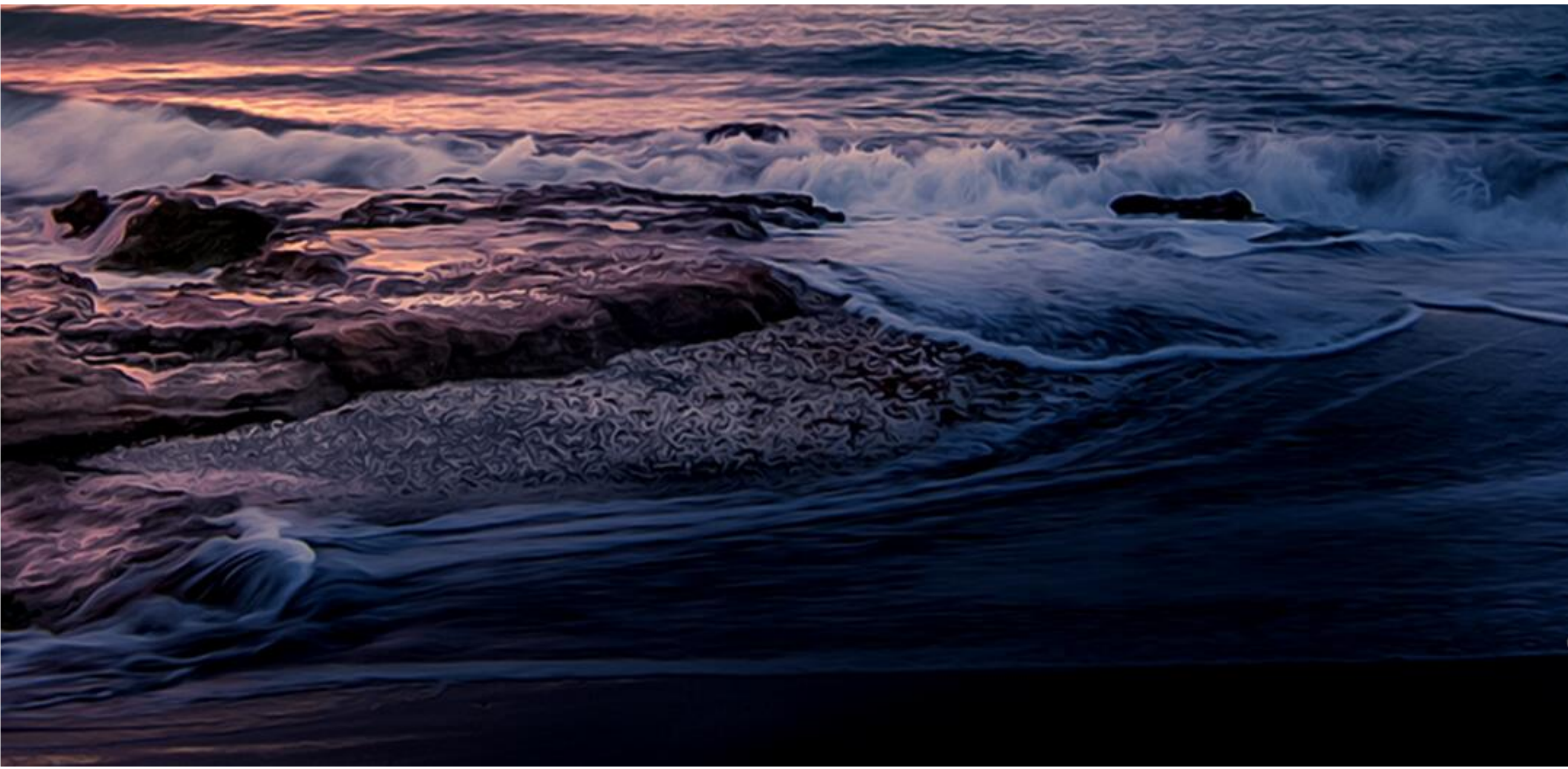
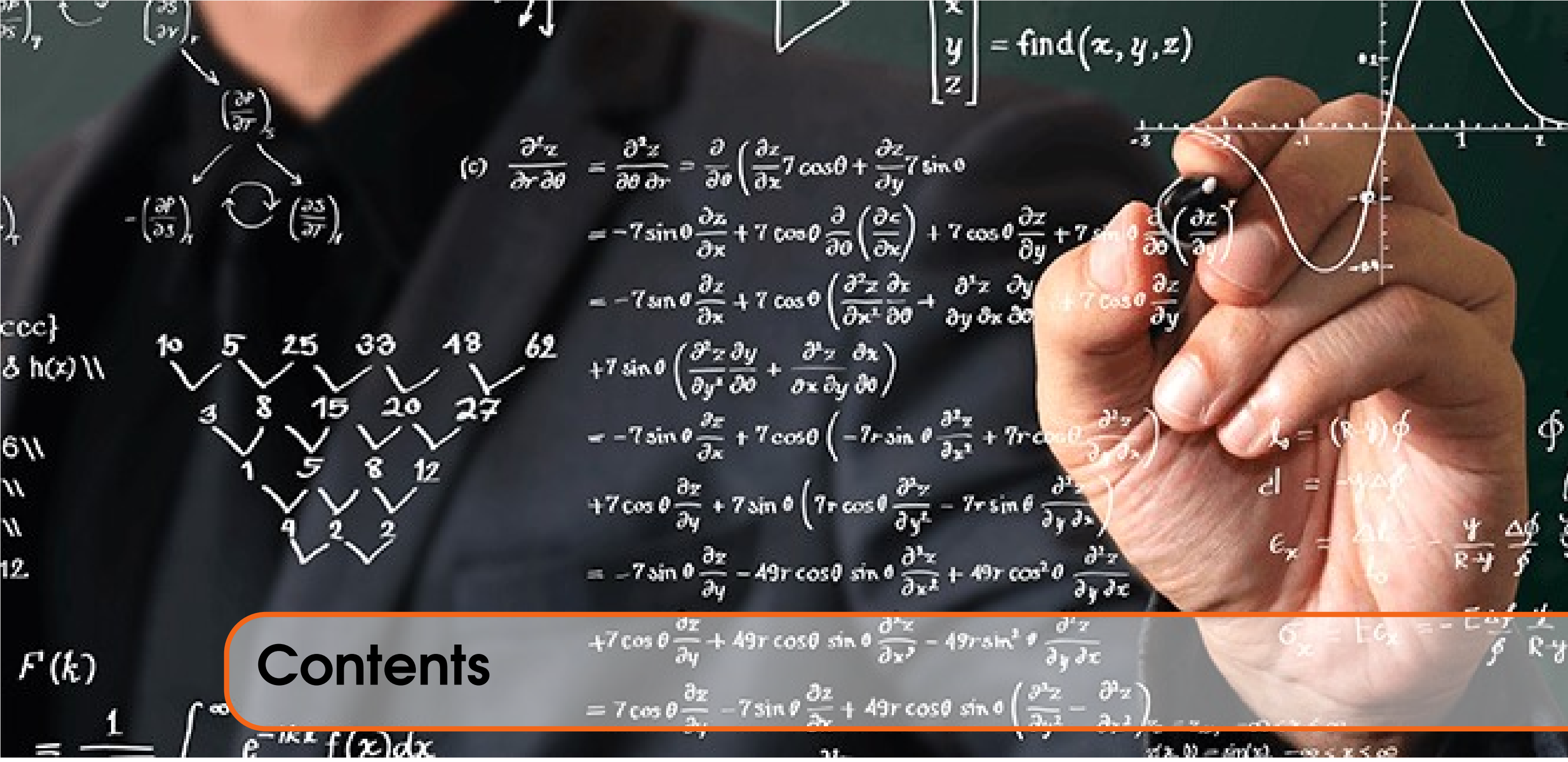
STAT 443 Course Notes

University of Waterloo

The One And Only
Waterloo 76er

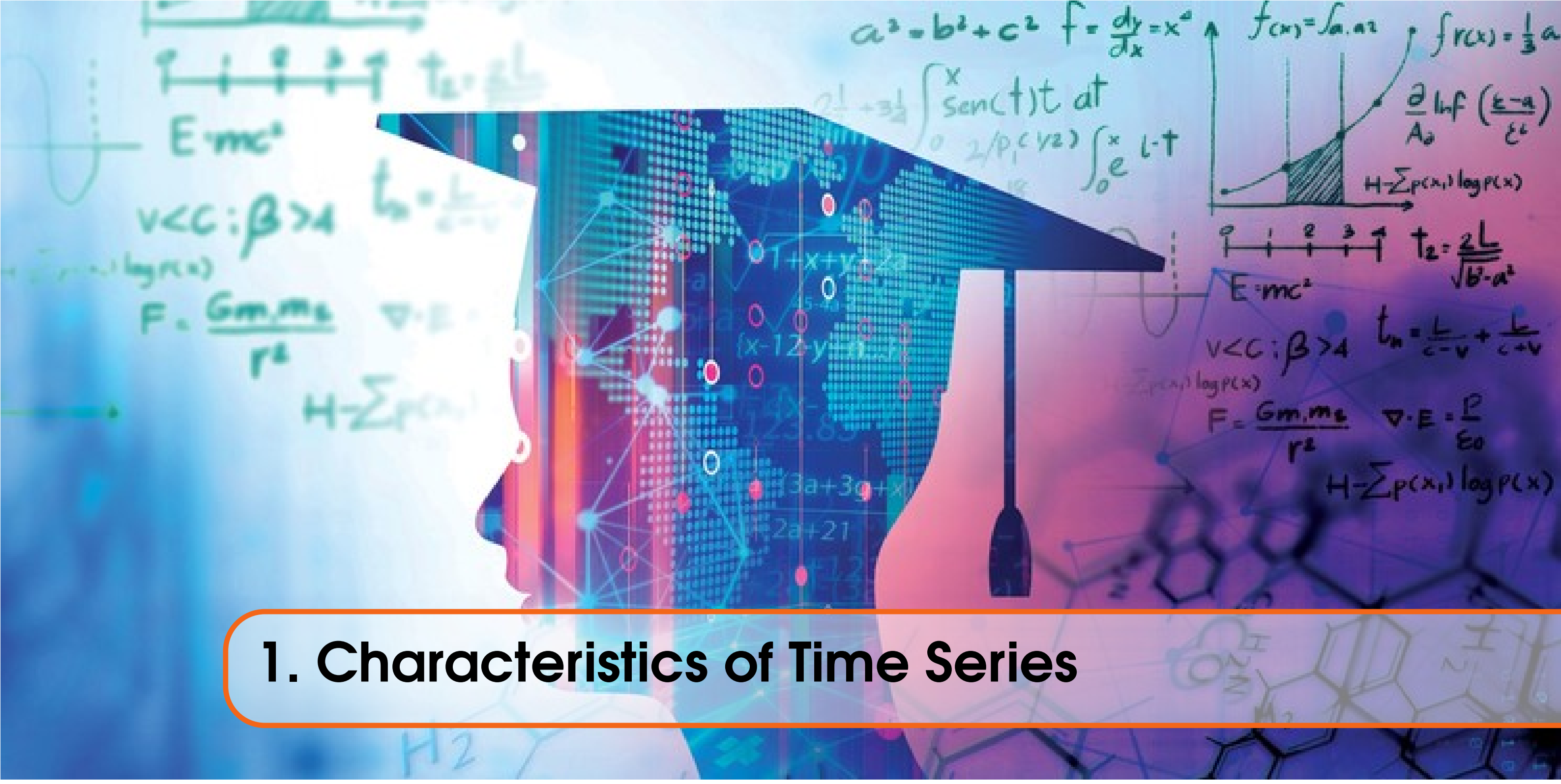
Bill Zhuo





Contents

I STAT443 Main Content



1. Characteristics of Time Series

1.1 What is a Time Series?

Definition 1.1.1 — Simple Random Sample. In classical statistics, we normally consider $X_1,\cdots,X_n \in \mathbb{R}^p$ to be a simple random sample. In particular,

1. **Independent and Identically Distributed (IID):** this describes $X_1,\cdots,X_n$
2. **Common Distribution Function F:** $X_i \sim F(\theta)$ with some parameter (could be a vector) θ.

■ **Example 1.1 — Classical Statistics.**

1. **Univariate Normal:** Let $X_i \overset{\text{iid}}{\sim} N\big(\mu,\sigma^2, i=1,\cdots,n\big)$ and we wish to estimate and perform inference on μ,σ^2.
2. **Multivariate Normal:** Let

$$X_i = \begin{bmatrix} \overset{\text{Dependent variable}}{Y_i} \\ \overset{\text{Independent variable}}{Z_i} \end{bmatrix}, i = 1,\cdots,n$$

Say $Y_i = \beta^\top Z_i + \varepsilon_i, \varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma_\varepsilon^2)$. One important thing in the classical statistics case, the index is not a variable to consider.

In such settings, one is typically interested in:

1. **Prediction:** how does the response variate behave with a future covariate?
2. **Inference:** how can we use the data to estimate parameters or unlock the underlying logic

■

Definition 1.1.2 — Time Series. We say $X_1,\cdots,X_T$ is an (observed) time series of length T if X_t denotes an observation obtained at time t. In particular, the observations are ordered in time (equal time interval). If $X_t \in \mathbb{R}$, we say $X_1,\cdots,X_T$, is a real-valued or scalar time series. If $X_t \in \mathbb{R}^p$, we say $X_1,\cdots,X_T$, is a multivariate or vector-valued time series.

■ **Example 1.2 — Stock Earnings.** Consider the following `R` output:

```
%library(astsa)
plot(jj, type="o", ylab="Quarterly Earnings per Share")
```

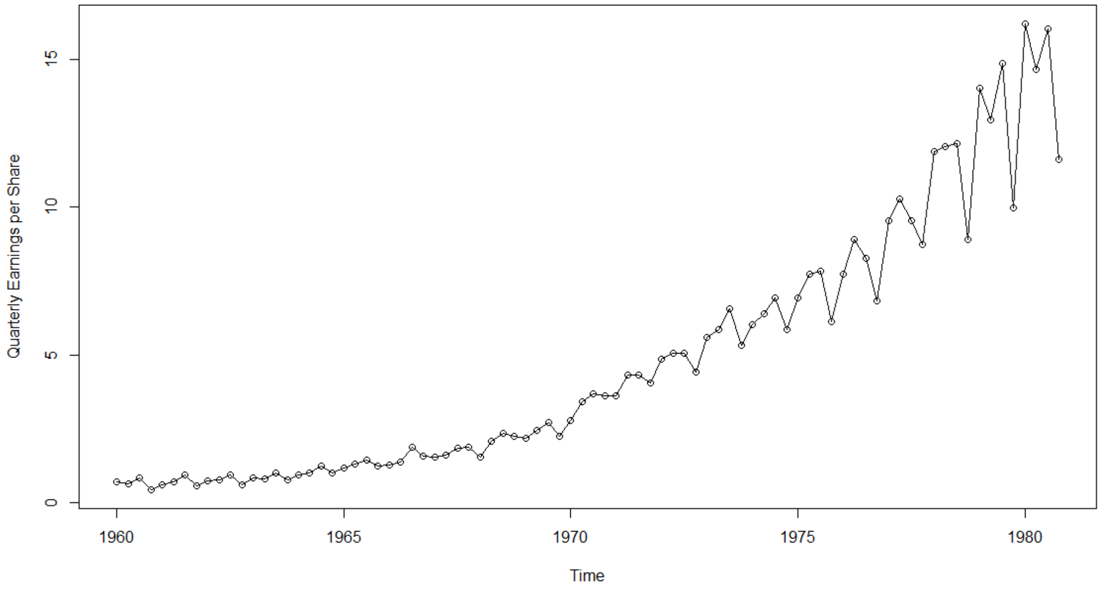


Figure 1.1.1: Quarterly Johnson and Johnson Earnings

Several observations that we can see from this time series are:

1. **Increasing trend over time**
2. **Heterogeneity in variance over time**

■

R In comparison to classical statistics, with time series data, we are typically concerned with the same goals as in classical case, prediction & inference. However, in contrast, the data often exhibit:

1. **Heterogeneity:**
   (a) Time trends: $\mathbb{E}[X_t] \neq \mathbb{E}[X_{t+h}]$
   (b) Heteroskedasticity: $\textbf{Var}(X_t) \neq \textbf{Var}(X_{t+h})$
2. **Serial Dependence (Serial Correlation)**: observations that are temporarily close appear to depend on each other

■ **Example 1.3 — Yearly Temperature.** Consider the following R output:

```
%plot(gtemp, type="o", ylab="Global Temperature Deviations")
```

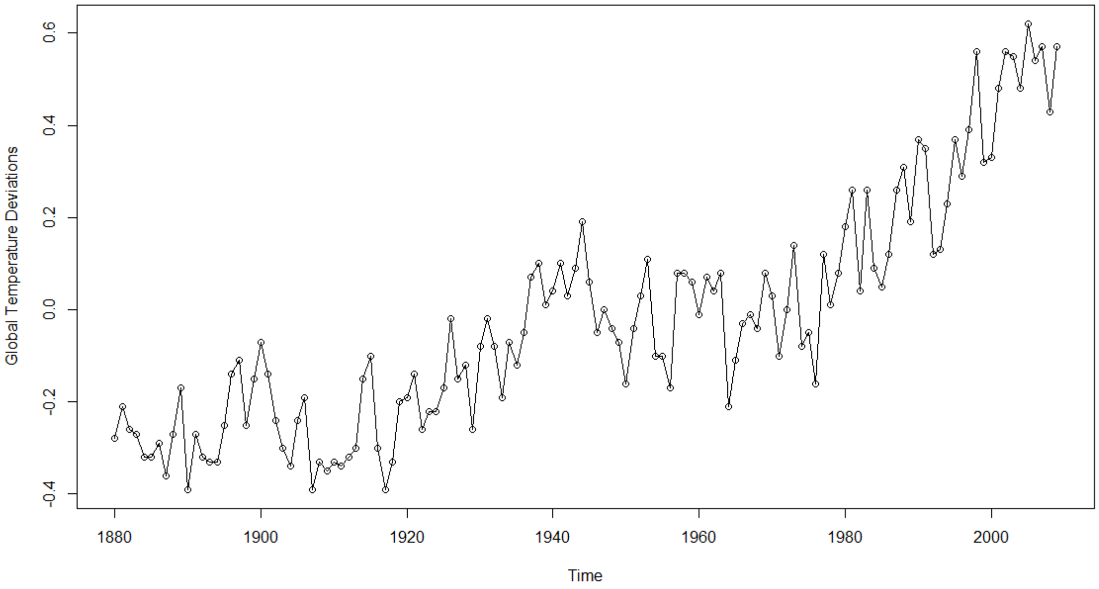


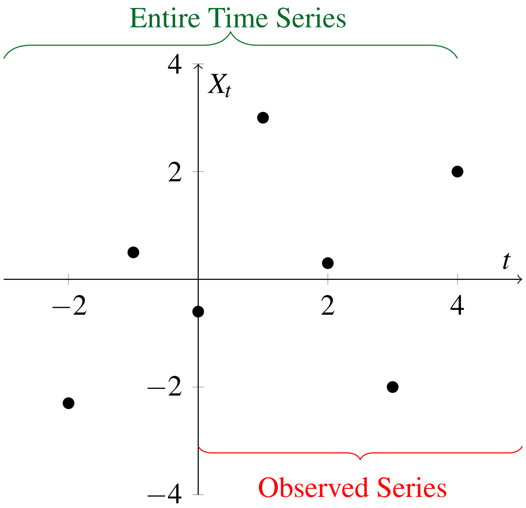
Figure 1.1.2: Deviation of global mean yearly temperature from the mean computed from 1951-1980

We can see the deviation is around 0 but the deviation is increasing over time.

> "Climate change is a hoax!"
>
> —Donald.J.Impeached.Twice.Trump

In this case, heteroskedasticity is not very apparent, thus, we may want to consider the trend. However, this may involve serial correlation, which will affect our forecast. ■

Definition 1.1.3 — Time Series (mAtHeMaticS). Formally, we say $\{X_t\}_{t\in\mathbb{Z}}$ is a time series if $\{X_t : t \in \mathbb{Z}\}$ is a stochastic process indexed by $\mathbb{Z}$. This means that there is a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ so that for all $t \in \mathbb{Z}$, $X_t : \Omega \to \mathbb{R}$ is a random variable. In relation to the original definition, we say that $X_1, \cdots, X_T$ is an observed stretch or a realization or a sample path of length T from $\{X_t\}_{t\in\mathbb{Z}}$.



Most of the time (come on, every time), we can only see the observed series (sample path), any inference on such a sample will incur some bias on the forecasting performance.

1.2 Basic Principles of Forecastings

So what is the problem of forecasting?

Definition 1.2.1 — Forecasting Problem. Let $\{X_t\}_{t\in\mathbb{Z}}$ be a time series. Based on $X_1,\cdots X_T$, we would like to produce a "best guess" for X_{T+h}, denoted by

$$\hat{X}_{T+h} = \hat{X}_{T+h|T}$$

Definition 1.2.2 — Forecast. For $h \geq 1$, our "best guess" $\hat{X}_{T+h} = f_h(X_T,\cdots,X_1)$ is called a forecast of X_{T+h} at horizon h.

Primary Goals in Forecasting:

1. Choose f_h "optimally". Normally, we or the practitioner have some measure, say $L(\cdot,\cdot)$, in mind for determining how "close", $\hat{X}_{T+h}$ is to X_{T+h}. We then wish to choose f_h so that

$$L(X_{T+h}, f_h(X_T,\cdots,X_1))$$

is minimized. Most common measure $L(\cdot,\cdot)$ is the mean-squared error (**MSE**), where

$$L(X,Y) = \mathbb{E}[(X-Y)^2]$$

2. **Quantify the uncertainty in the forecast**. This entails providing some description of how close we expect $\hat{X}_{T+h}$ to be X_{T+h}.

■ **Example 1.4** Suppose every minute, we flip a coin. Let $X_t : \mathbb{Z} \to \{H,T\} \cong \{1,-1\}$ for $t = 1,\cdots,T$. This produces a time series of length T, which is a random sequence of 1s and -1s. Note that $\mathbb{E}[X_t] = 0$ for $h \geq 1$, consider $\hat{X}_{T+h} = f(X_T,\cdots,X_1)$. Let

$$\begin{aligned}
L(X_{T+h}, \hat{X}_{T+h}) &= \mathbb{E}[(X_{T+h} - \hat{X}_{T+h})^2] \\
&= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] - 2\mathbb{E}[X_{T+h}\hat{X}_{T+h}] \\
&= \mathbf{Var}\,(X_t) + \mathbb{E}[\hat{X}_{T+h}^2] - 2\mathbb{E}[X_{T+h}f(X_T,\cdots,X_1)] \\
&= \mathbf{Var}\,(X_t) + \mathbb{E}[\hat{X}_{T+h}^2] - 2f(X_T,\cdots,X_1)\mathbb{E}[X_{T+h}] \\
&= \mathbf{Var}\,(X_t) + \hat{X}_{T+h}^2 && \mathbb{E}[X_t] = 0, \forall t
\end{aligned}$$

This is minimized by taking $\hat{X}_{T+h} = 0$. There is nothing "wrong" with this forecast. But empirically, if you are bound to perform this experiment, you are more than certain that none of the future observations will be 0! (This is not a factorial sign). But ideally, we would also be able to say that the sequence appears to be totally random, which means the "best" forecast cannot distinguish the outcomes clearly. ■

How to quantify the uncertainty in forecasting?

1. **Ideal:**have a predictive distribution of the form

$$X_{T+h}|X_T,\cdots,X_1$$

but this is pretty hard given that point estimate is not even that attainable.

2. **Excellent:** predictive intervals/sets, for some $\alpha \in (0,1)$, we find I_α so that

$$\mathbb{P}(X_{T+h} \in I_\alpha | X_T,\cdots,X_1) = \alpha$$

for example $\alpha = 0.95$ is typically used. Often such intervals take the form of

$$I_\alpha = \left(\hat{X}_{T+h} - \hat{\sigma}_h, \hat{X}_{T+h} + \hat{\sigma}_h\right)$$

R I don’t know about you. But this seems pretty confident to me.

R

1. In order to perform estimation of predictive distributions leads to one to estimate the joint distribution of

$$X_{T+h}, X_T,\cdots,X_1 \Longrightarrow ARMA, ARIMA, etc.$$

which will lead to later models.

2. It is important that we acknowledge that some things cannot be predicted! (But remember stonks always go up!)

> "It is tough to make predictions, especially about the future"
>
> —Yogi Berra

1.3 Stationarity

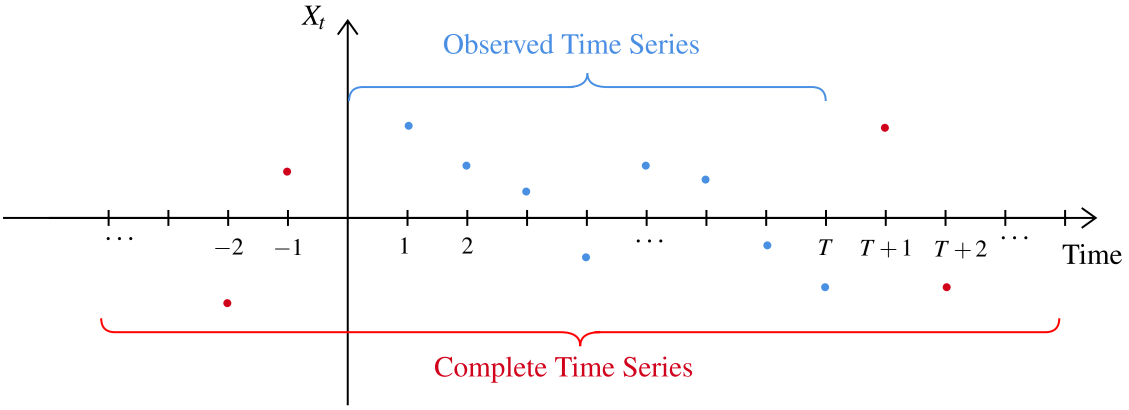
Given an observed time series $X_1,\cdots,X_T$, we are frequently interested in estimating the joint distribution of

$$X_{T+h}, X_T,\cdots,X_1$$

for forecasting and inference purposes. The joint distribution is a feature of the process $\{X_t\}_{t\in\mathbb{Z}}$

$$\underbrace{X_1,\cdots,X_T}_{\text{Time Series}} \overset{\text{Infer}}{\Longrightarrow} \underbrace{\{X_t\}_{t\in\mathbb{Z}}}_{\text{Stochastic Process}}$$

■ **Example 1.5 — Worst Case Scenario.** What we have is usually the following:



Say $X_t \sim F_t$ where F_t is a changing function of t. If so, it is hard to pool the data $X_1, \cdots, X_T$ to estimate F_t. This leads to **serial dependence**, if the distribution of (X_t, X_{t+h}) depends strongly on t, we have a similar problem in estimating, such as $\mathbf{Cov}(X_t, X_{t+h})$. We hope the distribution of data is the same in each data window. ■

Definition 1.3.1 — Strong Stationarity. We say that a time series $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary or strongly stationary if for each $k \geq 1, i_1, \cdots, i_k, h \in \mathbb{Z}$,

$$(X_{i_1}, \cdots, X_{i_k}) \overset{D}{=} (X_{i_1+h}, \cdots, X_{i_k+h})$$

this means the joint distribution at these points are shift-invariant.

R In other words, shifting the window on which you view the data does not change its distribution. This implies that if F_t is the CDF of X_t, then

$$F_t = F_{t+h} = F$$

such that all variables have a common joint distribution.

Definition 1.3.2 — Mean function and autocovariance function. For a time series $\{X_t\}_{t\in\mathbb{Z}}$ with $\mathbb{E}(X_t^2) < \infty$ for all $t \in \mathbb{Z}$, we denote the mean function of the series as

$$\mu_t = \mathbb{E}[X_t]$$

and the autocovariance function of the series is

$$\gamma(t,s) = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)] = \mathbf{Cov}(X_t, X_s)$$

These two essentially describe the first and second moment behaviours of the series.

Definition 1.3.3 — Stationarity (Weak). We say that a series $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary if $\mathbb{E}[X_t] = \mu$ does not depend on t, and if

$$\gamma(t,s) = f(|t-s|)$$

i.e., γ is a function of $|t-s|$. In this case, we usually write,

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

and we call the h as the lag parameter.

Additionally, the property when $\mathbb{E}[X_t] = \mu$ does not depend on t is often called **first order stationarity**. The property when $\gamma(t,s) = f(|t-s|)$ only depends on the lag $|t-s|$ is called **second order stationarity**. For a second order stationary process,

$$\gamma(h) = \mathbf{Cov}(X_t, X_{t+h}) \overset{t \mapsto t-h}{=} \mathbf{Cov}(X_{t-h}, X_{t-h+h}) = \mathbf{Cov}(X_t, X_{t-h}) = \gamma(-h)$$

Thus, normally, we only record $\gamma(h)$ with $h \geq 0$.

1.4 White Noise and Stationarity Examples

Definition 1.4.1 — Strong white noise. We say $\{X_t\}_{t\in\mathbb{Z}}$ is a strong white nosie if $\mathbb{E}[X_t] = 0$ and the $\{X_t\}_{t\in\mathbb{Z}}$ are independent and identically distributed (iid).

Definition 1.4.2 — Weak white noise. We say $\{X_t\}_{t\in\mathbb{Z}}$ is a weak white noise if $\mathbb{E}[X_t] = 0$ and

$$\gamma(t,s) = \mathbf{Cov}(X_t, X_s) = \begin{cases} \sigma^2 & |s-t| = 0 \\ 0 & |t-s| > 0 \end{cases}$$

Definition 1.4.3 — Gaussian white noise. We say $\{X_t\}_{t\in\mathbb{Z}}$ is a Gaussian white noise if $X_t \overset{\text{iid}}{\sim} N(0,\sigma^2)$

■ **Example 1.6 — Gaussian white noise.** The term "white" comes from spectral analysis, in which a white noise series shares the same spectral properties as white light: all periodicities occur with equal strength.

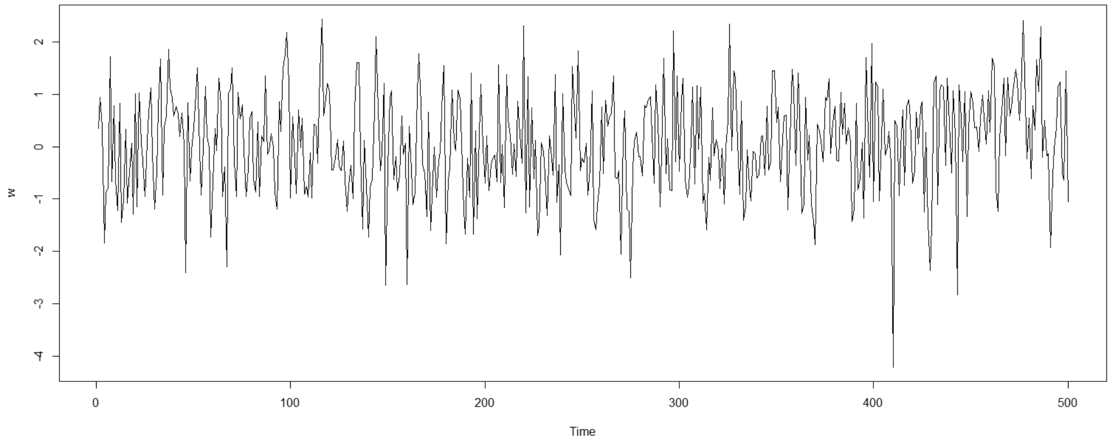


Figure 1.4.1: Gaussian white noise series

■

White noises are examples of stationary processes

Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise, then $\mathbb{E}[W_t] = 0$. This does not depend on t. Meanwhile,

$$\gamma(t,s) = \mathbf{Cov}(W_t, W_s) = \mathbb{E}[W_t W_s] = \begin{cases} \sigma_w^2 & |t-s| = 0 \\ 0 & |t-s| > 0 \end{cases}$$

this only depends on $|t-s|$. Thus, $\{W_t\}_{t\in\mathbb{Z}}$ is weakly stationary.

R $\{W_t\}_{t\in\mathbb{Z}}$ is also strictly stationary. Let $k \geq 1$ and

$$i_1 < i_2 < \cdots < i_k, h \in \mathbb{Z}$$

and by independence, we have

$$\begin{aligned} F(t_1,\cdots,t_k) &= \mathbb{P}(W_{i_1} < t_1,\cdots,W_{i_k} \leq t_k) \\ &= \prod_{j=1}^{k} \mathbb{P}(W_{i_j} \leq t_j) && \text{independence} \\ &= \prod_{j=1}^{k} \mathbb{P}(W_{i_{j+h}} \leq t_j) \\ &= \mathbb{P}(W_{i_1+h} < t_1,\cdots,W_{i_k+h} \leq t_k) \end{aligned}$$

Thus, also shift-invariant.

■ **Example 1.7 — Time series derived from white noises.** Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Define

$$X_t = W_t + \theta W_{t-1}, \theta \in \mathbb{R}$$

then, $\mathbb{E}[X_t] = \mathbb{E}[W_t] + \theta\mathbb{E}[W_{t-1}] = 0$. Thus, first order stationary. Then, for $\gamma(t,s)$, we consider several cases:

1. When $|t-s| = 0$,

$$\begin{aligned} \mathbb{E}[(W_t+\theta W_{t-1})^2] &= \mathbb{E}[W_t^2] + \theta^2\mathbb{E}[W_{t-1}^2] + 2\theta\underbrace{\mathbb{E}[W_tW_{t-1}]}_{=0} \\ &= (1+\theta^2)\sigma_w^2 \end{aligned}$$

2. when $t = s+1$ or $s = t+1$,

$$\mathbb{E}[(W_{s+1}+\theta W_s)(W_s+\theta W_{s-1})] = \theta\mathbb{E}[W_s^2] = \theta\sigma_w^2$$

3. when $|t-s| > 1$, we will have independence kick in again.

To summarize, we have

$$\gamma(t,s) = \begin{cases} (1+\theta^2)\sigma_w^2 & |t-s| = 0 \\ \theta\sigma_w^2 & |t-s| = 1 \\ 0 & |t-s| > 1 \end{cases}$$

$\{X_t\}_{t\in\mathbb{Z}}$ is also strictly stationary. Suppose $k \geq 1$, $i_1 < \cdots < i_k, h \in \mathbb{Z}$. Then,

$$\begin{aligned} \mathbb{P}(X_{i_1} \leq t_1,\cdots,X_{i_k} \leq t_k) &= \mathbb{P}(W_{i_1}+\theta W_{i_1-1} \leq t_1,\cdots,W_{i_k}+\theta W_{i_k-1} \leq t_k) \\ &= \mathbb{P}\left(\begin{bmatrix} W_{i_1-1} \\ W_{i_1} \\ \vdots \\ W_{i_k} \end{bmatrix} \in B\right) = \mathbb{P}\left(\begin{bmatrix} W_{i_1-1+h} \\ W_{i_1+h} \\ \vdots \\ W_{i_k+h} \end{bmatrix} \in B\right) \\ &= \mathbb{P}(W_{i_1+h}+\theta W_{i_1-1+h} \leq t_1,\cdots,W_{i_k+h}+\theta W_{i_k-1+h} \leq t_k) \\ &= \mathbb{P}(X_{i_1+h} \leq t_1,\cdots,X_{i_k+h} \leq t_k) \end{aligned}$$

where B is some subset of $\mathbb{R}^{i_k-i_1+1}$. Thus, shift-invariant. ■

Definition 1.4.4 — Bernoulli shift. Suppose $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Then if $X_t = g(\varepsilon_t, \varepsilon_{t-1}, \cdots)$ for some function $g : \mathbb{R}^\infty \to \mathbb{R}$, we say that $\{X_t\}_{t\in\mathbb{Z}}$ is a Bernoulli shift.

Note that g is not independent of t.

Theorem 1 If $\{X_t\}_{t\in\mathbb{Z}}$ is a Bernoulli shift, then $\{X_t\})_{t\in\mathbb{Z}}$ is strictly stationary.

R Norbert Wiener conjectured that every stationary sequence is a Bernoulli shift, which is not generally true.

Exercise 1.1 Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Let

$$X_t = \sum_{i=0}^{t} W_i + \sum_{i=t}^{-1} W_i$$

This is called a two-sided random walk. Show that $\{X_t\}_{t\in\mathbb{Z}}$ is first order stationary but not second order stationary. ■

1.5 Theoretical L^2 Framework for Time Series (Advanced)

Sometimes the process itself is a limiting process. For example, $X_t = \lim_{h\to\infty} X_{h,t}$, in what sense does this limit exist? (Metric, norm, topology, you name it...) We might want to know how "close" two random variables X, Y are, or is there a random variable achieves

$$\inf_{y\in S} d(Y,Z)$$

Definition 1.5.1 — L^2 space. Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The space L^2 is the set of random variables $X : \Omega \to \mathbb{R}$ measurable so that $\mathbb{E}[X^2] < \infty$.

Definition 1.5.2 — L^2 time series. We say that $\{X_t\}_{t\in\mathbb{Z}}$ is an L^2 time series if $X_t \in L^2$ for all $t \in \mathbb{Z}$.

We know that L^2 is a Hilbert space when equipped with the inner-product $X, Y \in L^2$ defined by

$$\langle X, Y\rangle = \mathbb{E}[X, Y]$$

R This is actually not an inner-product, but if you ignore the measure zero sets or take the equivalent classes, things are just fine.

since it satisfies:

1. **Bilinearity:** $\langle aX + bY, Z\rangle = a\langle X, Z\rangle + b\langle Y, Z\rangle$
2. **"Almost" positive definite:**

$$\langle X, X\rangle = \mathbb{E}[X^2] = 0 \iff X = 0\text{a.s.}$$

which implies $\mathbb{P}(X = 0) = 1$.

3. **Symmetric:** $\langle X, Y\rangle = \langle Y, X\rangle$

As a known fact, every Hilbert space is complete, so is L^2 space. This means, whenever $X_n \in L^2$ so that $\mathbb{E}[(X_n - X_m)^2] \to 0$ as $n, m \to \infty$, then there exists $X \in L^2$ so that $X_n \to X$.i.e, $\mathbb{E}[(X_n - X)^2] \to 0$. This results follows from the famous Riesz-Fischer Theorem for L^p spaces. You can take a look at this note for some reference:

```
https://student.cs.uwaterloo.ca/~w3zhuo/Public_Notes/PMATH450/PMATH450_
                    Lecture_Notes__Book_Version_.pdf
```

Useful tools for time series:

1. **Existence of Limits:**

$$X_{t,n} = \sum_{j=0}^{n} \psi_j \varepsilon_{t-j}$$

$\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Since, for $n > m$,

$$\begin{aligned}\mathbb{E}[(X_{t,n} - X_{t,m})^2] =& \mathbb{E}\left[\left(\sum_{j=m+1}^{n} \psi_j \varepsilon_{t-j}\right)^2\right]\\ &= \sum_{m+1}^{n} \psi_j^2 \sigma_\varepsilon^2 \xrightarrow{n,m\to\infty} 0\end{aligned}$$

when $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ (square summable), then by completeness, this series converges to some random variable X_t such that

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

2. **Projection Theorem in forecasting**: forecasting can often be cast as finding a random variable Y among a collection of possible forecasts $\mathscr{M}$, such as $\mathscr{M} = \text{span}\{X_T, \cdots, X_1\}$ so that

$$Y = \arg\inf_{Z\in\mathscr{M}} \mathbb{E}[(X_{T+h} - Z)^2]$$

When $\mathscr{M}$ is a closed linear subspace of L^2, the Projection Theorem guarantees that such a Y exist, and it must satisfy

$$\langle X_{T+h} - Y, Z\rangle = 0, \forall Z \in \mathscr{M}$$

must be in the orthogonal complement.

1.6 Weak vs. Strong Stationarity

Unfortunately,

$$\{X_t\}_{t\in\mathbb{Z}} \textbf{ strictly stationary} \not\Longrightarrow \{X_t\}_{t\in\mathbb{Z}} \textbf{ weakly stationary}$$

■ **Example 1.8 — Non-existing Expectation.** Suppose $X_t \overset{\text{iid}}{\sim}$ Cauchy random variables, i.e.,

$$\mathbb{P}(X_t \le s) = \int_{-\infty}^{s} \frac{1}{\pi(1+x^2)} dx$$

Then, $\mathbb{E}[X_t]$ is DNE, hence $\{X_t\}_{t\in\mathbb{Z}}$ is not weakly stationary. However, $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary since it is translation-invariant and iid. ■

So how to make this implication true? The second moment seems to be the problem.

Theorem 2 If $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary and $\mathbb{E}[X_0^2] < \infty$, then $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary.

Proof. Note that if $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary, we have $(X_t) \overset{D}{=} (X_0)$ so $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ does not depend on t. Similarly, $\mathbf{Var}(X_t) = \mathbf{Var}(X_0)$. Then, by Cauchy-Schwarz inequality,

$$\gamma(t,s) = \mathbf{Cov}(X_t, X_s) \le \mathbf{Var}(X_t) < \infty$$

and suppose $t < s$, by strict stationarity,

$$\mathbf{Cov}(X_t, X_s) = \mathbf{Cov}(X_0, X_{s-t})$$

since

$$(X_t, X_s) \overset{D}{=} (X_{t-t}, X_{s-t}) \overset{D}{=} (X_0, X_{s-t})$$

■

1.6.1 Equivalence of Stationarity for Guassians

Definition 1.6.1 — Gaussian process (time series). $\{X_t\}_{t\in\mathbb{Z}}$ is said to be a Guassian process or Gaussian time series if for each $k \ge 1$ $i_1 < i_2 < \cdots < i_k$,

$$\begin{aligned}(X_{i_1}, \cdots, X_{i_k}) &\sim \mathbf{MVN}(\mu(i_1, \cdots, i_k), \Sigma(i_1, \cdots, i_k)) \\ &\sim N_k(\mu_k, \Sigma_k)\end{aligned}$$

where

$$\mu_k = \begin{bmatrix}\mathbb{E}[X_{i_1}] \\ \vdots \\ \mathbb{E}[X_{i_k}]\end{bmatrix}, \qquad \Sigma_k = \left[\mathbf{Cov}(X_{i_j}, X_{i_r})\right]_{1\le j,r\le k}$$

Proposition 1.6.1 If $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary and Gaussian, then $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary.

Proof. If $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary, $\mathbb{E}[X_t] = \mu, \forall t$. We really just need to show the parameters are shift-invariant. First,

$$\mu_k = \begin{bmatrix}\mathbb{E}[X_{i_1}] \\ \vdots \\ \mathbb{E}[X_{i_k}]\end{bmatrix} = \begin{bmatrix}\mu \\ \vdots \\ \mu\end{bmatrix} = \begin{bmatrix}\mathbb{E}[X_{i_1+h}] \\ \vdots \\ \mathbb{E}[X_{i_k+h}]\end{bmatrix} = \vec{\mu}$$

on the other hand,

$$\begin{aligned}\mathbf{Var}(X_{i_1}, \cdots, X_{i_k}) &= \left[\mathbf{Cov}(X_{i_j}, X_{i_r})\right]_{1\le j,r\le k} \\ &= \left[\mathbf{Cov}(X_0, X_{i_r - i_j})\right]_{1\le j,r\le k} \\ &= \left[\mathbf{Cov}(X_0, X_{(i_r+h)-(i_j+h)})\right]_{1\le j,r\le k} \\ &= \left[\mathbf{Cov}(X_{i_j+h}, X_{i_r+h})\right]_{1\le j,r\le k} \\ &= \mathbf{Var}(X_{i_1+h}, \cdots, X_{i_k+h})\end{aligned}$$

Using the Gaussian assumption,

$$(X_{i_1}, \cdots, X_{i_k}) \overset{D}{=} N_k(\vec{\mu}, \Sigma_k) \overset{D}{=} (X_{i_1+h}, \cdots, X_{i_k+h})$$

Hence, $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary. ■

Exercise 1.2 Prove that if $\{X_t\}_{t\in\mathbb{Z}}$ is not weakly stationary (i.e., either $\mathbb{E}[X_t]$ depends on t or $\gamma(t,s)$ does not only depend on the lag), then $\{X_t\}_{t\in\mathbb{Z}}$ is not strictly stationary. ■

1.7 Signal & Noise Models

Ideally, a time series that we are considering was generated from a stationary process. If so, we can pool data to estimate the process underlying structure. For example, we can estimate its marginal distribution and serial dependence structure.

Foreseeably and unfortunately, most time series are evidently not stationary. What do we do?

■ **Example 1.9 — Stock Earnings Revisit.** Consider the following R output:

```
%library(astsa)
plot(jj, type="o", ylab="Quarterly Earnings per Share")
```

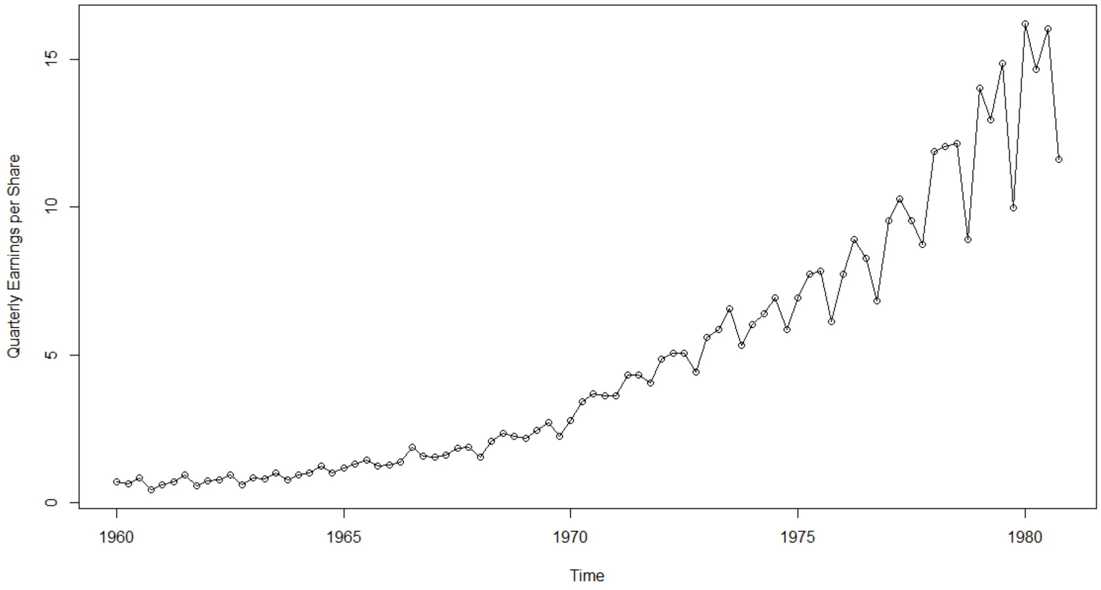


Figure 1.7.1: Quarterly Johnson and Johnson Earnings

We can observed that the average seems to be growing overtime. This is a sign of non-first-order stationarity. We can also see the variability seems to be increasing over time as well. This is a sign of non-second-order stationarity. ■

The *Signal + Noise Model* suggest:

$$X_t = S_t + \varepsilon_t$$

where

1. S_t is the deterministic signal or trend of the series
2. ε_t is the noise added to the signal satisfying $\mathbb{E}[\varepsilon_t] = 0$. There exists a (strong) white noise $\{W_t\}_{t\in\mathbb{Z}}$ so that

$$\varepsilon_t = g(W_t, W_{t-1}, \cdots) \qquad \textbf{[Stationary Noise]}$$
$$\varepsilon_t = g_t(W_t, W_{t-1}, \cdots) \quad \textbf{[Non-stationary Noise]}$$

The terms $\{W_t\}_{t\in\mathbb{Z}}$ are often called the innovations or shocks.

$\{W_t\}_{t\in\mathbb{Z}}$ and the function g can capture the serial dependence in $\{X_t\}_{t\in\mathbb{Z}}$'s noise.

Example 1.10 — Function $g = g_t(W_t, W_{t-1}, \cdots)$. 1. **Random walk:**

$$\varepsilon_t = \sum_{j=0}^{t} W_j$$

2. **Changing variance model:**

$$\varepsilon_t = \sigma(t)W_t$$

The goal is to estimate S_t, and infer the structure of the noise term ε_t. ■

Example 1.11 — Yearly Temperature Revisit. Consider the following R output:

```
%plot(gtemp, type="o", ylab="Global Temperature Deviations")
```

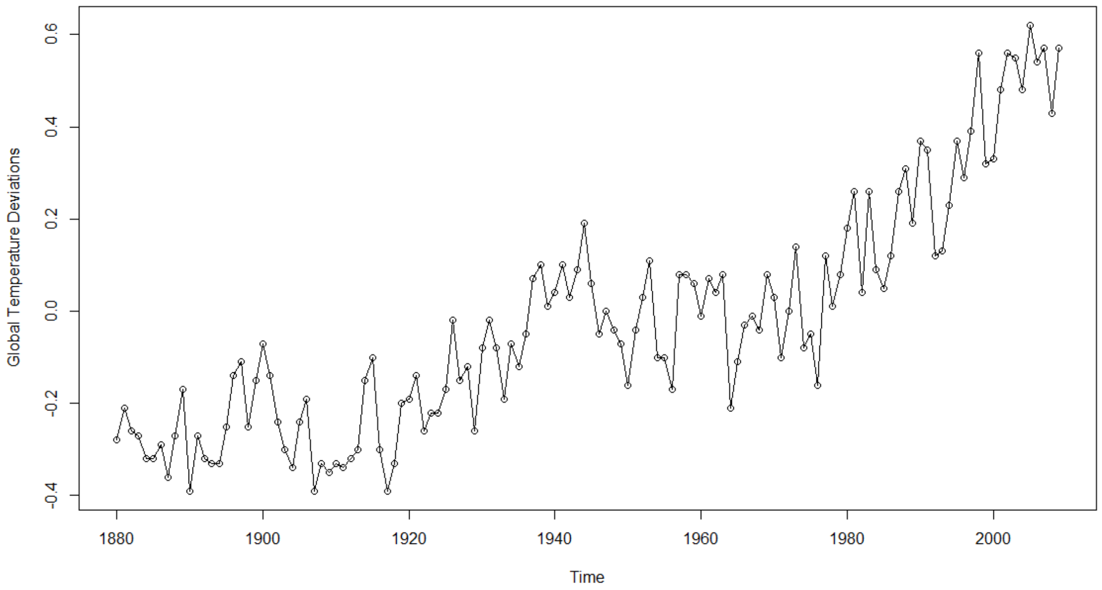


Figure 1.7.2: Deviation of global mean yearly temperature from the mean computed from 1951-1980

We see that the data appears to be non-stationary. This might be a good example to apply the signal + noise model. We compute the moving average (with some parameters) to identify the signal/trend. We may propose a linear trend,

$$S_t = \beta_0 + \beta_1 t$$

we can estimate parameters using ordinary least squares (OLS). Such β_0, β_1 can minimize

$$\sum_{t=1}^{T} (X_t - [\beta_0 + \beta_1 t])^2$$

■

Definition 1.7.1 — Detrending. Detrending a time series constitutes computing residuals based on an estimate for the signal/trend. A detrended time series is a time series of such residuals.

The process of detrending goes like this:

1. Estimate $\hat{S}_t$ for S_t
2. Detrend series by computing $\hat{\varepsilon}_t = X_t - \hat{S}_t$. We call $\{\hat{\varepsilon}_t\}_{t\in\mathbb{Z}}$ the detrended series.

Continuing previous example,

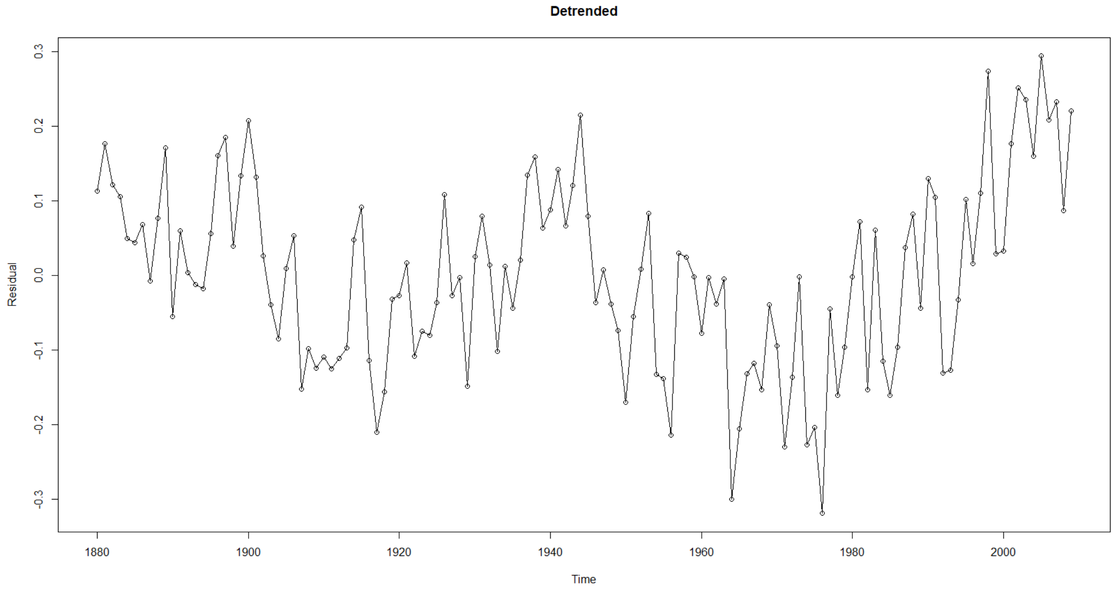


Figure 1.7.3: Detrended Time Series with OLS Fit

If the trend is now 0, there appears to be substantial serial dependence remaining in the series. We would hope this detrended series to look reasonably stationary.

1.8 Time Series Differencing

If the detrended series is reasonably stationary, we might proceed in estimating the structure of $\{\hat{\varepsilon}_t\}_{t\in\mathbb{Z}}$ as if it were stationary.

1.8.1 Random Walk Model with a Drift

Consider

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

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a strong white noise, δ is a constant called drift. The idea is that each step of the process is built upon its previous step similar to a random walk. By this iterative formulation, we can write

$$\begin{aligned} X_t &= 2\delta + X_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &= \underbrace{\cdots}_{t \text{ times}} \\ &= \underbrace{t\delta + X_0}_{S_t \sim \text{linear signal}} + \underbrace{\sum_{j=1}^{t} \varepsilon_j}_{randomwalknoise} \end{aligned}$$

R Notice that under the random walk model, we can compute the so-called **first difference**

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

so if $\{X_t\}_{t\in\mathbb{Z}}$ follows a random walk model, the series $\{\nabla X_t\}_{t\in\mathbb{Z}}$ should behave like a white noise shifted by δ.

Recall the temperature example,

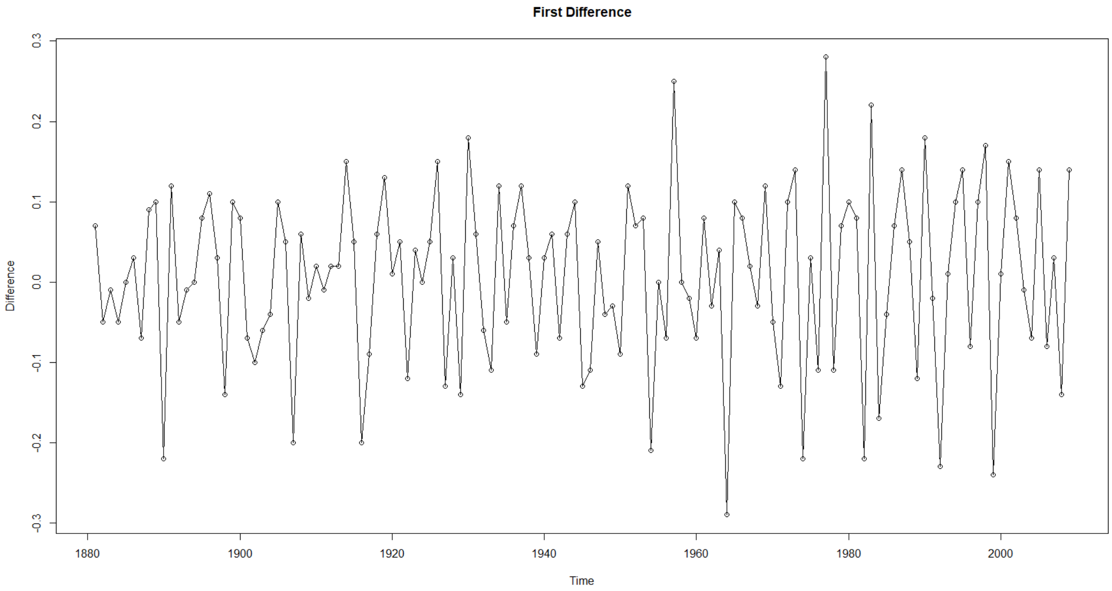


Figure 1.8.1: First differenced series ∇X_t. Average of first differenced series is $\hat{\delta} \approx 0.0066$

Compared to the OLS detrended plot shown before, this looks much more stationary.

Definition 1.8.1 — Differencing. Differencing a time series constitutes computing the difference between successive terms. A differenced time series is a time series of such differences. The first differenced series is denoted by

$$\nabla X_t = X_t - X_{t-1}$$

and is the series

$$\underbrace{X_2 - X_1, X_3 - X_2, \cdots, X_T - X_{T-1}}_{\text{Length of } T-1}$$

Higher order differences are calculated recursively, so

$$\nabla^d X_t = \nabla^{d-1}\nabla X_t \qquad (\nabla^0 X_t = X_t)$$

Detrending vs. Differencing

So far, detrending and differencing are both ways of reducing a potentially non-stationary time series to an approximately stationary series.

	Pros	**Cons**
Differencing	1. Does not require parameter estimations 2. Higher-order differencing can reduce even very "trendy" series to look more like noise	1. Can wash away the features of the series, and introduce more complicated structures 2. The trend is often of interest, and good estimates of the trend lend to improved long-range forecasts
Detrending	Vice Versa	Vice Versa

■ **Example 1.12 — When differencing complicates series.** Consider $X_t = W_t$, where $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Then,

$$\nabla X_t = W_t - W_{t-1} = Y_t$$

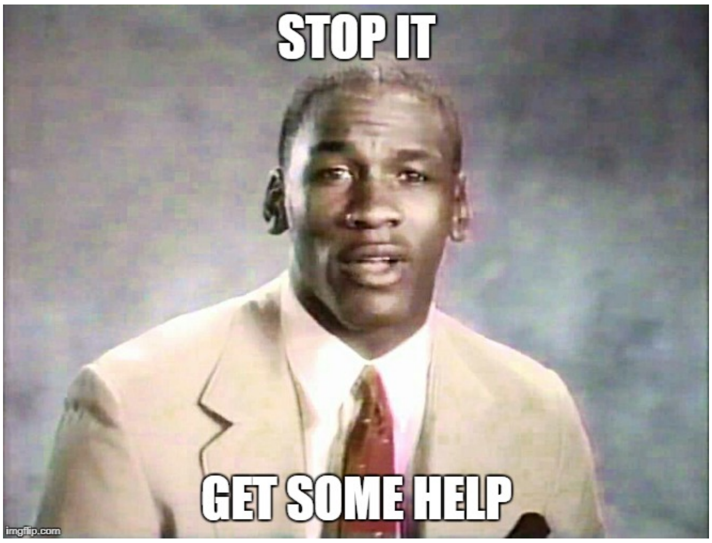
This is clearly a strictly stationary/weakly stationary process by Bernoulli shift and previous knowledge. We have

$$\gamma_X(h) = \mathbf{Cov}(X_t, X_{t+h}) = \begin{cases} \sigma_w^2 & h = 0 \\ 0 & h \geq 1 \end{cases}$$

while

$$\gamma_Y(h) = \mathbf{Cov}(Y_t, Y_{t+h}) = \begin{cases} 2\sigma_w^2 & h = 0 \\ -\sigma_w^2 & h = 1 \\ 0 & h \geq 2 \end{cases}$$

aha. More complicated...



■

1.9 Autocorrelation & Empirical Autocorrelation

Usually through either detrending or differencing, we arrive at a time series $\{X_t\}_{t\in\mathbb{Z}}$ that we may consider as reasonably stationary. Given such a series, we wish to estimate g, so that $X_t = g(W_t, W_{t-1}, \cdots)$ where $\{W_t\}_{t\in\mathbb{Z}}$ is a innovation sequence (strong white noise). We shall first assume g is linear as a starting point.

Definition 1.9.1 — Linear process. A time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be a linear process, if there exists a strong white noise $\{W_t\}_{t\in\mathbb{Z}}$, and coefficients $\{\psi_l\}_{l\in\mathbb{Z}} \subseteq \mathbb{R}$, so that $\sum_{l=-\infty}^{\infty}|\psi_l| < \infty$ (absolutely summable), and

$$X_t = \sum_{l=-\infty}^{\infty} \psi_l W_{t-l}$$

where $\mathbf{Var}(W_{t-l}) < \infty, \forall l$.

R This sum is well defined as a limit in L^2. Moreover, X_t depends on $W_t, W_{t-1}, \cdots$, but it also depends on the future term. This is bizarre. This is not exactly what we want for g.

Definition 1.9.2 — Causal Linear Process. We say $\{X_t\}_{t\in\mathbb{Z}}$ is a causal linear process, if

$$X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$$

This only depends on the past.

■ **Example 1.13 — Easy linear process.** Strong white noise sequence $\{X_t\}_{t\in\mathbb{Z}} = \{W_t\}_{t\in\mathbb{Z}}$ is a linear process trivially. ■

R Linear processes are strictly stationary since they can be written as Bernoulli shifts.

■ **Example 1.14 — MA(1) Process.** Recall the $\{X_t\}_{t\in\mathbb{Z}}$ given by

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

where $\{X_t\}_{t\in\mathbb{Z}}$ is a strong white noise. $\{X_t\}_{t\in\mathbb{Z}}$ is a linear process. We know that

$$\gamma_X(h) = \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0 \\ \theta\sigma_w^2 & h = 1 \\ 0 & h > 1 \end{cases}$$

when $h = 0$, the auto covariance function is always non-zero. From this example, we see that $\gamma_X(h)$ is non-zero for $h \geq 1$ only when the lagged terms in the linear process are non-zero. ■

This example shows us a way of figuring out what

$$g(W_t, W_{t-1}, \cdots) = \sum_{l=0}^{\infty} \psi_l W_{t-l}$$

must look like, by looking at the autocovariance function.
Before considering an estimator of the autocovariance function, we first introduce a standardized version of it.

Definition 1.9.3 — Autocorrelation Function. Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary. The autocorrelation function of $\{X_t\}_{t\in\mathbb{Z}}$ (ACF) is

$$\rho_X(h) = \frac{\gamma(h)}{\gamma(0)}, h \geq 0$$

Note that since $\gamma(0) = \mathbf{Var}(X_t) = \mathbf{Var}(X_0)$, we look at

$$|\gamma(h)| = |\mathbf{Cov}(X_t, X_{t+h})| \overset{\text{C-S Inequality}}{\leq} \sqrt{\mathbf{Var}(X_t)\,\mathbf{Var}(X_{t+h})} = \mathbf{Var}(X_0)$$

Thus, $|\rho_X(h)| \leq 1 \iff -1 \leq |\rho(h)| \leq 1$. Just like your ordinary correlation coefficient, but ACF means the correlation between X_t and its $h-$lagged counterpart.

Estimating $\gamma(h)$ and $\rho(h)$

Recall that

$$\gamma(h) = \mathbf{Cov}(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)], \mu = \mathbb{E}[X_t]$$

Hence, a sensible estimator is $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t = \bar{X}$, the sample mean/time series average. From this, we can estimate

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h} \underbrace{(X_t - \bar{X})(X_{t+h} - \bar{X})}_{\substack{\text{averaging over centred terms} \\ \text{h-time steps apart}}} \approx \frac{1}{T-h}\sum_{t=1}^{T-h}(X_t - \bar{X})(X_{t+h} - \bar{X})$$

Note that there are only $T-h$ terms in the sum since there are only $T-h$ pairs of terms that are h-lagged for observed time series of length T. Then,

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

■ **Example 1.15 — Estimation.** Consider $\{X_t\}_{t\in\mathbb{Z}} = \{W_t\}_{t\in\mathbb{Z}}$ where $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise with $\mathbf{Var}(W_t) = \sigma_W^2 < \infty$. Recall that

$$\gamma_X(h) = \begin{cases} \sigma_W^2 & h = 0 \\ 0 & h \geq 1 \end{cases}$$

then clearly,

$$\rho_X(h) = \begin{cases} 1 & h = 0 \\ 0 & h \geq 1 \end{cases}$$

in particular, $\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$ always. We generate a Gaussian white noise of length 500 and sample 130 out of it to plot ACF.

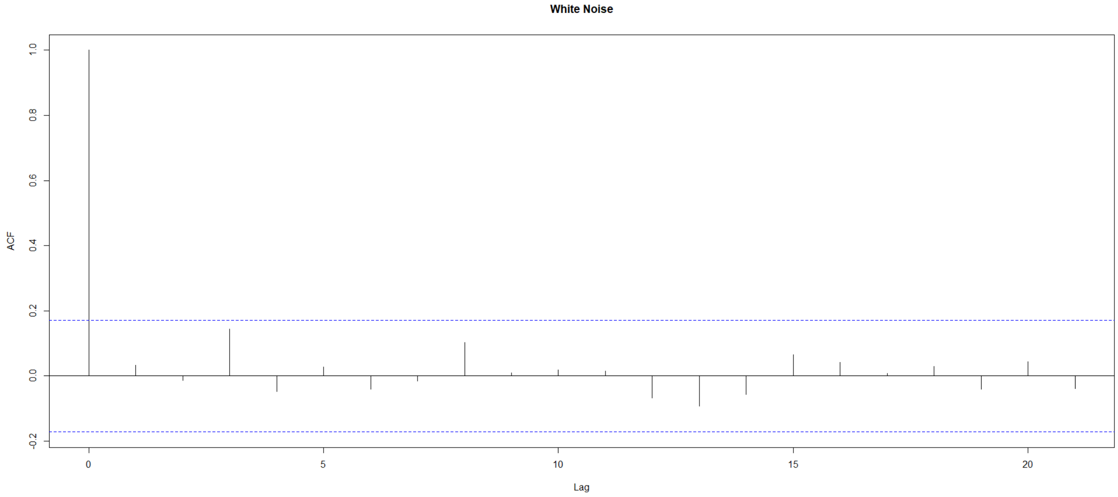


Figure 1.9.1: ACF of white noise, sample length 130

We see that $\rho(0) = 1$ and most of the terms after that stay within the region bounded by the blue lines. This is an indication of the process being strong white noise. Okay, so what are these magical blue boundaries? ■

1.10 Modes of Convergence of Random Variables (Review)

We have seen that $\hat{\gamma}(h)$ is an estimator of $\gamma(h)$, and we want to discuss the asymptotic properties of this estimator. Before doing this, we need to do some review on:

1. **Stochastic boundedness**
2. **Convergence in probability**
3. **Convergence in distribution**

Definition 1.10.1 — Bounded in probability. Suppose $\{X_n\}_n$ is a sequence of random variables. We say that $\{X_n\}_n$ is bounded in probability by $\{Y_n\}$ if for all $\varepsilon > 0$, there exists $M, N \in \mathbb{R}$ such that for all $n \geq N$,

$$\mathbb{P}\left(\left|\frac{X_n}{Y_n}\right| > M\right) \leq \varepsilon$$

In shorthand, $X_n = \mathcal{O}_p(Y_n)$ and say X_n is on the order of Y_n.

Definition 1.10.2 — Convergence in probability. We say $\{X_n\}_n$ converges in probability to X if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

If $\{a_n\}_n$ is a sequence of scalars, we abbreviate $\left\{\frac{X_n}{a_n}\right\}_n$ converges in probability to 0 as $X_n = o_p(a_n)$ if and only if

$$\mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0$$

Hence X_n converging in probability to 0 is denoted as $o_p(1)$. We also write $X_n \xrightarrow{P} X$ for convergence in probability.

Definition 1.10.3 We say that the sequence of scalar random variables $\{X_n\}_n$ white respective CDF's $F_n(x)$ converges in distribution to X with CDF $F(x)$ if for all continuity points y of F,

$$\lim_{n\to\infty} |F_n(y) - F(y)| = 0$$

R When $F(x)$ is the CDF of a continuous random variable, then

$$\lim_{n\to\infty} |F_n(y) - F(y)| = 0, \forall y \in Dom(F)$$

Recall a useful tool from STAT330.

Theorem 3 — Chebyshev's Inequality. If $\mathbb{E}(Y^2) < \infty$, then

$$\begin{aligned}\mathbb{E}[Y^2] &= \mathbb{E}\left[Y^2\mathbf{1}_{\{|Y|\geq\mu\}} + Y^2\mathbf{1}_{\{|Y|<\mu\}}\right]\\ &= \mathbb{E}\left[Y^2\mathbf{1}_{\{|Y|\geq\mu\}}\right] + \mathbb{E}\left[+Y^2\mathbf{1}_{\{|Y|<\mu\}}\right]\\ &\geq \mathbb{E}\left[Y^2\mathbf{1}_{\{|Y|\geq\mu\}}\right] \geq \mu^2\mathbb{E}[\mathbf{1}_{\{|Y|\geq\mu\}}]\\ &= \mu^2\mathbb{P}(|Y| \geq \mu)\end{aligned}$$

This gives the Chebyshev's inequality,

$$\mathbb{P}(|Y| \geq \mu) \leq \frac{\mathbb{E}[Y^2]}{\mu^2}$$

This is useful since it provides us a bound on the probability distribution when we want to discuss about different convergences. We simply look at the second moment. This can be generalized to higher-order moments. When $\mathbb{E}[|Y|^k] < \infty$, then

$$\mathbb{P}(|Y| \geq \mu) \leq \frac{\mathbb{E}[|Y|^k]}{\mu^k}$$

■ **Example 1.16** Suppose X_n is a strong white noise in L^2 ($\mathbb{E}[X_0^2] < \infty$), and let

$$\bar{X}_T = \frac{1}{T}\sum_{t=1}^{t} X_t$$

then,

1. $|\bar{X}_T| = o_p(1)$:

   Proof. Let $\varepsilon > 0$. We compute

   $$\textbf{Var}(\bar{X}_T) = \mathbb{E}[\bar{X}_T^2] = \frac{1}{T^2}\mathbb{E}\left[\left(\sum_{t=1}^{T} X_t\right)^2\right] == \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}[X_t X_s]$$

   Note the only non-zero terms are of the form $E[X_t^2]$. Thus,

   $$\textbf{Var}(\bar{X}_T) = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}[X_t^2] = \frac{\sigma^2}{T}$$

   By Chebyshev's inequality,

   $$\mathbb{P}(|\bar{X}_T| > \varepsilon) \leq \frac{\mathbb{E}[\bar{X}_T^2]}{\varepsilon^2} = \frac{\sigma^2/T}{\varepsilon^2} \xrightarrow{T\to\infty} 0$$

   This means $|\bar{X}_T| \xrightarrow{P} 0$ and $|\bar{X}_T| = o_p(1)$. ■

2. $\bar{X}_T = \mathcal{O}_p\left(\frac{1}{\sqrt{T}}\right)$:

   Proof. As before

   $$\textbf{Var}\left(\frac{\bar{X}_T}{1/\sqrt{T}}\right) = T\textbf{Var}(\bar{X}_T) = \sigma^2$$

   By Chebyshev's inequality, for $M > 0$,

   $$\mathbb{P}\left(\left|\frac{\bar{X}_T}{1/\sqrt{T}}\right| > M\right) \leq \frac{\sigma^2}{M^2} \xrightarrow{M\to\infty} 0$$

   Thus, $\sqrt{T}\bar{X}_T = \mathcal{O}_p(1)$ and $\bar{X}_T = \mathcal{O}_p\left(\frac{1}{\sqrt{T}}\right)$. ■

■

R Alternatively, we can show this using the Central Limit Theorem, which usually gives you tighter bounds than the Chebyshev's inequality but we need to deal with convergence in distribution first. By the CLT,

$$\sqrt{T}\bar{X}_T \xrightarrow{D} N(0,\sigma^2)$$

Therefore, if F_T is the CDF of $\sqrt{T}\bar{X}_T$ and Φ is the CDF of $N(0,1)$ random variable. Then,

$$\left|F_T(x)-\Phi\left(\frac{x}{\sigma}\right)\right| \xrightarrow{T\to\infty} 0, \forall x\in\mathbb{R}$$

For $\varepsilon>0$, we can choose M large enough so that

$$\Phi\left(\frac{x}{\sigma}\right)=1-\Phi\left(\frac{M}{\sigma}\right)\leq\frac{\varepsilon}{4}$$

For this particular M, we can choose T_0 such that $T\geq T_0$ implies

$$\left|F_T(-M)-\Phi\left(\frac{-M}{\sigma}\right)\right|\leq\frac{\varepsilon}{4}$$

and by flipping the CDFs, we also have

$$\left|F_T(M)-\Phi\left(\frac{M}{\sigma}\right)\right|\leq\frac{\varepsilon}{4}$$

Then,

$$\mathbb{P}(\left|\sqrt{T}\bar{X}_T\right|\geq M)=F_T(-M)+(1-F_T(M))\leq\frac{\varepsilon}{4}+\frac{\varepsilon}{4}+\frac{\varepsilon}{4}+\frac{\varepsilon}{4}=\varepsilon$$

You see this is a bit harder but CLT can use convergence in distribution to confirm boundedness in probability.

R In general, if $\frac{X_n}{a_n}$ converges to some non-degenerate (not Dirac-like) random variable in distribution, then $X_n=\mathcal{O}_p(a_n)$. This is interesting since we do not need convergence in probability to do this proof.

Algebra of $\mathcal{O}_p$ and o_p Notation

1. If $X_n=\mathcal{O}_p(a_n)$ and $Y_n=\mathcal{O}_p(b_n)$, then $X_n+Y_n=\mathcal{O}_p(a_n\vee b_n)$ where $a_n\vee b_n=\max\{a_n,b_n\}$
2. If $X_n=o_p(1)$ and $Y_n=o_p(1)$, then $X_n+Y_n=o_p(1)$
3. If $X_n=o_p(1)$ and $Y_n=o_p(1)$, then $X_nY_n=o_p(1)$

■ **Example 1.17 — Autocovariance estimator is consistent.** Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise in L^2 with $\mathbb{E}[W_t^4]<\infty$. Let $X_t=W_t+\theta W_{t-1},\theta\in\mathbb{R}$. Show that $\hat{\gamma}(1)\xrightarrow{P}\theta\sigma_W^2$.

Proof. Let $\bar{X}_T=\bar{X}=\frac{1}{T}\sum_{t=1}^T(W_t+\theta W_{t-1})=\frac{1}{T}\sum_{t=1}^T W_t+\frac{\theta}{T}\sum_{t=1}^T W_{t-1}=o_p(1)$. Then,

$$\hat{\gamma}(1)=\frac{1}{T}\sum_{t=1}^{T-1}(X_t-\bar{X})(X_{t+1}-\bar{X})=\frac{1}{T-1}\sum_{t=1}^{T}X_tX_{t+1}+\underbrace{\frac{T-1}{T}(\bar{X})^2}_{R_{1,T}}\underbrace{-\bar{X}\frac{1}{T}\sum_{t=1}^{T-1}X_t}_{R_{2,T}}\underbrace{-\bar{X}\frac{1}{T}\sum_{t=2}^{T}X_{t+1}}_{R_{3,T}}$$

Note that $R_{i,T}=o_p(1)$ for $i=1,2,3$. Then we focus on

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}X_tX_{t+1}&=\frac{1}{T}\sum_{t=1}^{T}(W_t+\theta W_{t-1})(W_{t+1}+\theta W_t)\\&=\frac{1}{T}\sum_{t=1}^{T}\theta W_t^2+G_{1,T}+G_{2,T}+G_{3,T}\end{aligned}$$

Note that, by the Strong Law of Large Number (SLLN), we have $\frac{1}{T}\sum_{t=1}^{T}\theta W_t^2 \xrightarrow{a.s} \theta\mathbb{E}[W_t^2] = \theta\sigma_W^2$. Now, $G_{1,T} = \frac{1}{T}\sum_{t=1}^{T} W_t W_{t+1}$: we calculate the variance to have probability bounds by Chebyshev. Note that $\mathbb{E}[G_{1,T}] = 0$ by independence. Then,

$$\mathbf{Var}\left(G_{1,T}\right) = \mathbb{E}[G_{1,T}^2] = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} \underbrace{\mathbb{E}[W_t W_{t+1} W_s W_{s+1}]}_{<\infty, \mathbb{E}[W_t^4]<\infty \text{ and C-S inequality}} = \frac{1}{T^2}\mathbb{E}[W_t^2 W_{t+1}^2] = \frac{1}{T}\sigma_W^2$$

Since $\mathbf{Var}\left(G_{1,T}\right) \xrightarrow{T\to\infty} 0$. Then, by similar argument as before, we know that $G_{1,T} = o_p(1)$. We can do this similarly for $G_{2,T}, G_{3,T}$. This finally gives us

$$\hat{\gamma}(1) \xrightarrow{P} \theta\sigma_W^2$$

■

■

"We distribute the sum... we distribute the sum... we distribute the sum......"

—Prof.Rice

1.11 *M* Dependent CLT (Advanced)

Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is a mean zero and strictly stationary time series with $\mathbb{E}[X_t^2] < \infty$. We are frequently faced with the problems:

1. What is the approximate distribution of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t = \sqrt{T}\bar{X}_T \overset{D}{\approx} N(0, \sigma_X^2)$?
2. If $\{X_t\}_{t\in\mathbb{Z}}$ is a strong white noise, what is the approximate distribution of $\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h} X_t X_{t+h} + o_p(1)$?

one of the main obstacles when trying to answer these two problems is the random variables in the summand are not necessarily iid. We know that $X_t X_{t+h}$'s are not iid but they are strictly stationary. Thus, the real question to ask is when the averaging estimator's approximate distribution with potential serial correlation but otherwise stationary is normal. This requires the generalized CLT with dependency.
Given an observed time series:

1. Only way to understand how $\{X_t\}_{t\in\mathbb{Z}}$ behaves, we have to observe replicates of the process
2. If process is suitably "weakly dependent", then we can observe replicates of the process by viewing on overlapping windows. In this case, we assume data observed on different windows are like independent replicates.

Definition 1.11.1 — $M-$dependent. We say a time series $\{X_t\}_{t\in\mathbb{Z}}$ is $M-$dependent for a positive integer M, if for all $t_1 < t_2 < \cdots < t_{d_1} < s_1 < s_2 < \cdots < s_{d_2} \in \mathbb{Z}$, so that $t_{d_1} + M \le s_1$ and

$$(X_{t_1}, \cdots, X_{t_{d_1}}) \perp\!\!\!\perp (X_{s_1}, \cdots, X_{s_{d_2}})$$

■ **Example 1.18 — $M-$dependent time series .** $\{W_t\}_{t\in\mathbb{Z}}$ strong white noise is 1-dependent and $X_t = W_t + \theta W_{t-1}$ is 2-dependent. ■

Definition 1.11.2 — Triangular array. We say $\left\{X_{i,j} : 1 \le j \le n_i, 1 \le i < \infty\right\}$ forms a triangular array of mean zero L^2 random variables if for each $i-$fixed, $X_{i,1}, \cdots, X_{i,n_i}$ are independent, and $n_i < n_{i+1}$.

This is basically saying row-wise random variables are independent as illustrated below.

$$\begin{array}{ccccc} X_{1,1} & \cdots & X_{1,n_1} & & \\ X_{2,1} & \cdots & \cdots & X_{2,n_2} & \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{array}$$

Theorem 4 — Lindeberg-Feller CLT for Triangular Arrays. Let $\left\{X_{i,j} : 1 \leq j \leq n_i, 1 \leq i < \infty\right\}$ be a triangular array of mean zero and L^2 random variables. Define

$$\sigma_i^2 = \sum_{j=1}^{n} \mathbf{Var}\left(X_{i,j}\right), \qquad S_i = \frac{1}{\sigma_i} \sum_{j=1}^{n_i} X_{i,j}$$

If for $\varepsilon > 0$,

$$\underbrace{\frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} \mathbb{E}\left[X_{i,j}^2 \mathbf{1}_{\left\{\left|X_{i,j}\right| > \varepsilon \sigma_i\right\}}\right] \xrightarrow{i \to \infty} 0}_{\textbf{Lindeberg's Condition}}$$

Then, $S_i \xrightarrow{D} N(0,1)$

The LHS of the Lindeberg's condition is like a percentage of row variance (between 0 and 1). It calculates the percentage of values, $X_{i,j}$, which contribute non-negligible amount to the row variance. We want to check this percentage goes to 0 as we move down the triangular array. This is also called **uniform asymptotic negligibility condition**.

Theorem 5 — M =dependent CLT. Suppose $\{X_t\}_{t \in \mathbb{Z}}$ is a strictly stationary, and $M-$dependent time series with $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[X_t^2] < \infty$. Then

$$S_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t = \sqrt{T}\bar{X} \xrightarrow{D} N(0, \sigma_M^2)$$

where

$$\sigma_M^2 = \sum_{h=-m}^{m} \gamma(h) = \gamma(0) + 2 \sum_{h=1}^{m} \gamma(h)$$

Need $\sigma_M \neq 0$.

Proof. **Bernstein Blocking Argument:** we take a given time series of length T.

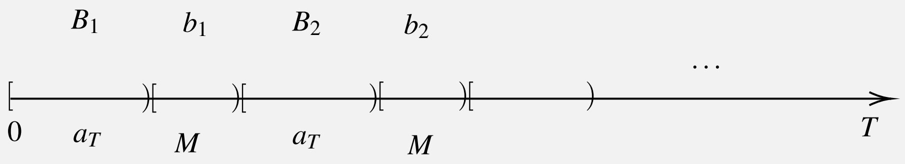


Figure 1.11.1: Bernstein blocking argument

Let a_T denote big block size and M denote little block size. We assume $a_T \to \infty$ as $T \to \infty$ but $\frac{a_T}{T} \to 0$. Then,

$$N = \text{number of blocks} = \left\lfloor \frac{T}{M + a_T} \right\rfloor$$

Let

$$B_j = \{i : (j-1)(a_T+M)+i \le i \le ja_T+(j-1)M\},$$
$$b_j = \{i : ja_T+(j-1)M+1 \le j \le j(a_T+M)\}$$

since $a_T \nearrow \infty$, for T sufficiently large, $a_T > M$, and so by $M-$dependence, $\sum_{t\in B_j} X_t$ is independent from $\sum_{t\in B_k} X_t$ whenever $j \neq k$. Similar for b_j, b_k.

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t &= \frac{1}{\sqrt{T}}\sum_{j=1}^{N}\sum_{t\in B_j} X_t + \frac{1}{\sqrt{T}}\sum_{j=1}^{N}\sum_{t\in b_j} X_t + R \\ &= G_{1,T} + G_{2,T} + G_{3,T}\end{aligned}$$

We want the show the big blocks dominate. Note that

$$\mathbf{Var}\left(G_{2,T}\right) = \frac{1}{T}\sum_{j=1}^{N}\mathbb{E}\left[\left(\sum_{t=1}^{M} X_t\right)^2\right] \overset{\text{strict stat}}{=} \frac{N}{T}\mathbb{E}\left[\left(\sum_{t=1}^{M} X_t\right)^2\right]$$

In particular,

$$\mathbb{E}\left[\left(\sum_{t=1}^{M} X_t\right)^2\right] = \sum_{t=1}^{M}\sum_{s=1}^{M}\mathbb{E}[X_t X_s] = \sum_{t=1}^{M}\sum_{s=1}^{M}\gamma(|t-s|)$$

let $h = t - s$, then

$$= \sum_{h=1-m}^{m-1}(m-|h|)\gamma(h) = C < \infty$$

Thus,

$$\mathbf{Var}\left(G_{2,T}\right) = \frac{NC}{T} \xrightarrow{a_T\to\infty} 0 \Longrightarrow G_{2,T} = o_p(1)$$

Notice

$$\begin{aligned}G_{1,T} &= \frac{1}{\sqrt{T}}\sum_{j=1}^{N}\sum_{t\in B_j} X_t \\ &= \sum_{j=1}^{N}\underbrace{\left[\frac{\sum_{t\in B_j} X_t}{\sqrt{T}}\right]}_{Y_{j,T}}\end{aligned}$$

Note that $\{Y_{j,T}\}$ forms a triangular array! Thus, we check the Lindeberg's condition.

$$\begin{aligned}
\mathbf{Var}(G_{1,T}) &= \sum_{j=1}^{N} \mathbf{Var}(Y_{j,T}) \\
\mathbf{Var}(Y_{j,T}) = \mathbf{Var}(Y_{1,T}) &= \frac{1}{T}\mathbb{E}\left[\left(\sum_{t=1}^{a_T} X_i\right)^2\right] \\
\frac{1}{T}\sum_{t=1}^{a_T}\sum_{s=1}^{a_T}\mathbb{E}[X_tX_s] &= \frac{1}{T}\sum_{h=1-a_T}^{a_T}(a_T-|h|)\gamma(h) \\
&\overset{\gamma(h)=0,|h|\geq M}{=} \frac{1}{T}\sum_{h=-M}^{M}(a_T-|h|)\gamma(h) \\
\implies \mathbf{Var}(G_{1,T}) &= \frac{N}{T}\sum_{h=-M}^{M}(a_T-|h|)\gamma(h) \\
\overset{\frac{N}{T}\approx\frac{1}{a_T}}{\implies} \mathbf{Var}(G_{1,T}) &\xrightarrow{T\to\infty} \sum_{h=-M}^{M}\gamma(h)
\end{aligned}$$

Thus, the variance of $G_{1,T}$ is bounded. So, we must show

$$\sum_{j=1}^{N}\mathbb{E}\left[Y_{j,T}^2\mathbf{1}_{\{|Y_{j,T}|>\varepsilon\sigma_N\}}\right] \xrightarrow{T\to\infty} 0$$

$$\sum_{j=1}^{N}\mathbb{E}\left[Y_{j,T}^2\mathbf{1}_{\{|Y_{j,T}|>\varepsilon\sigma_N\}}\right] \overset{\text{iid}}{=} N\mathbb{E}\left[Y_{1,T}^2\mathbf{1}_{\{|Y_{1,T}|>\varepsilon\sigma_N\}}\right]$$

Note that $\mathbb{E}[|Y|^{2+\delta}] \geq \mathbb{E}[|Y|^{2+\delta}\mathbf{1}_{\{|Y|>\varepsilon\}}] \geq \varepsilon^{\delta}\mathbb{E}[|Y|^2\mathbf{1}_{\{|Y|>\varepsilon\}}]$ (non-trivial). Then,

$$\mathbb{E}[|Y|^2\mathbf{1}_{\{|Y|>\varepsilon\}}] \leq \frac{\mathbb{E}[|Y|^{2+\delta}]}{\varepsilon^{\delta}}$$

It may be shown that $\mathbb{E}\left[\left|Y_{j,T}\right|^{2+\delta}\right] \leq C\left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}}$ for some positive constant C. Thus,

$$N\mathbb{E}\left[Y_{1,T}^2\mathbf{1}_{\{|Y_{1,T}|>\varepsilon\sigma_N\}}\right] \leq \frac{N}{(\varepsilon\sigma_N)^{\delta}}C\left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}} = \frac{C}{(\varepsilon\sigma_N)^{\delta}}\frac{Na_T}{T}\left(\frac{a_T}{T}\right)^{\frac{\delta}{2}} \xrightarrow{T\to\infty} 0$$

Lindeberg's condition holds. This implies

$$\frac{G_{1,T}}{\sigma_N} \xrightarrow{D} N(0,1)$$

by Lindeberg-Feller theorem and since $\sigma_N^2 \xrightarrow{T\to\infty} \sum_{j=-M}^{M}\gamma(j)$, we have

$$G_{1,T} \xrightarrow{D} N\left(0, \sum_{j=-M}^{M}\gamma(j)\right)$$

Since $G_{2,T} = o_p(1)$, we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \xrightarrow{D} N\left(0, \sum_{j=-M}^{M} \gamma(j)\right)$$

■

Non-tivial Aside: Calculating $\mathbb{E}[|Y_{1,T}|^{2+\delta}]$

We want to show

$$\mathbb{E}[|Y_{1,T}|^{2+\delta}] \leq C\left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}}$$

where all variables are as in the proof of the $M-$dependent CLT.

Theorem 6 — Rosenthal's Inequality. If $X_1, \cdots, X_n$ are independent random variables with $\mathbb{E}[|X_i|^{2+\delta}] < \infty$ with $\delta > 0$. Then,

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^{2+\delta}\right] \leq c_q n^{\delta/2} \sum_{i=1}^{n} \mathbb{E}[|X_i|^{2+\delta}`]$$

for some constant c_p. In particular, when X_i's are iid, then

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^{2+\delta}\right] \leq c_q n^{\frac{2+\delta}{2}} \mathbb{E}[|X_1|^{2+\delta}]$$

Proof. See Petrov, *Limit Theorems of Probability Theory*, page 59. ■

Proposition 1.11.1 For arbitrary random variables $X_1, \cdots, X_n$,

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^{2+\delta}\right] \leq n^{(2+\delta)-1} \sum_{i=1}^{n} \mathbb{E}\left[|X_i|^{2+\delta}\right]$$

Proof. By Jensen's Inequality, for all real numbers $a_1, \cdots, a_n$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} a_i\right|^{2+\delta} \leq \frac{1}{n}\sum_{i=1}^{n} |a_i|^{2+\delta}$$

since $f(x) = x^{2+\delta}$ is convex. By rearrangement,

$$\left|\sum_{i=1}^{n} a_i\right|^{2+\delta} \leq n^{(2+\delta)-1} \sum_{i=1}^{n} |a_i|^{2+\delta}$$

take $a_i \sim X_i$ and take expectations. ■

Proof. **for the non-trivial fact**

$$\sum_{t=1}^{a_T} X_t = \sum_{j=0}^{M} \sum_{\substack{t=j \mod M+1 \\ 1 \leq t \leq a_T}} X_t$$

the variables in the most inner summand are separated by at least M time steps and are hence iid. Thus,

$$\mathbb{E}\left[\left|\sum_{t=1}^{a_T} X_t\right|^{2+\delta}\right] \overset{\text{Prop'N 1.11.1}}{\leq} (M+1)^{(2+\delta)-1}\mathbb{E}\left[\left|\sum_{\substack{t=j \mod M+1\\ 1\leq t\leq a_T}} X_t\right|^{2+\delta}\right]$$

$$\overset{\text{Rosent' Ineq}}{\leq} (M+1)^{(2+\delta)-1}\left(\frac{a_T}{M+1}\right)^{\frac{2+\delta}{2}} c_p\mathbb{E}\left[|X_1|^{2+\delta}\right]$$

$$= Ca_T^{\frac{2+\delta}{2}}$$

where C is the constant required in the proof of $M-$dependent CLT. ■

1.12 Linear Process CLT (Advanced)

We have seen if X_t is a $M-$dependent, strictly stationary with zero mean and in L^2, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t \xrightarrow{D} N\left(0, \sum_{h=-M}^{M}\gamma(h)\right)$$

■ **Example 1.19 — $M-$dependent time series.** The time series given by $X_t = \sum_{l=0}^{M}\psi_l W_{t-l}$ where $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise in L^2 is a $M-$dependent process. ■

However, a general (causal) linear process $X_t = \sum_{l=0}^{\infty}\psi_l W_{t-l}$ is not $M-$dependent. How do we establish a CLT for this kind of infinite dependency back into the past?

Figure 1.12.1: When you are looking into the void, the void is also looking back at you.

Theorem 7 — Basic Approximation Theorem (BAT). Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is a sequence of random variables so that there exists an array $\{Y_{m,n} : m,n \geq 1\}$ so that

1. For each fixed m, $Y_{m,n} \xrightarrow{D} Y_m$ as $n \to \infty$
2. $Y_m \xrightarrow{D} Y$ as $m \to \infty$ for some random variable Y
3. $\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}(|X_n - Y_{m,n}| > \varepsilon) 0$ for all $\varepsilon > 0$

Then, $X_n \xrightarrow{D} Y$ as $n \to \infty$.

Proof. Using characteristic functions in Shumway & Stoffer Appendix. ■

R This looks hella like uniform boundedness principle or simple function approximation theorem.

Theorem 8 — Linear Process CLT. Suppose $X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$ is a causal linear process with absolutely summable coefficient sequence and $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise in L^2. Then if $S_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t$, then

$$S_T \xrightarrow{D} N\left(0, \underbrace{\sum_{l=-\infty}^{\infty} \gamma(l)}_{\text{Long-run variance of } X_t}\right)$$

Proof. In this case, $\{X_t\}_{t\in\mathbb{Z}}$ is strictly (and weakly stationary). And we have the following nice observation of $\gamma(h)$.

$$\gamma(h) = \mathbb{E}[X_t X_{t+h}] = \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \psi_l W_{t-l}\right)\left(\sum_{j=0}^{\infty} \psi_j W_{t+h-j}\right)\right]$$

$$\overset{\text{Fubini}}{=} \sum_{l=0}^{\infty}\sum_{j=0}^{\infty} \psi_l \psi_j \underbrace{\mathbb{E}[W_{t-l} W_{t+h-j}]}_{\neq 0, j=l+h} = \sum_{l=0}^{\infty} \psi_l \psi_{l+h} \sigma_W^2$$

Then,

$$\sum_{h=-\infty}^{\infty} \gamma(h) = \sum_{h=-\infty}^{\infty} \left|\sum_{l=0}^{\infty} \psi_l \psi_{l+h} \sigma_W^2\right| \leq \sum_{l=0}^{\infty} |\psi_l| \sum_{h=-\infty}^{\infty} |\psi_l| \sigma_W^2 < \infty$$

and it is well-defined. Note that $\mathbb{E}[S_T] = 0$ and

$$\mathbf{Var}\,(S_T) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathbb{E}[X_t X_s] = \frac{1}{T}\sum_{h=1-T}^{T-1} (T - |h|)\gamma(h) = \sum_{h=1-T}^{T-1} \left(1 - \frac{|h|}{T}\right)\gamma(h)$$

Note that $\left(1 - \frac{|h|}{T}\right)\gamma(h) \leq |\gamma(h)|$. Since $\{|\gamma(h)|\}$ is summable, by Dominated Convergence Theorem (DCT),

$$\mathbf{Var}\,(S_T) \xrightarrow{T\to\infty} \sum_{h=-\infty}^{\infty} \gamma(h)$$

Define $X_{t,m} = \sum_{l=0}^{m} \psi_l W_{t-l}$ and $S_{T,m} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_{t,m}$ forms a $m-$dependent approximation to S_T.

1. By the $m-$dependent CLT,

$$S_{T,m} \xrightarrow{D} N\left(0, \sum_{h=-m}^{m} \gamma_m(h)\right) =: S'_m$$

   and $\gamma_m(h) = \mathbb{E}[X_{t,m}X_{t+h,m}]$.
2. By DCT, $\sum_{h=-m}^{m} \gamma_m(h) \xrightarrow{m\to\infty} \sum_{h=-\infty}^{\infty} \gamma(h)$. And hence, $S'_m \xrightarrow{m\to\infty,D} N\big(0, \sum_{h=-\infty}^{\infty} \gamma(h)\big)$
3. Note that

$$\begin{aligned}\mathbb{E}\left[(S_{T,m} - S_T)^2\right] &= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(\underbrace{X_t - X_{t,m}}_{\sum_{l=m+1}^{\infty}\psi_l W_{t-l}})^2\right] \\ &\leq \sum_{h=1-T}^{T-1}\left(1-\frac{|h|}{T}\right)\sum_{l=m+1}^{\infty}|\psi_l||\psi_{l+h}|\sigma_W^2 \\ &\leq \sum_{l=m+1}^{\infty}|\psi_l|\left(\sum_{h=-\infty}^{\infty}|\psi_h|\right)\sigma_W^2 \xrightarrow{m\to\infty} 0\end{aligned}$$

   This will give us the third condition of BAT using Chebyshev's inequality.

Nice, we got all the BAT conditions. Hence,

$$S_T \xrightarrow{D} N\left(0, \sum_{l=-\infty}^{\infty} \gamma(l)\right)$$

■

1.13 Asymptotic Properties of Empirical ACF

If $X_1, \cdots, X_T$ is an observed time series that we think it was generated by a stationary process. Then, $\mathbf{Cov}(X_t, X_{t+h})$ does not depend on t. Then, recall

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_t - \bar{X})(X_{t+h} - \bar{X})$$

and

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

Questions:

1. Are $\hat{\gamma}$ and $\hat{\rho}$ consistent?
2. What is the approximate distribution of $\hat{\gamma}(h)$ or $\hat{\rho}(h)$

Consistency

By adding and subtracting μ in the definition of $\hat{\gamma}(h)$, we may assume without loss of generality that $\mathbb{E}[X_t] = 0$. Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary, and $X_t = g(W_t, W_{t-1}, \cdots)$ (Bernoulli-shift). We first need to show consistency of $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$. This is not as easy as we think since X_t's are not necessarily iid as in classical statistics. So, we cannot use Law of Large Number to bombard this problem. There is an extension in Ergodic theorem. We shall ignore it. Hence,

$$\bar{X} \xrightarrow{P} 0$$

by the Ergodic theorem. Furthermore,

$$\begin{aligned}\hat{\gamma}(h) &= \frac{1}{T}\sum_{t=1}^{T-h}(X_t-\bar{X})(X_{t+h}-\bar{X}) \\ &= \frac{1}{T}\sum_{t=1}^{T-h}X_tX_{t+h} - \underbrace{\bar{X}\frac{1}{T}\sum_{t=1}^{T-h}X_t}_{\xrightarrow{P}0} - \underbrace{\bar{X}\frac{1}{T}\sum_{t=1}^{T-h}X_{t+h}}_{\xrightarrow{P}0} + \underbrace{\frac{T-h}{T}(\bar{X})^2}_{\xrightarrow{P}0}\end{aligned}$$

Note that

$$\mathbb{E}[X_tX_{t+h}] = \gamma(h)$$

and $X_tX_{t+h} = g_h(W_{t+h}, W_{t+h-1}, \cdots)$ is another ergodic sequence. Again, by the Ergodic theorem,

$$\frac{1}{T}\sum_{t=1}^{T-h}X_tX_{t+h} \xrightarrow{P} \gamma(h)$$

Thus, $\hat{\gamma}(h) \xrightarrow{P} \gamma(h)$ and $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \xrightarrow{P} \rho(h)$ under strict stationarity and ergodicity on finite second moment.

Distribution

Consider simple (but perhaps most important) case, $\{X_t\}_{t\in\mathbb{Z}}$ is a strong white noise in L^4. Note that the finite 4-th moment assumption is not really needed here but this will be explained why it is classically assumed. We know that

$$\hat{\gamma}(h) \xrightarrow{P} 0$$

similarly as before,

$$\hat{\gamma}(h) = \underbrace{\frac{1}{T}\sum_{t=1}^{T-h}X_tX_{t+h}}_{\tilde{\gamma}(h)} + R$$

Note that $\mathbb{E}[\tilde{\gamma}(h)] = 0$ for $h \geq 1$ while

$$\mathbf{Var}\,(\tilde{\gamma}(h)) = \mathbb{E}[\tilde{\gamma}^2(h)] = \frac{1}{T^2}\sum_{t=1}^{T-h}\sum_{s=1}^{T-h}\mathbb{E}[X_tX_{t+h}X_sX_{s+h}]$$

well-defined since we are dealing with white noises (or say finite 4-th moment). The summand is non-zero only when $t = s$, hence,

$$\mathbf{Var}\,(\tilde{\gamma}(h)) = \frac{1}{T^2}\sum_{t=1}^{T-h}\mathbb{E}[X_t^2X_{t+h}^2] = \frac{T-h}{T^2}\sigma_X^4$$

Therefore,

$$\mathbf{Var}\left(\sqrt{T}\tilde{\gamma}(h)\right) \xrightarrow{T\to\infty} \sigma_X^4$$

Theorem 9 If $\{X_t\}_{t\in\mathbb{Z}}$ is a strong white noise with finite 4-th moment, then

$$\sqrt{T}\tilde{\gamma}(h) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T-h}X_tX_{t+h} \xrightarrow{D} N\left(0, \sigma_X^4\right)$$

Proof. Direct application of Martingale CLT which is dervied from $M-$depenent CLT. ■

Corollary 1.13.1 Given that $\sqrt{T}\hat{\gamma}(h) \xrightarrow{D} N(0,\sigma_X^4)$ and $\hat{\gamma}(0) \xrightarrow{P} \sigma_X^2$ by Strong Law of Large Number. By Slutsky's Theorem,

$$\sqrt{T}\frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \sqrt{T}\hat{\rho}(h) \xrightarrow{D} N(0,1)$$

Proposition 1.13.2 If $\{X_t\}_{t\in\mathbb{Z}}$ is a strong white noise,

$$\left(-\frac{Z_{\alpha/2}}{\sqrt{T}}, \frac{Z_{\alpha/2}}{\sqrt{T}}\right)$$

is a $(1-\alpha)$ prediction interval for $\hat{\rho}(h)$ for all h with T large where $\Phi(Z_{\alpha/2}) = 1-\frac{\alpha}{2}$.

In particular, $\left(-\frac{1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right)$ is an approximate 95% prediction interval for $\hat{\rho}(h)$ assuming the data is generated by a strong white noise. Moreover, we know that $\hat{\rho}(h)$ is consistent. Thus, for a non-zero $\rho(h)$, when T gets larger, eventually, we can see this interval will exclude this $\rho(h)$. This gives us a way to examine at what lag there exists substantial serial correlation in the time series.

Recall that

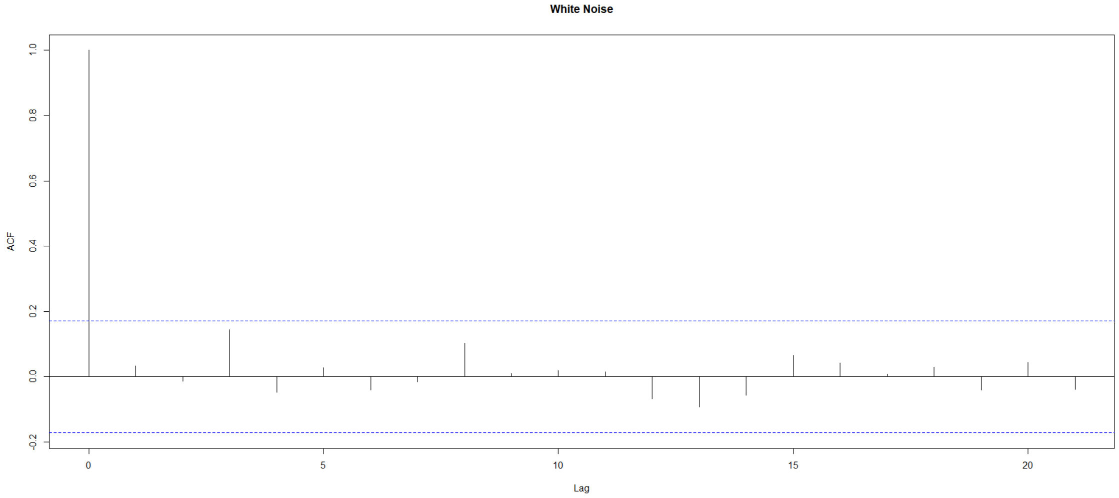


Figure 1.13.1: ACF of white noise, sample length 130

Now, we know these blue boundaries are $\pm\frac{1.96}{\sqrt{T}}$. We see that most of the lines are within the boundaries but there could be some instances outside the boundaries since we are not 100% confidence.

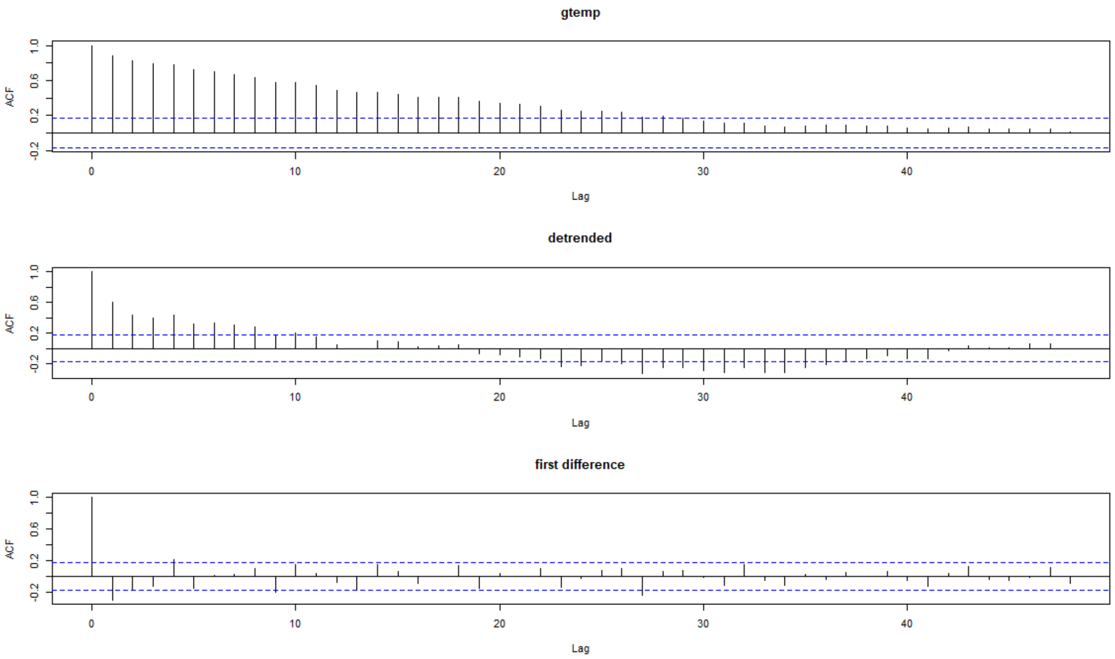


Figure 1.13.2: ACF/Detrending

■ **Example 1.20 — Global temperature.** Based on this comparison, we might be able to say there is mild serial correlation at lag 1 since we take the first difference. ■

1.13.1 Interpretting the ACF: non-stationarity

We have an excellent understanding of how $\hat{\rho}(h)$ behaves when $X_1, \cdots, X_T$ is a strong white noise including consistency and distribution. What happens when we calculate the Empiricial ACF for non-stationary data?

■ **Example 1.21 — ACF for non-stationary time series.** Consider $X_t = t + W_t$ where W_t is a strong white noise. This is clearly not stationary since there is a linear trend t. Consider

$$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} t + W_t = \frac{1}{T}\frac{T(T+1)}{2} + \bar{W} = \frac{T+1}{2} + \bar{W}$$

while

$$\begin{aligned}
\hat{\gamma}(h) &= \frac{1}{T}\sum_{t=1}^{T-h}\left(t + W_t - \frac{T+1}{2} - \bar{W}\right)\left(t + h + W_{t+h} - \frac{T+1}{2} - \bar{W}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T-h}\left(t - \frac{T+1}{2}\right)(t + h - \frac{T+1}{2}) + \text{smaller terms} \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(t - \frac{T+1}{2}\right)^2}_{\approx\mathcal{O}(T^2)} + \underbrace{\frac{1}{T}\sum_{t=1}^{T-h} h\left(t - \frac{T+1}{2}\right)}_{\frac{h}{T}\left[\frac{(T-h)(T-h+1)}{2} - \frac{(T+1)(T-h)}{2}\right]\approx\mathcal{O}(T)}
\end{aligned}$$

Thus, the first part is the dominant term. It follows in this case that

$$\frac{\hat{\gamma}(h)}{T^2} \to C$$

for some constant C for all h when $T \to \infty$. Hence,

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)T^2}{\hat{\gamma}(0)T^2} \xrightarrow{P} 1, \forall h$$

In general, for a time series with a trend in it, the empirical autocorrelation will be big and close to 1. ■

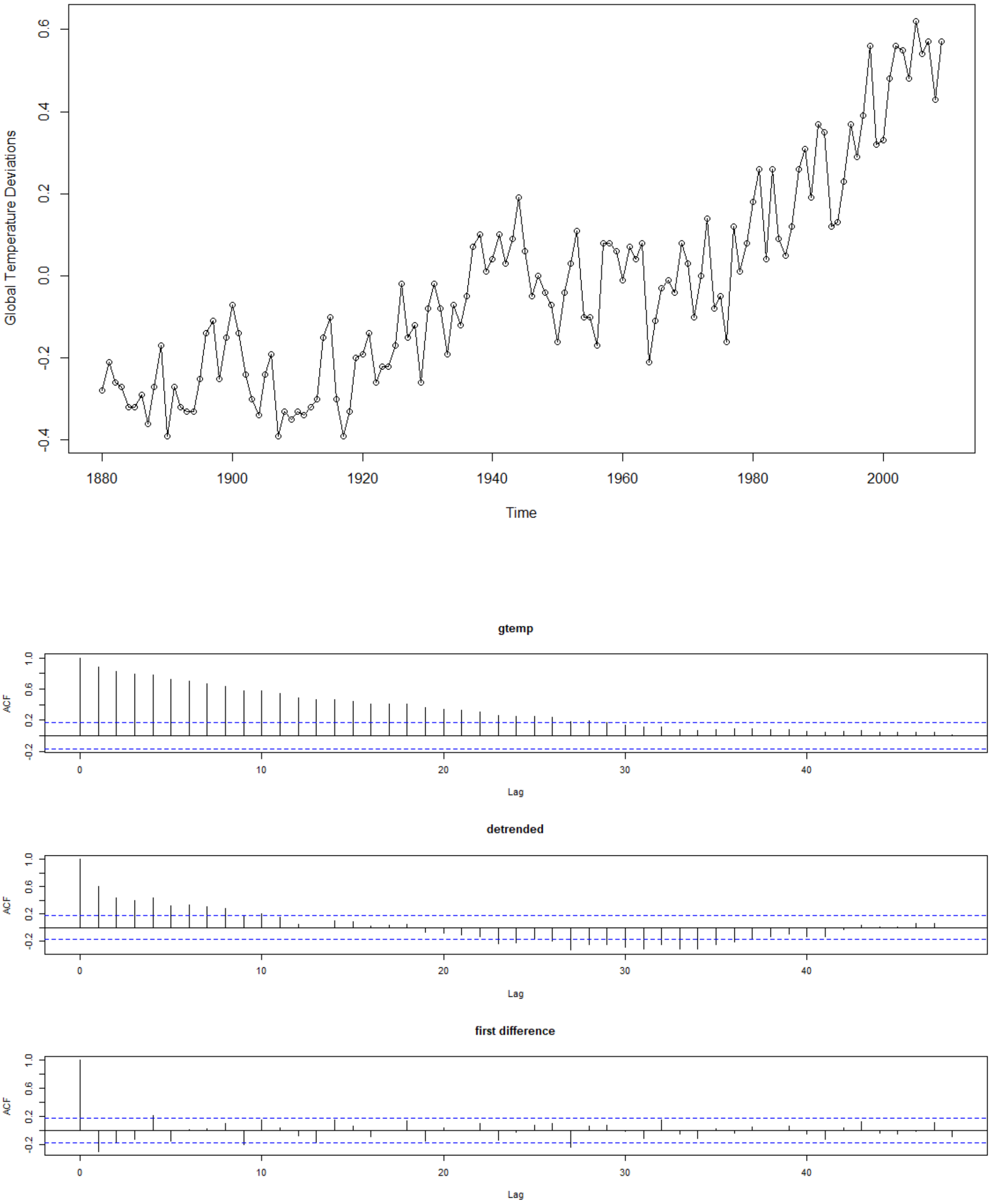


Figure 1.13.3: ACF/Detrending

■ **Example 1.22 — Global temperature.** We see this time series has a positive trend and the ACF on the original data shows large autocorrelation. ■

1.14 Moving Average Processes

Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is stationary. To identify serial dependence using ACF $\hat{\rho}(h)$, we can look for those $\hat{\rho}(h)$ with values outside the $\pm\frac{1.96}{\sqrt{T}}$ blue boundaries. How can we model such a serial dependence? We have seen that most of the processes can be characterized as some types of Bernoulli shift, say $X_t = g(W_t, W_{t-1}, \cdots) = \sum_{l=0}^{\infty}\psi_l W_{t-l}$ (causal linear process) where the parameters are $\{\psi_l\}$. However, it is not feasible to estimate infinitely many parameters. Thus, we need to assume these coefficients arise from a parsimonious, simple and not depending on too many parameter, model.

Definition 1.14.1 — Moving average $MA(q)$. Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong WN with finite variance σ_W. We say $\{X_t\}_{t\in\mathbb{Z}}$ is a moving average process of order q, denoted by $MA(q)$ if there exists coefficients $\theta_1, \cdots, \theta_q \in \mathbb{R}$ with $\theta_q \neq 0$ so that

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q} = \sum_{l=0}^{q}\theta_l W_{t-l}, (\theta_0 = 1)$$

This is like a truncated linear process.

■ **Example 1.23 — Parsimonious time series.** Moving average of order q is an example. ■

Definition 1.14.2 — Backshift operator. The backshift operator, B, is defined by

$$B^j X_t = X_{t-j}$$

B is assumed further to be linear in the sense that for $a, b \in \mathbb{R}$,

$$(aB^i + bB^k) = aB^j X_t + bB^k X_t = aX_{t-j} + bX_{t-k}$$

■ **Example 1.24**

1. First difference: $\nabla X_t = (1-B)X_t$

■

Definition 1.14.3 — Moving average polynomial. We say $\theta(x) = 1 + \theta_1 x + \cdots + \theta_q x^q$ is the moving average polynomial. If $X_t \sim MA(q)$,

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

This provides a succinct way of notation.

Proposition 1.14.1 — Properties of $MA(q)$ Processes.

1. $MA(0)$ is a strong WN
2. If $X_t \sim MA(q)$, then

$$\mathbb{E}[X_t] = \mathbb{E}\left[\sum_{l=0}^{q}\theta_l W_{t-l}\right] = 0$$

and

$$\mathbf{Var}(X_t) = \mathbb{E}\left[\left(\sum_{l=0}^{q}\theta_l W_{t-l}\right)^2\right] = \sum_{l=0}^{q}\theta_l^2\sigma_W^2$$

$$\gamma(h) = \mathbf{Cov}(X_t, X_{t+h}) = \mathbb{E}\left[\left(\sum_{l=0}^{q}\theta_l W_{t-l}\right)\left(\sum_{k=0}^{q}\theta_l k_{t+h-k}\right)\right] \underset{k=l+h}{=} \begin{cases}\sum_{j=0}^{q-|h|}\theta_j\theta_{j+h}\sigma_W^2 & 0 \leq |h| \leq q\\ 0 & |h| > q\end{cases}$$

then,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \frac{\sum_{j=0}^{q-|h|}\theta_j\theta_{j+h}\sigma_W^2}{\sum_{j=0}\theta_j^2\sigma_W^2} & 0 \le |h| \le q \\ 0 & |h| > q \end{cases} \begin{cases} \frac{\sum_{j=0}^{q-|h|}\theta_j\theta_{j+h}}{\sum_{j=0}\theta_j^2} & 0 \le |h| \le q \\ 0 & |h| > q \end{cases}$$

Note that: by choosing $\theta_1, \cdots, \theta_q$ appropriately, we can achieve any ACF we want $\rho(h), 1 \le h \le q$.

3. If $X_t \sim MA(q)$, then X_t is $q-$dependent.

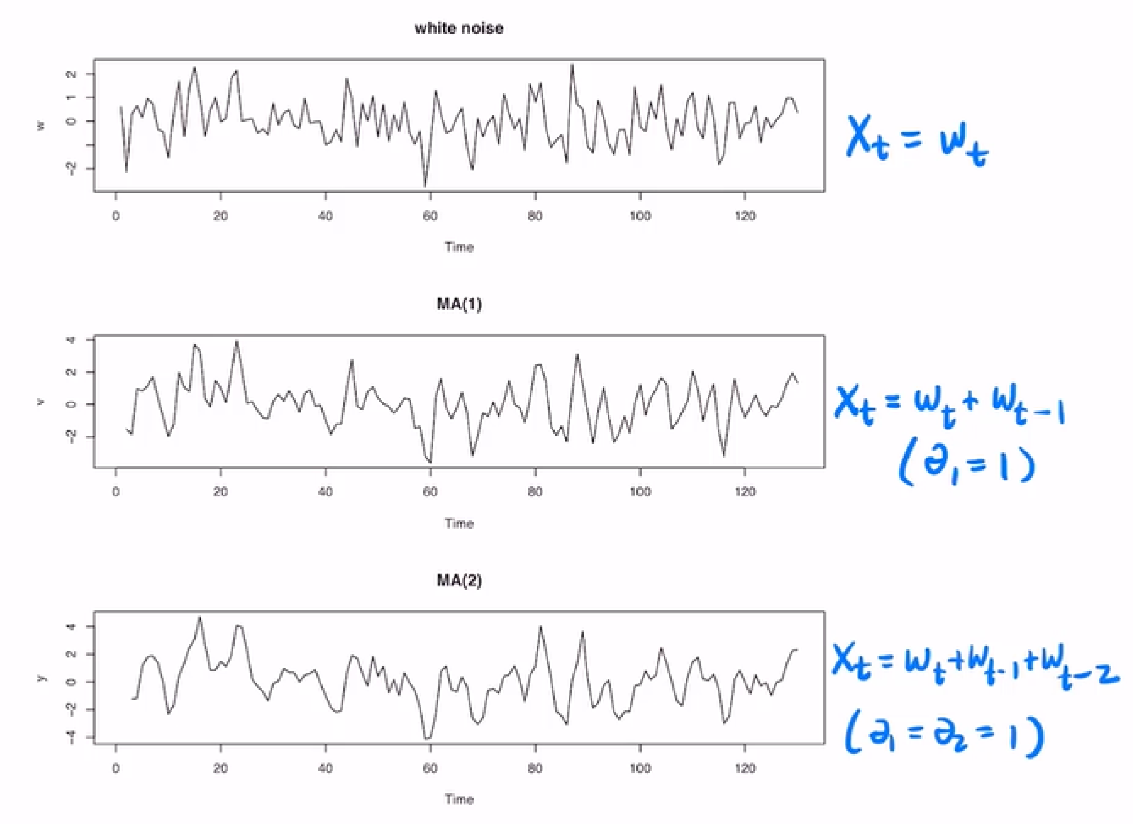


Figure 1.14.1: Realizations of MA processes with coefficients equal to 1, smoother version of strong white noise

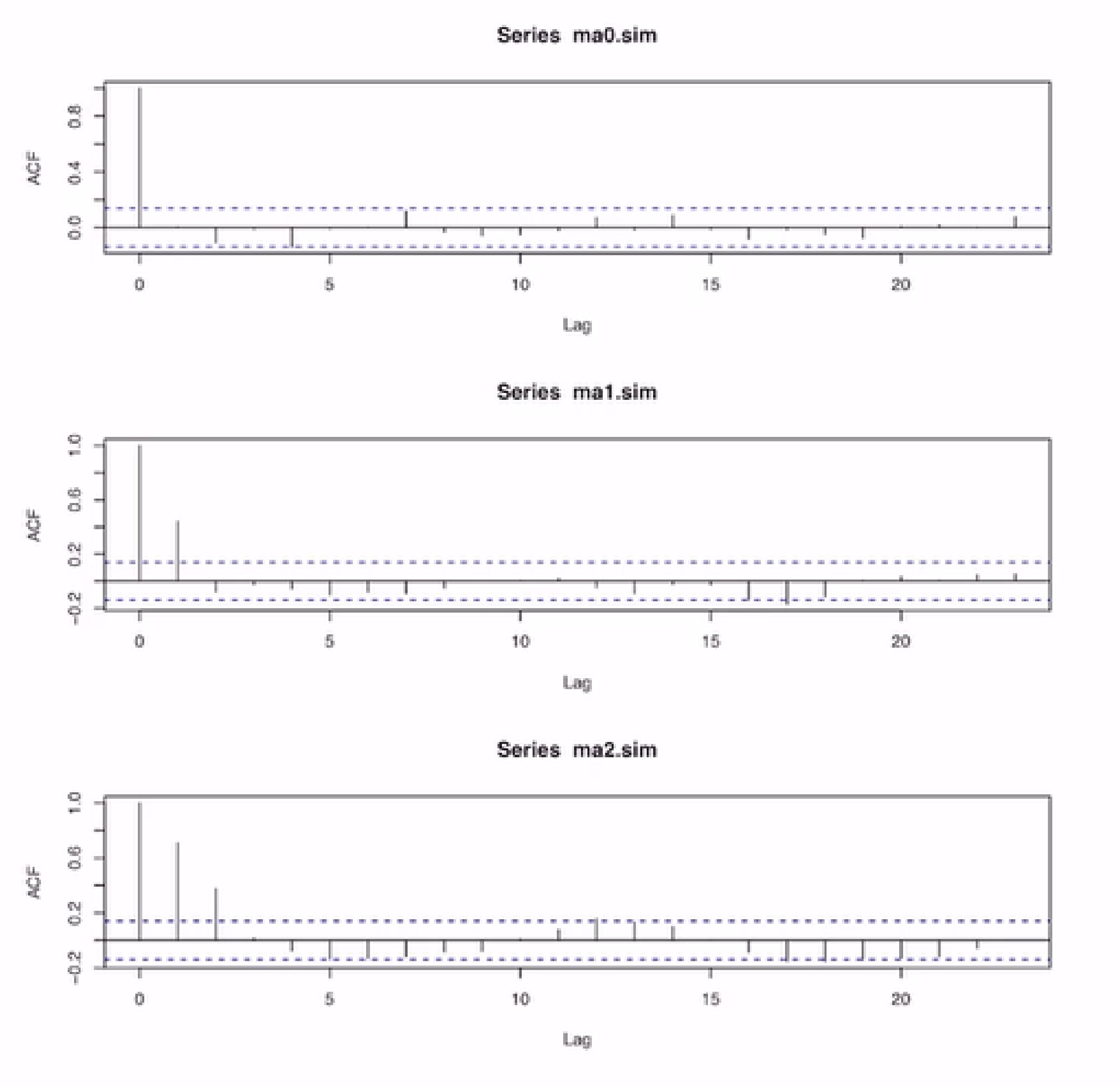


Figure 1.14.2: ACF plots of corresponding moving average series

1.15 Autoregressive Processes

Definition 1.15.1 — Autoregressive process $AR(1)$. Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong WN with finite variance σ_W^2. We say X_t is a autoregressive process of order 1, denoted by $AR(1)$, if there exists a constant ϕ so that

$$X_t = \phi X_{t-1} + W_t, t \in \mathbb{Z}$$

using the backshift operator, this may also be expressed as

$$(1-\phi B)X_t = W_t$$

Interpretation:

1. **Prediction:** form a linear model (regression) for predicting X_t as $X_t = \phi X_{t-1} + W_t$ by using observed covariate/independent variable.
2. **Markovian Property:**

$$X_t|X_{t-1}, X_{t-2}, \cdots = X_t|X_{t-1}$$

One important question is, does there exist a stationary process X_t satisfying

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

Then,

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + W_{t-1}) + W_t \\ &= \phi^2 X_{t-2} + \phi W_{t-1} + W_T \\ &\vdots k \text{ times} \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j W_{t-j} \end{aligned}$$

Note that if $|\phi| > 1$, the sum blows up. Suppose $|\phi| < 1$, then $\phi^k X_{t-k} \to 0$ in the L^2-sense and the second term converges to

$$\sum_{j=0}^{\infty} \phi^j W_{t-j}$$

which is a causal linear process. Moreover, if $X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}$, X_t is strictly stationary, and

$$\begin{aligned} X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j} &= \sum_{j=1}^{\infty} \phi^j W_{t-j} + W_t \\ &= \phi\left(\sum_{j=1}^{\infty} \phi^{j-1} W_{t-j}\right) + W_t \\ &\overset{j \to j-1}{=} \phi\left(\sum_{j=0}^{\infty} \phi^j W_{t-1-j}\right) + W_t = \phi X_{t-1} + W_T \end{aligned}$$

Thus, X_t satisfies $AR(1)$ equation.

Theorem 10 If $|\phi| < 1$, then there exists a strictly stationary and causal linear process X_t so that

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

What if $|\phi| > 1$? If $X_t = \phi X_{t-1} + W_t, t \in \mathbb{Z}$. Then, reversely,

$$X_t = \frac{X_{t+1}}{\phi} - \frac{W_{t+1}}{\phi} = \cdots = \frac{X_{t+k}}{\phi^k} - \sum_{j=1}^{k} \frac{W_{t+j}}{\phi^j}$$

This converges to (in L^2 sense)

$$-\sum_{j=1}^{\infty} \frac{W_{t+j}}{\phi^j}$$

This sequence is strictly stationary since it is a Bernoulli shift. This is not as desirable since it is future dependent. Normally, we try to avoid this.
What if $|\phi| = 1$? In this case, there is no stationary process X_t so that

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

Proof. When $\phi = 1$. If $X_t = X_{t-1} + W_t$, then

$$X_t = \sum_{j=1}^{t} W_j + X_0 \implies X_t - X_0 = \sum_{j=1}^{t} W_j$$

$$\mathbf{Var}\,(X_t - X_0) = \mathbf{Var}\,(X_t) + \mathbf{Var}\,(X_0) - 2\mathbf{Cov}(X_t, X_0) \leq 4\mathbf{Var}\,(X_0)$$

but

$$\mathbf{Var}\left(\sum_{j=1}^{t} W_j\right) = t\sigma_W^2 \xrightarrow{t\to\infty} \infty$$

This yields a contradiction. ■

Now, we have the full picture of the stationary solutions satisfying the autoregressive recursion.

Proposition 1.15.1 — Properties of Causal $AR(1)$ $(|\phi| < 1)$. 1. The span of dependence of X_t is infinite

$$X_t = \sum_{l=0}^{\infty} \phi^l W_{t-l}$$

2. The ACF: note that

$$\mathbf{Var}\,(X_t) = \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \phi^l W_{t-l}\right)^2\right] = \sum_{l=0}^{\infty} \phi^{2l}\sigma_W^2 = \frac{\sigma_W^2}{(1-\phi^2)}$$

similarly,

$$\begin{aligned}\gamma(h) &= \mathbf{Cov}(X_t, X_{t+h}) \\ &= \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \phi^l W_{t-l}\right)\left(\sum_{k=0}^{\infty} \phi^l W_{t+h-k}\right)\right] \\ &\overset{k=l+h}{=} \sum_{l=0}^{\infty} \phi^l \phi^{l+h} \sigma_W^2 \\ &= \phi^h \sum_{l=0}^{\infty} \phi^{2l} \sigma_W^2 \\ &= \phi^h \frac{\sigma_W^2}{1-\phi^2} = \phi^h \mathbf{Var}\,(X_t)\end{aligned}$$

Hence,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, h \geq 0$$

Note: this decays geometrically in the lag parameter.

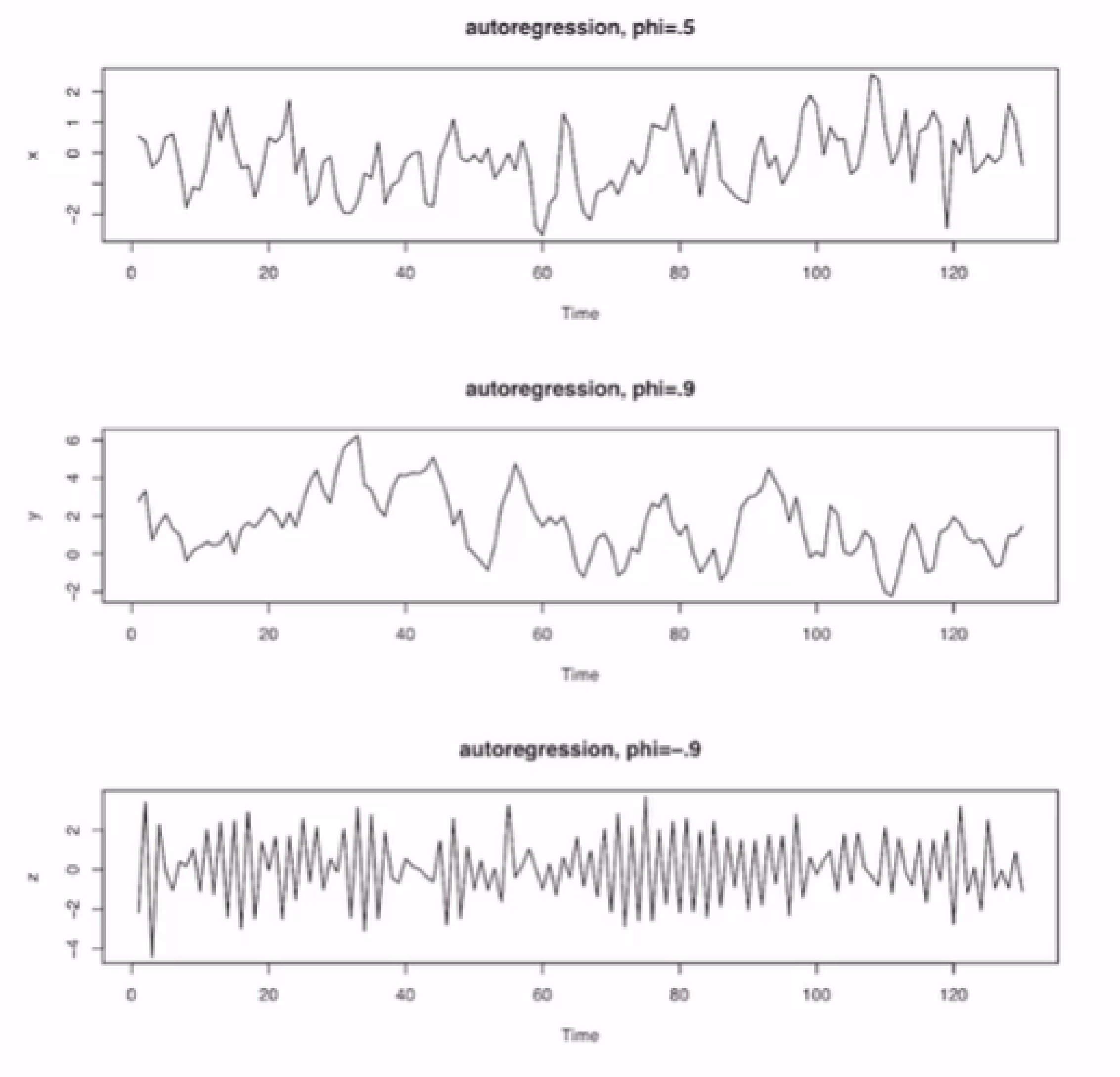


Figure 1.15.1: Realizations of $AR(1)$ processes

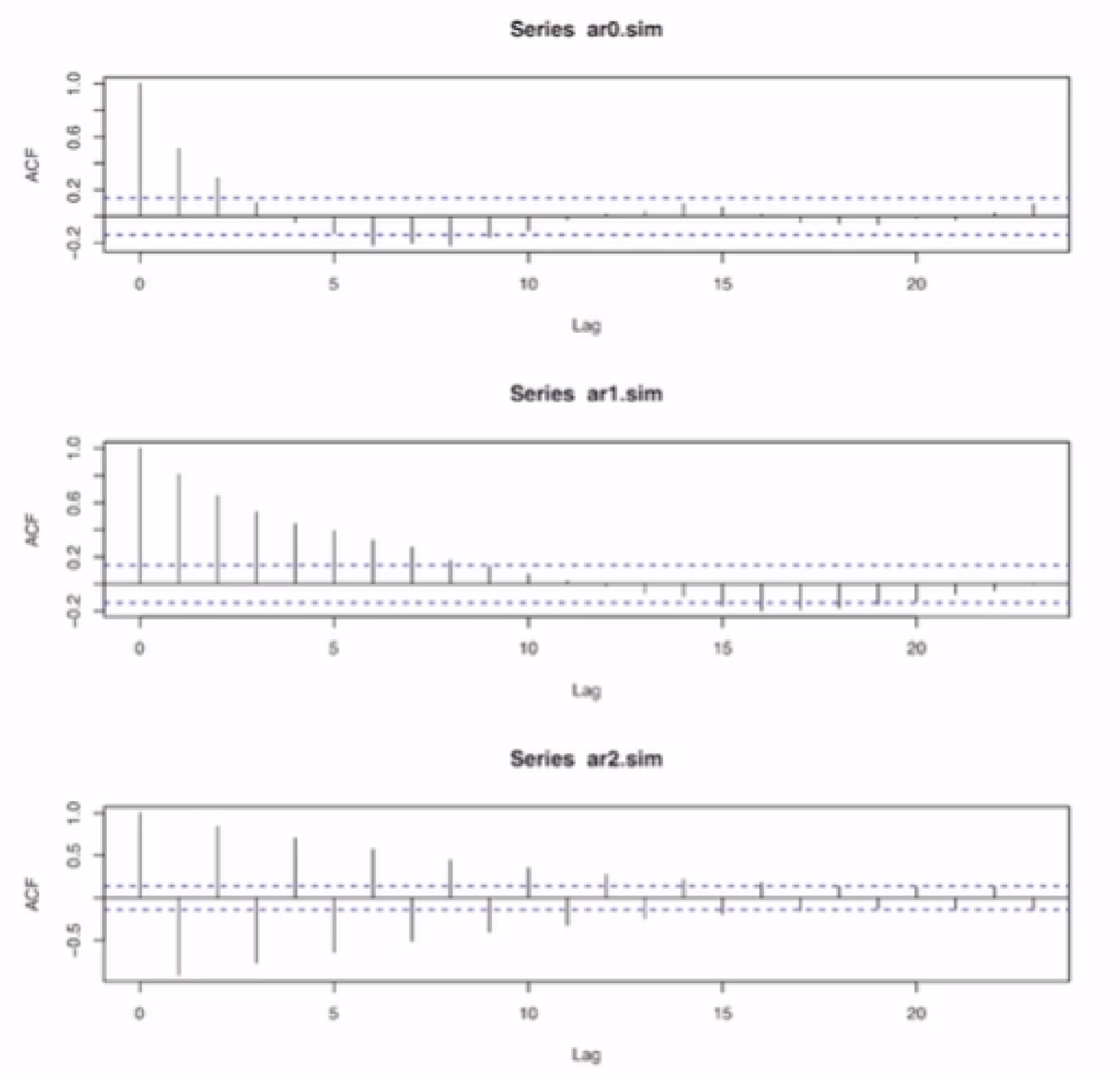


Figure 1.15.2: Corresponding ACF plots

Definition 1.15.2 — $AR(p)$**.** We say X_t follows an autoregressive of order p, denoted by $AR(p)$ if there exists coefficients $\phi_1, \cdots, \phi_p \in \mathbb{R}$ where $\phi_p \neq 0$ so that

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t$$

We define

$$\phi(x) = 1 - \phi_1 - \cdots - \phi_p x^p$$

to be the autoregressive polynomial. By backshift operator,

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

1.16 Autoregressive Moving Average Processes

We have seen the moving average polynomial

$$\theta(x) = 1 + \theta_1 x + \cdots + \theta_q x^q, (\theta_q \neq 0)$$

and autoregressive polynomial

$$\phi(x) = 1 - \phi_1 x - \cdots - \phi_p x^p, (\phi_p \neq 0)$$

Why not combine these two (*MA* and *AR*) together?

I got a MA(q), I got a AR(p). Ummm. ARMA(p,q)!

Definition 1.16.1 — $ARMA(p,q)$. Given a strong WN sequence W_t, we say that X_t is an autoregressive moving average process of orders p,q if

$$\phi(B)X_t = \theta(B)W_t$$

This implies the model

$$X_t = \underbrace{\phi_1 X_{t-1} + \cdots + \phi_p X_{t-p}}_{\text{regression on previous p values}} + \underbrace{W_t + \theta_1 w_{t-1} + \cdots + \theta_q W_{t-q}}_{\text{moving average error}}$$

Using ARMA model to model autocorrelation

Note that $MA(q)$ can be specified by ACF's first q lags. Whereas $AR(p)$ has a geometric decay/oscillations in ACF. ARMA combines these two behaviours.

R **Parameter redundancy** consider $X_t = W_t$ where $X_t \sim MA(0)$. Then,

$$0.5X_{t-1} = 0.5W_{t-1} \implies X_t - 0.5X_{t-1} = W_t - 0.5W_{t-1} \implies X_t \sim ARMA(1,1)$$

In this case, $\phi(z) = 1 - 0.5z$ and the root is $z_0 = 2$ while $\theta(z) = 1 - 0.5z$ has the root $z_0 = 2$ as well. Parameter redundancy manifests as shared zeroes in ϕ and θ. We always assume models are "reduced" by factoring and dividing away common zeroes in $\phi(z), \theta(z)$.

Definition 1.16.2 — Causal. We say an $ARMA(p,q)$ model is causal if there exists X_t satisfying $\phi(B)X_t = \theta(B)W_t$, and

$$X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$$

is a causal linear process solution.

Definition 1.16.3 — Invertible. We say an $ARMA(p,q)$ model is invertible if there exists X_t satisfying $\phi(B)X_t = \theta(B)W_t$ and

$$W_t = \sum_{l=0}^{\infty} \pi_l X_{t-l}$$

W_t can be expressed as a linear function of X_t.

R Causality and invertibility implies the information in $\{X_t\}_{t \leq T}$, observed series, is the same as the information in $\{W_t\}_{t \leq T}$. We can basically go back and forth between X_t and W_t.

Theorem 11 — Causality. By the fundamental theorem of algebra, the autoregressive polynomial $\phi(z)$ has exactly p roots, say $z_1, \cdots, z_p \in \mathbb{C}$. If

$$\rho := \min_{1 \leq j \leq p} \|z_j\| > 1$$

then there exists a stationary and causal time series X_t that satisfies the $ARMA$ equation $\phi(B)X_t = \theta(B)W_t$ and

$$X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$$

the coefficients sequence is absolutely summable. In fact, $|\psi_l| \leq \frac{1}{\rho} l$ shows a geometric decay and

$$\psi(z) = \sum_{l=0}^{\infty} \psi_l z^l = \frac{\theta(z)}{\phi(z)}, |z| \leq 1$$

In essence, $X_t = \frac{\theta(B)}{\phi(B)} W_t = \sum_{j=0}^{\infty} \psi_j B^j W_t$. The key is that

$$\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \phi_j z^j, |z| \leq 1$$

has a convergent power series in the unit disk in the complex plane.

Theorem 12 — Invertibility. If $z_1, \cdots, z_q$ are the zeros of $\theta(z)$, and $\min_{1 \leq i \leq} \|z_i\| > 1$, then X_t is invertible,

$$W_t = \sum_{l=0}^{\infty} \pi_l X_{t-l}$$

coefficients $\{\pi_l\}_{l=0}^{\infty}$ satisfy

$$\pi(z) = \sum_{l=0}^{\infty} \pi_l z^l = \frac{\phi(z)}{\theta(z)}, |z| \leq 1$$

R When we look for coefficients $\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q$, we want to do so in such a way that

$$\phi(z), \theta(z) \neq 0, |z| \leq 1$$

1.16.1 ARMA Process: Examples

■ **Example 1.25 —** $ARMA(2,2)$**?.** Consider the $ARMA(2,2)$ model

$$X_t = \frac{1}{4} X_{t-1} + \frac{1}{8} X_{t-2} + W_t - \frac{5}{6} W_{t-1} + \frac{1}{6} W_{t-2}$$

1. Is there a stationary and causal solution X_t?
2. Is it invertible?
3. Is there parameter redundancy?

The AR polynomial is $\phi(z) = 1 - \frac{1}{4}z - \frac{1}{8}z^2$ while the MA polynomial is $\theta(z) = 1 - \frac{5}{6}z + \frac{1}{6}z^2$. The roots of ϕ are

$$\frac{z \pm \sqrt{4 + 4 \times 8}}{-2} = -1 \pm 3 = -4, 2$$

while the roots of θ are $2, 3$. Then,

$$\phi(z) = -\frac{1}{8}(z+4)(z-2), \qquad \theta(z) = \frac{1}{6}(z-2)(z-3)$$

This suggest parameter redundancy. Thus, X_t satisfies an $ARMA(1,1)$ with $\phi(z) = -\frac{1}{8}(z+4)$ and $\phi(z) = \frac{1}{6}(z-3)$. Since the roots of ϕ and θ are outside of the unit circle in $\mathbb{Z}$. Thus, X_t is stationary causal and invertible.

■

■ **Example 1.26** Suppose

$$X_t = -\frac{1}{4}X_{t-1} + W_t - \frac{1}{3}W_{t-1}$$

$X_t \sim ARMA(1,1)$. Note that $\phi(z) = 1 + \frac{1}{4}z$ and the root is -4. So X_t is stationary and causal, and can be represented as a linear process

$$X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$$

how can we calculate the coefficients ψ_l?
We know that

$$\psi(z) = \sum_{l=0}^{\infty} \psi_l z^l = \frac{\theta(z)}{\phi(z)}, |z| \leq 1 \implies \psi(z)\phi(z) = \theta(z)$$

We can match the coefficients of these power series.

$$\begin{array}{lcc}
z^0 & & \psi_0 = 1 \\
z^1 & & \frac{\psi_0}{4} + \psi_1 = -\frac{1}{3} \implies \psi = -\frac{7}{12} \\
z^2 & & \frac{\psi_1}{4} + \psi_2 = 0 \implies \psi_2 = \frac{-7}{12}\frac{1}{4} \\
\vdots & & \vdots \\
z^l & \underbrace{\frac{\psi_{l-1}}{4} + \psi_l}_{\text{Finite liner difference equation}} & = 0 \implies \psi_l = \frac{-7}{12}\left(\frac{1}{4}\right)^{l-1}
\end{array}$$

Automated in the `ARMAtoMA` function in `R`. ■

If X_t is a stationary and causal solution to the $ARMA(p,q)$ model

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

and

$$\gamma_X(h) = \mathbb{E}[X_t X_{t+h}] = \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \psi_j W_{t-j}\right)\left(\sum_{k=0}^{\infty} \psi_k W_{t+h-k}\right)\right] = \sum_{j=0}^{\infty} \underbrace{\psi_j \psi_{j+h}}_{\text{can be solved as before}} \sigma_W^2$$

This has been automated by the `ARMAacf` function in `R`.

2. Forecasting

2.1 Forecasting Stationary Processes Using L^2-Loss

Suppose that we have an observed time series $X_1, \cdots, X_T$ that we believe it has been generated by an underlying stationary process. We would like to produce an $h-$step ahead forecast

$$\hat{X}_{T+h} = \hat{X}_{T+h|T} = f(X_T, \cdots, X_1)$$

Ideally, $\hat{X}_{T+h}$ would minimize the prediction error

$$L(X_{T+h}, \hat{X}_{T+h}) = \min_f L(X_{T+h}, f(X_T, \cdots, X_1))$$

where L is a loss function.
Frequently, the loss function is taken to be the Mean-squared error (**MSE**),

$$\mathbf{MSE}(X_{T+h}, \hat{X}_{T+h}) = L(X_{T+h}, \hat{X}_{T+h}) = \mathbb{E}\left[(X_{T+h} - \hat{X}_{T+h})^2\right]$$

when using **MSE**, it is natural to consider

$$L^2 = \left\{\text{random variables } X : \mathbb{E}[X^2] < \infty\right\}$$

It is known that L^2 is a Hilbert space equipped with the inner product

$$\langle X, Y \rangle = \mathbb{E}[XY]$$

Hilbert spaces can be considered as generalizations of the Euclidean spaces $(\mathbb{R}^d)$ in which the geometry and notion of projection are preserved. For example,

$$\mathbf{Proj}(X \to Y) = \langle X, Y \rangle Y$$

Definition 2.1.1 — Closed linear subspace of L^2. We say $M \subseteq L^2$ is a closed linear subspace, if it satisfies:

1. **Linearity:** $X, Y \in M$, $\alpha, \beta \in \mathbb{R}$, then $\alpha X + \beta Y \in M$
2. **Closed:** if $X_n \to X$ (in the sense that $\mathbf{MSE}(X_n, X) \to 0$, and $X_n \in M$, then $X \in M$.

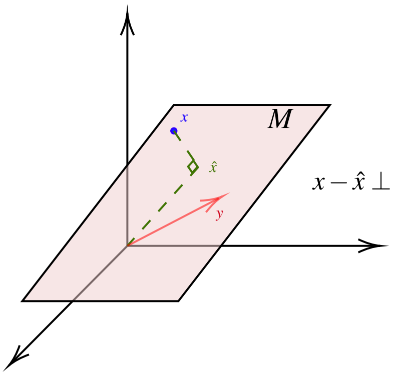
Theorem 13 — Projection theorem. Let $M \subseteq L^2$ be a closed linear subspace and $x \in L^2$, then there exists a unique $\hat{X} \in M$ so that

$$\mathbb{E}[(X-\hat{X})^2] = \inf_{y \in M} \mathbb{E}[(X-Y)^2]$$

Moreover, such optimal $\hat{X}$ must satisfies the following corollary:

Corollary 2.1.1 — Prediction Equations/Normal Equations. $X - \hat{X} \in M^{\perp} \Longrightarrow \mathbb{E}[(X-\hat{X})Y] = 0, \forall y \in M$

This provides us a systematic way to find the $\hat{X}$ since we can have a set of testing Y and solve for the system of expectations.



In MSE forecasting, we want to choose $\hat{X}_{T+h}$ such that

$$\mathbb{E}[(X_{T+h} - \hat{X}_{T+h})^2] = \inf_{y \in M} \mathbb{E}[(X_{T+h} - y)^2]$$

where M is a closed linear subspace based on the available data. So, it is really about the M that we consider. Two candidates:

1. $M_1 = \{Z : Z = f(X_T, \cdots, X_1), f$ is any Borel measurable function$\}$. In this case,
$$\hat{X}_{T+h} = \mathbb{E}[X_{T+h} | X_T, \cdots, X_1]$$
Unfortunately, M_1 is enormous and complicated! (Guess the cardinality)
2. $M_2 = \overline{\text{span}}\{1, X_T, \cdots, X_1\}$, which is the space of all the linear functions of $X_1, \cdots, X_T$ including its closure. $\hat{X}_{T+h}$ is called the **Best Linear Prediction** (BLP).

2.1.1 Best Linear Prediction

Suppose X_t is a (weakly) stationary time series. The best linear prediction details finding $\hat{X}_{T+h}$ so that

$$\mathbb{E}\left[(X_{T+h} - \hat{X}_{T+h})^2\right] = \inf_{y \in M_2} \mathbb{E}[(X_{T+h} - Y)^2]$$

where $M_2 = \overline{\text{span}}\{1, X_T, \cdots, X_1\}$. By the projection theorem, $\hat{X}_{T+h}$ exists uniquely and it is the best predictor among M_2.

Definition 2.1.2 — Projection onto. If $\hat{X}$ satisfies

$$\mathbb{E}[(X-\hat{X})^2] = \inf_{y \in M} \mathbb{E}[(X-Y)^2]$$

we say that $\hat{X}$ is the projection of X onto M and write

$$\hat{X} = \mathbf{Proj}(X|M)$$

In particular, the BLP is

$$\hat{X}_{T+h} = \mathbf{Proj}(X_{T+h}|M_2)$$

Now, we want to find the actual form of the BLP. Consider the case when $h=1$.The BLP is of the form

$$\hat{X}_{T+1} = \phi_{T,0} + \sum_{j=1}^{T} \phi_{T,j} X_j \cong \phi_{T,0} + \sum_{j=0}^{T} \phi_{T,j}(X_j - \mu)$$

where $\mu = \mathbb{E}[X_t]$. By the projection theorem, $\hat{X}_{T+1}$ must satisfy the prediction equations.

$$\mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})Y\right] = 0, \forall Y \in M_2$$

in particular,

$$\begin{aligned} \mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})1\right] &= 0 & [Y=1] \\ \mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})X_j\right] &= 0 & 1 \le j \le T, [Y = X_j] \end{aligned}$$

since $\mathbb{E}[X_j - \mu] = 0$. Then,

$$0 = \mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})1\right] = \mu - \phi_{T,0} \implies \phi_{T,0} = \mu$$

Before proceeding, note that this implies

$$\mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})X_j\right] = \mathbb{E}\left[(X_{T+1} - \mu - (\hat{X}_{T+1} - \mu))(X_j - \mu)\right]$$

so, we could assume WLOG $\mu = 0$ by this shifting. Thus, $\mathbb{E}[X_i X_j] = \gamma(|j-i|)$. Therefore,

$$\begin{aligned} 0 = \mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})X_j\right] &= \gamma(|T+1-k|) - \sum_{j=1}^{T} \phi_{T,j}\gamma(|j-k|) \\ \implies \sum_{j=1}^{T} \phi_{T,j}\gamma(|j-k|) &= \gamma(|T+1-k|) \end{aligned}$$

this defines a linear system of equations for $\phi_{T,1}, \cdots, \phi_{T,T}$. If

$$\gamma_T = \begin{bmatrix} \gamma(T) \\ \vdots \\ \gamma(1) \end{bmatrix} \in \mathbb{R}^T, \qquad \Gamma_T = [\gamma(j-k) : 1 \le j,k \le T] \in \mathbb{R}^{T \times T}$$

$\phi_T = (\phi_{T,1}, \cdots, \phi_{T,T})^\top \in \mathbb{R}^T$. This linear system may be expressed as

$$\Gamma_T \phi_T = \gamma_T \underset{\Gamma_T \text{ invertible}}{\implies} \phi_T = \Gamma_T^{-1} \gamma_T$$

the BLP is then of the form

$$\hat{X}_{T+1} = \phi_T^\top \vec{X}_T = \left(\Gamma_T^{-1}\gamma_T\right)^\top \vec{X}_T$$

where $\vec{X}_T = (X_T, \cdots, X_1)^\top$. You see, this Γ_T needs to be non-singular.

Theorem 14 If $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, then Γ_T is non-singular.

Most stationary processes, those whose serial dependence decays over time, have non-singular Γ_T, which will be the common ones that we deal with.

Note that

$$\hat{X}_{T+1}^2 = \gamma_T^\top \Gamma_T^{-1} \underbrace{\vec{X}_T \vec{X}_T^\top}_{\mathbb{E}[\vec{X}_T \vec{X}_T^\top] = \Gamma_T} \Gamma_T^{-1} \gamma_T$$

thus,

$$\mathbb{E}\left[\hat{X}_{T+1}^2\right] = \gamma_T^\top \Gamma_T^{-1} \gamma_T$$

also, since $\mathbb{E}[X_{T+1} \vec{X}_T] = \gamma_T$, which implies

$$\mathbb{E}\left[\hat{X}_{T+1} \hat{X}_{T+1}\right] = \gamma_T^\top \Gamma_T^{-1} \gamma_T$$

It follows that the Mean-squared prediction error is

$$\begin{aligned} P_{T+1}^T &:= \mathbb{E}\left[(X_{T+1} - \hat{X}_{T+1})^2\right] \\ &= \mathbb{E}\left[X_{T+1}^2 - 2X_{T+1}\hat{X}_{T+1} + \hat{X}_{T+1}^2\right] \\ &= \gamma(0) - 2\gamma_T \Gamma_T^{-1} \gamma_T + \gamma_T \Gamma_T^{-1} \gamma_T \\ &= \gamma(0) - \gamma_T \Gamma_T^{-1} \gamma_T \end{aligned}$$

The mean squared prediction error has a simple and computable form depending on $\gamma(h), 1 \leq h \leq T$.

2.2 Partial ACF

Suppose $X_t \sim ARMA(p,q)$, we might be able to identify p,q by looking at the ACF.

1. If $X_t \sim AR(p)$, then ACF has a geometric decay
2. If $X_t \sim MA(q)$, then ACF is non-zero at first q lags, then zero beyond

An ACF of an $ARMA(p,q)$ model can be calculated by calculating the linear process coefficients $\{\psi_l\}_{l=0}^{\infty}$, which is automated by `ARMAacf` in `R`.

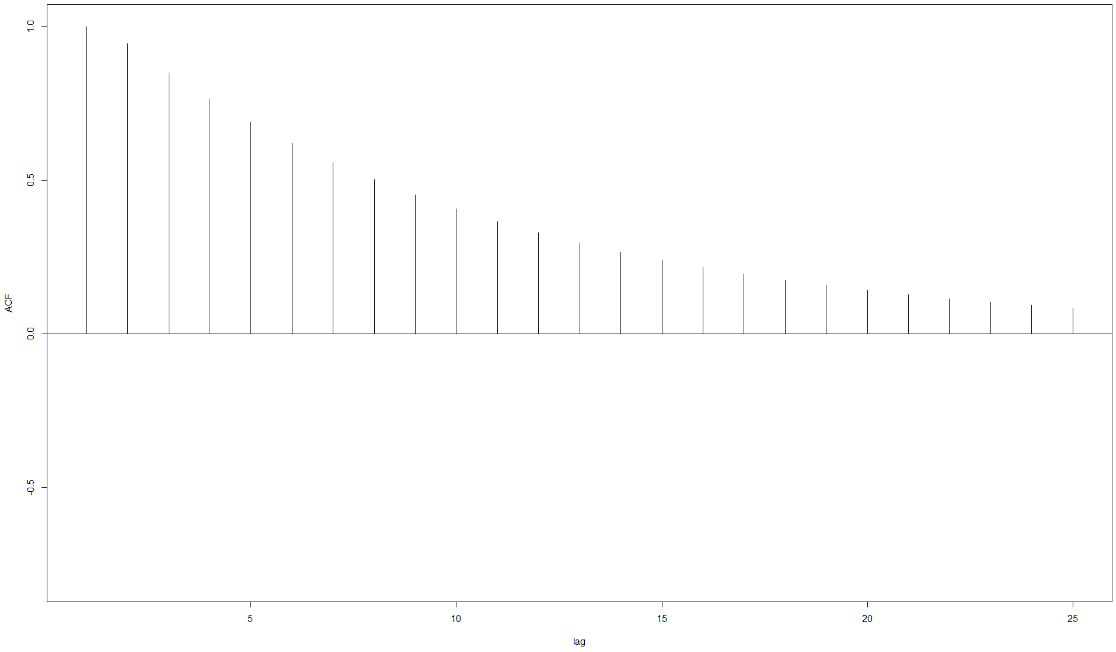


Figure 2.2.1: ACF of $ARMA(1,1): x_t = 0.9x_{t-1} + w_t + 0.5w_{t-1}$

Note that this $ARMA(1,1)$ has a geometric decay pattern, which is hard to distinguish from a AR model. This suggests an alternative.

Definition 2.2.1 — Partial ACF (PACF). The partial ACF of a stationary process $\{X_t\}_{t\in\mathbb{Z}}$ is

$$\phi_{h,h} = \mathbf{Corr}(X_{t+h} - \mathbf{Proj}(X_{t+h}|X_{t+h-1},\cdots,X_{t+1}), X_t - \mathbf{Proj}(X_{t+h}|X_{t+h-1},\cdots,X_{t+1}))$$

The normal interpretation of PACF is the autocorrelation between X_t and X_{t+h} after removing the linear dependence on the intervening variables $X_{t+h-1},\cdots,X_{t+1}$.

R If $X_t \sim AR(p)$ (causal), then

$$\phi_{h,h} = 0, \forall h \geq p+1$$

Proof. If $X_t \sim AR(p)$, then

$$X_{t+h} = \sum_{j=1}^{p} \phi_j X_{t+h-j} + W_{t+h}$$

$$\mathbf{Proj}(X_{t+h}|X_{t+h-1},\cdots,X_{t+1}) = \sum_{k=1}^{h-1} \beta_k X_{t+h-k}$$

and minimizes

$$\begin{aligned}
&\mathbb{E}\left[\left(X_{t+h} - \sum_{k=1}^{h-1}\beta_k X_{t+h-k}\right)^2\right] \\
&= \mathbb{E}\left[\left(W_{t+h} + \sum_{j=1}^{p}\phi_j X_{t+h-j} - \sum_{k=1}^{h-1}\beta_k X_{t+h-k}\right)^2\right] \\
&\overset{\text{independence}}{=} \sigma_W^2 + \mathbb{E}\left[\left(\sum_{j=1}^{p}\phi_j X_{t+h-j} - \sum_{k=1}^{h-1}\beta_k X_{t+h-k}\right)^2\right]
\end{aligned}$$

which is minimized by setting $\beta_j = \phi_j, 1 \leq j \leq p$ and $\beta_j = 0$ for $j \geq p+1$. Hence,

$$X_{t+h} - \mathbf{Proj}(X_{t+h}|X_{t+h-1}, \cdots, X_{t+1}) = W_{t+h}, \qquad (h \geq p+1)$$

this implies,

$$\phi(h,h) = \mathbf{Corr}(W_{t+h}, X_t - \mathbf{Proj}(X_{t+h}|X_{t+h-1}, \cdots, X_{t+1})) = 0$$

due to independence by causality. ■

It can be shown that if $X_t \sim MA(q)$ (invertible), then

$$\phi_{h,h} \neq 0, |\phi_{h,h}| = \mathcal{O}(r^h), 0 < r < 1$$

with a geometric decay.

We summarize our findings into a reference table below.

	ACF	PACF
MA(q)	Cuts off after q	Geometric decay
AR(p)	Geometric decay	Cuts off after p

Estimating the PACF

Using the BLP theory,

$$\hat{\phi}_{h,h} = \left(\hat{\Gamma}_h^{-1}\hat{\gamma}_h\right)[h]$$

(this notation is very CS-triggering)
where

$$\hat{\Gamma}_h = [\hat{\gamma}(j-k), 1 \leq i, k \leq h] \in \mathbb{R}^{h \times h}$$

and

$$\hat{\gamma}_h = [\hat{\gamma}(1), \cdots, \hat{\gamma}(h)] \in \mathbb{R}^h$$

2.3 Forecasting Causal & Invertible ARMA Processes

Suppose X_t follows a causal & invertible $ARMA(p,q)$ model so that $\phi(B)X_t = \theta(B)W_t$. Having observed $X_T, \cdots, X_1$, we wish to predict X_{T+h}

$$\hat{X}_{T+h} = \mathbf{Proj}(X_{T+h}|M_2) \approx \mathbb{E}[X_{T+h}|X_T, \cdots, X_1]$$

this similarity is due to causality and invertability of X_t when following some linear function of W_t. Further, $\hat{X}_{T+h} \approx \tilde{X}_{T+h} = \mathbb{E}[X_{T+h}|X_T, \cdots, X_1, X_0, \cdots]$. This is true due to the geometric decay of the dependence on past values. We shall calculate $X_{\widetilde{T+h}}$. Since X_t is causal and invertible,

$$X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}, \qquad W_t = \sum_{l=0}^{\infty} \pi_l X_{t-l}$$

where $\psi_0 = \pi_0 = 1$. Note ψ_l and π_l are computable by solving homogeneous linear difference equations. These representations imply that the information in the observed times series of length T is just as much as the information in the first T WN. So,

$$\tilde{X}_{T+h} = \mathbb{E}[X_{T+h}|X_T, \cdots, X_1, X_0, \cdots] = \mathbb{E}[X_{T+h}|W_T, \cdots, W_1, W_0, \cdots]$$

1.
$$\begin{aligned}\tilde{X}_{T+h} &= \mathbb{E}\left[\sum_{l=0}^{\infty}\psi_l W_{T+h-l}|W_T,\cdots,W_1,W_0,\cdots\right]\\ &= \mathbb{E}\left[\sum_{l=0}^{h-1}\underbrace{\psi_l W_{T+h-l}}_{=0}|W_T,\cdots\right]+\mathbb{E}\left[\sum_{l=h}^{\infty}\psi_l W_{T+h-l}|W_T,\cdots\right]\\ &= \sum_{l=h}^{\infty}\psi_l W_{T+h-l}\end{aligned}$$

2. Using invertiblity,
$$\begin{aligned}0=\mathbb{E}[W_{T+h}|X_T,\cdots] &= \mathbb{E}\left[\sum_{l=0}^{\infty}\pi_l X_{T+h-l}|X_T\right]\\ &\overset{\pi_0=1}{=} \tilde{X}_{T+h}+\sum_{l=1}^{h-1}\pi_l\tilde{X}_{T+h-l}+\sum_{l=h}^{\infty}\pi_l X_{T+h-l}\end{aligned}$$

   thus,
$$\tilde{X}_{T+h}=-\sum_{l=1}^{h-1}\pi_l\tilde{X}_{T+h-l}-\sum_{l=h}^{\infty}\pi_l X_{T+h-l}$$

Truncated ARMA Prediction

this is given by
$$\hat{X}_{T+h}=-\sum_{l=1}^{h-1}\pi_l\hat{X}_{T+h-l}-\sum_{l=h}^{T+h-1}\pi_l X_{T+h-l}$$
which is truncated to observed values. This can be computed recursively. We can also estimate the residuals:
$$\hat{\omega}_t=\phi(B)\hat{X}_t-\theta_1\hat{\omega}_{t-1}-\cdots\theta_q\hat{\omega}_{t-q}$$
we conduct a mean initialization, $\hat{\omega}_t=0, t\le 0, t\ge T$ and $\hat{X}_t=0, t\le 0, t\ge T+1, \hat{X}_t=X_t, 1\le t\le T$. Estimator for σ_W^2 is given by
$$\hat{\sigma}_W^2=\frac{1}{T}\sum_{t=1}^{T}\hat{\omega}_t^2$$

Mean-squared prediction error

Since $\hat{X}_{T+h}\approx\sum_{j=h}^{\infty}\psi_j W_{t-j}$, then
$$P_{T+h}^T=\mathbb{E}\left[\left(X_{T+h}-\hat{X}_{T+h}\right)^2\right]=\mathbb{E}\left[\left(\sum_{j=0}^{h-1}\psi_j W_{t-j}\right)^2\right]=\sigma_W^2\sum_{j=0}^{h-1}\psi_j^2$$
The estimated MSPE is
$$\hat{p}_{T+h}^T=\hat{\sigma}_W^2\sum_{j=0}^{h-1}\psi_j^2$$
Nice, we can construct some prediction intervals for our forecasts. Since $\hat{X}_{T+h}\approx\mathbb{E}[X_{T+h}|X_T,\cdots]$, we have $\mathbb{E}[\hat{X}_{T+h}-X_{T+h}]=0$ by the Tower property and the variance is then $\mathbb{E}\left[\left(\hat{X}_{T+h}-X_{T+h}\right)^2\right]=P_{T+h}^T$. Then, $\frac{\hat{X}_{T+h}-X_{T+h}}{\sqrt{\hat{P}_{T+h}^T}}$ is an approximate random variable with mean 0 and unit variance. If we apply the naive symmetric assumption, then the $1-\alpha$ prediction interval for X_{T+h} is given by
$$\hat{X}_{T+h}\pm c_{\alpha/2}\sqrt{\hat{P}_{T+h}^T}$$
There are some choices for $c_{\alpha/2}$:

1. Standard normal critical value: the motivation is that if W_t is Gaussian, then $X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$ is Gaussian
2. Empirical critical value of residuals (standardized):
$$\frac{\hat{\omega}_t}{\sigma_W}, 1 \leq t \leq T$$
3. Student-t distribution, Pareto, or skewed distributions fitted to standardized residuals.

Long Range Behaviour of ARMA forecasts

Suppose $Y_t = S_t + X_t, X_t \sim ARMA(p,q)$.

$$\hat{Y}_{T+h} = \hat{S}_{T+h} + \hat{X}_{T+h} = \hat{S}_{T+h} + \underbrace{\sum_{j=h}^{\infty} \psi_j W_{T+h-j}}_{\to 0 \text{ geometrically!}}$$

Thus, in the long run, $\hat{Y}_{T+h}$ is converging fast to $\hat{S}_{T+h}$. Thus, we better get the trend for long range forecasts. Moreover,

$$P_{T+h}^{T} = \hat{\sigma}_W^2 \sum_{j=0}^{h-1} \psi_j^2 \to \hat{\sigma}_W^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma_X(0)$$

In the long run, the MSE is the variance of X_t.

■ **Example 2.1 — ARMA Forecasting.** Let's try ARMA forecasting on a real dataset.

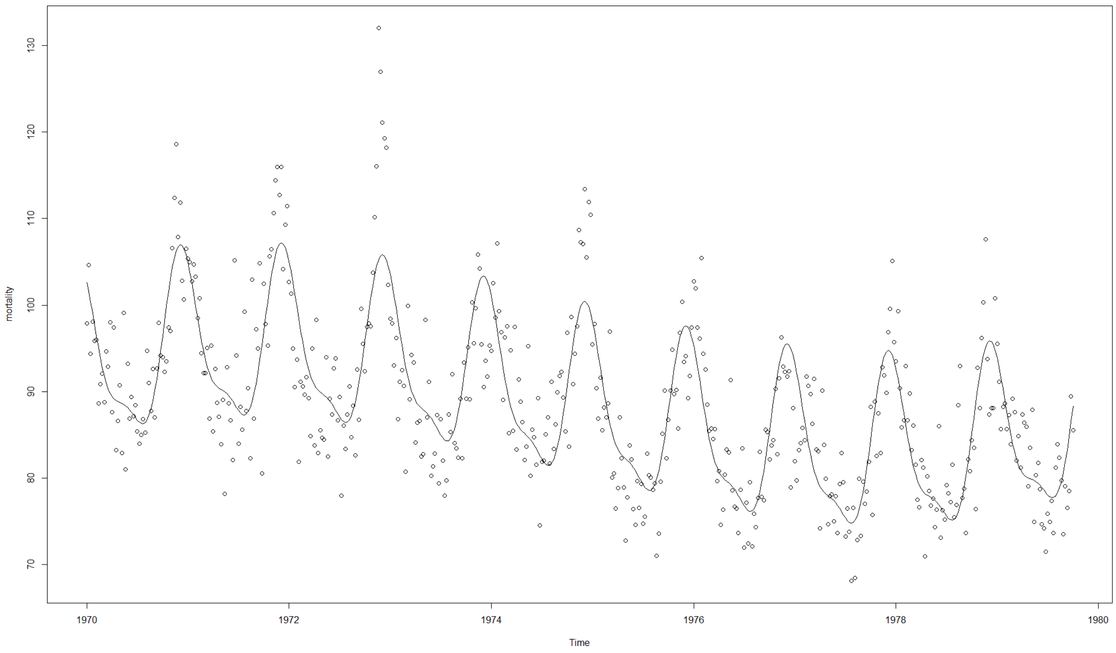


Figure 2.3.1: Weekly cardiovascular mortality

Let X_t the time series and model it as

$$X_t = S_t + Y_t$$

where $Y_t \sim ARMA(p,q)$ process. S_t is the seasonal and polynomial trend.

$$S_t = \underbrace{\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3}_{\text{polynomial}} + \underbrace{\beta_4 \sin\left(\frac{2\pi}{52}t\right) + \beta_5 \cos\left(\frac{2\pi}{52}t\right)}_{\text{yearly cycle}} + \underbrace{\beta_6 \sin\left(\frac{2\pi}{26}t\right) + \beta_7 \cos\left(\frac{2\pi}{26}t\right)}_{\text{half-yearly cycle}}$$

Decided on this trend using AIC, which will be discussed later. To check whether this model fits well, we have

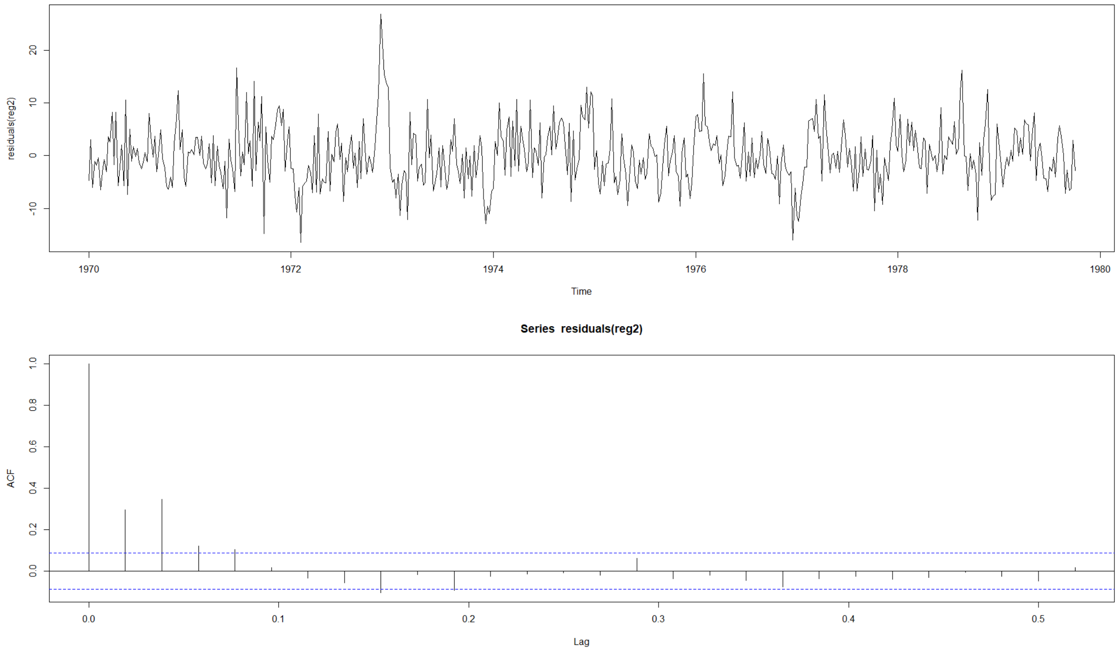


Figure 2.3.2: Residual plot and ACF

We note that the residual is reasonably around 0 and ACF signals stationarity. ■

2.4 Estimation of $ARMA(p,q)$ Parameters

2.4.1 AR Case

Suppose we observe a time series $X_1,\cdots,X_T \sim ARMA(p,q)$ where

$$\phi(B)X_t = \theta(B)W - T$$

Our goal is to estimate $\phi_1,\cdots,\phi_p$ (AR parameters) and $\theta_1,\cdots,\theta_q$ (MA parameters), and σ_W^2 (the white noise variance).

AR(1) Case:
consider $X_t = \phi X_{t-1} + W_t$ and $\mathbb{E}[W_t^2] = \sigma_W^2$. The idea is to use OLS such that

$$\hat{\phi} = \arg\min_{|\phi|<1} \sum_{t=2}^{T} (X_t - \phi X_{t-1})^2$$

This leads to upon some calculus:

$$\hat{\phi} = \frac{\sum_{t=2}^{T} X_t X_{t-1}}{\sum_{t=2}^{T} X_t^2} \approx \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \hat{\rho}(1) \xrightarrow{p,T\to\infty} \phi$$

similarly,

$$\hat{\sigma}_W^2 = \frac{1}{T-1} \sum_{t=2}^{T} (X_t - \hat{\phi} X_{t-1})^2$$

which is the sample variance of the residuals.

General AR(p)

consider $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t$. Again, by OLS,

$$\vec{\phi} = (\phi_1, \cdots, \phi_p)^\top \in \mathbb{R}^p$$

$$\hat{\vec{\phi}} = \arg \min_{\substack{\vec{\phi}: X_t \text{ admits} \\ \text{a stationary and} \\ \text{causal solution}}} \sum_{t=p+1}^{T} (X_t - \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t)^2$$

solve using calculus, this leads to a system of p linear equations of the form

$$\hat{\Gamma}_p \hat{\vec{\phi}} = \hat{\vec{\gamma}}_p,\ \hat{\Gamma}_p = (\hat{\gamma}(i-j) : 1 \le i,k \le p) \in \mathbb{R}^{p\times p}, \hat{\vec{\gamma}}_p = (\hat{\gamma}(1), \cdots, \hat{\gamma}(p))^\top$$

The resulting OLS estimator takes the approximate form:

$$\hat{\vec{\phi}} = \hat{\Gamma}_p^{-1} \hat{\vec{\gamma}}_p,\ \hat{\sigma}_W^2 = \hat{\gamma}(0) - \hat{\vec{\gamma}}_p^\top \hat{\Gamma}_p^{-1} \hat{\vec{\gamma}}_p$$

Method of Moments Set parameters so that empirical moments match theoretical moments induced by the model. If $X_t \sim AR(p)$, then for all $1 \le h \le p$ lags,

$$\begin{aligned} \gamma(h) &= \mathbb{E}[X_t X_{t+h}] = \mathbb{E}\big[X_t(\phi_1 X_{t+h-1} + \cdots + \phi_p X_{t+h-p} + W_{t+h})\big] \\ &= \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p) + \underbrace{0}_{\text{causality}: X_t \perp W_{t+h}} \end{aligned}$$

This implies the linear system:

$$\vec{\gamma}_p = \Gamma_p \vec{\phi}$$

Note that

$$X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}, \psi_0 = 1$$

and $W_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$. Since $\psi_0 = 1$, we actually have

$$\sigma_W^2 = \mathbb{E}[X_t W_t] = \gamma(0) - \phi_1 \gamma(1) - \cdots - \phi_p \gamma(p)$$

this provides one more linear equation. Adding this to the previous system, we have

Theorem 15 — Yule-Walker Equations.

$$\begin{cases} \sigma_W^2 = \gamma(0) - \phi_1 \gamma(1) - \cdots - \phi_p \gamma(p) \\ \vec{\gamma}_p = \Gamma_p \vec{\phi} \end{cases}$$

Resulting Yule-Walker estimators:

$$\hat{\vec{\phi}} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p,\ \hat{\sigma}_W^2 = \hat{\gamma}(0) - \hat{\vec{\gamma}}_p^\top \hat{\Gamma}_p^{-1} \hat{\gamma}_p$$

■ **Example 2.2 — AR(1).** The YW estimators are

$$\hat{\phi} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \hat{\rho}(1),\ \hat{\sigma}_W^2 = \hat{\gamma}(0) - \frac{\hat{\gamma}^2(1)}{\hat{\gamma}(0)}$$

■

Theorem 16 — Asymptotic equivalence between OLS and YW. If $X_t \sim AR(p)$ is causal, then

$$\frac{\hat{\phi}_{OLS,i}}{\hat{\phi}_{YW,i}} \xrightarrow{p,T\to\infty} 1$$

OLS and YW estimates are asymptotically equivalent.

Theorem 17 — CLT for YW and OLS.

$$\sqrt{T}\left(\hat{\vec{\phi}}_{YW} - \vec{\phi}\right) \xrightarrow{D,T\to\infty} N_p\left(0, \sigma_W^2 \Gamma_p^{-1}\right)$$

where $\sigma_W^2 \Gamma_p^{-1}$ is the optimal variance among all possible (asymptotically) unbiased estimators [efficients].

Results can be used to obtain confidence intervals for ϕ.

2.4.2 Estimating ARMA Parameters (MLE Approach)

OLS and YW estimators are effective in estimating the $AR(p)$ parameters, but are difficult to apply to fitting $MA(q)$ and general $ARMA(p,q)$ models since the white noises W_t are not observable, and YW equations are not linear in the MA parameters.

Definition 2.4.1 — Latent Variables. Variables associated with the noise, unobserved quantities.

When we have latent variables, MLE tends to work the best.
Suppose $X_t \sim AR(1)$ causal, $X_t = \phi X_{t-1} + W_t$, $W_t \overset{iid}{\sim} N(0, \sigma_W^2)$ with Gaussian distributional assumption. Then,

$$X_t = \sum_{l=0}^{\infty} \phi_l W_{t-l}$$

is Gaussian. In fact, L^2-limits of Gaussian random variables are Gaussian. This can be shown using MGF or characteristic function. Moreover, $X_1, \cdots, X_T$ are jointly Gaussian, since

$$a_1 X_1 + \cdots + a_T X_T = \sum_{l=0}^{\infty} \phi_l (a_1 W_{1-l} + \cdots + a_T W_{T-l})$$

the likelihood function is the joint distribution of $X_T, \cdots, X_1$,

$$L(\phi, \sigma_W^2) = f(X_T, \cdots, X_1; \phi, \sigma_W^2)$$

which is $T-$variate Gaussian density.

Key Idea in Time Series

to evaluate the likelihood, condition on the path/past!

$$\begin{aligned} f(X_T, \cdots, X_1) &= f(X_T | X_{T-1}, \cdots, X_1) f(X_{T-1}, \cdots, X_1) \\ &= f(X_T | X_{T_1}, \cdots, X_1) f(X_{T-1} | X_{T-2}, \cdots, X_1) \cdots f(X_1) \\ &= \prod_{i=1}^{T} \underbrace{f(X_i | X_{i-1}, \cdots, X_1)}_{\text{Gaussian}} \end{aligned}$$

According to A2, we have

$$AR(1) \Longrightarrow X_i | X_{i-1} = X_i | X_{i-1}, \cdots, X_1 \sim N\left(\phi X_{i=1}, \sigma_W^2\right)$$

Thus,

$$L(\phi,\sigma_W^2) = \prod_{i=2}^{T} \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(x_i-\phi x_{i-1})^2}{2\sigma_W^2}} f(x_i)$$

$$= (2\pi\sigma_W^2)^{-\frac{T-1}{2}} e^{-\sum_{i=2}^{T}\frac{(x_i-\phi x_{i-1})^2}{2\sigma_W^2}} f(x_1;\phi,\sigma_W^2)$$

Maximizing $L(\phi,\sigma_W^2)$ in this case leads to a similar estimator as OLS/YW.

General ARMA(p,q) Case

Again, $X_T,\cdots,X_1$ are jointly Gaussian if W_t ∼Gaussian,

$$L(\underbrace{\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q,\sigma_W^2}_{\vec{\theta}\in\mathbb{R}^{p+q+1}}) = \prod_{i=1}^{T} f(X_i|X_{i-1},\cdots,X_1)$$

in particular,

$$X_i|X_{i-1},\cdots,X_1 \sim N\left(\tilde{X}_{i|i-1}(\vec{\theta}), P_{i-1}^{i}(\vec{\theta})\right)$$

This complicated likelihood can be maximized using numerical optimization techniques, such as Newton-Raphson or conjugate gradient.

Theorem 18 The MLE's of $\phi_1,\cdots\phi_p,\theta_1,\cdots,\theta_q,\sigma_W^2$ are $\sqrt{T}$ consistent and asymptotically Normal with asymptotic covariance equal to the inverse of the information matrix. In this sense, they are asymptotically optimal.

R

1. MLE estimation reduces to OLS, YW equations for AR(p) models
2. For general ARMA estimation MLE is thought to be optimal in most situations. (used as a default/benchmark)

2.5 Model Selection Diagnostic Tests

Using MLE, we can fit an ARMA(p,q) to an observed time series $X_1,\cdots,X_T$. But how to select the orders of an ARMA(p,q) model?

Usual Methods

1. Examine ACF and PACF.
2. Model diagnostics/Goodness-of-fit tests: examine the residuals of the ARMA(p,q) model to check for the plausibility of the white noise assumption.
3. Model selection methods: information criteria, cross-validation scheme.

2.5.1 Model Diagnostic

If the ARMA(p,q) model fits the data well, then the estimated residuals

$$\hat{W}_t = \frac{X_t - \tilde{X}_{t|t-1}}{\sqrt{\hat{P}_t^{t-1}}}$$

where $\tilde{X}_{t|t-1}$ is the truncated predictor of X_t based on $X_{t-1},\cdots,X_1$ and $\hat{P}_t^{t-1}$ is the estimated MSE. These standardized residuals should behave like white noise. This can be investigated by considering the empirical ACF $\hat{\rho}_W(h)$.

As a measure of how "white" the residuals are, it is common to evaluate the cumulative significance of $\hat{\rho}_W(h), 1 \leq h \leq H$ by applying a "white noise test". Suppose $W_1, \cdots, W_T$ is a strong WN, and $\hat{\rho}_W(h)$ is the empirical ACF of this series, we know that

$$\sqrt{T}\hat{\rho}_W(h) \xrightarrow{D} N(0,1)$$

for each fixed h. Also, for $j \neq h$,

$$\begin{aligned}
&\textbf{Cov}(\sqrt{T}\hat{\gamma}_W(h), \sqrt{T}\hat{\gamma}_W(j)) \\
\approx& T\mathbb{E}\left[\sum_{t=1}^{T} W_t W_{t+h}\right]\left[\sum_{s=1}^{T} W_s W_{s+j}\right] \\
=& T\sum_{t=1}^{T}\sum_{s=1}^{T} \underbrace{\mathbb{E}[W_t W_{t+h} W_s W_{s+j}]}_{\text{alway zero!}} = 0
\end{aligned}$$

Using the Martingale or m-dependent CLT's, it can be shown that

$$\begin{bmatrix} \sqrt{T}\hat{\rho}_W(1) \\ \vdots \\ \sqrt{T}\hat{\rho}_W(H) \end{bmatrix} \xrightarrow{D} N_H(0, I_{H\times H})$$

then,

$$T\sum_{h=1}^{H} \hat{\rho}_W^2(h) \xrightarrow{D} \chi^2(H)$$

This motivate the following white noise test for ARMA.

Box-Ljung-Pierce Test

If $X_t \sim ARMA(p,q)$ model, and $\hat{W}_t$ are the model residuals with empirical ACF $\hat{\rho}_W(h)$, then if

$$Q(T,H) = T(T+2)\sum_{h=1}^{H} \frac{\hat{\rho}_W^2(h)}{T-h} \approx T\sum_{h=1}^{H} \hat{\rho}_W^2(h)$$

where H is the maximal lag parameter. Then,

$$Q(T,H) \xrightarrow{D, T\to\infty} \chi^2(H-(p+q))$$

where we lose $p+q$ degrees of freedom for fitting model. The BLP test p-value is then computed as

$$P_{BLP} = \mathbb{P}\big(\chi^2(H-(p+q)) > Q(T,H)\big)$$

R If $X_t \sim ARMA(p,q)$, and $\hat{W}_t$ are calculated based on an ARMA(p',q') model where $p' < p$ or $q' < q$ (model is under-specified), then

$$Q(T,H) \xrightarrow{p, T\to\infty} \infty$$

The interpretation is that, if P_{BLP} are small, the model is ill-fitting or under specified.

2.5.2 Model Selection:Information Criteria

Suppose we are trying to select the orders p,q of an ARMA(p,q) model to fit to $X_1,\cdots,X_T$. A natural idea is to maximize the likelihood of the data as a function of p,q. However, the problem is that the likelihood is monotonically increasing as a function of p,q. Maximizing would lead to overfitting. The usual solution is to maximize the likelihood subject to a penalty term on the number of parameters (complexity) of the the model. Let the number of parameters in the ARMA(p,q) model be denoted by $k = p+q+1$. Then,

$$-2\log\left[L(X_1,\cdots,X_T;\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}_W^2)\right]+P(T,k)$$

where $P(T,k)$ is some increasing function k to penalize using too many parameters. Optimal p,q should balance the model fit with the penalty for complexity.

Common penalty terms

Definition 2.5.1 — AIC.

$$\mathbf{AIC}(p,q) = -2\log\left[L(X_1,\cdots,X_T;\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}_W^2)\right]+\frac{2k+T}{T}$$

comes from estimate the Kullback-Liebler distance (KL-divergence) from the fitted model to the "true" model

Definition 2.5.2 — BIC.

$$\mathbf{BIC}(p,q) = -2\log\left[L(X_1,\cdots,X_T;\hat{\vec{\phi}},\hat{\vec{\theta}},\hat{\sigma}_W^2)\right]+\frac{2\log(T)}{T}$$

comes from approximating and maximizing the posterior distribution of the model given the data.

For interpretation, smaller AIC/BIC, the better the model. Information criteria are also used in trend fitting:
Suppose

$$X_t = S_t + Y_t = \overbrace{f_t(\vec{\beta})} + Y_t$$

where β is a vector of parameters in $\mathbb{R}^k$. We can estimate $\vec{\beta}$ with $\hat{\vec{\beta}}$ using OLS. We can look at

$$\mathbf{SS(Res)}_T = \sum_{t=1}^{T}(X_t - f_t(\hat{\vec{\beta}}))^2$$

Information criteria typically calculated assuming Y_t is Gaussian WN and are of the form

$$\mathbf{SS(Res)}_T + P(T,k)$$

where $P(T,k)$ could be AIC or BIC.

R

1. In trend fitting, the assumption of Gaussian WN residuals is often in doubt.
2. AIC/BIC are not perfect! They are but one of many tools useful in model selection.
   (a) **Strengths:**
      i. Easy to compute
      ii. Facilitates comparing many models quickly
   (b) **Weaknesses:**
      i. Likelihood must be specified
      ii. There is a degree of arbitrariness to the choice of penalty
3. It can be shown that minimizing the AIC is related to minimizing the 1-step forecast MSE, and so when the application is forecasting, AIC is more common.

2.6 ARIMA Models

We have seen that many time series appear stationary after differencing.

Definition 2.6.1 — Integrated time series. We say a time series $\{X_t\}_{t\in\mathbb{Z}}$ is integrated to order d if $\nabla^d X_t$ is stationary, but $\nabla^j X_t, 1 \leq j \leq d$ is not stationary.

R The motivation is that if Y_t is stationary and $X_t = \sum_{j=1}^t Y_j$, then X_t is integrated to order 1. Similarly, $Z_t = \sum_{j=1}^t X_j$ is integrated to order 2.

Definition 2.6.2 — ARIMA(p,d,q). We say $\{X_t\}_{t\in\mathbb{Z}}$ follows an autoregressive integrated moving average process of orders p,d,q, denoted by ARIMA(p,d,q). If

$$\phi(B) \quad \overbrace{(1-B)^d X_t}^{X_t \text{ is integrated to order } d} \quad = \theta(B)W_t$$

so basically, $\nabla^d X_t$ follows an ARMA(p,q).

Forecasting ARIMA(p,d,q) processes

1. $Y_t = \nabla^d X_t$ follows ARMA(p,q) model, and so can be forecasted using truncated ARMA prediction.
2. Forecasts $\hat{y}_{T+h|T}$ can be used to forecast X_{T+h} by reversing the differencing.

■ **Example 2.3** Suppose $d = 1$. Then,

$$Y_{T+1} = X_{T+1} - X_T$$

so, $\hat{X}_{T+1|T} = X_T + \hat{Y}_{T+1|T}$. This can be iterated to produce longer horizon forecast. ■

For our forecasts, the prediction MSE is approximately of the form

$$P_{T+1}^T \cong \sigma_W^2 \sum_{j=0}^{h-1} \psi_{j,*}^2$$

where $\psi_{j,*}$ is the coefficients of z^j in the power series expansion (centred at zero) of $\frac{\theta(z)}{\phi(z)(1-z)^d}$. (I don't know about you, this looks hella Laurent to me) The idea is that

$$X_t \approx \frac{\theta(B)}{\phi(B)(1-B)^d} W_t$$

■ **Example 2.4** Suppose $X_t \sim ARIMA(0,1,0)$. Then,

$$X_t - X_{t-1} = (1-B)X_t = W_t \implies X_t = \sum_{j=1}^t W_j$$

If $Y_t = \nabla X_t$,

$$T + \hat{h}|T = 0$$

since we are just forecasting the W_t. Then,

$$\hat{X}_{T+1|T} = X_T + \hat{Y}_{T+1|T} = X_T$$

similarly, $\hat{X}_{T+h|T} = X_T$. This means the best predictor of a random walk is the last known location. We can also calculate the prediction MSE,

$$\frac{\theta(z)}{\phi(z)(1-z)^d} = \frac{1}{1-z} = \sum_{j=0}^{\infty} z^j, |z| < 1$$

Thus, $\psi_{j,*} = 1, \forall j \geq 0$ and $P_{T+1}^{T} \cong \sigma_W^2 \sum_{j=0}^{h-1} \psi_{j,*}^2 = h\sigma_W^2$. Note that

$$\mathbb{E}\left[(\hat{X}_{T+h|T} - X_{T+h})^2\right] = \mathbb{E}\left[\left(\sum_{j=T+1}^{T+h} W_j\right)^2\right] = h\sigma_W^2$$

in this case, our approximation is actually exact. If we forecast this result, we will see something as follow:

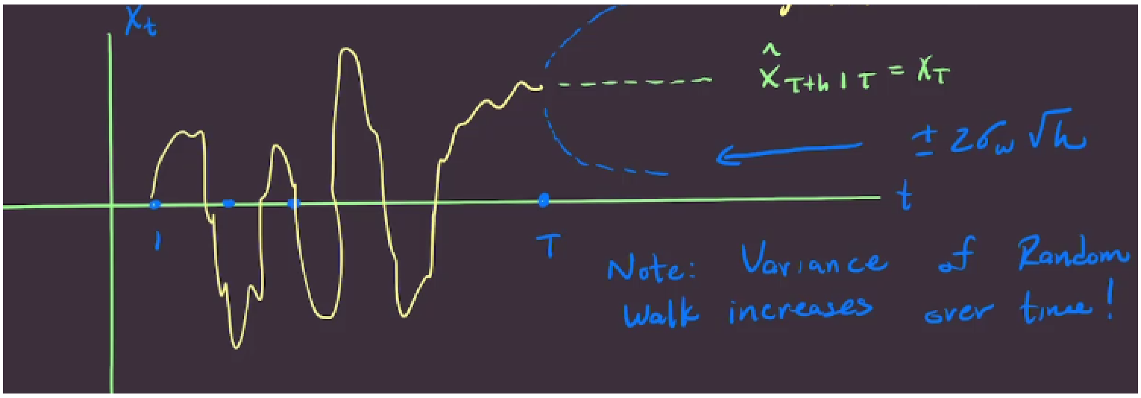


Figure 2.6.1: Forecasting a random walk

■

One intuition that we can get from this simple example is that the prediction MSE tends to increase as lag increases.

How to decide in practice on degree of differencing?

1. **Eye-ball test:** start differencing until starts to look like a stationary process
2. **Dickey-Fuller, KPSS test:** formal stationary tests
3. **Cross-validation scheme**

■ **Example 2.5 — R Demo - ARIMA.** Incomplete codes with mysterious `dat` data. XD ■

From the note taker, recent days have been rough. Now, I am finally back to fully committing to studies. Sorry for the delay of notes update.

2.7 SARIMA Models

Frequently, time series exhibit "seasonality", the **rough definition** is, a time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be "seasonal" if it exhibits regular variation so that for some lag s, X_t is "similar" to X_{t-s}. The sources of seasonality could be weather, scheduled events, or agricultural activities. Typically, this lead to yearly, monthly, weekly, or quarterly cycles.

R ARIMA models that we have seen are not ideal for modelling seasonality since they are derived upon random walk with stationary errors. It is like you are taking baby steps but cannot see the whole picture to capture the seasonality.

Let’s take a look at an example.

■ Example 2.6 — Mortality

```
%
library(astsa)
cmort2=cmort[seq(1,508,by=4)]
cmort2=ts(cmort2, frequency =13)
plot(cmort2)
```

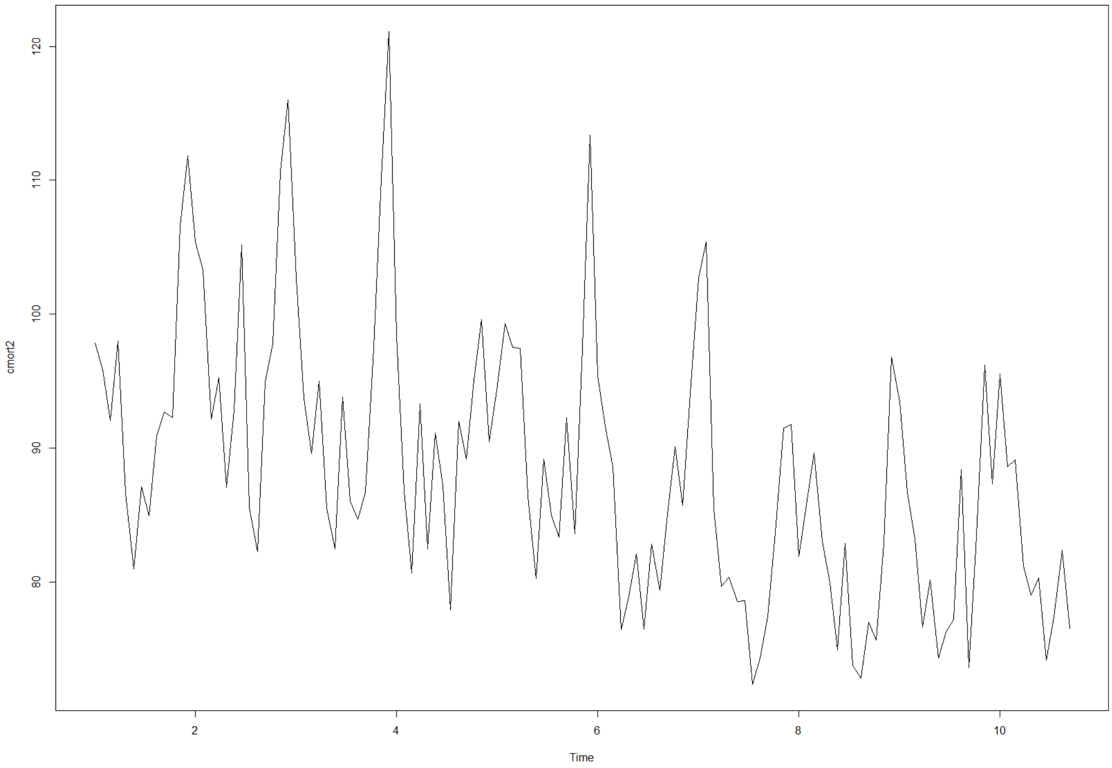


Figure 2.7.1: Mortality time series (weekly)

as we can see that this shows some seasonality on a quaterly basis. One simple justification for that mortality cases are affected by the season and its corresponding weather condition. If we look at the first difference and its corresponding ACF, PACF, we can see that

```
%
par(mfrow = c(1, 3))
plot(diff(cmort2), main = "First Difference")
acf(diff(cmort2))
pacf(diff(cmort2))
```

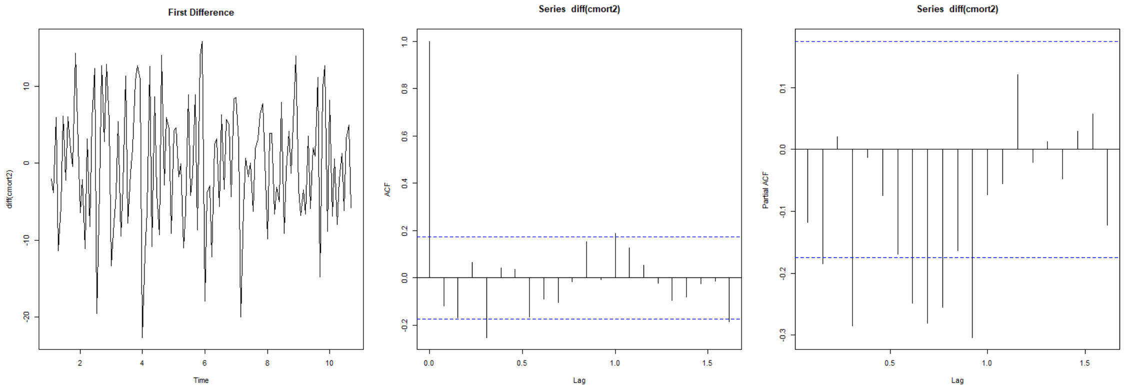
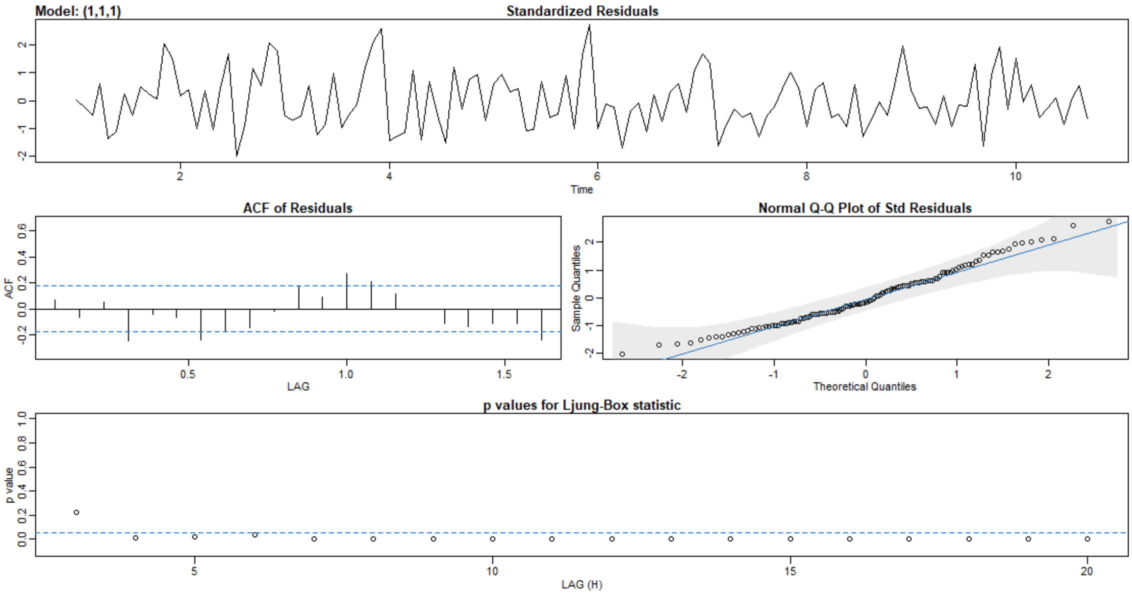


Figure 2.7.2: Mortality time series: first difference, ACF, PACF

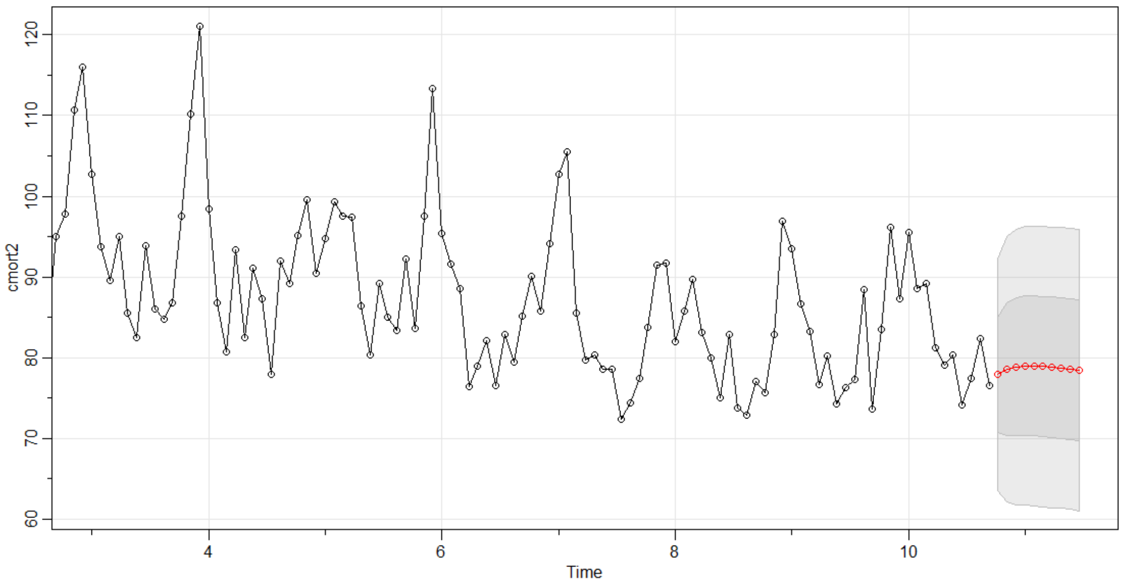
the lags are suggesting a ARIMA(1,1,1) model. We run the following model diagnostics.

```
%
sarima(cmort2, 1, 1, 1)
```



But this is not very white (my output is a bit different from what Greg had in the lecture). We can check the forecasted series and see that

```
%
sarima.for(cmort2,10, 1, 1, 1)
```



you have obtain a garbage that should probably dropped and receive an WF. Notice the ARIMA model bases on random walk model, then the best estimate will be the last one. ■

Definition 2.7.1 — Seasonal ARIMA (SARIMA). $\{X_t\}_{t\in\mathbb{Z}}$ is said to follow a seasonal ARIMA model or order p,d,q, ARIMA parameters, and P,D,Q, seasonal parameters, and a seasonal period s, denoted by $SARIMA(p,d,)\times(P,D,Q)_s$, if

$$\Phi(B^s)\underbrace{\phi(B)}_{\text{normal AR}}\underbrace{(1-B^s)^D(1-B)^d}_{\text{regular \& seasonal differencing}}X_t=\Theta(B^s)\underbrace{\theta(B)}_{\text{normal MA}}W_t$$

where

$$\Phi(z)=1-\Phi_1 z-\cdots-\Phi_P z^P,\ \Theta(z)=1+\Theta_1 z+\cdots+\Theta_Q z^Q$$

are called the seasonal AR and MA polynomials.

■ **Example 2.7** Suppose $X_t\sim SARIMA(1,1,1)\times(1,1,1)_{13}$. This is similar to the mortality dataset. Then,

$$\begin{cases}\Phi(z)=1-\Phi_1 z,\ \Theta(z)=1+\Theta_1 z\\ \phi(z)=1-\phi_1 z,\ \theta(z)=1+\theta_1 z\end{cases}$$

then,

$$(1-\Phi_1 B^{13})(1-\phi_1 B)\underbrace{(1-B^{13})(1-B)X_t}_{=:y_t}=(1+\Theta_1 B^{13})(1+\theta_1 B)W_t$$

we can write (sure we can)

$$y_t-\Phi_1 y_{t-13}-\phi_1 y_{t-1}+\phi_1\Phi_1 y_{t-14}=\text{some MA terms}$$

This means $y_t=f(y_{t-13},y_{t-1},\text{MA noise})$ where f is some linear function. You see that y_{t-13} in the function input. That's how SARIMA can capture the seasonality. ■

R

1. $y_t=(1-B^s)^D(1-B)^d X_t$, a SARIMA model is just one big ARIMA model for y_t. *Note taker's opinion:* the lower case parameters are trying to tell us the behaviour within a cycle while the upper case parameters are trying to ignore those micro behaviour and obtain the larger picture.

2. Advantage over standard ARMA and ARIMA models is parsimony. Since seasonal series have the feature that X_t is similar to X_{t-s}, we introduce just a few additional terms to model X_t as a function of X_{t-s}.

Fitting SARIMA Models

1. Usually the seasonal lag, s, is known from expertise or preliminary data analysis.
2. Differencing and seasonal differencing can be decided upon by
   (a) **Eye-ball test**: examing ACF and PACF.

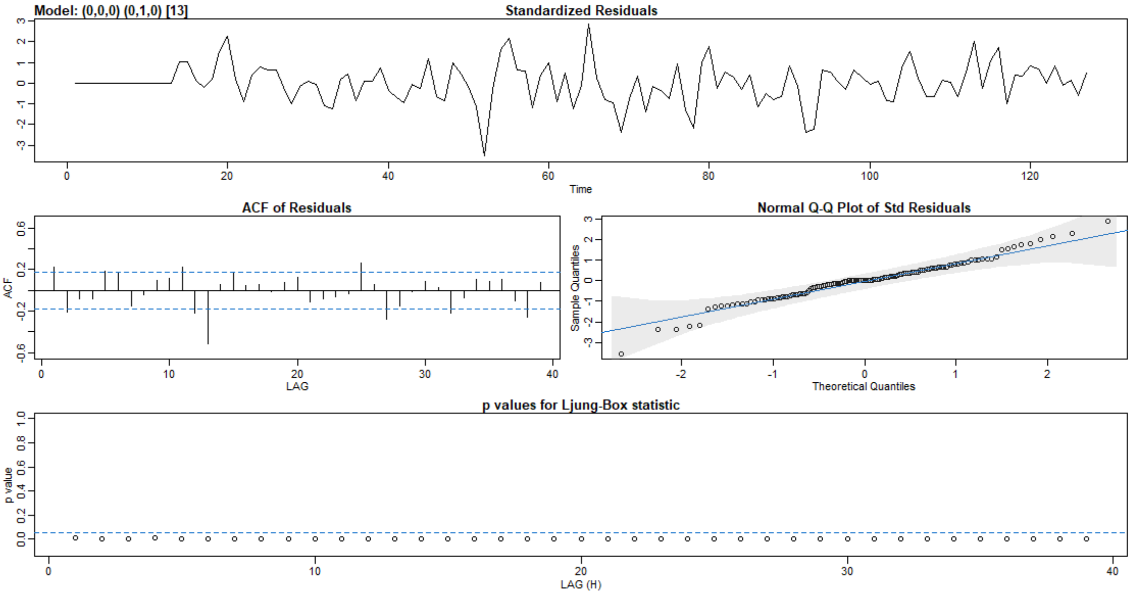
   > When I was studying in China. Teacher barely proves anything because solutions are too "obvious" by eye-balling.

   (b) **Stationarity tests**
   (c) **Cross-validation scheme**
3. Choosing the order and estimating the components of $\Phi, \phi, \Theta, \theta$ can be done in the same way as with ARMA models, such as MLE method and residual analysis.

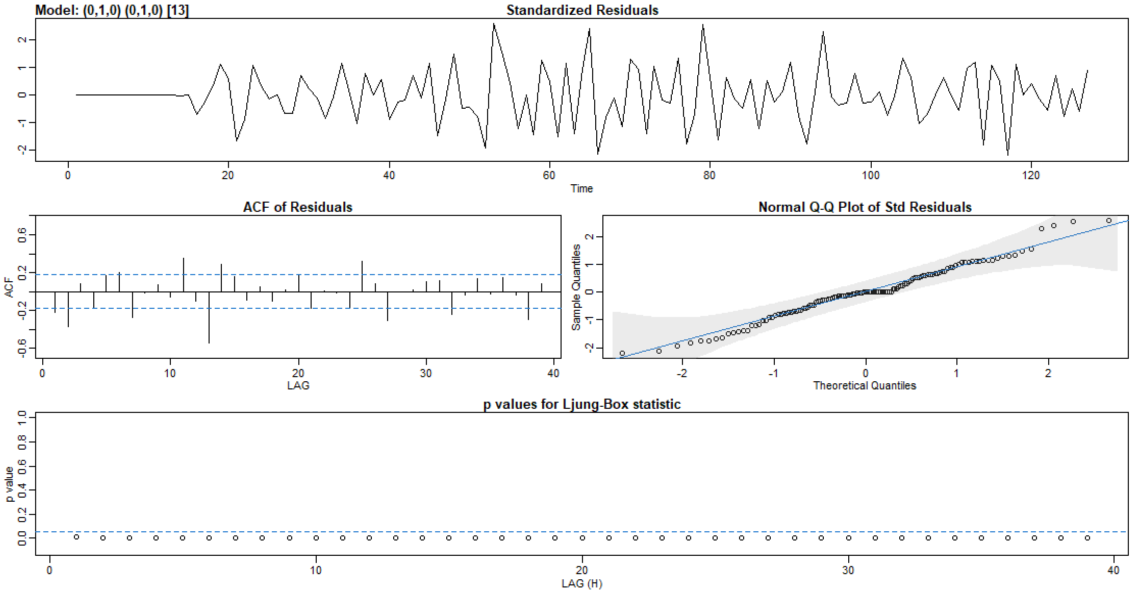
■ Example 2.8 — Back to the mortality example. We would expect to see some significant serial correlations at the seasonal lags, this means the ACF of the seasonal lags will probably be outside the prediction interval drawn. Now, we need to do some fitting:

```
%
sarima(cmort2, 0,0,0,0,1,0,13)
sarima(cmort2, 0,1,0,0,1,0,13)
sarima(cmort2, 1,1,1,1,1,1,13)
sarima(cmort2, 2,1,1,1,1,3,13)
sarima(cmort2, 2,1,2,1,1,3,13)
```

we start by augmenting the model by increasing the order.



This looks to be reasonable stationary but serial correlations at the seasonal lag in the ACF is still apparent.



After first differencing, the serial correlation at the seasonal lag is till prominent. Now, we just need to fit more models to see a better diagnostic result.

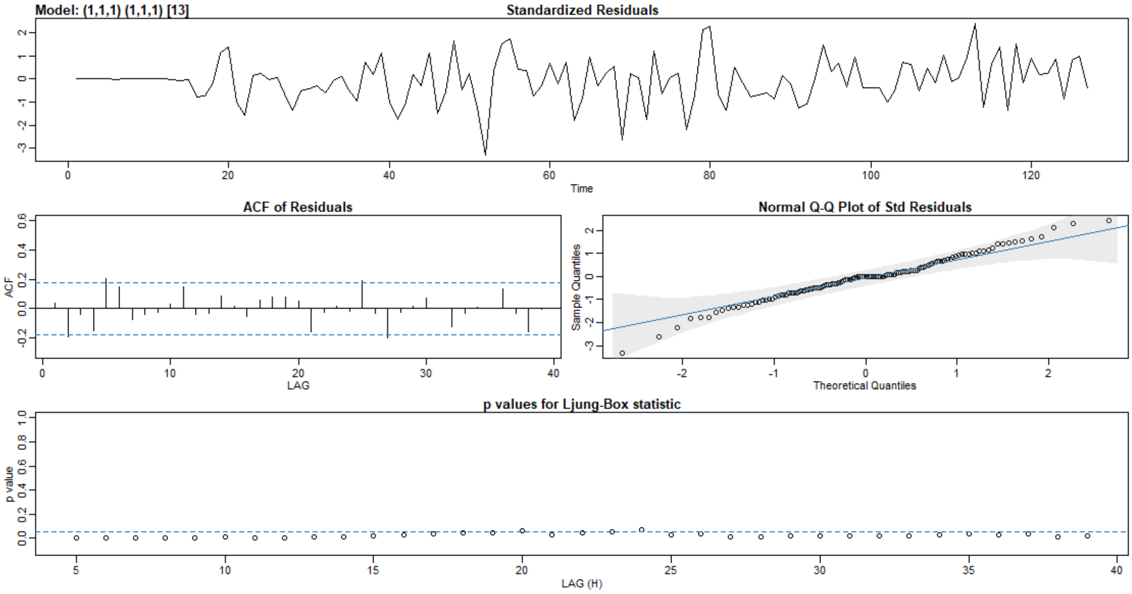


Figure 2.7.3: AIC:4.716147, BIC:3.786075

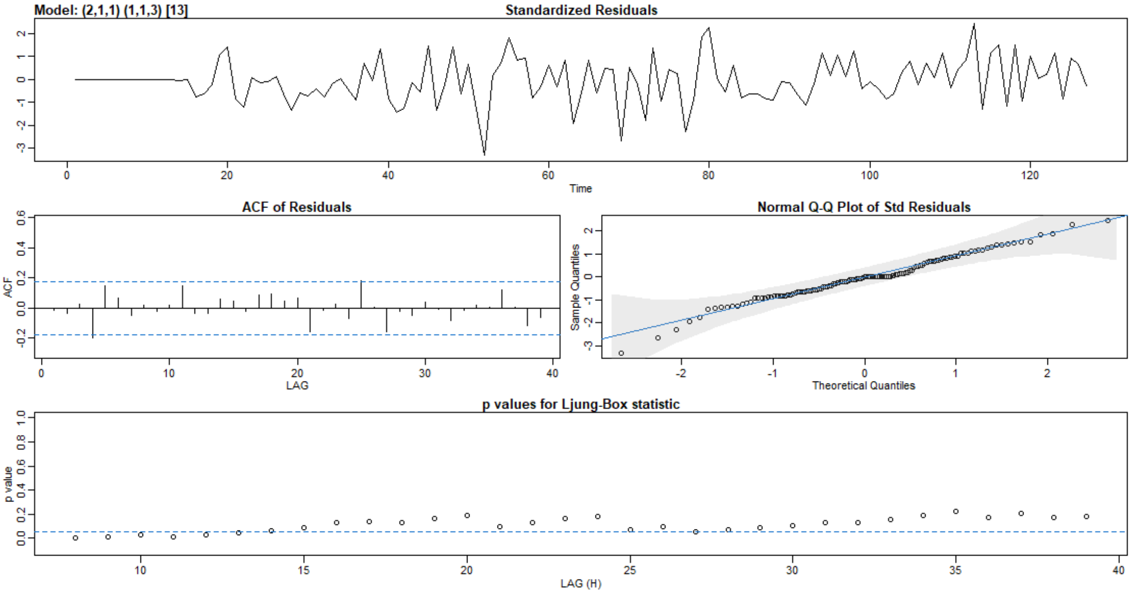


Figure 2.7.4: AIC:4.649194, BIC:3.790603

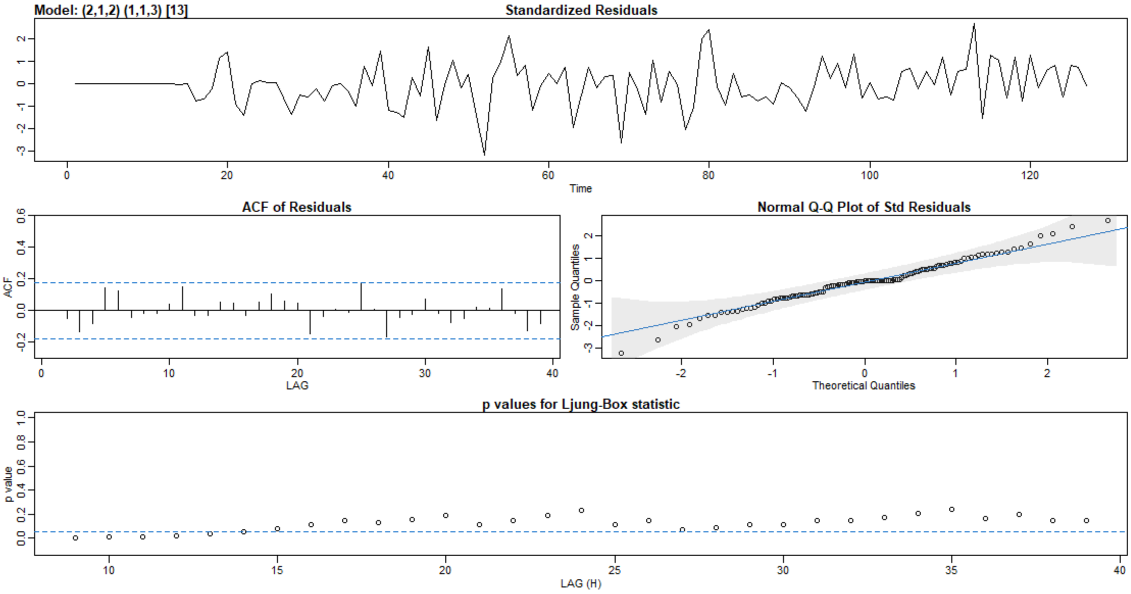


Figure 2.7.5: AIC:4.67116, BIC:3.82246

This suggests we go with $SARIMA(2,1,1) \times (1,1,3)_{13}$ and have the following forecast.

```
%sarima.for(cmort2, 20, 2,1,2,1,1,3,13)
```

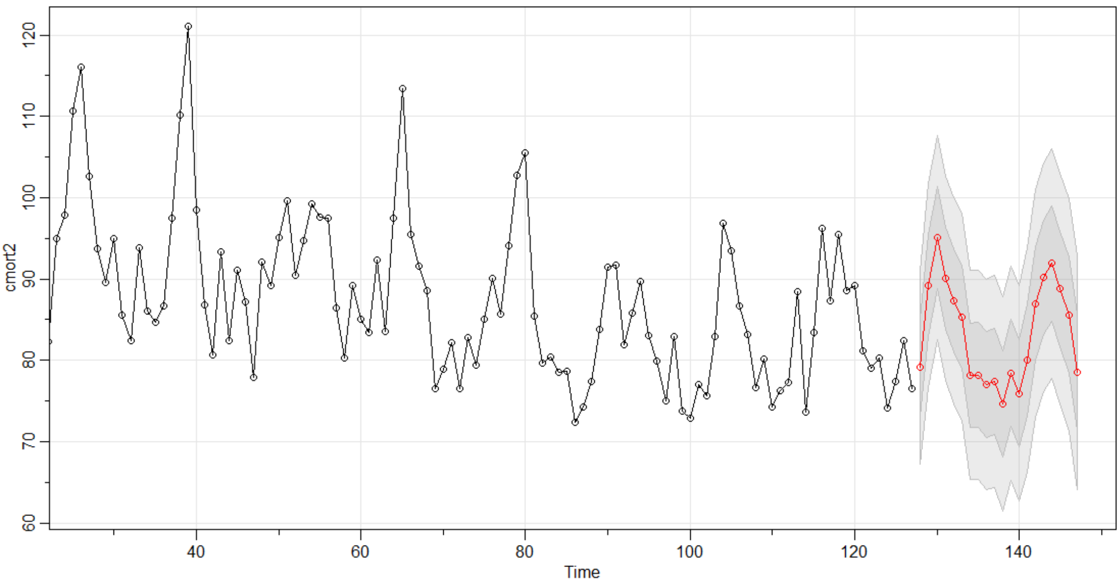


Figure 2.7.6: $SARIMA(2,1,1) \times (1,1,3)_{13}$ at 20 steps

we can even look at the long range forecast.

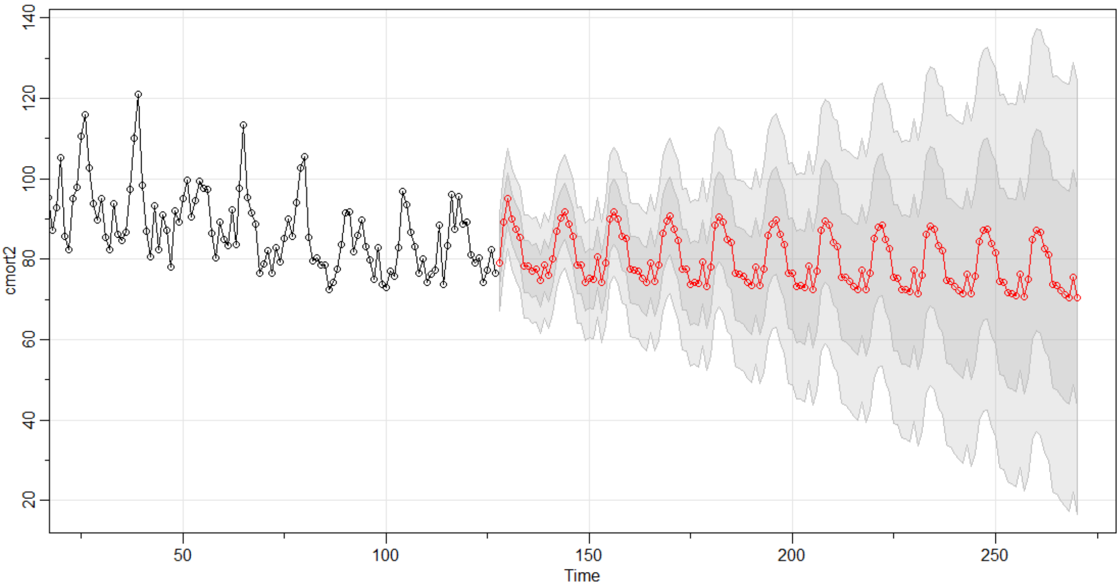


Figure 2.7.7: $SARIMA(2,1,1) \times (1,1,3)_{13}$ in 11 years

One thing to keep in mind that there could be a lot of different possible models that look "good". To actually back-testing (yeah quant) them, we usually need to introduce some CV scheme and evaluation metric. ■

2.7.1 R Demo - Cardiovascular Mortality

Would recommend watch the demo video. But one sentence to summarize what is going below is a grid parameter search for the "optimal" SARIMA model fit.

```
%
library(astsa)
```

```

cmort2=cmort[seq(1,508,by=4)]

cmort2=ts(cmort2, frequency =13)

plot(cmort2)

par(mfrow = c(1, 3))
plot(diff(cmort2), main = "First Difference")
acf(diff(cmort2))
pacf(diff(cmort2))

sarima(diff(cmort2), 1, 0, 1)
par(mfrow = c(1, 1))
sarima.for(cmort2,10, 1, 0, 1)


sarima(cmort2, 0,0,0,0,1,0,13)
par(mfrow = c(1, 1))
sarima.for(cmort2, 13*11, 2,1,1,1,1,3,13)

num=3
aic.mat=rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)
bic.mat=rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)

for(ar in 0:num){
    for(ma in 0:num){
        for(AR in 0:num){
            for(MA in 0:num){
                x=tryCatch(sarima(cmort2,ar,0,ma,AR,0,MA,13),error=function(e){x=list(); x$AIC=10^100; x$BIC=10^100 ; return(x)})
                aic.mat[ar,ma,AR,MA]=x$AIC
                bic.mat[ar,ma,AR,MA]=x$BIC
            }}}}
y.aic=which(aic.mat==min(aic.mat), arr.ind=TRUE)
y.bic=which(bic.mat==min(bic.mat), arr.ind=TRUE)
sarima(cmort2,y.aic[1],0,y.aic[2],y.aic[3],0,y.aic[4],13)
sarima.for(cmort2,20,y.aic[1],0,y.aic[2],y.aic[3],0,y.aic[4],13)

y.aic00=y.aic

library(astsa)
cmort2=cmort[seq(1,508,by=4)]
num=3
aic.mat=rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)
bic.mat=rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)

for(ar in 0:num){
    for(ma in 0:num){
        for(AR in 0:num){
            for(MA in 0:num){
                x=tryCatch(sarima(cmort2,ar,1,ma,AR,1,MA,13),error=function(e){x=list(); x$AIC=10^100; x$BIC=10^100 ; return(x)})
                aic.mat[ar,ma,AR,MA]=x$AIC
                bic.mat[ar,ma,AR,MA]=x$BIC
            }}}}
y.aic=which(aic.mat==min(aic.mat), arr.ind=TRUE)
```

```
y.bic=which(bic.mat==min(bic.mat), arr.ind=TRUE)
sarima(cmort2,y.aic[1],1,y.aic[2],y.aic[3],1,y.aic[4],13)
sarima.for(cmort2,20,y.aic[1],1,y.aic[2],y.aic[3],1,y.aic[4],13)

y.aic11=y.aic



for(ar in 0:num){
    for(ma in 0:num){
        for(AR in 0:num){
            for(MA in 0:num){
                x=tryCatch(sarima(cmort2,ar,1,ma,AR,0,MA,13),error=function(e){x=list(); x$AIC=10^100; x$BIC=10^100 ; return(x)})
                aic.mat[ar,ma,AR,MA]=x$AIC
                bic.mat[ar,ma,AR,MA]=x$BIC
            }}}}
y.aic=which(aic.mat==min(aic.mat), arr.ind=TRUE)
y.bic=which(bic.mat==min(bic.mat), arr.ind=TRUE)
sarima(cmort2,y.aic[1],1,y.aic[2],y.aic[3],1,y.aic[4],13)
sarima.for(cmort2,20,y.aic[1],1,y.aic[2],y.aic[3],1,y.aic[4],13)

y.aic10=y.aic


library(astsa)
cmort2=cmort[seq(1,508,by=4)]
num=2
aic.mat=rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)
bic.mat=rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)%o%rep(10^100,num)

for(ar in 0:num){
    for(ma in 0:num){
        for(AR in 0:num){
            for(MA in 0:num){
                x=tryCatch(sarima(cmort2,ar,0,ma,AR,1,MA,13),error=function(e){x=list(); x$AIC=10^100; x$BIC=10^100 ; return(x)})
                aic.mat[ar,ma,AR,MA]=x$AIC
                bic.mat[ar,ma,AR,MA]=x$BIC
            }}}}
y.aic=which(aic.mat==min(aic.mat), arr.ind=TRUE)
y.bic=which(bic.mat==min(bic.mat), arr.ind=TRUE)
sarima(cmort2,y.aic[1],1,y.aic[2],y.aic[3],1,y.aic[4],13)
sarima.for(cmort2,20,y.aic[1],1,y.aic[2],y.aic[3],1,y.aic[4],13)

y.aic01=y.aic

y.aic00
y.aic01
y.aic10
y.aic11

sarima(cmort2,y.aic10[1],1,y.aic10[2],y.aic10[3],0,y.aic10[4],13)

sarima(cmort2,y.aic01[1],0,y.aic01[2],y.aic01[3],1,y.aic01[4],13)

sarima(cmort2,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13)

sarima(cmort2,y.aic11[1],1,y.aic11[2],y.aic11[3],1,y.aic11[4],13)
```

```

sarima.for(cmort2,20, y.aic10[1],1,y.aic10[2],y.aic10[3],0,y.aic10[4],13)

sarima.for(cmort2,20,y.aic01[1],0,y.aic01[2],y.aic01[3],1,y.aic01[4],13)

sarima.for(cmort2,20,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13)

sarima.for(cmort2,20,y.aic11[1],1,y.aic11[2],y.aic11[3],1,y.aic11[4],13)


library("fpp2")
library("forecast")

fitAR = auto.arima(cmort2) ##explained in Section
#8.7 of HH book

fitAR

checkresiduals(fitAR)
fcast <- forecast(fitAR, PI=TRUE, h=20)
autoplot(fcast)
```

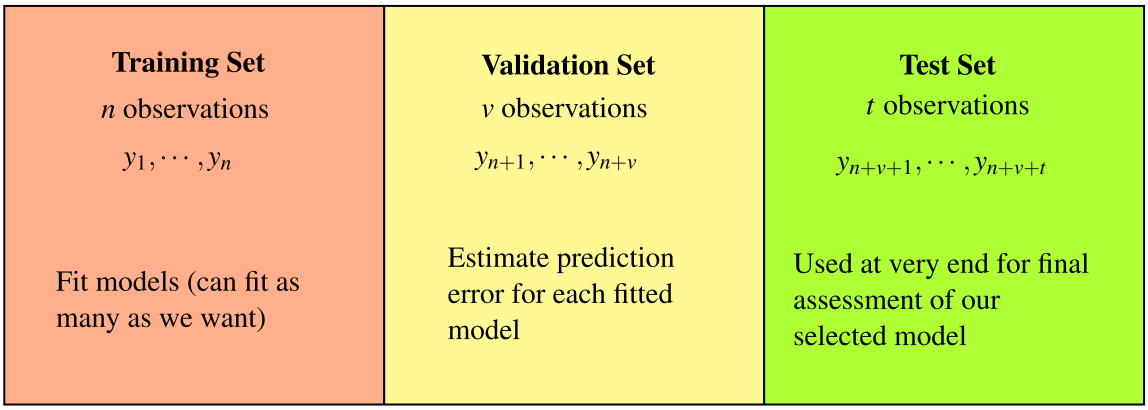
2.8 Time Series Cross Validation

Definition 2.8.1 — Cross Validation (CV). Cross-validation is a data driven model evaluation and selection tool for predictive models that entails:

1. Splitting the available data into training and testing sets
2. Fitting a model(s) on the training sets
3. Evaluating predictions of the model on the test sets as an overall evaluation of model quality

Standard CV

Ideally, we have lots of data, conceptualize having 3 parts:



Training Set
n observations
$y_1,\cdots,y_n$
Fit models (can fit as many as we want)

Validation Set
v observations
$y_{n+1},\cdots,y_{n+v}$
Estimate prediction error for each fitted model

Test Set
t observations
$y_{n+v+1},\cdots,y_{n+v+t}$
Used at very end for final assessment of our selected model

Figure 2.8.1: Train-Validation-Test

For example, we can use **MSE** as a metric (could be any loss function) and compute and compare:

1. **Training MSE**$=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ We know that our usual estimate $\hat{\sigma}^2=\frac{\textbf{SS(Res)}}{n-p-1}$ is a scaled version of this to compensate for the number of predictors.
2. **Validation MSE**$=\frac{1}{b}\sum_{i=n+1}^{n+v}(y_i-\hat{y}_i)^2$ This is considered as an estimate of *MSPE* on new data
3. **Testing MSE**$=\frac{1}{t}\sum_{i=n+v+1}^{n+v+t}(y_i-\hat{y}_i)^2$ This is the actual test of prediction *MSPE*

R The idea here is that: **MSE** on validation set should approximate **MSE** on testing set since neither set of observation used to fit model.

■ **Example 2.9 — oVeRFItInG?!.** So if we are using MSE/RMSE as metric (as related to $\hat{\sigma}^2, \hat{\sigma}$) and it is significantly larger on validation compared to training set. We probably overfitted and can't expect model to generalize well to new data.

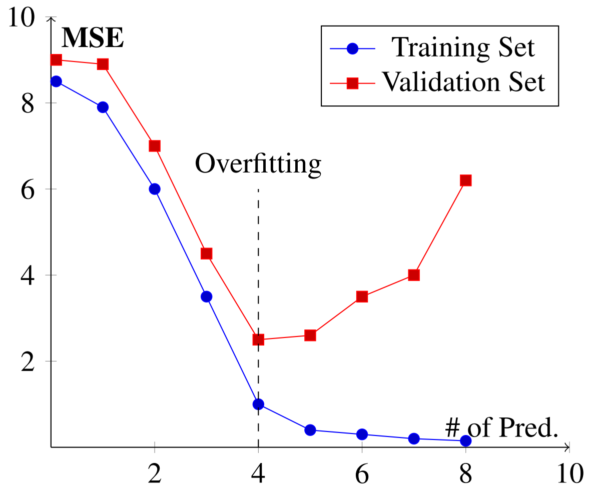


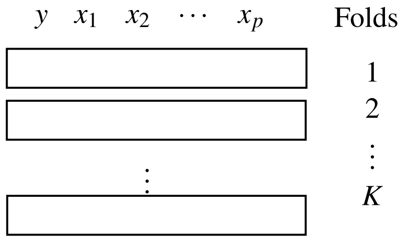
Figure 2.8.2: Error metric vs. # of predictors on the 3-set

■

Practical Applications

1. **Simplest:** randomly divide available data into train/validation, such as 80/20 split for training/validation. It has the following **weaknesses:**
   (a) we do not use all data for training
   (b) we only get estimate of prediction error
2. **Better:** use cross-validation scheme (CV)

How to do CV with K folds?



For example, using **RMSE**, we have **RMSE**$_1, \cdots,$ **RMSE**$_K$. We can take the average

$$\frac{1}{K}\sum_{k=1}^{K} \textbf{RMSE}_k$$

as an estimate of $RMSPE$.

1. Divide available data for training and validation into K roughly equal-sized sets (folds) randomly
2. For CV fold k, use data in fold k as validation and train on the rest of data.

Thus, to estimate prediction error for a given model, we fit it K times, each time treating data in fold $1, 2, \cdots, K$ as validation. Thus, we have K estimates of prediction error.

Say we are given some candidate models (e.g. with different variables included). We can do $K-$fold CV using each of them, and choose the one with lowest average predictor error across folds.

R

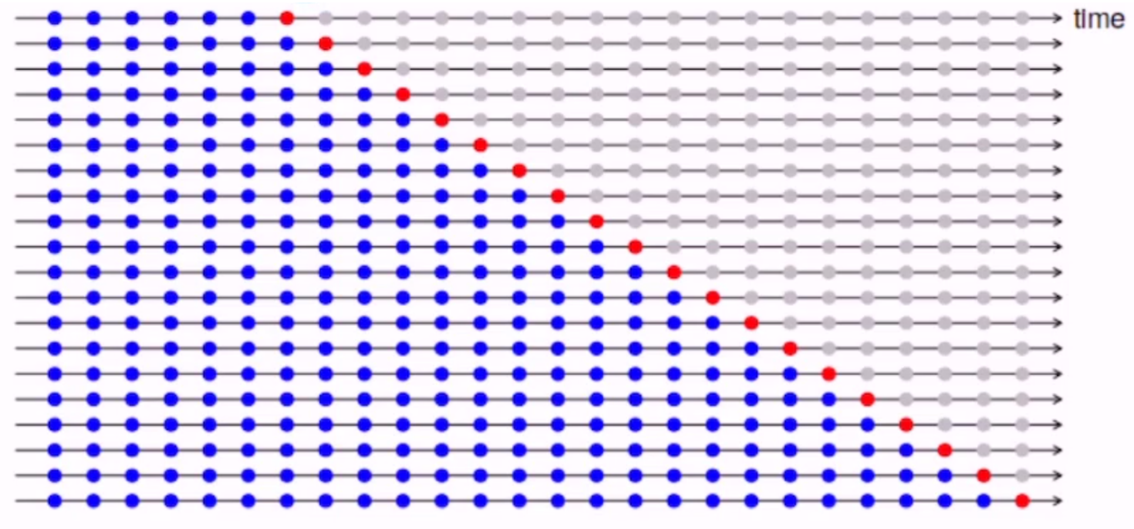
1. K is often called the number of folds

2. $K = n$, the procedure is often called "leave-one-out" (LOO) CV.
3. $K = 10$ is quite common

Problems with time series CV

1. Randomly splitting the data scramble up any serial dependence relationship
2. In time-series forecasting, it is most natural to use the past (recent past) to predict future values

The most common time series cross-validation scheme is the **Expanding window**.



Algorithm 2.1 — Time Series Cross-Validation Algorithm.

1. Split data into training and testing ranges $1 \leq t_r \leq T$. The common choice is $t_r \approx 75\%T$. The test sample is $X_{t_r+1}, \cdots, X_T$
2. For each j in $t_r+1, \cdots, T$, use model to forecast $\hat{X}_{j+1|j}$ on $X_1, \cdots, X_j$. Calculate the loss
$$L(\hat{X}_{j+1|j}; X_{j+1}) = L_j$$
3. CV score of the model is
$$CV(M) = \sum_{j=t_r+1}^{T} L_j$$

R

1. If interested in longer horizon forecasting, you can compare
$$\hat{X}_{j+1|j}, \cdots, \hat{X}_{j+h|j} \text{ to } X_{j+1}, \cdots, X_{j+h}$$
2. **"Stationarity"** is crucial in time series CV since the model errors in the present must be similar to the errors in the future!
3. One normally cannot cross-validate everything! Too much computation required maybe.

2.8.1 R Demo - CV

```
%
cmort2=ts(cmort2, frequency =13)
plot(cmort2)
```

```
tr= round(length(cmort2)*0.75)
cv00=1:length( tr:(length(cmort2)-1) )
cv11=1:length( tr:(length(cmort2)-1) )
cv01=1:length( tr:(length(cmort2)-1) )
cv10=1:length( tr:(length(cmort2)-1) )
cvAUTOAR=1:length( tr:(length(cmort2)-1) )

for(j in (tr:(length(cmort2)-1))){
  print(j)
  #x1=sarima.for(cmort2[1:j],1,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13)
  x00= tryCatch(sarima.for(cmort2[1:j],1,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13),error=function(e){x=list(); x$pred=cmort2[j] ; return(x)})
  x10= tryCatch(sarima.for(cmort2[1:j],1,y.aic10[1],1,y.aic10[2],y.aic10[3],0,y.aic10[4],13),error=function(e){x=list(); x$pred=cmort2[j] ; return(x)})
  x01= tryCatch(sarima.for(cmort2[1:j],1,y.aic01[1],0,y.aic01[2],y.aic01[3],1,y.aic01[4],13),error=function(e){x=list(); x$pred=cmort2[j] ; return(x)})
  x11= tryCatch(sarima.for(cmort2[1:j],1,y.aic11[1],1,y.aic11[2],y.aic11[3],1,y.aic11[4],13),error=function(e){x=list(); x$pred=cmort2[j] ; return(x)})
  cmort2j=ts(cmort2[1:j],frequency=13)
  xAAR= forecast(auto.arima(cmort2j),h=1)$mean
  cvAUTOAR[j-tr+1]=xAAR-cmort2[j+1]
  cv00[j-tr+1]=x00$pred-cmort2[j+1]
  cv10[j-tr+1]=x10$pred-cmort2[j+1]
  cv01[j-tr+1]=x01$pred-cmort2[j+1]
  cv11[j-tr+1]=x11$pred-cmort2[j+1]
}

#One year ahead
h=13
tr= round(length(cmort2)*0.70)
cv00h=1:length( tr:(length(cmort2)-1-h) )
cv11h=1:length( tr:(length(cmort2)-1-h) )
cv01h=1:length( tr:(length(cmort2)-1-h) )
cv10h=1:length( tr:(length(cmort2)-1-h) )

for(j in (tr:(length(cmort2)-1-h))){
  print(j)
  #x1=sarima.for(cmort2[1:j],1,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13)
  x00= tryCatch(sarima.for(cmort2[1:j],h,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13),error=function(e){x=list(); x$pred=rep(cmort2[j],h) ; return(x)})
  x10= tryCatch(sarima.for(cmort2[1:j],h,y.aic10[1],1,y.aic10[2],y.aic10[3],0,y.aic10[4],13),error=function(e){x=list(); x$pred=rep(cmort2[j],h) ; return(x)})
  x01= tryCatch(sarima.for(cmort2[1:j],h,y.aic01[1],0,y.aic01[2],y.aic01[3],1,y.aic01[4],13),error=function(e){x=list(); x$pred=rep(cmort2[j],h) ; return(x)})
  x11= tryCatch(sarima.for(cmort2[1:j],h,y.aic11[1],1,y.aic11[2],y.aic11[3],1,y.aic11[4],13),error=function(e){x=list(); x$pred=rep(cmort2[j],h) ; return(x)})
  cv00h[j-tr+1]=sum((x00$pred-cmort2[(j+1):(j+h)])^2)
  cv10h[j-tr+1]=sum((x10$pred-cmort2[(j+1):(j+h)])^2)
  cv01h[j-tr+1]=sum((x01$pred-cmort2[(j+1):(j+h)])^2)
  cv11h[j-tr+1]=sum((x11$pred-cmort2[(j+1):(j+h)])^2)
}


mean(cv00)
mean(cv10)
mean(cv01)
mean(cv11)
mean(cvAUTOAR)
sum(cv00^2)
sum(cv10^2)
sum(cv01^2)
sum(cv11^2)
sum(cvAUTOAR^2)




mean(cv00h)
mean(cv10h)
mean(cv01h)
mean(cv11h)





sarima.for(cmort2,20,y.aic10[1],1,y.aic10[2],y.aic10[3],0,y.aic10[4],13)

sarima.for(cmort2,20,y.aic10[1],1,y.aic10[2],y.aic10[3],0,y.aic10[4],13)

sarima.for(cmort2,20,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00[4],13)

sarima.for(cmort2,20,y.aic11[1],1,y.aic11[2],y.aic11[3],1,y.aic11[4],13)

##Exponential Smoothing and Feed Forward NN


#install.packages("forecast")
#install.packages("fpp2")
library("fpp2")
library("forecast")

cmort2=ts(cmort2,frequency = 13)

x=forecast(fit,h=30)

fitAR = auto.arima(cmort2)
#fcast <- forecast(fitAR, PI=TRUE, h=30,bootstrap = TRUE)
fcast <- forecast(fitAR, PI=TRUE, h=30)

autoplot(fcast)

fit <- nnetar(cmort2, lambda=0)
autoplot(forecast(fit,h=30))
fcast <- forecast(fit, PI=TRUE, h=30)
autoplot(fcast)

fitET=ets(cmort2)
fcast <- forecast(fitET, h=30,PI=TRUE)
autoplot(fcast)
```

```



tr= round(length(cmort2)*0.75)
cvSARIMA=1:length( tr:(length(cmort2)-1) )
cvAUTOAR=1:length( tr:(length(cmort2)-1) )

cvETS=1:length( tr:(length(cmort2)-1) )
cvNNAR=1:length( tr:(length(cmort2)-1) )


for(j in (tr:(length(cmort2)-1))){
  print(j)
  #x1=sarima.for(cmort2[1:j],1,y.aic00[1],0,y.aic00[2],y.aic00[3],0,y.aic00
    [4],13)
  xSARIMA= tryCatch(sarima.for(cmort2[1:j],1,y.aic00[1],0,y.aic00[2],y.
    aic00[3],0,y.aic00[4],13),error=function(e){x=list(); x$pred=cmort2[j]
    ; return(x)})
  cmort2j=ts(cmort2[1:j],frequency=13)
  xAAR= forecast(auto.arima(cmort2j),h=1)$mean
  xETS = forecast(ets(cmort2j),h=1)$mean
  autoplot(forecast(ets(cmort2j),h=1))
  xNNAR = forecast(nnetar(cmort2j, lambda=0),h=1)$mean
  cvSARIMA[j-tr+1]=xSARIMA$pred-cmort2[j+1]
  cvAUTOAR[j-tr+1]=xAAR-cmort2[j+1]
  cvETS[j-tr+1]=xETS-cmort2[j+1]
  cvNNAR[j-tr+1]=xNNAR-cmort2[j+1]
  print(xETS-cmort2[j+1])
  print(xNNAR-cmort2[j+1])

}

sum(cvSARIMA^2)
sum(cvETS^2)
sum(cvNNAR^2)
sum(cvAUTOAR^2)
```

2.8.2 Simulated or Boostrapped Prediction Intervals

Usually forecasts are of the form:

$$\hat{X}_{T+1|T} = g(X_T, X_{T-1}, \cdots, X_1, W_{T+1})$$

where W_{T+1} is a strong WN innovation. Often even models are additive so that

$$\hat{X}_{T+1|T} = g(X_T, X_{T-1}, \cdots, X_1) + W_{T+1}$$

Simple and powerful method to produce prediction intervals is to use simulation!

Algorithm 2.2 — Simulated Prediction Intervals.

1. Choose a distribution for $\{W_t\}_{t\in\mathbb{Z}}$. For example, Gaussian and $\hat{\sigma}^2_W$ could be estimated from the time series residuals.
2. For $b = 1, \cdots, B$, B is a large number ($B = 10^5$), simulate $\left\{W^{(b)}_{T+1}\right\}$
3. Compute $\hat{X}^{(b)}_{T+1|t} = g(X_T, \cdots, X_1) + W^{(b)}_{T+1}, b = 1, \cdots, B$.
4. Denote the empirical $q-$th quantile of $\left\{X^{(b)}_{T+1} : b = 1, \cdots, B\right\}$ by $\hat{Q}_{T+1}(q)$. We set the

$1-\alpha$ prediction interval as

$$\left(\hat{Q}_{T+1}(\alpha/2), \hat{Q}_{T+1}(1-\alpha/2)\right)$$

Note that for longer horizon forecasts, prediction intervals can be obtained using iteration:

$$\hat{X}^{(b)}_{T+h|T} = g(\hat{X}^{(b)}_{T+h-1|T}, \cdots, \hat{X}^{(b)}_{T+1|T}, X_T, \cdots, X_1) + W^{(b)}_{T+h}$$

then the prediction interval is given by

$$\left(\hat{Q}_{T+h}(\alpha/2), \hat{Q}_{T+h}(1-\alpha/2)\right)$$

Distributions to choose for W_t

1. $W_t \sim N(0, \hat{\sigma}^2_W)$ with the residual estimated. This leads to approximately the same well-known prediction intervals
2. A distribution fit to the estimated residuals $\hat{W}_t$, such as $t-$distribution, Pareto, etc.
3. The empirical distribution of the residuals $\hat{W}_t$, which means randomly drawing from $\{\hat{W}_1, \cdots, \hat{W}_T\}$. Well, this is called **Bootstrapping**. Usually the most preferred way to produce prediction interval. However, an important consideration of the bootstrap is that the residuals should be white. Otherwise, how can we choose randomly? We can check this using ACF or a WN test.

2.8.3 R Demo - Prediction Intervals

```
%
###Bootstrapped PI's

plot(cmort2)

fitAR = auto.arima(cmort2)

fitAR

checkresiduals(fitAR)

fcast <- forecast(fitAR, PI=TRUE, h=20)
autoplot(fcast)

fcast <- forecast(fitAR, PI=TRUE, h=20,bootstrap = TRUE)
autoplot(fcast)
```

2.9 Exponential Smoothing

ARIMA models aim to model a time series, potentially after differencing towards stationarity, in terms of its autocorrelation (linear process).
Exponential smoothing aims to propose flexible model of the trend and seasonality observed in a time series.

2.9.1 Simple Exponential Smoothing

Suppose we wish to forecast a time series $X_1, \cdots, X_T$. Two extreme forecasts are $\hat{X}_{T+1|T} = X_T$, random walk, and $\hat{X}_{T+1|T} = \bar{X}$, IID sequuence. Exponential smoothing provides a compromise between these two.

$$\hat{X}_{T+1|T} = \alpha X_T + \alpha(1-\alpha)X_{T-1} + \cdots + \alpha(1-\alpha)^{T-1}X_1, \alpha \in [0,1]$$

the weights applied to past observations decrease exponentially quickly. Simple exponential smoothing can be stated as a recursive system of equations:

1. **Prediction equation:** $\hat{X}_{T+1} = l_T$
2. **Smoothing equation (Level equation):** $l_T = \alpha X_T + (1-\alpha)l_{T-1}$, which is a convex combination of the last observed value and last prediction.
3. **Initial condition:** l_0

The parameters defining the model are $\alpha \in [0,1]$ and l_0. Estimation of these parameters may be conducted using MLE or OLS:

$$(\hat{\alpha}, \hat{l}_0) = \arg\min_{0\leq\alpha\leq 1, l_0\in\mathbb{R}} \sum_{i=2}^{T} (x_i - l_i(\alpha, l_0))^2$$

then,

$$\hat{X}_{T+1} = \hat{\alpha} + (1-\hat{\alpha})l_T(\hat{\alpha}, \hat{l}_0)$$

which can be computed recursively.

Linear Trend Exponential Smoothing

The prediction equation is

$$\hat{X}_{T+h} = \underbrace{l_T}_{\text{level}} + h \underbrace{b_T}_{\text{slope}}$$

the level equation is

$$l_T = \alpha X_T + (1-\alpha)(l_{T-1} + b_{T-1})$$

the trend/slope equation is

$$b_T = \beta \underbrace{(l_T - l_{T-1})}_{\text{last observed slope}} + (1-\beta) \underbrace{b_{T-1}}_{\text{last slope}}$$

the initial conditions are: l_0, b_0 and the parameters are $\alpha, \beta \in [0,1], l_0, b_0 \in \mathbb{R}$, which can be estimated using MLE or OLS.

Trend+Seasonal Exponential Smoothing

Suppose h is the forecast horizon of interest, and the time series has seasonal period p. Set $k = \left\lfloor \frac{h-1}{p} \right\rfloor$.

The prediction equation is

$$\hat{X}_{T+h} = \underbrace{l_T}_{\text{level}} + h \underbrace{b_T}_{\text{slope}} + \underbrace{S_{T+h-p(k+1)}}_{\text{seasonal effect}}$$

the level equation is

$$l_T = \alpha(X_T - S_{T-p}) + (1-\alpha)(l_{T-1} + b_{T-1})$$

the slope equation is

$$b_T = \beta(l_T - l_{T-1}) + (1-\beta)b_{T-1}$$

seasonal equation is

$$S_T = \gamma(X_T - l_{T-1} - b_{T-1}) + (1-\gamma)S_{T-p}$$

the initial conditions are $l_0, b_0, S_0, \cdots, S_{-p+1}$. The parameters are $\alpha, \beta, \gamma \in [0,1], l_0, b_0, S_0, \cdots, S_{-p+1} \in \mathbb{R}$.

2.9.2 Exponential Smoothing as a State Space Model

Rewriting the level equation in simple exponential smoothing gives us

$$l_t = l_{t-1} + \alpha\underbrace{(X_t - l_{t-1})}_{=:\varepsilon_t} = l_{t-1} + \alpha\varepsilon_t$$

also $X_t = l_{t-1} + \varepsilon_t$. This means we can write the prediction equations as functions of these errors. Therefore, these equations can be reformulated as

1. **prediction equation:** $X_t = l_{t-1} + \varepsilon_t$
2. **Level equation:** $l_t = l_{t-1} + \alpha\varepsilon_t$

Why is this useful? If we make a parametric assumption on ε_t, then we can use Likelihood techniques (MLE, AIC, simulation based prediction intervals). Such equations are examples of "State Space" models.

Definition 2.9.1 — State Space Model. We say $\{X_t\}_{t\in\mathbb{Z}}$ follows a general state space model if

1. **Observation equation:** $X_t = A_t y_t + \varepsilon_t$, where A is a measurement matrix and y_t is a state vector and ε_t is an observed error. The state vector is unknown.
2. **State equation:** $y_t = \Phi y_{t-1} + w_t$

ε_t, w_t might be dependent.

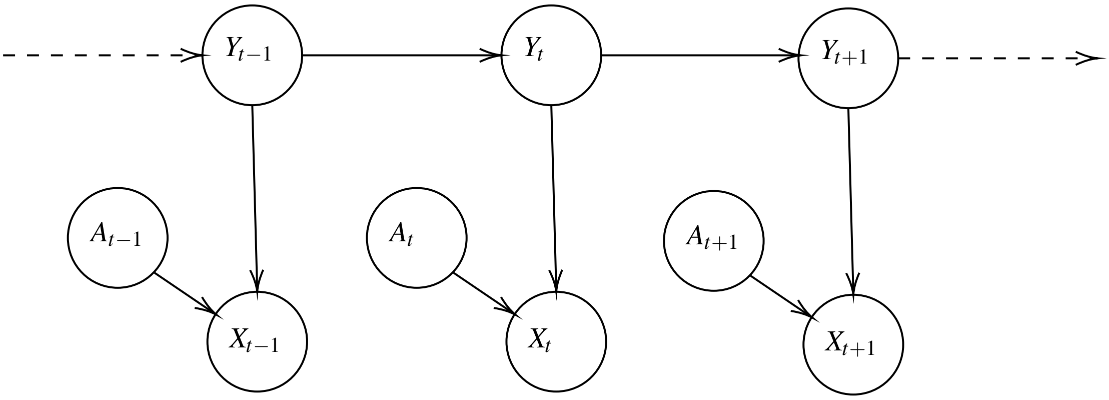


Figure 2.9.1: Intuition of the state space model

■ **Example 2.10** 1. AR(1):

$$X_t = y_t,\ \ y_t = \phi y_{t-1} + w_t$$

where w_t is a strong WN.

2. Simple exponential smoothing:

$$X_t = y_{t-1} + \varepsilon_t,\ \ y_t = y_{t-1} + \alpha\varepsilon_t$$

where ε_t is a strong WN.

All ARMA and exponential smoothing models can be written in state-space model form. ■

Parameter estimation and model selection using state-model

Consider

$$X_t = l_{t-1} + \varepsilon_t,\ \ l_t = l_{t-1} + \alpha\varepsilon_t,\ \ \varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

with initial condition: l_0. We can use the traditional likelihood technique

$$L(X_1, \cdots, X_T; \alpha, l_0, \sigma_\varepsilon^2) = \prod_{i=1}^{T} L(\underbrace{X_i | X_{i-1}, \cdots, X_1}_{\sim N(l_{i-1}(\alpha, l_0), \sigma_\varepsilon^2)}; \alpha, l_0, \sigma_\varepsilon^2)$$

we can maximize likelihood numerically. Use this to calculate AIC or BIC.

2.9.3 Multiplicative Exponential Smoothing Models

Standard exponential smoothing has "additive" errors, in the sense that

$$X_t = l_{t-1} + \varepsilon_t, \;\; l_t = \alpha X_t + (1-\alpha) l_{t-1} \implies \varepsilon_t = X_t - l_{t-1}$$

we can also formulate exponential smoothing in terms of "multiplicative" errors, so $\varepsilon_t = \frac{X_t - l_{t-1}}{l_{t-1}}$ which is relative to the previous level. Then,

$$X_t = l_{t-1}(1+\varepsilon_t), \;\; l_t = \alpha X_t + (1-\alpha) l_{t-1} = l_{t-1}(1+\alpha\varepsilon_t)$$

Why consider multiplicative errors: it is important to note that since the level follows the same exponential smoothing equation, the forecasts from multiplicative and additive error models will be the same.

The difference arises from how uncertainty/error propogates in the model.

1. **Additive:** $\hat{X}_{T+h} = l_T + \sum_{j=T+1}^{T+h} \varepsilon_j$, where the MSE scales like h.
2. **Multiplicative:** $\hat{X}_{T+h} = l_T \prod_{j=T+1}^{T+h} (1+\varepsilon_j)$, where the MSE is scaling like $\left(\mathbb{E}[(1+\varepsilon_0)^2]\right)^h$, which could grow very quickly as $h \to \infty$.

Multiplicative Linear + Trend and Holt Winters Exponential Smoothing

For linear+trend state space formulation, we now consider

$$\varepsilon_t = \frac{X_t - (l_{t-1} + b_{t-1})}{l_{t-1} + b_{t-1}}, \;\; X_t = (l_{t-1} + b_{t-1})(1+\varepsilon_t), \;\; l_t = (l_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$$

and $b_t = b_{t-1} + \beta(l_{t-1} + b_{t-1})\varepsilon_t$ where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.
For seasonal exponential smoothing, we now consider

$$X_t = (l_{t-1} + b_{t-1}) S_{t-p}(1+\varepsilon_t), \;\; l_t = (l_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$$

and

$$b_t = b_{t-1} + \beta(l_{t-1} + b_{t-1})\varepsilon_t, \;\; S_t = S_{t-p}(1+\gamma\varepsilon_t)$$

When to use additive vs multiplicative?

Seasonal exponential smoothing models:

1. Multiplicative models implies that as the level increases (decreases) the seasonal fluctuations increase (decrease). Additive models suggest seasonal fluctuations remain constant as trend fluctuates. If the seasonal fluctuation increases as the level increases, we should consider a multiplicative model
2. Use AIC/BIC: the AIC can be evaluated for each state-space model, and compared
3. Cross-validation

2.9.4 Exponential Smoothing Model Selection and Prediction Intervals

Given the state-space formulation of exponential smoothing and the use of MLE to estimate the parameters, it is common to use AIC to choose among competing exponential smoothing models. Other options include cross-validation or residual analysis (white noise testing).

Prediction Intervals: using the state space formulation, valid prediction intervals may be computed using simulation.

■ **Example 2.11 — Simple ES.** Suppose $\hat{X}_{T+1|T} = \hat{l}_T$ and $X_{T+1} \cong \hat{l}_T + \varepsilon_{T+1}$ by state space formula and $\varepsilon_{T+1} \sim N(0, \sigma_\varepsilon^2)$.

1. $\hat{\sigma}_\varepsilon^2 = \frac{1}{T-1}\sum_{j=2}^{T}(X_j - \hat{l}_{T-1})^2$
2. simulate $\hat{X}_{T+1|T}^{(b)} = \hat{l}_T + \varepsilon_{T+1}^{(b)}$ where $\varepsilon_{T+1}^{(b)} \sim N(0, \hat{\sigma}_\varepsilon^2)$.
3. Use 5% and 95% sample quantiles of $\varepsilon_{T+1}^{(b)}$ as a prediction interval.

■

R In many casese the prediction MSE assuming $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ can be computed explicitly. See Section 7.7 of Hyndman and A.

An important consideration in applying this approach is that ε_t should behave like Gaussian WN. We can check using a residual analysis

1. White noise test, ACF plots
2. QQ plot (normality)

2.10 Neural Network Autoregression

A simple Neural Network Architecture has an input layer, covariates and predictors, a hidden layer(s), and an output layer, response or prediction. At any particular layer, the inputs are mapped to the j-th neuron linearly. The value taken on the jth neuron is

$$z_j = b_j + \sum_{i=1}^{m} w_{i,j} x_i$$

where x_i is the ith input. To calculate the input to the next layer, it requires an activation function (non-linear). For example, sigmoid function or ReLu. If the activation function is linear, then the NN is really just a linear regression.

2.10.1 Neural Network AR

Definition 2.10.1 — NNAR. The input layer is consisted of $X_t, \cdots, X_{t-p}$ and the output layer is consisted of X_{t+1}. A NN model with k hidden states, assuming one hidden layer, we call such a model NNAR(p,k) model. We note that if $k = 0$, $NNAR(p,0) = AR(p)$.

Definition 2.10.2 — Seasonal NNAR. THe inpout layer is consisted of $X_t, \cdots, X_{t-p}, X_{t-m}, \cdots, X_{t-P_m}$ where m is the seasonal lag. The output layer is consisted of X_{t+1}. We call this a NNSAR(p,k,P).

Model Selection

The process of choosing k, p, P can be carried out using CV and weights estimated using OLS. We can formulate it as

$$X_{t+1} = \underbrace{f(\underline{X}_t)}_{\text{Neural Network}} + \varepsilon_{t+1}$$

By calculating the residuals using the bootstrap samples, we have

$$X_{T+1}^{(b)} = \hat{f}(\underline{X}_T) + \hat{\varepsilon}_{T+1}^{(b)}, \quad b = 1, \cdots, B$$

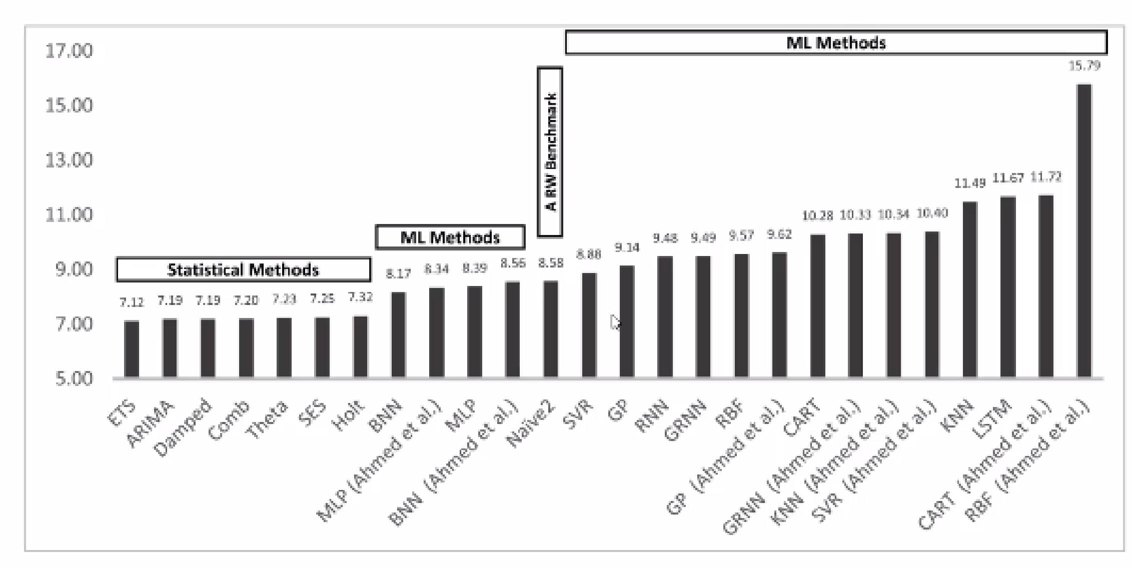


Figure 2.10.1: Duhhhhhhhhh

2.11 Conditional Heteroscedasticity

Definition 2.11.1 — Heteroscedasticity. Changing variance over time.

For example, if $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary, then $\{X_t\}_{t\in\mathbb{Z}}$ is "homoscedastic" in the sense that $\mathbf{Var}(X_t) = \sigma_X^2$ does not change over time.

Definition 2.11.2 — Heteroscedastic. We say a time series $\{X_t\}_{t\in\mathbb{Z}}$ is heteroscedastic if $\mathbf{Var}(X_t) = \sigma_{X,t}^2$ depends on t and changes at some points.

In particular, these are not stationary.
In the context of conditional heteroscedastic time series, we often consdier asset price or "financial" time series.
Suppose X_t is the price of an asset at time t. The returns of X_t is the time series $y_t = \nabla X_t$. The log-returns of a positive asset price X_t are $y_t = \log\left(\frac{X_t}{X_{t-1}}\right)$. The volatility (VIX go brrrrrrrr) is the standard deviation.
A common observation, especially prominent with financial and asset price data, is that periods of volatility or heteroscedasticity tend to cluster. The big shocks cause volatile periods, that further propagate volatility until things calm down. ARMA and linear time series models are not useful for capturing this phenomenon since they tend to model the conditional mean while leaving the variance untouched.

Definition 2.11.3 — Conditional Heteroscedasticity. We say a time seris $\{X_t\}_{t\in\mathbb{Z}}$ is conditional heteroscedastic if

$$\mathbf{Var}(X_t|X_{t-1},\cdots) = \sigma_{X,t}^2$$

changes with t.

2.12 ARCH & GARCH Models

Definition 2.12.1 — ARCH. Let $\{W_t\}_{t\in\mathbb{Z}}$ be a unit variance strong white noise with $\mathbb{E}[W_t] = 0$ and $\mathbf{Var}(W_t) = 1$. We say $\{X_t\}_{t\in\mathbb{Z}}$ follows an Autoregressive, Conditionally, Heteroscedastic

(ARCH) model if there exists parameters $w \geq 0, \alpha_1 \geq 0$ so that

$$X_t = \sigma_t W_t, \quad \sigma_t^2 = \omega + \alpha_1 X_{t-1}^2$$

R This model is due to Robert Engle, 1982.

Definition 2.12.2 — ARCH(p). We say $\{X_t\}_{t\in\mathbb{Z}}$ follows a ARCH(p) model if $\{W_t\}_{t\in\mathbb{Z}}$ is a strong WN with unit variance and

$$X_t =\sim_t W_t, \quad \sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2, \; \omega \geq 0, \alpha_j \geq 0, i = 1, \cdots, p$$

R

1. σ_t^2 is called the conditional variance or volatility. Imagine that there exists a representation
$$X_t = g(W_t, W_{t-1}, \cdots)$$
a stationary process satisfying the ARCH model, then in the case of ARCH(1), we have
$$\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 = g_\sigma(W_{t-1}, \cdots)$$
so,
$$\mathbf{Var}(X_t|W_{t-1},\cdots) = \mathbf{Var}(\sigma_t W_t|W_{t-1},\cdots) = \sigma_t^2 \mathbf{Var}(W_t) = \sigma_t^2$$
which is indeed the conditional variance that we are trying to model.
2. Engle won the Nobel prize in economics in part for the "methods of analyzing economic series with time varying volatility (ARCH)" in 2003.
3. One problem noticed early on was that ARCH(p) models required large orders p to model asset returns.

Definition 2.12.3 — GARCH(p,q). We say $\{X_t\}_{t\in\mathbb{Z}}$ follows a Generalized ARCH (GARCH) model if $\{W_t\}_{t\in\mathbb{Z}}$ is a unit variance strong WN, and

$$X_t = \sigma_t W_t, \quad \sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2$$

where $\omega, \alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q \geq 0$, then $X_t \sim GARCH(p,q)$.
Model proposed by Bollerslev, 1986.

Proposition 2.12.1 — Properties of GARCH. Suppose for the moment that there exists a stationary and causal time series X_t satisfying the GARCH(p,q) model

$$X_t = g(W_t, \cdots) \implies \sigma_t^2 = g_\sigma(W_{t-1}, \cdots)$$

1. $\mathbb{E}[X_t] = \mathbb{E}[\sigma_t]\mathbb{E}[W_t] = 0$
$$\gamma_X(h) = \mathbb{E}[X_{t+h}X_t] = \mathbb{E}[\sigma_{t+h}W_{t+h}\sigma_t W_t] = 0$$
GARCH series have mean 0 and are serially uncorrelated in this case.
2. Suppose $X_t \sim ARCH(1)$,
$$\begin{aligned} X_t^2 &= \sigma_t^2 W_t^2 = \sigma_t^2(W_t^2 + 1 - 1) = \sigma_t^2 + \sigma_t^2(W_t^2 - 1) \\ &= \omega + \alpha_1 X_{t-1}^2 + \underbrace{\sigma_t^2}_{g(W_{t-1},\cdots)} \underbrace{W_t^2 - 1}_{\text{mean } 0} && \text{last part is a weak WN} \\ \implies & X_t^2 \sim AR(1) \end{aligned}$$

In general if $X_t \sim GARCH(p,q)$, then X_t^2 follows an ARMA model with weak WN innovations.

$\{X_t\}_{t\in\mathbb{Z}} \sim GARCH(p,q)$, then $\{X_t\}_{t\in\mathbb{Z}}^2$ is serially correlated.

2.12.1 Stationarity Conditions of GARCH Models

Suppose $X_t \sim GARCH(p,q)$ model

$$X_t = \sigma_t W_t, \mathbb{E}[W_t] = 0, \mathbf{Var}\,(W_t) = \mathbb{E}[W_t^2] = 1,\ \sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2$$

where $\omega, \alpha_i, \beta_j \geq 0$. The question is, under what conditions on $\alpha_i, \beta_j, \omega$, does a stationary process $\{X_t\}_{t\in\mathbb{Z}}$ satisfyies these equations?

R Suppose a stationary solution exists that is a causal Bernoulli shift,

$$X_t = g(W_t, W_{t-1}, \cdots)$$

then $\sigma_t^2 = g_\sigma(W_{t-1}, W_{t-2}, \cdots)$. If $\mathbf{Var}\,(X_t) < \infty$, note that $\mathbf{Var}\,(X_t) = \mathbf{Var}\,(\sigma_t W_t) = \mathbb{E}[\sigma_t^2 W_t^2] = \mathbb{E}[\sigma_t^2] = \sigma_X^2$. Using the GARCH recursion,

$$\mathbb{E}[\sigma_t^2] = \omega + \sum_{j=1}^{p} \alpha_j \mathbb{E}[X_{t-j}^2] + \sum_{k=1}^{q} \beta_k \mathbb{E}[\sigma_{t-k}^2]$$

$$\sigma_X^2 = \omega + \sum_{j=1}^{p} \alpha_j \sigma_X^2 + \sum_{k=1}^{q} \beta_k \sigma_X^2$$

solving this gives

$$\sigma_X^2 = \frac{\omega}{1 - \sum_{j=1}^{p} \alpha_j - \sum_{k=1}^{q} \beta_k}$$

this suggests that in order for a solution to exists in L^2, we need

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$$

proposed by Bollerslew, 1986.

Consider GARCH(1,1) case,

$$X_t = \sigma_t W_t,\ \ \sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2$$

In order to get stationary solution, we just need a stationary conditional variance process. Let $a(z) = \alpha z^2 + \beta$. Iterate GARCH recursion,

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2 = \omega + \alpha(\sigma_{t-1}^2 W_{t-1}^2) + \beta \sigma_{t-1}^2 \\
&= \omega t + \left[\alpha W_{t-1}^2 + \beta\right] \sigma_{t-1}^2 = \omega + a(W_{t-1}) \sigma_{t-1}^2 \\
&= \omega + \omega a(W_{t-1}) + \omega a(W_{t-1}) a(W_{t-2}) + \cdots \\
&= \omega \left(1 + \sum_{j=1}^{\infty} \prod_{j=1}^{i} a(W_{t-j}) \right) \\
&= g_\sigma(W_{t-1}, \cdots)
\end{aligned}$$

The solution must be of the form

$$\sigma_t^2 = \omega \left(1 + \sum_{j=1}^{\infty} \prod_{j=1}^{i} a(W_{t-j}) \right)$$

when is this well-defined (analysis time! who's excited?)? Note that

$$\prod_{j=1}^{i} a(W_{t-j}) = \exp\left\{ \underbrace{\sum_{i=1}^{j} \log(a(W_{t-i}))}_{\text{random walk}} \right\}$$

if $\mathbb{E}[\log(a(W_0))] > 0$, then $\sum_{i=1}^{j} \log(a(W_{t-i})) \to +\infty$ with probability 1. If $\mathbb{E}[\log(a(W_0))] < 0$, then $\sum_{i=1}^{j} \log(a(W_{t-i})) \to +\infty < -\infty$, this is good. If $\mathbb{E}[\log(a(W_0))] = 0$, we have oscillation, which is not good.

Theorem 19 A stationary solution $\{X_t\}_{t\in\mathbb{Z}}$ exist to the GARCH(1,1) equations if and only if

$$\gamma = \mathbb{E}[\log\left(\alpha W_0^2 + \beta\right)] < 0$$

the top Lyapounov Exponent. The solution is of the form

$$X_t = \sigma_t W_t, \ \ \sigma^2 = \omega\left(1 + \sum_{j=1}^{\infty} \prod_{j=1}^{i} \alpha W_{t-j}^2 + \beta\right))$$

which is a Bernoulli shift but not linear.

R

1. If $\gamma < 0$, $\omega = 0$ forces $X_t \equiv 0$, so normally, we assume $\omega > 0$.
2. Condition $\gamma = \mathbb{E}[\log\left(\alpha W_0^2 + \beta\right)] < 0$ depends on the distribution of W_t.
3. A sufficient condition (existence) is $\alpha_1 + \beta_1 < 1$.

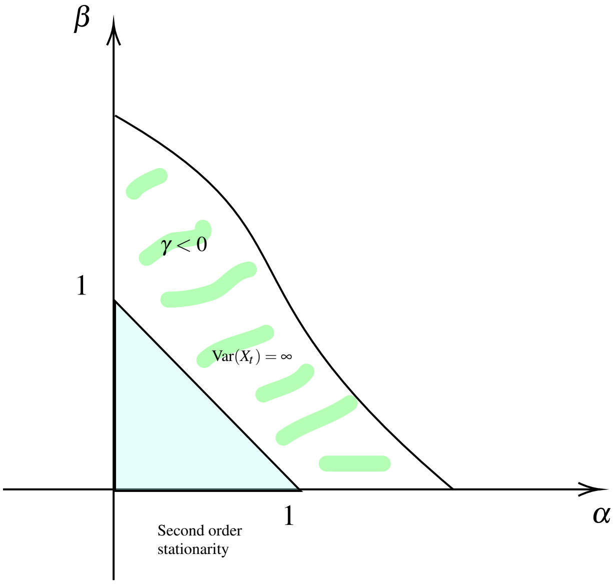
Proof. By Jensen's Inequality on convex/concave function, then

$$\mathbb{E}[\log\left(\alpha W_0^2 + \beta\right)] \leq \log\left(\mathbb{E}[\alpha W_0^2] + \beta\right) = \log(\alpha + \beta) < 0 \Longrightarrow \alpha + \beta < 1$$

■

4. on second order stationarity: if $\alpha_1 + \beta_1 > 1$, we have seen that **Var** (X_t) is not well-defined. If $\alpha_1 + \beta_1 < 1$, then $\mathbb{E}[\sigma^2] < \infty$ and it equals to $\frac{\omega}{1-\alpha-\beta}$. Assuming $\alpha_1 + \beta_1 < 1$, X_t is a weak WN since $\gamma_X(h) = 0$.

Region of Stationarity for GARCH(1,1)



Stationarity of General GARCH(p,q)

General conditions exist for when a GARCH(p,q) process has a strictly stationary solution: let

$$\begin{aligned}
\tau_t &= (\beta_1 + \alpha_1 W_t^2, \beta_2, \cdots, \beta_{q-1}) \in \mathbb{R}^{q-1} \\
\xi_t &= (X_t^2, 0, \cdots, 0) \in \mathbb{R}^{q-1} \\
\alpha &= (\alpha_2, \cdots, \alpha_{p-1}) \in \mathbb{R}^{p-2} \\
I_c &= c \times c \text{ Identity} \\
N &= (\omega, 0, \cdots, 0) \in \mathbb{R}^{p+q-1} \\
Y_t &= (\sigma_t^2, \cdots, \sigma_{t-q+1}^2, X_t^2, \cdots, X_{t-p+1}^2) \in \mathbb{R}^{p+q-1}
\end{aligned}$$

and

$$M_t = \begin{bmatrix} \tau_t & \beta_q & \alpha & \alpha_p \\ I_{q-1} & 0 & 0 & 0 \\ \xi_t & 0 & 0 & 0 \\ 0 & 0 & I_{p-2} & 0 \end{bmatrix} \in \mathbb{R}^{p+q-1 \times p+q-1}$$

Theorem 20 X_t solves the GARCH(p,q) equations if and only if

$$Y_t = M_t Y_{t-1} + N$$

This representation is known as the Markov representation of the GARCH equations. This defines a first order vector AR for Y_t with random matrix coefficients M_t. Let A_t be a stationary sequence of random $p+q-1 \times p+q-1$ matrices, and define, for an arbitrary norm on matrices $\|\cdot\|$, the scalar random variables

$$r_t = \|A_t A_{t-1} \cdots A_t\|$$

under some relative mild conditions (ergodicity)

$$\gamma = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[\log(r_t)]$$

is well defined and is called the top Lyapounov exponent of the sequence $A_t, t \in \mathbb{Z}$. This result is coming from Ergodic theory in the 1970s.

Theorem 21 A stationary solution to the GARCH(p,q) equation exists if and only if $\gamma < 0$ where γ is the top Lyapounov exponent of sequence $M_t, t \in \mathbb{Z}$ appearing in the Markov representation. When a stationary solution exists, it is causal and unique.

Theorem 22 A necessary and sufficient condition for there to exist a second order stationary solution to the GARCH(p,q) equations is that

$$\sum_{j=1}^{p} \alpha_j + \sum_{l=1}^{q} \beta_l < 1$$

2.12.2 Identifying GARCH Models

The decision to fit a volatility (GARCH) model to a time series often arises from

1. Observing volatility (conditional heteroskedasticity)
2. Conditional variance forecasting is of specific interest (risk analysis, financial time series analysis)

If strong serial correlation is observed in the series, one often fits initially an ARMA model, and then fits a GARCH model to the residual.

Identifying Serial Correlation:

Recall that the normal ACF bounds (blue lines) are constructed based on the assumption that the series is a strong WN. A GARCH series is a weak WN.
ACF bounds for weak white noise: suppose $X_t \sim GARCH(1,1)$, then $\gamma_X(h) = 0, h \geq 1$ and

$$\hat{\gamma}_X(h) \approx \frac{1}{T} \sum_{j=1}^{T-h} X_t X_{t+h} \Longrightarrow \mathbb{E}[\hat{\gamma}_X(h)] = 0$$

note that (after some calculations)

$$\mathbf{Var}\left(\sqrt{T}\hat{\gamma}_X(h)\right) = \frac{1}{T} \sum_{j=1}^{T-h} \mathbb{E}[X_{j+h}^2 X_j^2] \approx \mathbb{E}[X_0^2 X_{-h}^2]$$

which does not simplify to σ_X^4.

Theorem 23 If X_t is a weak WN (suitably weakly dependent), then

$$\sqrt{T}\hat{\gamma}_X(h) \xrightarrow{D} N(0, \mathbb{E}[X_0^2 X_{-h}^2])$$

as $T \to \infty$.

R

1. $\mathbb{E}[X_0^2 X_{-h}^2]$ can be consistently estimated by

$$\hat{\sigma}_h^2 = \frac{1}{T} \sum_{j=1}^{T-h} X_{j+h}^2 X_j^2$$

an approximate $1-\alpha$ prediction interval for $\hat{\rho}(h)$ under the assumption of a weak white noise is

$$\pm \frac{1}{\sqrt{T}} Z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_h}{\hat{\gamma}(0)}$$

the blue lines depend on h now!

2. Note that
$$\mathbb{E}[X_0^2 X_{-h}^2] = \mathbb{E}^2[X_0^2] + \underbrace{\mathbf{Cov}(X_0^2, X_{-h}^2)}_{>0}$$
Thus, in GARCH setting, the weak white noise intervals for ACF are often wider.

2.12.3 Tests for Homoscedasticity

Conditional heteroscedasticity is characterized by correlation in the squared sequence X_t^2.

Method to evaluate conditional heteroscedasticity

We can apply a WN test to sequence X_t^2.

Theorem 24 — Portmenteau. Let $\hat{\rho}_{X^2}(h)$ denote the empirical ACF of the series $X_t^2, t = 1, \cdots, T$. Then if X_t is a strong WN with $\mathbb{E}[X_t^4] < \infty$,

$$Q(T,H) = T\sum_{h=1}^{H} \hat{\rho}_{X^2}(h) \xrightarrow{D} \chi^2(H)$$

If $X_t \sim GARCH(p,q)$ model, then $Q(T,H) \xrightarrow{P} \infty$. The p-value of test of homoscedasticity vs. conditional heteroscedasticity is

$$p = \mathbb{P}(\chi^2(H) \geq Q(T,H))$$

R

1. This test has several names in the literature, including Li-Mcleod Test.
2. Often it is applied to the GARCH residuals in order to evaluate goodness-of-fit of a GARCH model (and decide on p,q).

2.12.4 Estimating GARCH parameters

When considering ARCH(1), we have

$$X_t^2 \sim AR(1), \;\; X_t^2 = \omega + \alpha X_{t-1}^2 + V_t, \;\; V_t = \sigma_t^2(W_t^2 - 1)$$

. This suggests estimating ω, α using least squares.

$$(\hat{\omega}, \hat{\alpha}) = \arg\min_{\omega \geq 0, 0 < \alpha < 1} \sum_{t=2}^{T} (X_t^2 - (\omega + \alpha X_{t-1}^2))^2$$

R This leads to consistent estimation.

For a general ARCH(p) model, we can also use least squares with the loss function

$$L(\vec{\alpha}) = \sum_{j=p+1}^{T} (X_j^2 - (\omega + \alpha_1 X_{j-1}^2 + \cdots + \alpha_p X_{j-p}^2))^2$$

where $\vec{\alpha} = (\omega, \alpha_1, \cdots, \alpha_p)^\top$. This is minimized by

$$\hat{\vec{\alpha}} = (X^\top X)^{-1} X^\top Y$$

where

$$X = \begin{bmatrix} 1 & X_p^2 & \cdots & X_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{T-1}^2 & \cdots & X_{T-p}^2 \end{bmatrix}, \; Y = [X_{p+1}^2, \cdots, X_T^2]^\top$$

This really looks like the design matrix in MLR.

Theorem 25 — Chapter 6, Francq & Zakoian. The OLS estimators of the ARCH(p) are consistent if $\mathbb{E}[X_t^4] < \infty$, and are $\sqrt{T}-$consistent and asymptotically Gaussian if $\mathbb{E}[X_t^8] < \infty$ under "regularity conditions" including:

1. The true ARCH parameters admit a stationary and causal solution
2. The innovations $\{W_t\}_{t\in\mathbb{Z}}$ have a non-degenerate distribution.

Quasi-Maximum Likelihood Estimation for GARCH(p,q)

Suppose $X_t \sim ARCH(1)$ with

$$X_t = \sigma_t W_t, \;\; \sigma_t^2 = \omega + \alpha X_{t-1}^2$$

with additional parametric assumption of $W_t \sim N(0,1)$. Assuming the model admits a stationary and causal solution with $\omega \geq 0, 0 \leq \alpha < 1$, then

$$\underbrace{X_t | X_{t-1}}_{\sigma_t^2 \text{ is known here}} \sim N(0, \omega + \alpha X_{t-1}^2)$$

Then,

$$L(\omega, \alpha) = \prod_{t=2}^{T} L(\omega, \alpha, X_t | X_{t-1}, \cdots, X_t)$$

we can find ω, α to maximize this with numerical methods.
In the general GARCH(p,q) case, we can define

$$X_t | X_{t-1}, \cdots, X_t \overset{P}{\approx} X_t | \underbrace{X_{t-1}, X_{t-2}, \cdots}_{\text{infinite past}} \sim N(0, \sigma_t^2)$$

where

$$\sigma_t^2(\omega, \vec{\alpha}, \vec{\beta}) = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{l=1}^{q} \beta_l \sigma_{t-l}^2$$

thus,

$$L(\omega, \vec{\alpha}, \vec{\beta}) = \prod_{j=\max\{p,q\}=1}^{T} \underbrace{f_{\omega,\vec{\alpha},\vec{\beta}}(X_j | X_{j-1}, \cdots, X_1)}_{\substack{\text{conditional Gaussian density with} \\ N(0, \sigma_j^2(\omega, \vec{\alpha}, \vec{\beta}))}}$$

The catch here is, as the equation $\sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{l=1}^{q} \beta_l \sigma_{t-l}^2$ is iterated to calculated the conditional likelihood. Eventually, things arise that are unknown are $\{X_j : j \leq 0\}$ and $\left\{\sigma_j^2 : j \leq 0\right\}$. This requires some initialization here. Two possible choices:

1. $\sigma_t^2 = \omega, \; X_t^2 = \omega, t \leq 0$
2. $\sigma_t^2 = \omega, \; X_t^2 = 0, t \leq 0$

Note that if the series is quite long, the initialization will not have much of an effect.
WARNING: be very careful when fitting a GARCH to a short series as the initialization has a huge impact.

Parameter Constraints

We need to find

$$(\hat{\omega},\hat{\vec{\alpha}},\hat{\vec{\beta}}) = \arg \min_{\text{admitting stationary solution}} L(\omega,\vec{\alpha},\vec{\beta})$$

How do we define this search space?

1. "Hyper-pyramid":

$$\left\{(\omega,\vec{\alpha},\vec{\beta}) : \omega > 0, \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1, \alpha_i, \beta_j \geq 0\right\}$$

   the solution is second order stationary.

2. $$\left\{(\omega,\vec{\alpha},\vec{\beta}) : \text{Top Lyapounov exponent} < 0\right\}$$

   this is the entire stationary parameter region. Much more harder to search.

Theorem 26 — Property of GARCH Estimators. If $X_t \sim GARCH(p,q)$ and it admits a stationary and causal solution. The Quasi-MLE (QMLE) estimators are consistent. If $W_t \sim N(0,1)$ in fact, QMLE becomes MLE, then the estimators are efficient, which means it achieves the smallest variance among consistent estimators. If $W_t \not\sim N(0,1)$, the QMLE may not be efficient, but it is in several special cases.

The takeaway here is that QMLE estimation is the benchmark of GARCH model parameter estimation.

2.12.5 Forecasting the Conditional Variance and GARCH Residuals

If $X_t \sim GARCH(p,q)$ model, $(\omega,\vec{\alpha},\vec{\beta})$ can be estimated using QMLE and get $(\hat{\omega},\hat{\vec{\alpha}},\hat{\vec{\beta}})$. These provide us estimates of conditional variance:

$$\hat{\sigma}_t^2 = \hat{\omega} + \sum_{j=1}^{p} \hat{\alpha}_j X_{t-j}^2 + \sum_{l=1}^{q} \hat{\beta}_l \hat{\sigma}_{t-l}^2,\ q+1 \leq t \leq T$$

for the prior variances,

$$\hat{\sigma}_j^2 = \hat{\omega} + \sum_{l=1}^{\min(j,p)} \hat{\alpha}_l X_{j-l}^2,\ 1 \leq t \leq q$$

The GARCH residuals can be estimated using

$$\hat{W}_t = \frac{X_t}{\hat{\sigma}_t}$$

we can use these to perform model diagnostics to check

1. Whiteness or squared correlation
2. Normality for interval estimations
3. These may also be used in bootstrap procedures

To forecast the conditional variance, we have

1. **1-step:**

$$\hat{\sigma}_{T+1}^2 = \hat{\omega} + \sum_{j=1}^{p} \hat{\alpha}_j X_{T+1-j}^2 + \sum_{l=1}^{q} \hat{\beta}_l \hat{\sigma}_{T+1-l}^2$$

   with initialization: $X_t^2 = \hat{\omega}, \hat{\sigma}_t^2 = \omega, t \leq 0$

2. $h-$**step:**

$$\hat{\sigma}^2_{T+h} = \hat{\omega} + \sum_{j=1}^{p} \hat{\alpha}_j X^2_{T+h-j} + \sum_{l=1}^{q} \hat{\beta}_l \hat{\sigma}^2_{T+h-l}$$

which is an iterative process. Moreover,

$$\hat{X}^2_t = \begin{cases} X^2_t & t \leq 0 \\ \hat{\omega} \text{ or } \frac{\hat{\omega}}{1-\sum_{j=1}^{p}\hat{\alpha}_j-\sum_{l=1}^{q}\hat{\beta}_l} & \end{cases}$$