Phi_0 A_1 \\ 0 &= \Phi_2 - \Phi_1 A_1 - \Phi_0 A_2 \\ &\vdots \\ 0 &= \phi_i - \sum_{j=1}^{i} \Phi_{i-j} A_j \vdots \end{aligned}$$

Corollary 21.3.0.1 (MA representation). *A stable* $VAR(p)$ *process has its moving average representation as*

$$y_t = \Phi(B)C + \Phi(B)u_t$$

where $\Phi(B) = (I_K - A_1 B - ... - A_p B^p)^{-1}$.

Remark 21.3.1 (connection with Wold decomposition). The MA representation is an example of the Wold decomposition for stationary multivariate time series [Theorem 21.2.7].

Lemma 21.3.10 (mean of VAR(p), and mean-adjusted form). *For a stable Var(p) process* X_t, *we have*

- $\mu = \Phi(B)C$, *where* $\Phi(B) = (I_K - A_1 B - ... - A_p B^p)^{-1}$.
- *It can be written as*

$$X_t - \mu = A_1(X_{t-1} - \mu) + A_2(X_{t-2} - \mu) + ... + A_p(X_{t-p} - \mu) + \eta_t.$$

21.3.3 Vector moving-average model

Definition 21.3.10 (VMA operator). *A VAR operator of order q is a matrix polynomial of lag operator given as*

$$\Theta_p(B) = I_k - \Theta_1 B - \Theta_2 B^2 - ... - \Theta_p B^p$$

where B is the lag operator, $\Theta_i \in \mathbb{R}^{k \times k}$

Definition 21.3.11 (VMA(q) model). *[9, p. 106] A VMA model of order q for a m dimensional time series z_t is given by*

$$z_t = \mu + a_t - \sum_{i=1}^{q} \theta_i a_{t-i},$$

where

- *$\mu \in \mathbb{R}^m$ is a constant vector;*
- *$\theta_i \in \mathbb{R}^{m \times m}, i = 1, 2, ..., q$ are matrices and $\theta_q \neq 0$;*
- *$\{a_t\}$ is a white noise process with $Var[a_t] = \Sigma_a \in \mathbb{R}^{m \times m}$.*

  Using VMA operator, the VMA(q) model can also be written as

$$z_t = \mu + \Theta_q a_t.$$

Lemma 21.3.11 (properties of VMA(q) model). *[9, p. 110] Consider a VMA(q) model of a time series z_t. It follows that*

- *$E[z_t] = \mu$.*
- *$Var[z_t] = \Gamma_0 = \Sigma_a + \sum_{i=1}^{q} \theta_i \Sigma_a \theta_i^T$.*
- $$Cov(z_t, z_{t-j}) = \Gamma_j = \sum_{i=j}^{q} \theta_j \Sigma_a \theta_{i-j}^T,$$

  where $\theta_0 = -I_k, \forall j = 1, ..., q$.
- *$\Gamma_j \triangleq Cov(z_t, z_{t-j}) = 0, \forall j > q$.*

Proof. (1)

$$E[z_t] = E[\mu + a_t - \sum_{i=1}^{q} \theta_i a_{t-i}] = \mu + 0.$$

(2)

$$\begin{aligned} Var[z_t] &= E[(z_t - \mu)(z_t - \mu)^T] \\ &= E[\Theta_q a_t a_t^T \Theta_q^T] \\ &= E[a_t a_t^T] + \sum_{i=1}^{q} \theta_i \Sigma_a \theta_i^T \end{aligned}$$

(3)(4) Note that

$$\begin{aligned} Var[z_t] &= E[(z_t - \mu)(z_t - \mu)^T] \\ &= E[\Theta_q B^j a_t a_t^T \Theta_q^T] \end{aligned}$$

□

21.4 Autoregressive conditional heteroscedastic model

21.4.1 ARCH models

21.4.1.1 *The motivation and the model*

In financial time series application, observations like daily returns of SP500 exhibit a phenomenon known as **volatility clustering**, where large daily returns tend to be followed by large daily returns and small daily returns tend to be followed by daily returns [Figure 21.4.1]. For example, during the 2008 financial crisis, stock price fluctuate strongly due to the sell-off and buying-dip behavior of investors.

Clearly, for linear processes we covered so far do not display this feature since they have constant magnitude of the shocks (i.e., constant σ). In this section, we add additional modeling on the σ and allow σ to be have its own dynamic features. These type of models are known as **autoregressive conditional heteroscedastic model (ARCH)**, meaning σ will follow an autoregressive process.

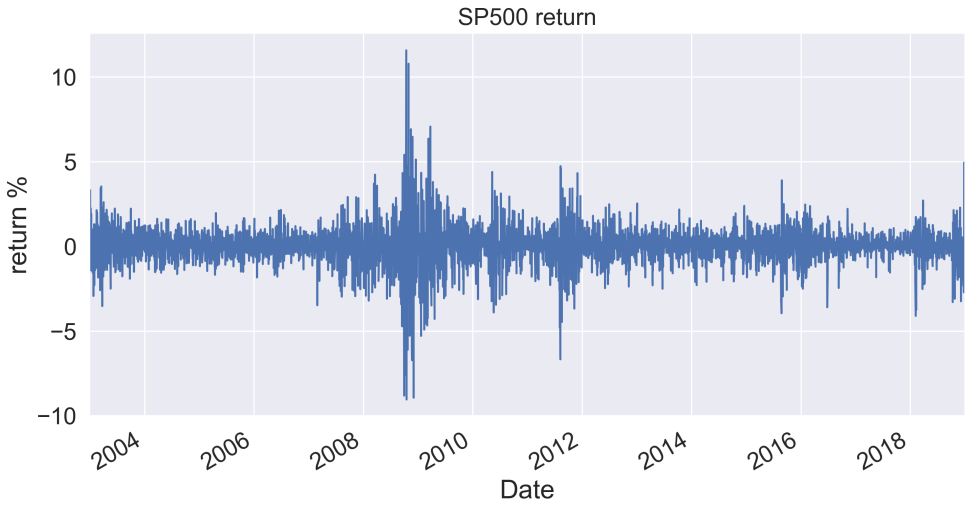

Figure 21.4.1: Stock index SP500 daily return between 2014 and 2019.

Definition 21.4.1 (autoregressive conditional heteroscedastic model of order q, ARCH(q)). *Let a_t denotes the error terms. The ARCH(q) model assumes*

$$a_t = \sigma_t \epsilon_t$$

where $\epsilon_t \sim WN(0,1)$*, and*

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2, \alpha_i \geq 0, \forall i = 0, ..., q$$

Definition 21.4.2 (autoregressive conditional heteroscedastic model of order 1, ARCH(1)). *Let* a_t *denotes the error terms. The ARCH(1) model assumes*

$$a_t = \sigma_t \epsilon_t$$

where $\epsilon_t \sim WN(0,1)$*, and*

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2, \alpha_0, \alpha_1 \geq 0.$$

We also require $\alpha_1 < 1$ *for stationarity.*

Example 21.4.1 (ARCH(1)). Figure 21.4.2 we show the simulated representative trajectories from ARCH(1) model with coefficients $a = 0.9$ and $a = 0.5$. Volatility peaks and clustering are more often For larger α.

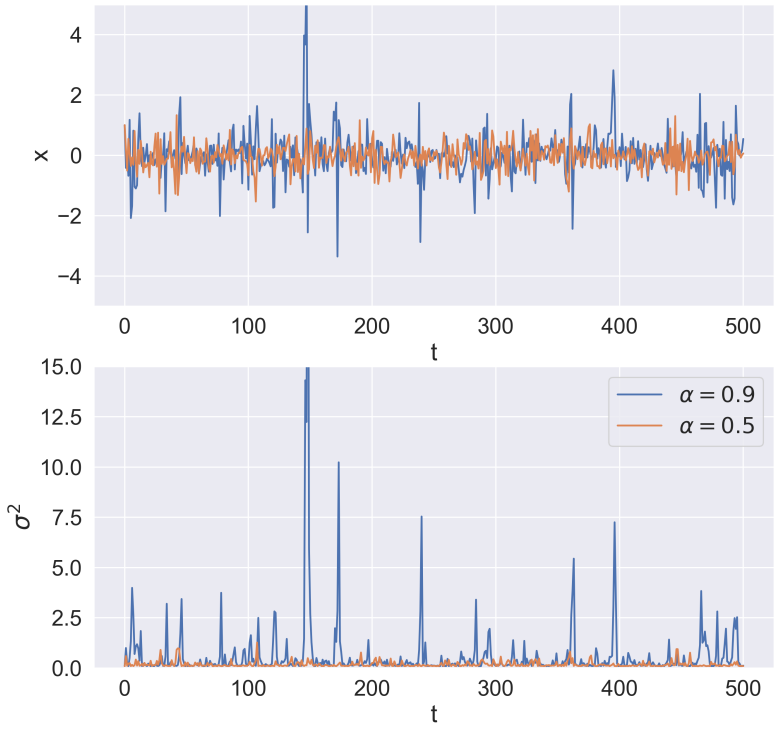

Figure 21.4.2: Simulated representative trajectories from ARCH(1) model with coefficients $a = 0.9$ and $a = 0.5$.

Remark 21.4.1 (intuition and characteristics).

- a_t can be simply viewed as a 'special' white noise process in which the variance $E[a_t^2]$ is correlated with previous history.
- Note that a_t is uncorrelated with previous history a_{t-1} even though σ_t is correlated, i.e.
$$E[a_t a_{t-1}] = E[\sigma_t \sigma_{t-1} \epsilon_t \epsilon_{t-1}] = E[\sigma_t \sigma_{t-1}] E[\epsilon_t \epsilon_{t-1}] = 0.$$
- a_t is stationary if σ_t is stationary. It can be showed that when $\alpha_0 > 0, 0 < \alpha_1 < 1$, σ_t is stationary. Since ϵ_t and σ_t are both stationary and independent, then $\sigma_t \epsilon_t$ is stationary. More resulted are showed in the following sections.

21.4.1.2 *Statistical properties*

Lemma 21.4.1 (weak stationarity of ARCH(1)). *The ARCH(1) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2, \alpha_1 \in [0, 1)$$

has the following properties:

- *It has the constant mean $E[a_t] = 0$, and*

$$Var[a_t] = \frac{\alpha_0}{1-\alpha_1}, Cov(a_t, a_s) = \delta(s,t)\frac{\alpha_0}{1-\alpha_1}.$$

  That is, ARCH(1) with $\alpha_1 \in [0,1)$ is a weakly stationary process.
- *(stationary fourth moment) If $\alpha_1 < \sqrt{1/3}$, then a_t has stationary fourth moment given by*

$$E[a_t^4] = 3\frac{\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

Proof. (1)(a)

$$E[a_t] = E[\sigma_t \epsilon_t] = E[\sigma_t]E[\epsilon_t] = 0.$$

(b) Note that

$$Cov(a_t, a_s) = E[a_t^2 a_s^2] = E[\sigma_t \sigma_s \epsilon_t \epsilon_s] = E[\sigma_t \sigma_s]E[\epsilon_t^2]E[\epsilon_s^2] = 0, \ if \ t \neq s.$$

If $t = s$, $Cov(a_t, a_s) = E[\sigma_t^2]$ In the following we will calculate $E[\sigma_t^2]$. Note that we can also write ARCH(1) as

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 \sigma_{t-1}^2.$$

Then

$$\begin{aligned} E[\sigma_t^2] &= \alpha_0 + \alpha_1 E[\sigma_{t-1}^2] \\ &= \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 E[\sigma_{t-2}^2]) \\ &= \alpha_0 + \alpha_1\alpha_0 + \alpha_1(\alpha_0 + \alpha_1 E[\sigma_{t-3}^2]) \\ &= \alpha_0(1 + \alpha_1 + \alpha_1^2 + ...) \\ &= \frac{\alpha_0}{1-\alpha_1} \end{aligned}$$

(2)

$$\begin{aligned}
E[a_t^4] &= E[E[a_t^4|\mathcal{F}_{t-1}]] \\
&= E[E[\epsilon_t^4(\alpha_0+\alpha_1 a_{t-1}^2)^2|\mathcal{F}_{t-1}]] \\
&= E[\epsilon_t^4 E[(\alpha_0+\alpha_1 a_{t-1}^2)^2|\mathcal{F}_{t-1}]] \\
&= 3E[E[(\alpha_0+\alpha_1 a_{t-1}^2)^2|\mathcal{F}_{t-1}]] \\
&= 3E[(\alpha_0^2+2\alpha_0\alpha_1 a_{t-1}^2+\alpha_1^2 a_{t-1}^4)] \\
&= 3(\alpha_0^2+2\alpha_0\alpha_1 E[a_{t-1}^2]+\alpha_1^2 E[a_{t-1}^4] \\
&= 3(\alpha_0^2+2\alpha_0\alpha_1\frac{\alpha_0}{1-\alpha_1}+\alpha_1^2 E[a_{t-1}^4] \\
&= 3K+\beta E[a_{t-1}^4], K=\alpha_0^2+2\alpha_0\alpha_1\frac{\alpha_0}{1-\alpha_1}=\frac{\alpha_0(1+\alpha_1)}{1-\alpha_1}, \beta=3\alpha_1^2 \\
&= 3K+3K\beta+\beta^2 E[a_{t-2}^4] \\
&= 3K(1+\beta+\beta^2+...) \\
&= \frac{3K}{1-\beta}
\end{aligned}$$

where we use $E[\epsilon_t^4]=3$. □

Theorem 21.4.1 (basic statistical properties of ARCH(1)). *[5, p. 118] Consider an ARCH(1) model given by*

$$a_t=\sigma_t\epsilon_t, \sigma_t^2=\alpha_0+\alpha_1 a_{t-1}^2, \alpha_1\in[0,1).$$

Let $\mathcal{F}_{t-1}$ denote the information up to time $t-1$. It follows that

- *(unconditional mean and variance)*

$$E[a_t]=0, Var[a_t]=E[a_t^2]=\frac{\alpha_0}{1-\alpha_1}, Cov(a_t,a_s)=\delta(s,t)\frac{\alpha_0}{1-\alpha_1}.$$

  As a result, we can write

$$\sigma_t^2=(1-\alpha_1)Var[a_t]+\alpha_1 a_{t-1}^2.$$

- *(conditional mean and variance) a_t has conditional mean and variance given by*

$$E[a_t|\mathcal{F}_{t-1}]=0$$

$$E[a_t^2|\mathcal{F}_{t-1}]=Var[a_t|\mathcal{F}_{t-1}]=\sigma_t^2=(\alpha_0+\alpha_1 a_{t-1}^2).$$
$$E[a_t^2|\mathcal{F}_{t-2}]=Var[a_t|\mathcal{F}_{t-1}]=\sigma_t^2=(\alpha_0+\alpha_1 a_{t-1}^2).$$

- *(covariance structure of squares)*

$$E(a_t^2, a_s^2) =$$

$$Cov(a_t^2, a_s^2) =$$

In particular $\alpha_0 = 0$,

- *(heavy tail property)*

$$kurt(a_t) = \frac{E[a_t^4]}{(E[a_t^2])^2} = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3.$$

Proof. (1)(a)

$$E[a_t] = E[\sigma_t \epsilon_t] = E[\sigma_t]E[\epsilon_t] = 0.$$

(b)

$$\begin{aligned} Var[a_t] &= E[a_t^2] - (E[a_t])^2 \\ &= E[a_t^2] \\ &= E[\sigma_t^2 \epsilon_t^2] \\ &= E[\sigma_t^2] \\ &= \alpha_0 + \alpha_1 E[a_{t-1}^2] \end{aligned}$$

use the fact that a_t is a stationary process, thus $E[a_{t-1}^2] = E[a_t^2]$, we have $Var[a_t] = \frac{\alpha_0}{1-\alpha_1}$.
(2) (a)

$$\begin{aligned} E[a_t|\mathcal{F}_{t-1}] &= E[\epsilon_t \sqrt{\alpha_0 + \alpha_1 a_{t-1}^2}|\mathcal{F}_{t-1}] \\ &= E[\epsilon_t|\mathcal{F}_{t-1}]E[\sqrt{\alpha_0 + \alpha_1 a_{t-1}^2}|\mathcal{F}_{t-1}] \\ &= 0 \end{aligned}$$

(b)

$$\begin{aligned} Var[a_t|\mathcal{F}_{t-1}] &= E[a_t^2|\mathcal{F}_{t-1}] - (E[a_t|\mathcal{F}_{t-1}])^2 \\ &= E[\epsilon_t^2|\mathcal{F}_{t-1}](\alpha_0 + \alpha_1 a_{t-1}^2) - 0 \\ &= (\alpha_0 + \alpha_1 a_{t-1}^2) \end{aligned}$$

(3)

$$\begin{aligned}
E[a_t^2 a_s^2] &= E[\epsilon_t^2 \epsilon_s^2 \sigma_t^2 \sigma_s^2] \\
&= E[\epsilon_t^2] E[\epsilon_s^2] E[\sigma_t^2 \sigma_s^2] \\
&= E[(\alpha_0 + \alpha_1 a_{t-1}^2)(\alpha_0 + \alpha_1 a_{s-1}^2)] \\
&= E[(\alpha_0^2 + \alpha_1^2 a_{t-1}^2 a_{s-1}^2 + \alpha_0 \alpha_1 a_{t-1}^2 + \alpha_0 \alpha_1 a_{s-1}^2)] \\
&= \alpha_0^2 + 2\alpha_0 \alpha_1 \frac{\alpha_0}{1-\alpha_1} + \alpha_1^2 E[a_{t-1}^2 a_{s-1}^2] \\
&= K + \alpha_1^2 E[a_{t-1}^2 a_{s-1}^2], K = \alpha_0^2 + 2\alpha_0 \alpha_1 \frac{\alpha_0}{1-\alpha_1} = \frac{\alpha_0^2(1+\alpha_1)}{1-\alpha_1} \\
&= \frac{K}{1-\alpha_1^2} \\
&= \frac{\alpha_0^2}{(1-\alpha_1)^2}
\end{aligned}$$

(4)Note that $E[a_t^2] = Var[a_t^2] = \alpha_0/(1-\alpha_1)$ and

$$E[a_t^4] = 3\frac{\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)},$$

from Lemma 21.4.1. □

Remark 21.4.2 (interpretation). [5, p. 116]

- (predictability) a_t cannot be predicted based on history because of zero covariance; a_t^2 can be predicted from history because

$$E[a_t^2 | \mathcal{F}_{t-1}] = Var[a_t | \mathcal{F}_{t-1}] = \sigma_t^2 = (\alpha_0 + \alpha_1 a_{t-1}^2).$$

- From

$$\sigma_t^2 = (1-\alpha_1) Var[a_t] + \alpha_1 a_{t-1}^2,$$

large variations in a_t tend to be followed by large variations and small variations tend to be followed by small variations.

- $$kurt(a_t) = \frac{E[a_t^4]}{(E[a_t^2])^2} = \frac{E[\sigma_t^4] E[\epsilon_t^4]}{(E[\sigma_t^2] E[\epsilon_t^2])^2} = \frac{3E[\sigma_t^4]}{(E[\sigma_t^2])^2} > 3$$

where we use the fact that $E[\sigma_t^4] > (E[\sigma_t^2])^2$ due to

$$Var[a_t^2] = E[a_t^4] - (E[a_t^2])^2 > 0.$$

- Intuitively, we can view a_t as a mixture of Gaussian, which has fat tails.

Theorem 21.4.2 (basic statistical properties of ARCH(q)). *[5, p. 118] Consider a weakly stationary ARCH(q) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + \cdots + \alpha_q a_{t-q}^2.$$

Let $\mathcal{F}_{t-1}$ denote the information up to time $t-1$. It follows that

- *(unconditional mean and variance)*

$$E[a_t] = 0, Var[a_t] = E[a_t^2] = \frac{\alpha_0}{1 - \sum_{i=1}^q \alpha_i}, Cov(a_t, a_s) = \delta(s,t) \frac{\alpha_0}{1 - \sum_{i=1}^q \alpha_i}.$$

- *(conditional mean and variance) a_t has conditional mean and variance given by*

$$E[a_t | \mathcal{F}_{t-1}] = 0, E[a_t^2 | \mathcal{F}_{t-1}] = Var[a_t | \mathcal{F}_{t-1}] = \sigma_t^2 = (\alpha_0 + \sum_{i=1}^q \alpha_i a_{t-i}^2).$$

- *(mean prediction) Let $t < s$, then*

$$E[a_t | \mathcal{F}_s] = 0,$$

  that is, a_t is unpredictable.

Proof. (1)(a)

$$E[a_t] = E[\sigma_t \epsilon_t] = E[\sigma_t] E[\epsilon_t] = 0.$$

(b)

$$\begin{aligned} Var[a_t] &= E[a_t^2] - (E[a_t])^2 \\ &= E[a_t^2] \\ &= E[\sigma_t^2 \epsilon_t^2] \\ &= E[\sigma_t^2] \\ &= \alpha_0 + \sum_{i=1}^q \alpha_i E[a_{t-i}^2] \end{aligned}$$

use the fact that a_t is a stationary process, thus $E[a_{t-i}^2] = E[a_t^2]$, we have $Var[a_t] = \frac{\alpha_0}{1-\sum_{i=1}^{q}\alpha_i}$.
(2) (a)

$$\begin{aligned} E[a_t|\mathcal{F}_{t-1}] &= E[\epsilon_t\sqrt{\alpha_0 + \sum_{i=1}^{q}\alpha_i a_{t-i}^2}|\mathcal{F}_{t-1}] \\ &= E[\epsilon_t|\mathcal{F}_{t-1}]E[\sqrt{\alpha_0\sum_{i=1}^{q}\alpha_i a_{t-i}^2}|\mathcal{F}_{t-1}] \\ &= 0 \end{aligned}$$

(b)

$$\begin{aligned} Var[a_t|\mathcal{F}_{t-1}] &= E[a_t^2|\mathcal{F}_{t-1}] - (E[a_t|\mathcal{F}_{t-1}])^2 \\ &= E[\epsilon_t^2|\mathcal{F}_{t-1}](\alpha_0 + \sum_{i=1}^{q}\alpha_i a_{t-i}^2) - 0 \\ &= (\alpha_0 + \sum_{i=1}^{q}\alpha_i a_{t-i}^2) \end{aligned}$$

(3)

$$E[a_t|\mathcal{F}_s] = E[\sigma_t\epsilon_t|\mathcal{F}_s] = E[\sigma_t|\mathcal{F}_s]E[\epsilon_t|\mathcal{F}_s] = 0,$$

note that σ_t and ϵ_t are independent because σ_t depends on the previous shocks that are independent of ϵ_t. □

21.4.1.3 *Variance forecasting*

Lemma 21.4.2 (conditional expectation equivalence between variance and square innovations for ARCH(q)). *Consider an ARCH(q) model given by*

$$a_t = \sigma_t\epsilon_t, \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i a_{t-i}^2, \sum_{i=1}^{q}\alpha_i \in [0,1).$$

Let $\mathcal{F}_t$ denote the information available up to time t. Then

$$E[\sigma_t^2|\mathcal{F}_s] = E[a_t^2|\mathcal{F}_s], s < t.$$

Proof. Note that

$$a_t^2 = \sigma_t^2\epsilon_t^2.$$

Take conditional expectation on both sides and get

$$E[a_t^2|\mathcal{F}_s] = E[\sigma_t^2\epsilon_t^2|\mathcal{F}_s] = E[\sigma_t^2|\mathcal{F}_s]E[\epsilon_t^2|\mathcal{F}_s] = E[\sigma_t^2|\mathcal{F}_s].$$

where we used the fact that $E[\epsilon_t^2|\mathcal{F}_s] = 1$ and the independence between σ_t and ϵ_t. □

Theorem 21.4.3 (variance prediction in ARCH(1)). *Consider an ARCH(1) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2, \alpha_1 \in [0, 1).$$

Let $\mathcal{F}_t$ denote the information available up to time t. Then,

$$\begin{aligned}
E[a_{t+1}^2|\mathcal{F}_t] = E[\sigma_{t+1}^2|\mathcal{F}_t] &\triangleq \sigma_{t+1}^2 = \alpha_0 + \alpha_1 a_t^2 \\
E[a_{t+2}^2|\mathcal{F}_t] = E[\sigma_{t+2}^2|\mathcal{F}_t] &\triangleq \hat{\sigma}_{t+2}^2 = \sigma^2 + (\alpha_1)(\sigma_{t+1}^2 - \sigma^2) \\
E[a_{t+3}^2|\mathcal{F}_t] = E[\sigma_{t+3}^2|\mathcal{F}_t] &\triangleq \hat{\sigma}_{t+3}^2 = \sigma^2 + (\alpha_1)^2(\sigma_{t+1}^2 - \sigma^2) \\
&\vdots \\
E[a_{t+l}^2|\mathcal{F}_t] = E[\sigma_{t+l}^2|\mathcal{F}_t] &\triangleq \hat{\sigma}_{t+l}^2 = \sigma^2 + (\alpha_1)^{l-1}(\sigma_t^2 - \sigma^2)
\end{aligned}$$

where σ^2 is the unconditional variance

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1}.$$

Proof. Note that the equivalence between square innovation and variance is discussed in Lemma 21.4.2. Further note that σ_{t+1}^2 given $\mathcal{F}_t$ is actually deterministic quantity. For the rest, we have

$$\begin{aligned}
\sigma_{t+1}^2 &= \alpha_0 + \alpha_1 a_t^2 \\
E[\sigma_{t+2}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+2}^2 &= \alpha_0 + \alpha_1 E[a_{t+1}^2|\mathcal{F}_t] \\
&= \alpha_0 + \alpha_1 \sigma_{t+1}^2 \\
&= \sigma^2 + (\alpha_1)(\sigma_{t+1}^2 - \sigma^2) \\
E[\sigma_{t+3}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+3}^2 &= \alpha_0 + \alpha_1 E[a_{t+2}^2|\mathcal{F}_t] \\
&= \alpha_0 + \alpha_1 \hat{\sigma}_{t+2}^2 \\
&= \sigma^2 + \alpha_1(\hat{\sigma}_{t+2}^2 - \sigma^2) \\
&= \sigma^2 + (\alpha_1)^2(\sigma_{t+1}^2 - \sigma^2)
\end{aligned}$$

where we use the fact of conditional variance [Theorem 21.4.1] that

$$E[a_{t+1}^2|\mathcal{F}_t] = \sigma_{t+1}^2 = (\alpha_0 + \alpha_1 a_t^2).$$

and the fact that

$$E[a_{t+2}^2|\mathcal{F}_t] = E[\sigma_{t+2}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+2}^2$$

.

Others can be proved similarly. □

Remark 21.4.3 (implication for convergence rate). We can see that $\hat{\sigma}^2_{t+l} \to \sigma$ as $l \to \infty$. If the variance spikes up during a crisis, then the number of the period before the variance restore to equilibrium value σ can be estimated using the value of $(\alpha_1 + \beta_1)$(the larger the longer).

Theorem 21.4.4 (variance prediction in ARCH(q)). *Consider an ARCH(q) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2, \sum_{i=1}^{q} \alpha_i \in [0, 1).$$

Let $\mathcal{F}_t$ denote the information available up to time t. Then,

$$E[a_{t+1}^2|\mathcal{F}_t] = E[\sigma_{t+1}^2|\mathcal{F}_t] \triangleq \sigma_{t+1}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i+1}^2$$

$$E[a_{t+2}^2|\mathcal{F}_t] = E[\sigma_{t+2}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+2}^2 = \alpha_0 + \alpha_1 \hat{\sigma}_{t+1}^2 + \sum_{i=2}^{q} \alpha_i a_{t-i+2}^2$$

$$E[a_{t+3}^2|\mathcal{F}_t] = E[\sigma_{t+3}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+3}^2 = \alpha_0 + \sum_{i=1}^{2} \alpha_i \hat{\sigma}_{t-i+3}^2 + \sum_{i=3}^{q} \alpha_i a_{t-i+3}^2$$

$$\vdots$$

$$E[a_{t+l}^2|\mathcal{F}_t] = E[\sigma_{t+l}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+l}^2 = \alpha_0 + \sum_{i=1}^{l-1} \alpha_i \hat{\sigma}_{t-i+l}^2 + \sum_{i=l}^{q} \alpha_i a_{t-i+l}^2$$

$$\vdots$$

$$E[a_{t+\infty}^2|\mathcal{F}_t] = E[\sigma_{t+\infty}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+\infty}^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i}$$

where note that the unconditional variance

$$\sigma^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i}.$$

Proof.

$$\begin{aligned}
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2 \\
E[\sigma_{t+1}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+1}^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i E[a_{t-i+1}^2|\mathcal{F}_t] \\
&= \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i+1}^2 \\
E[\sigma_{t+2}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+2}^2 &= \alpha_0 + \alpha_1 E[a_{t-i+2}^2|\mathcal{F}_t] + \sum_{i=2}^{q} \alpha_i E[a_{t-i+2}^2|\mathcal{F}_t] \\
&= \alpha_0 + \alpha_1 \hat{a}_{t-i+1}^2 + \sum_{i=2}^{q} \alpha_i a_{t-i+1}^2 \\
E[\sigma_{t+3}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+3}^2 &= \alpha_0 + \sum_{i=1}^{2} \alpha_i E[a_{t-i+3}^2|\mathcal{F}_t] + \sum_{i=3}^{q} \alpha_i E[a_{t-i+3}^2|\mathcal{F}_t] \\
&= \alpha_0 + \sum_{i=1}^{2} \alpha_i \hat{\sigma}_{t-i+3}^2 + \sum_{i=3}^{q} \alpha_i a_{t-i+3}^2
\end{aligned}$$

Others can be proved similarly.

To calculate the variance prediction for $l \to \infty$, we use the result of unconditional variance. □

21.4.1.4 *Detect ARCH effect*

Lemma 21.4.3 (order determination in ARCH process via linear regression).

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2$$

Proof. Note that in our $ARCH(m)$ model, we have

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2.$$

Because a_t^2 is the unbiased estimator of σ_t^2, i.e.

$$E[a_t^2] = E[\sigma_t^2 \epsilon_t^2] = E[\sigma_t^2]$$

□

21.4.1.5 *Parameter estimation*

Lemma 21.4.4 (conditional likelihood estimation). *[5, p. 120]*

The conditional log-likelihood function is given by

$$L = \sum_{t=m+1}^{T} [-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_t^2) - \frac{1}{2}\frac{a_t^2}{\sigma_t^2}]$$

21.4.2 GARCH models

21.4.2.1 *The model*

ARCH models can be extended by including volatility/variance history. The extended model is **known as generalized autoregressive conditional heteroscedastic (GARCH)** model.

Definition 21.4.3 (generalized autoregressive conditional heteroscedastic model, general case). *[5, p. 132] Let a_t denotes the error terms. The GARCH(p,q) model assumes*

$$a_t = \sigma_t \epsilon_t$$

where $\epsilon_t \sim WN(0,1)$, and

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2, \alpha_i, \beta_j \geq 0, \forall i, j \geq 1$$

Definition 21.4.4 (generalized autoregressive conditional heteroscedastic model, order 1, GARCH(1,1)). *Let a_t denotes the error terms. The GARCH(q) model assumes*

$$a_t = \sigma_t \epsilon_t$$

where $\epsilon_t \sim WN(0,1)$, and

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \alpha_0, \alpha_1, \beta_1 \geq 0.$$

We also require $\alpha_1 + \beta_1 < 1$ for stationarity.

Lemma 21.4.5 (weak stationarity of GARCH(1,1)). *Consider a GARCH(1,1) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \alpha_1 + \beta_1 \in [0,1).$$

It follows that

- *It has the constant mean $E[a_t] = 0$, and*

$$Cov(a_t, a_s) = \delta(s,t)\frac{\alpha_0}{1-\alpha_1-\beta_1}.$$

  That is, GARCH(1,1) with $\alpha_1 + \beta_1 \in [0,1)$ is a weakly stationary process.
- *(stationary fourth moment)*

Proof. (1)

$$E[a_t] = E[\sigma_t \epsilon_t] = E[\sigma_t]E[\epsilon_t] = 0.$$

(2) Note that

$$Cov(a_t, a_s) = E[a_t a_s] = E[\sigma_t \sigma_s \epsilon_t \epsilon_s] = E[\sigma_t \sigma_s]E[\epsilon_t]E[\epsilon_s] = 0, \ if \ t \neq s.$$

where we use the independence between ϵ_t and σ_t. If $t = s$, $Cov(a_t, a_s) = E[\sigma_t^2]E[\epsilon_t^2] = E[\sigma_t^2]$. In the following we will calculate $E[\sigma_t^2]$. Note that we can also write GARCH(1,1) as

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 \sigma_{t-1}^2 + \beta_1 \epsilon_{t-1}^2 \sigma_{t-1}^2.$$

Then

$$\begin{aligned} E[\sigma_t^2] &= \alpha_0 + (\alpha_1 + \beta_1)E[\sigma_{t-1}^2] \\ &= \alpha_0 + (\alpha_1 + \beta_1)(\alpha_0 + (\alpha_1 + \beta_1)E[\sigma_{t-2}^2]) \\ &= \alpha_0 + (\alpha_1 + \beta_1)\alpha_0 + (\alpha_1 + \beta_1)(\alpha_0 + (\alpha_1 + \beta_1)E[\sigma_{t-3}^2]) \\ &= \alpha_0(1 + (\alpha_1 + \beta_1) + (\alpha_1 + \beta_1)^2 + ...) \\ &= \frac{\alpha_0}{1-\alpha_1-\beta_1} \end{aligned}$$

□

Theorem 21.4.5 (basic statistical properties of GARCH(1,1)). *[5, p. 132] The GARCH(1,1) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \alpha_1 + \beta_1 \in [0,1)$$

has the following property:

- *(unconditional mean and variance)*

$$E[a_t] = 0$$

and

$$Var[a_t] = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}.$$

- *(conditional mean and variance) a_t has conditional mean and variance given by*

$$E[a_t|\mathcal{F}_{t-1}] = 0,$$

$$E[a_t^2|\mathcal{F}_{t-1}] = Var[a_t|\mathcal{F}_{t-1}] = \sigma_t^2 = (\alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2).$$

- *(mean prediction) Let $t < s$, then*

$$E[a_t|\mathcal{F}_s] = 0,$$

that is, a_t is unpredictable.

Proof. (1)

$$E[a_t] = E[\sigma_t \epsilon_t] = E[\sigma_t]E[\epsilon_t] = 0.$$

$$\begin{aligned} Var[a_t] &= E[a_t^2] - (E[a_t])^2 \\ &= E[a_t^2] \\ &= E[\sigma_t^2 \epsilon_t^2] \\ &= E[\sigma_t^2] \\ &= \alpha_0 + \alpha_1 E[a_{t-1}^2] + \beta_1 \sigma_{t-1}^2 \\ &= \alpha_0 + (\alpha_1 + \beta_1) E[a_{t-1}^2] \end{aligned}$$

use the fact that a_t is a stationary process, thus $E[a_{t-1}^2] = E[a_t^2]$, we have $Var[a_t] = \frac{\alpha_0}{1-\alpha_1-\beta_1}$.
(2) (a)

$$\begin{aligned} E[a_t|\mathcal{F}_{t-1}] &= E[\epsilon_t \sqrt{\alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2}|\mathcal{F}_{t-1}] \\ &= E[\epsilon_t|\mathcal{F}_{t-1}] E[\sqrt{\alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2}|\mathcal{F}_{t-1}] \\ &= 0 \end{aligned}$$

(b)

$$\begin{aligned} Var[a_t|\mathcal{F}_{t-1}] &= E[a_t^2|\mathcal{F}_{t-1}] - (E[a_t|\mathcal{F}_{t-1}])^2 \\ &= E[\epsilon_t^2|\mathcal{F}_{t-1}](\alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2) - 0 \\ &= (\alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2) \end{aligned}$$

(3)

$$E[a_t|\mathcal{F}_s] = E[\sigma_t\epsilon_t|\mathcal{F}_s] = E[\sigma_t|\mathcal{F}_s]E[\epsilon_t|\mathcal{F}_s] = 0,$$

note that σ_t and ϵ_t are independent because σ_t depends on the previous shocks that are independent of ϵ_t. □

21.4.2.2 *Connecting GARCH to ARCH*

Similar to MA representation of AR processes, GARCH can also be represented by ARCH. Below, we show $ARCH(\infty)$ representation of $GARCH(1,1)$.

$$\begin{aligned}
\sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1\sigma_{t-1}^2 \\
&= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1(\alpha_0 + \alpha_1 a_{t-2}^2 + \beta_1\sigma_{t-2}^2) \\
&= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1\alpha_0 + \beta_1\alpha_1 a_{t-2}^2 + \beta_1^2\sigma_{t-2}^2 \\
&= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1\alpha_0 + \beta_1\alpha_1 a_{t-2}^2 + \beta_1^2(\alpha_0 + \alpha_1 a_{t-2}^2 + \beta_1\sigma_{t-2}^2) \\
&\vdots \\
&= \frac{\alpha_0}{1-\beta_1} + \alpha_1\sum_{i=0}^{\infty} a_{t-1-i}^2\beta_1^i
\end{aligned}$$

Therefore, σ_t^2 contains history back to infinity, even though the strength decreases geometrically.

21.4.2.3 *Variance forecasting*

Lemma 21.4.6 (conditional expectation equivalence between variance and square innovations for ARCH(q)). *Consider an GARCH(p,q) model given by*

$$a_t = \sigma_t\epsilon_t$$

where $\epsilon_t \sim WN(0,1)$, *and*

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i a_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2, \alpha_i, \beta_j \geq 0, \forall i, j \geq 1.$$

Let $\mathcal{F}_t$ *denote the information available up to time t. Then*

$$E[\sigma_t^2|\mathcal{F}_s] = E[a_t^2|\mathcal{F}_s], s < t.$$

Proof. Note that

$$a_t^2 = \sigma_t^2 \epsilon_t^2.$$

Take conditional expectation on both sides and get

$$E[a_t^2|\mathcal{F}_s] = E[\sigma_t^2 \epsilon_t^2|\mathcal{F}_s] = E[\sigma_t^2|\mathcal{F}_s]E[\epsilon_t^2|\mathcal{F}_s] = E[\sigma_t^2|\mathcal{F}_s].$$

where we used the fact that $E[\epsilon_t^2|\mathcal{F}_s] = 1$ and the independence between σ_t and ϵ_t. □

Theorem 21.4.6 (variance prediction in GARCH(1,1)). *Consider a GARCH(1,1) model given by*

$$a_t = \sigma_t \epsilon_t, \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \alpha_1 + \beta_1 \in [0, 1).$$

Let $\mathcal{F}_t$ denote the information available up to time t. Then,

$$\begin{aligned}
E[\sigma_{t+1}^2|\mathcal{F}_t] &\triangleq \sigma_{t+1}^2 = \alpha_0 + \alpha_1 a_t^2 + \beta_1 \sigma_t^2 \\
E[\sigma_{t+2}^2|\mathcal{F}_t] &\triangleq \hat{\sigma}_{t+2}^2 = \sigma^2 + (\alpha_1 + \beta_1)(\sigma_{t+1}^2 - \sigma^2) \\
E[\sigma_{t+3}^2|\mathcal{F}_t] &\triangleq \hat{\sigma}_{t+2}^2 = \sigma^2 + (\alpha_1 + \beta_1)^2(\sigma_{t+1}^2 - \sigma^2) \\
&\vdots \\
E[\sigma_{t+l}^2|\mathcal{F}_t] &\triangleq \hat{\sigma}_{t+l}^2 = \sigma^2 + (\alpha_1 + \beta_1)^{l-1}(\sigma_{t+1}^2 - \sigma^2)
\end{aligned}$$

where σ^2 is the unconditional variance

$$\sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}.$$

Proof. Note that the equivalence between square innovation and variance is discussed in Lemma 21.4.2.

Further note that σ_{t+1}^2 given $\mathcal{F}_t$ is actually deterministic quantity. For the rest, we have

$$\begin{aligned}
\sigma_{t+1}^2 &= \alpha_0 + \alpha_1 a_t^2 + \beta_1 \sigma_t^2 \\
E[\sigma_{t+2}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+2}^2 &= \alpha_0 + \alpha_1 E[a_{t+1}^2|\mathcal{F}_t] + \beta_1 \sigma_{t+1}^2 \\
&= \alpha_0 + \alpha_1 \sigma_{t+1}^2 + \beta_1 \sigma_{t+1}^2 \\
&= \alpha_0 + (\alpha_1 + \beta_1)\sigma_{t+1}^2 \\
&= \sigma^2 + (\alpha_1 + \beta_1)(\sigma_{t+1}^2 - \sigma^2) \\
E[\sigma_{t+3}^2|\mathcal{F}_t] \triangleq \hat{\sigma}_{t+3}^2 &= \alpha_0 + \alpha_1 E[a_{t+2}^2|\mathcal{F}_t] + \beta_1 \hat{\sigma}_{t+2}^2 \\
&= \alpha_0 + \alpha_1 \hat{\sigma}_{t+2}^2 + \beta_1 \hat{\sigma}_{t+2}^2 \\
&= \alpha_0 + (\alpha_1 + \beta_1)\hat{\sigma}_{t+2}^2 \\
&= \sigma^2 + (\alpha_1 + \beta_1)(\hat{\sigma}_{t+2} - \sigma^2) \\
&= \sigma^2 + (\alpha_1 + \beta_1)^2(\sigma_{t+1}^2 - \sigma^2)
\end{aligned}$$

where we use the fact [Theorem 21.4.5] that

$$E[a_{t+1}^2|\mathcal{F}_t] = \sigma_{t+1}^2 = (\alpha_0 + \alpha_1 a_t^2 + \beta_1 \sigma_t^2).$$

Others can be proved similarly. □

Remark 21.4.4 (implication for convergence rate). We can see that $\hat{\sigma}_{t+l}^2 \to \sigma$ as $l \to \infty$. If the variance spikes up during a crisis, then the number of the period before the variance restore to equilibrium value σ can be estimated using the value of $(\alpha_1 + \beta_1)$(the larger the longer).

21.5 Notes on Bibliography

Introductory level treatment, see [1]. Intermediate level treatment, see [2][11][6][12][8]. Advanced level treatment, see [13]

For multivariate time series and cointegration, see [9][14][10].

For finanical time series, see [5][9][7].

Analysis of Integrated and Cointegrated Time Series with R

Popular python libraries on time sereis analysis including statemodels and ARCH.

BIBLIOGRAPHY

1. Chatfield, C. The Analysis of Time Series: An Introduction, (Chapman & Hall/CRC Texts in Statistical Science) (2003).
2. Brockwell, P. J. & Davis, R. A. *Introduction to time series and forecasting* (Springer Science & Business Media, 2002).
3. Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. STL: A seasonal-trend decomposition. *Journal of official statistics* **6,** 3–73 (1990).
4. Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. *Time series analysis: forecasting and control* (John Wiley & Sons, 2015).
5. Tsay, R. S. *Analysis of financial time series* (John Wiley & Sons, 2005).
6. Shumway, R. H. & Stoffer, D. S. *Time series analysis and its applications: with R examples* (Springer Science & Business Media, 2010).
7. Hayashi, F. Econometrics. Princeton. *New Jersey, USA: Princeton University* (2000).
8. Subba Rao, S. *A course in time series analysis lecture notes* (Texas A & M, 2017).
9. Tsay, R. S. *Multivariate time series analysis: with R and financial applications* (John Wiley & Sons, 2013).
10. Lütkepohl, H. *New introduction to multiple time series analysis* (Springer Science & Business Media, 2005).
11. Hamilton, J. D. *Time series analysis* (Princeton university press Princeton, 1994).
12. Enders, W. *Applied Econometric Time Series, 4th Edition* ISBN: 9781118918661 (Wiley, 2014).
13. Brockwell, P. J. & Davis, R. A. *Time series: theory and methods* (Springer Science & Business Media, 1991).
14. Pfaff, B. *Analysis of integrated and cointegrated time series with R* (Springer Science & Business Media, 2008).

Part V

STATISTICAL LEARNING METHODS