STAT 443 - Forecasting

Cameron Roopnarine

Last updated: January 30, 2021

Contents

Chapter 1

Characteristics of Time Series

1.1 What is a time series?

In classical statistics, we normally consider $X_1, \dots, X_n \in \mathbf{R}^p$, a **simple random sample**.

In particular,

(1) $X_1, \dots, X_n$ are i.i.d. (independent and identically distributed)

(2) $X_i \sim F_\theta$ which is a common distribution characterized by θ.

Examples:

1. $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and we wish to estimate and perform inference on μ and σ^2.
2. $X_i = \begin{bmatrix} Y_i \\ Z_i \end{bmatrix}$ where Y_i is a dependent variable, and Z_i is an independent variable. Perhaps we happen to observe Y_i and Z_i in pairs, and we posit a model:

$$Y_i = \beta^\top Z_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$$

REMARK 1.1.1

The relationship between Y_i and Z_i doesn't depend on i, it only depends through the common parameter β, and it assumes that ε_i has fixed variance for each i.

3. In such settings, one is typically interested in:
   (a) Prediction: based on the data, how can we predict the behaviour of these variables in the future?
   (b) Inference: how do we use the data to try to estimate and better understand the underlying mechanism which generates the data? For example, a linear model or simple Gaussian model.

DEFINITION 1.1.2: Time series

We say $X_1, \dots, X_T$ 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.

DEFINITION 1.1.3: Real-valued time series

If $X_t \in \mathbf{R}$, we say $X_1, \dots, X_T$ is a **real-valued (scalar) time series**.

DEFINITION 1.1.4: Multivariate time series

If $X_t \in \mathbf{R}^p$, we say $X_1, \dots, X_T$ is a **multivariate (vector-valued) time series**.

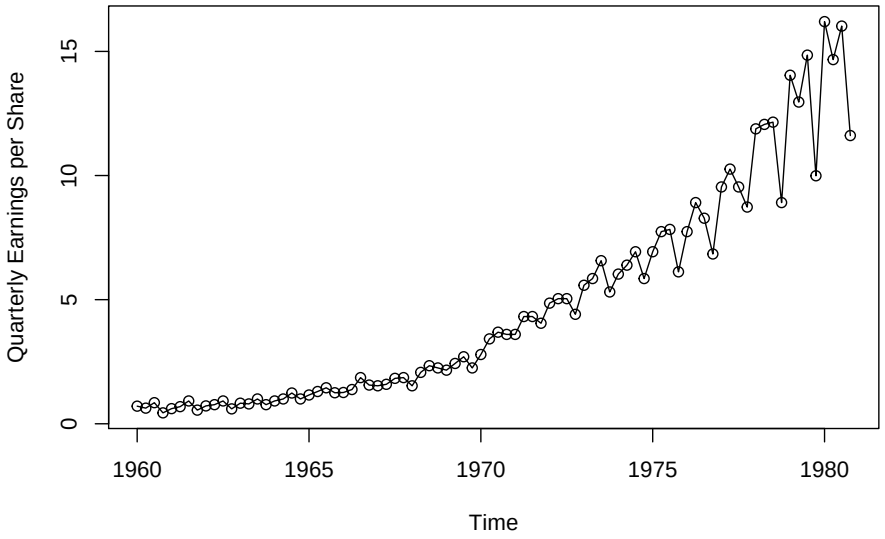


Figure 1.1: Quarterly Johnson and Johnson Earnings

```
# Figure 1.1
plot(jj, type = "o", ylab = "Quarterly Earnings per Share")
```

Observe that in Figure 1.1:

- The earnings are steadily increasing over time.
- There is heterogeneity in the variance over time.

With time series data, we are typically concerned with the same goals as in classical statistics (prediction and inference). However, in contrast with time series, the data often exhibit:

(1) **Heterogeneity**

- Time trends $\rightarrow \mathbb{E}[X_t] \neq \mathbb{E}[X_{t+h}]$.
- Heteroskedasticity $\rightarrow \mathbb{V}(X_t) \neq \mathbb{V}(X_{t+h})$.

In classical statistics, it's assumed that all the observations have the same distribution which is clearly not the case in time series.

(2) **Serial Dependence (Serial Correlation)**

- Observations that are temporally close appear to depend on each other.

In classical statistics, each successive observation is assumed to be independent which is clearly not the case in time series.

```
# Figure 1.2
plot(gtemp, type = "o", ylab = "Global Temperature Deviations")
```

Observe that in Figure 1.2:

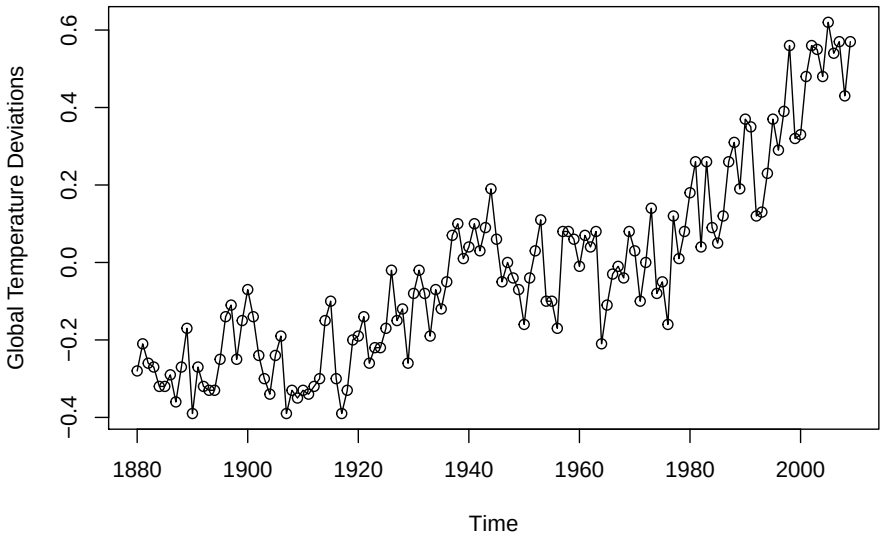


Figure 1.2: x_t is the deviation of global mean yearly temperature from the mean computed from 1951–1980

- The global temperature is steadily increasing over time.
- Heterogeneity exists within the mean over time.
- Heterogeneity exists within the variance over time, although it is not very apparent.
- Serial dependence occurs.

Let's formally define a time series.

DEFINITION 1.1.5: Time series

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **time series** if $\{X_t : t \in \mathbf{Z}\}$ is a stochastic process indexed by **Z**. In other words, there is a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_t : \Omega \to \mathbf{R}$ is a random variable for all t.
In relation to the original definition, we say $X_1, \dots, X_T$ is an **observed stretch** (**realization**, **simple path**) of length T from $\{X_t\}_{t\in\mathbf{Z}}$.

Formally speaking, we think of a time series as being a little snippet of one long sample path the stochastic process for which would characterize all the serial dependence, time trends, and heteroskedasticity that exist within a time series.

1.2 Basic Principles of Forecasting

Consider a time series of length T, namely $X_1, \dots, X_T$. Based on $X_1, \dots, X_T$, we would like to produce a "best guess" for X_{T+h}:

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

DEFINITION 1.2.1: Forecast, Horizon

For $h \geq 1$, our "best guess"

$$\hat{X}_{T+h} = f_h(X_T, \dots, X_1)$$

is called a **forecast** of X_{T+h} at **horizon** h.

Goals of Forecasting

Goal 1

- Choose f_n "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 the true value, X_{T+h}. We then wish to choose f_h so that $L(X_{T+h}, f_h(X_T, \dots, X_1))$ is minimized, where $L(\cdot, \cdot)$ is a loss function.

EXAMPLE 1.2.2

The most common measure of $L(\cdot, \cdot)$ is the **mean-squared error** (MSE), defined by

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

Goal 2

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

EXAMPLE 1.2.3: Why is it important to quantify uncertainty?

Suppose every minute, we flip a coin and denote
- (Heads): $H \to 1$
- (Tails): $T \to -1$
- $X_t =$ outcome in minute t, where $t = 1, \dots, T$.

This produces a time series of length T, which is a random sequence of (1)'s and (-1)'s. Note $\mathbb{E}[X_t] = 0$ for all t. If we wish to forecast for $h \geq 1$, consider $\hat{X}_{T+h} = f(X_T, \dots, X_1)$, thus

$$\begin{aligned} L(X_{T+h}, \hat{X}_{T+h}) &= \mathbb{E}[(X_{T+h} - \hat{X}_{T+h})^2] \\ &= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] - 2\mathbb{E}[X_{T+h}\hat{X}_{T+h}] \\ &= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] - 2\mathbb{E}[X_{T+h}]\mathbb{E}[\hat{X}_{T+h}] \\ &= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] \end{aligned}$$

Note that we can write $\mathbb{E}[X_{T+h}\hat{X}_{T+h}] = \mathbb{E}[X_{T+h}]\mathbb{E}[\hat{X}_{T+h}]$ since $\hat{X}_{T+h}$ is a function of the data $X_T, \dots, X_1$, and hence independent of X_{T+h}.
Furthermore, note that $\mathbb{E}[X_{T+h}^2] = \mathbb{V}(X_t)$ since $\mathbb{E}[X_{T+h}] = 0$.
We can minimize this by taking $\hat{X}_{T+h} = 0$. There's nothing "wrong" with this forecast, but ideally we would also be able to say that the sequence appears to be random, and that we don't expect this forecast to be close to the actual value.
Furthermore, for this basic reason, one can always argue that any forecast that's not accompanied with some type of quantification of how close we expect the forecast to be, is at very least hard to interpret; at worst, meaningless because it doesn't describe the accuracy for which we expect the forecast to perform.

How can we quantify the uncertainty in forecasting?

Ideal: The predictive distribution, that is,

$$X_{T+h} \mid X_T, \dots, X_1$$

Excellent: Predictive intervals/sets, that is, for some $\alpha \in (0,1)$ find an interval I_α such that

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

A common example is with $\alpha = 0.95$. Often times, such intervals take the form

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

Concluding Remarks:

1. Estimating predictive distribution leads one towards *estimating* the joint distribution of

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

   For example, the ARMA and ARIMA models.

2. It is important that we acknowledge that some things cannot be predicted!

"It's tough to make predictions, especially about the future."—Yogi Berra

1.3 Definitions of Stationary

Given a time series $X_1, \dots, X_T$, we are frequently interested in estimating the joint distribution of

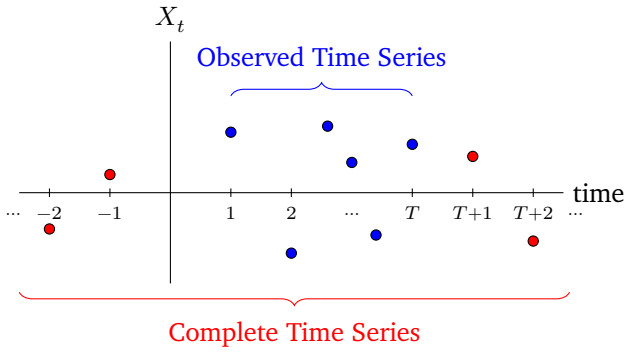
$$X_{T+h}, X_T, \dots, X_1$$

which is useful for forecasting and inference.

The joint distribution is a feature of the process $\{X_t\}_{t \in \mathbf{Z}}$

$$X_1, \dots, X_T \xrightarrow[\text{infer}]{} \{X_t\}_{t \in \mathbf{Z}}$$

- $X_1, \dots, X_T$: Observed data.
- $\{X_t\}_{t \in \mathbf{Z}}$: Stochastic process.



Worst case: $X_t \sim F_t$, where F_t is a *changing* function of t. If so, it is hard to pool the data $X_1, \dots, X_T$ to estimate F_t. If **serial dependence** occurs; that is, if the distribution of (X_t, X_{t+h}) depends strongly on t, then we have a similar problem in estimating e.g. $\text{Cov}(X_t, X_{t+h})$.

DEFINITION 1.3.1: Strictly stationary

We say that a time series $\{X_t\}_{t\in\mathbf{Z}}$ is **strictly stationary** (**strongly stationary**) if for each $k \geq 1$, $i_1, \dots, i_k, h \in \mathbf{Z}$,

$$(X_{i_1}, \dots, X_{i_k}) \overset{\mathrm{d}}{=} (X_{i_1+h}, \dots, X_{i_k+h})$$

If we look at the k-dimensional joint distribution $(X_{i_1}, \dots, X_{i_k})$ of the series at points $i_1, \dots, i_k$, then strict stationary means this is shift-invariant. That is, shifting the window on which you view the data, does not change its distribution. This implies that if F_t = CDF of X_t, then $F_t = F_{t+h} = F$ that is, all variables have a common distribution function.

DEFINITION 1.3.2: Mean function

For a time series $\{X_t\}_{t\in\mathbf{Z}}$, with $\mathbb{E}[X_t^2] < \infty$ for all $t \in \mathbf{Z}$, we denote the **mean function** of the time series as

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

DEFINITION 1.3.3: Autocovariance function

The **autocovariance** function of the time series $\{X_t\}_{t\in\mathbf{Z}}$ is defined as

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

DEFINITION 1.3.4: Weakly stationary, Lag

We say that a time series $\{X_t\}_{t\in\mathbf{Z}}$ is **weakly stationary** if $\mathbb{E}[X_t] = \mu$ which does not depend on t, and if

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

that is, $\gamma(t,s)$ is a function of $|t-s|$. In this case, we usually write

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

and we call the input h the **lag** parameter.

Additional terminology:

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

$$\begin{aligned} \gamma(h) &= \mathrm{Cov}(X_t, X_{t+h}) \\ &= \mathrm{Cov}(X_{t-h}, X_{t-h+h}) \qquad\qquad t \to (t-h) \\ &= \mathrm{Cov}(X_t, X_{t-h}) \\ &= \gamma(-h) \end{aligned}$$

  Since $\gamma(h) = \gamma(-h)$, we normally, we only record $\gamma(h)$ for $h \geq 0$.

1.4 White Noise and Stationary Examples

DEFINITION 1.4.1: Strong white noise

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **strong white noise** if $\mathbb{E}[X_t] = 0$ and the $\{X_t\}_{t\in\mathbf{Z}}$ are i.i.d.

DEFINITION 1.4.2: Weak white noise

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **weak white noise** if $\mathbb{E}[X_t] = 0$ and

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

DEFINITION 1.4.3: Gaussian white noise

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

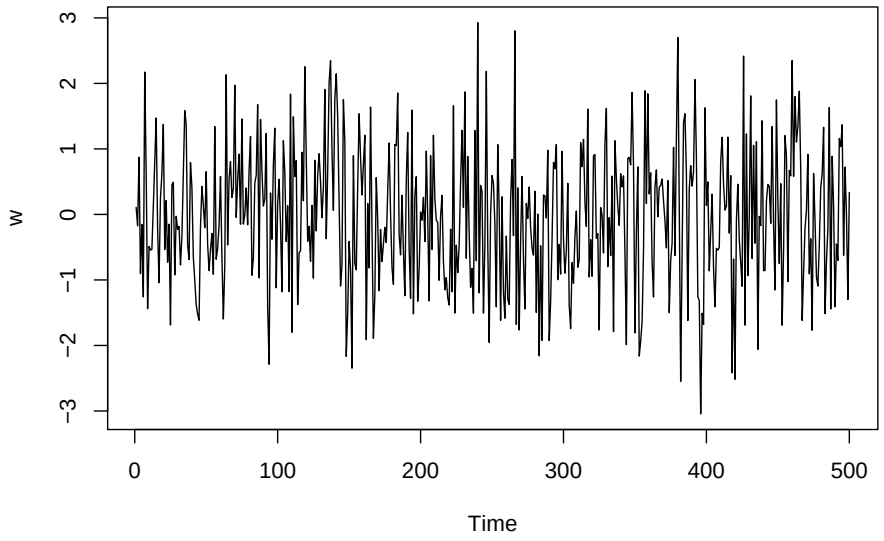


Figure 1.3: Gaussian White Noise of Length 500

```
# Figure 1.3
plot.ts(rnorm(500), main = "Gaussian White Noise", ylab = "w")
```

Figure 1.3 is a Gaussian *white* noise series. **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.

EXAMPLE 1.4.4

Suppose $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise, then $\mathbb{E}[W_t] = 0$; that is, it doesn't depend on t.

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

only depends on $|t-s|$. Therefore, $\{W_t\}_{t\in\mathbf{Z}}$ is **weakly stationary**. Furthermore, we claim that $\{W_t\}_{t\in\mathbf{Z}}$

is **strictly stationary**. Let $k \geq 1$, $i_1, \dots, i_k, h \in \mathbf{Z}$ with $i_1 < \cdots < i_k$, then

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

EXAMPLE 1.4.5

Suppose $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise. Define $X_t = W_t + \theta W_{t-1}$ for $\theta \in \mathbf{R}$. Since $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise, we have $\mathbb{E}[W_t] = 0$ for all t, hence we have $\mathbb{E}[X_t] = \mathbb{E}[W_t + \theta W_{t-1}] = \mathbb{E}[W_t] + \theta\mathbb{E}[W_{t-1}] = 0$ which is first order stationary.

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

We obtain these calculations as follows:

- $|t-s| = 0$.

$$\mathbb{E}[(W_t + \theta W_{t-1})^2] = \mathbb{E}[W_t^2] + \theta^2\mathbb{E}[W_{t-1}^2] + 2\mathbb{E}[\theta W_t W_{t-1}] = (1+\theta^2)\sigma_W^2$$

  since W_t is independent of W_{t-1}. The calculation is easy to verify.
- $t = s+1$ (or $s = t+1$).

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

  since W_{s+1} is independent of W_s and W_{s-1}. The calculation is easy to verify.
- $|t-s| > 1$. $W_t + \theta W_{t-1}$ is independent of $W_s + \theta W_{s-1}$.

We claim that $\{X_t\}_{t\in\mathbf{Z}}$ is also strictly stationary. Let $k \geq 1$, $i_1, \dots, i_k, h \in \mathbf{Z}$ with $i_1 < \cdots < i_k$, then

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

where B is some subset of $\mathbf{R}^{i_k - i_1 + 1}$, and hence is shift-invariant.

DEFINITION 1.4.6: Bernoulli shift

Suppose $\{\varepsilon_t\}_{t\in\mathbf{Z}}$ is a strong white noise. If $X_t = g(\varepsilon_t, \varepsilon_{t-1}, \dots)$ for some function $g : \mathbf{R}^\infty \to \mathbf{R}$, we say that $\{X_t\}_{t\in\mathbf{Z}}$ is a **Bernoulli shift**.

THEOREM 1.4.7

If $\{X_t\}_{t\in\mathbf{Z}}$ is a Bernoulli shift, then $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary.

REMARK 1.4.8

Norbert Wiener conjectured that **every** stationary sequence is a Bernoulli shift, which is not true. The truth is, almost every one is.

EXERCISE 1.4.9

Suppose $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise. The **two-sided random walk** is defined as

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

Show that $\{X_t\}_{t\in\mathbf{Z}}$ is first order stationary, but $\{X_t\}_{t\in\mathbf{Z}}$ is not second order stationary.
Solution. $\{X_t\}_{t\in\mathbf{Z}}$ is first order stationary since

$$\begin{aligned}
\mathbb{E}[X_t] &= \mathbb{E}\left[\sum_{i=0}^{t} W_i + \sum_{i=t}^{-1} W_i\right] \\
&= \mathbb{E}[W_0 + W_1 + \cdots + W_{t-1} + W_t + W_t + W_{t-1} + \cdots + W_0 + W_{-1}] \\
&= \mathbb{E}[W_{-1}] + \mathbb{E}[2W_0] + \mathbb{E}[2W_1] + \cdots + \mathbb{E}[2W_{t-1}] \\
&= 0 + 2(0) + \cdots + 2(0) \\
&= 0
\end{aligned}$$

since $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise; that is, $\mathbb{E}[W_t] = 0$ for all t.
$\{X_t\}_{t\in\mathbf{Z}}$ is not second order stationary since if $t > 0$ the second sum is simply $\sum_{i=t}^{-1} W_i = 0$ and we have

$$\begin{aligned}
\mathbb{E}[(X_t - \mu_t)(X_t - \mu_t)] &= \mathbb{E}[X_t^2] \\
&= \mathbb{E}\left[\left(\sum_{i=0}^{t} W_i\right)^2\right] \\
&= \mathbb{E}[W_0^2] + \cdots + \mathbb{E}[W_t^2] && \text{since } W_i \perp\!\!\!\perp W_j \text{ for } i \neq j \\
&= t\sigma_W^2
\end{aligned}$$

which depends on t.

1.5 Weak versus Strong Stationary

Sadly, $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary does not imply $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary.

EXAMPLE 1.5.1

Suppose $X_t \overset{\text{iid}}{\sim}$ Cauchy Random Variables; that is,

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

Then, $\mathbb{E}[X_t]$ does not exist, and hence $\{X_t\}_{t\in\mathbf{Z}}$ cannot be weakly stationary. However, $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary in this case since $\{X_t\}_{t\in\mathbf{Z}}$ is a strong white noise.

THEOREM 1.5.2

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

Proof of: Theorem 1.5.2

Note that if $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary, then

$$(X_t) \overset{\mathrm{d}}{=} (X_0)$$

so that $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ which does not depend on t, and also

$$\mathbb{V}(X_t) = \mathbb{V}(X_0)$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\gamma(t,s) &= \mathrm{Cov}(X_t, X_s) \\
&= \mathbb{E}[(X_s-\mu)(X_t-\mu)] \\
&\le \{\mathbb{E}[(X_s-\mu)^2]\}^{1/2}\{\mathbb{E}[(X_t-\mu)^2]\}^{1/2} \\
&= \sqrt{\mathbb{V}(X_s)}\sqrt{\mathbb{V}(X_t)} \\
&= \mathbb{V}(X_t) < \infty
\end{aligned}$$

If $t < s$, then

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

since it is shift-invariant, and hence if we shift everything over by t,

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

DEFINITION 1.5.3: Gaussian process

$\{X_t\}_{t\in\mathbf{Z}}$ is said to be a **Gaussian process** (**Gaussian time series**) if for each $k \ge 1$, $i_1 < i_2 < \cdots < i_k$ we have

$$(X_{i_1}, \ldots, X_{i_k}) \sim \mathrm{MVN}(\boldsymbol{\mu}_k(i_1,\ldots,i_k), \boldsymbol{\Sigma}_{k\times k}(i_1,\ldots,i_k))$$

$$\boldsymbol{\mu}_k = \begin{bmatrix} \mathbb{E}[X_{i_1}] \\ \vdots \\ \mathbb{E}[X_{i_k}] \end{bmatrix} \qquad \boldsymbol{\Sigma}_{k\times k} = \mathrm{Cov}(X_{i_j}, X_{i_r})_{1\le j,r\le k}$$

THEOREM 1.5.4

If $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary and is a Gaussian process, then $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary.

Proof of: Theorem 1.5.4

If $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary, then $\mathbb{E}[X_t] = \mu$ for all t.

$$(X_{i_1}, \ldots, X_{i_k}) \to \begin{bmatrix} \mathbb{E}[X_{i_1}] \\ \vdots \\ \mathbb{E}[X_{i_k}] \end{bmatrix} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}[X_{i_1+h}] \\ \vdots \\ \mathbb{E}[X_{i_k+h}] \end{bmatrix}$$

Also,

$$\begin{aligned}\mathbb{V}(X_{i_1},\dots,X_{i_k}) &= \text{Cov}(X_{i_j},X_{i_r})_{1\le j,\,r\le k}\\ &= \text{Cov}(X_0,X_{i_r-i_j})_{1\le j,\,r\le k}\\ &= \text{Cov}(X_0,X_{i_r+h-(i_j+h)})_{1\le j,\,r\le k}\\ &= \text{Cov}(X_{i_j+h},X_{i_r+h})_{1\le j,\,r\le k}\\ &= \mathbb{V}(X_{i_1+h},\dots,X_{i_k+h})\end{aligned}$$

EXAMPLE 1.5.5

Using the Gaussian assumption

$$(X_{i_1},\dots,X_{i_k}) \overset{\text{d}}{=} \text{MVN}(\boldsymbol{\mu},\boldsymbol{\Sigma}_{k\times k}) \overset{\text{d}}{=} (X_{i_1+h},\dots,X_{i_k+h})$$

Hence $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary in this case.

EXERCISE 1.5.6

Prove that if $\{X_t\}_{t\in\mathbf{Z}}$ is <u>not</u> weakly stationary; that is, either $\mathbb{E}[X_t]$ depends on t or $\gamma(t,s)$ does not depend on the lag, and has a finite mean and variance, then $\{X_t\}_{t\in\mathbf{Z}}$ is <u>not</u> strictly stationary.

1.6 † Theoretical L2 Framework for Time Series

- $X_t = \lim_{h\to\infty} X_{h,t}$. In what sense does this limit exist?
- How "close" are two random variables X and Y?
- Is there a random variable that achieves
$$\inf_{y\in S} d(Y,S)$$

DEFINITION 1.6.1: L^2 space

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

DEFINITION 1.6.2: L^2-time series

We say that $\{X_t\}_{t\in\mathbf{Z}}$ is and L^2-time series if $X_t\in L^2$ for all $t\in\mathbf{Z}$.

L^2 is a Hilbert space when equipped with inner product, $X,Y\in L^2$.

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

$\langle\cdot,\cdot\rangle$ is an inner product since it is

(1) Linear: $\langle aX+bY,Z\rangle = a\langle X,Z\rangle + b\langle Y,Z\rangle$.

(2) "Almost" Positive Definite: $\langle X,X\rangle = \mathbb{E}[X^2]=0 \iff X=0$ almost surely. Which implies $\mathbb{P}(X=0)=1$.

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

L^2 is complete with this inner product; that is, 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 $X_n\to X$; that is, $\mathbb{E}[(X_n-X)^2]\to 0$. This follows from the "famous" Riesz-Fischer Theorem.

Useful Tools for Time Series

(1) **Existence of Limits**

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

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

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

only if $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, then there **must** exist a random variable X_t (by the completeness of L^2), so that

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

(2) **Projection Theorem and Forecasting**. Forecasting can be often cast as finding a random variable Y among a collection of possible forecasts $\mathcal{M}$ (e.g. $\mathcal{M} = \text{Span}(X_T, \dots, X_1)$) so that

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

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

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

must be in the orthogonal complement.

1.7 Signal and 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 processes underlying structure (e.g. its marginal distribution and serial dependence structure)

Most time series are evidently not stationary.

Looking back at Figure 1.1:

- Mean appears to increase, so it is not first order stationary;
- Variability also appears to increase, so it is not second order stationary;
- Therefore, it is not strictly stationary.

Signal and Noise Model: $X_t = S_t + \varepsilon_t$

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

$$\varepsilon_t = g(W_t, W_{t-1}, \dots) \quad \text{[Stationary Noise]}$$

$$\varepsilon_t = g_t(W_t, W_{t-1}, \dots) \quad \text{[Non-stationary Noise]}$$

The terms $\{W_t\}_{t\in\mathbf{Z}}$ are often called the "innovations" or "shocks" during the random behaviour of X_t.

g is used to try to capture noise that can potentially have serial dependence.

EXAMPLE 1.7.1

An example of a function g so that $\varepsilon_t = g_t(W_t, W_{t-1}, ...)$ might be a **random walk**; that is, $\varepsilon_t = \sum_{j=0}^{t} W_j$. Another example could be the **changing variance models**; that is, $\varepsilon_t = \sigma(t)W_t$.

Our goal is to estimate S_t, and then infer the structure of ε_t.

In Figure 1.2, the model appears to be non-stationary (trending upwards over time), so we might try the signal and noise model. We might posit a linear trend, or even higher order functions.

For the temperature data, we may posit that

$$S_t = \beta_0 + \beta_1 t \quad \text{[Linear Trend]}$$

The trend may be estimated by ordinary least squares (OLS). We choose β_0 and β_1 to minimize

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

This can be done in R using the `lm()` command, and can easily be computed with calculus. Figure 1.4 is a small example of the global temperature data superimposed with `lm()`'s estimate.

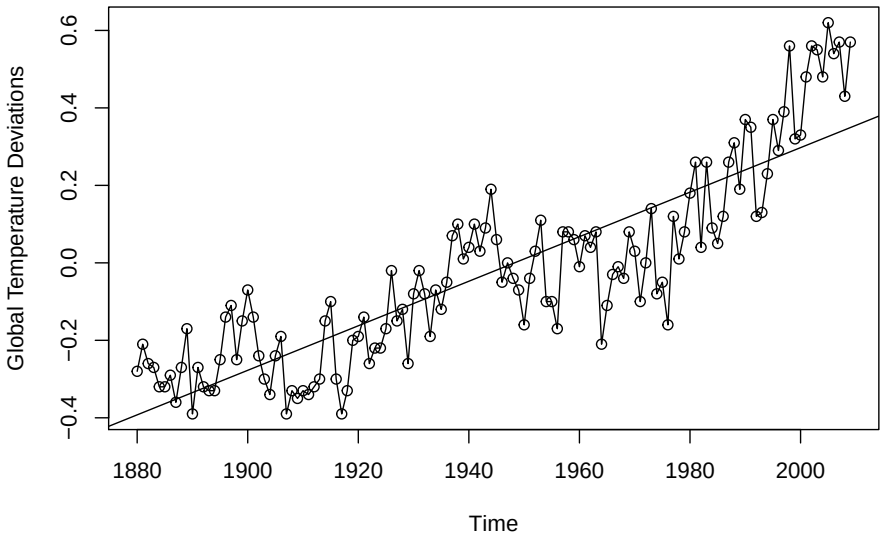


Figure 1.4: OLS estimate of linear trend

```
# Figure 1.4
fit <- lm(gtemp ~ time(gtemp), na.action = NULL)
plot.ts(gtemp, type = "o", ylab = "Global Temperature Deviations")
abline(fit)
```

Let's introduce some terminology about trends.

DEFINITION 1.7.2: Detrended time series

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

1. Estimate $S_t \to \hat{S}_t$
2. Detrend series: $X_t - \hat{S}_t = Y_t$ where Y_t is the "detrended" series.

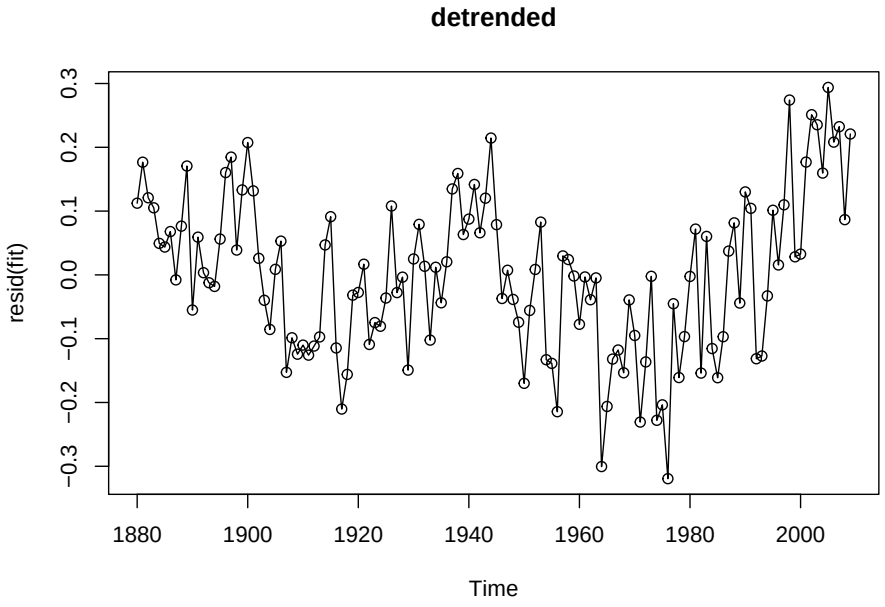


Figure 1.5: Residuals of OLS fit.

```
# Figure 1.5
plot(resid(fit), type = "o", main = "detrended")
```

In Figure 1.5: If trend is now zero, there appears to be a substantial serial dependence remaining in the time series.

1.8 Time Series Differencing

Signal and Noise Model: $X_t = S_t + \varepsilon_t$. Hopefully, upon estimating S_t with $\hat{S}_t$, we find $X_t - \hat{S}_t = \hat{\varepsilon}_t$ (detrended series) looks reasonably stationary. If the residuals would be reasonably stationary, we might proceed in estimating their underlying structure of $\{\hat{\varepsilon}_t\}_{t=1,\ldots,T}$. as if it were stationary. In particular, we might try to estimate their marginal distributions and/or their serial dependence structure. If we thought those estimates were reasonably good, we would have a good idea of how the time series X_t behaves.

Random Walk with Drift Model. Let ε_t be a strong white noise.

$$\begin{aligned} X_t &= \delta + X_{t-1} + \varepsilon_t \\ &= \delta + \delta + X_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &= \delta + \delta + \delta + X_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \text{ times} \\ &= t\delta + X_0 + \sum_{j=1}^{t} \varepsilon_j \end{aligned}$$

where we note that $t\delta + X_0 = S_t$ is a linear signal, and $\sum_{j=1}^{t} \varepsilon_j$ is a random walk noise.

Notice that under the Random Walk Model.

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

So, if X_t follows a random walk model, the series $Y_t = \nabla X_t$ should behave like a white noise shifted by δ.

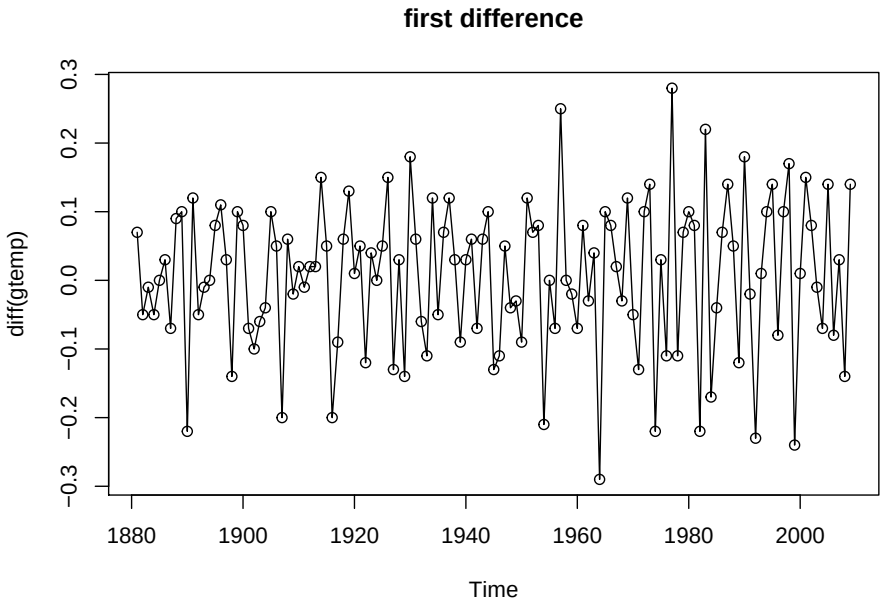


Figure 1.6: First differenced series. Average of first differenced series is $\hat{\delta} \approx 0.0066$

```
# Figure 1.6
plot(diff(gtemp), type = "o", main = "first difference")
```

In Figure 1.6: To see what this looks like in this temperature example, here is a plot of $\nabla X_t = X_t - X_{t-1}$ for the temperature deviation data. As you can see if you look at this compared to the detrended series using linear trend, I would say this series looks much more like a white noise (there does not appear to be any discernible patterns in this first difference). If you calculate the mean of this first difference series, that would be an estimator for the drift term in the random walk model which here is ≈ 0.0066.

DEFINITION 1.8.1: Differenced time series

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

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

and is the series of length $T-1$, namely

$$X_2 - X_1, X_3 - X_2, \dots, X_T - X_{T-1}$$

Higher order differences are calculated recursively, so

$$\nabla^d X_t = \nabla^{d-1} \nabla X_t$$

where ∇^d is the d^{th} order difference and we define $\nabla^0 X_t = X_t$.

Detrending and Differencing are both ways of reducing a (potentially non-stationary) time series to an approximately stationary series.

Differencing vs. Detrending

Pros:

- Differencing does not require the parameter estimation (don't need to estimate S_t).
- Higher order differencing can reduce even very "trendy" series to look more like noise.

Cons:

- Differencing can "wash away" features of the series, and introduce more complicated structures.
- The trend is often of interest, and good estimates of the trend lead to improved long-range forecasts.

EXAMPLE 1.8.2: Potentially Complicating Series with Differencing

$X_t = W_t$ where W_t is a strong white noise.

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

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

More complicated:

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

```
# Figure 1.7
par(mfrow = c(2, 1))
plot(diff(gtemp), main = "first difference Temp data")
plot(rnorm(gtemp),
     type = "l",
     main = "white noise",
     ylab = "w")
```

In Figure 1.7: If these two series behave in the same way, then it stands to reason that

$$g(\varepsilon_t, \varepsilon_{t-1}, \dots) = \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\text{temp}})$$

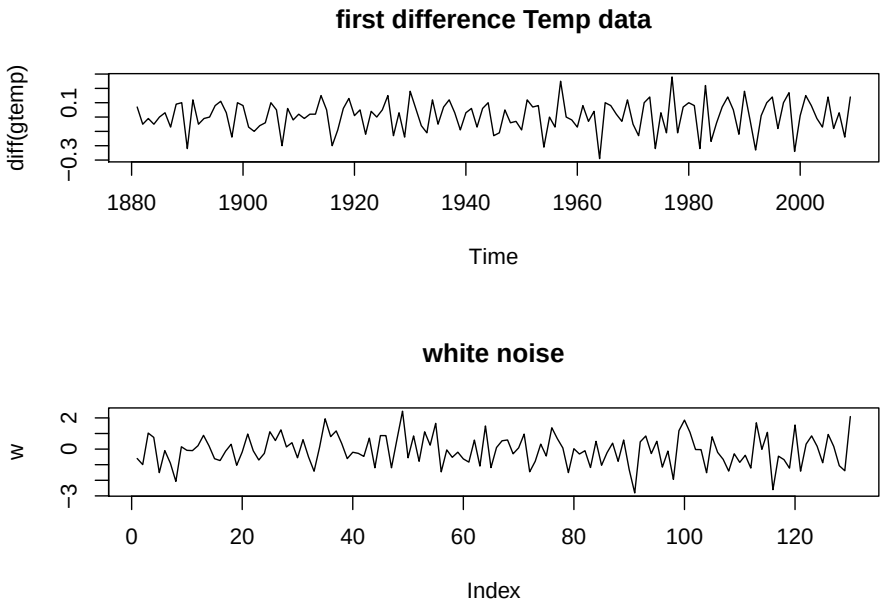


Figure 1.7: First Difference and White Noise

Chapter 2

Time Series Regression and Exploratory Data Analysis

2.1 Autocorrelation and Empirical Autocorrelation

Usually through either detrending or differencing, we arrive at a series $\{X_t\}_{t\in\mathbf{Z}}$ that we may consider as stationary.

Given such a series, we wish to estimate a function g, so that

$$X_t = g(W_t, W_{t-1}, ...)$$

$\{W_t\}_{t\in\mathbf{Z}}$ is a "innovation" sequence (strong white noise) which could admit serial dependence, etc.

In a first pass, it's reasonable to assume that g is a linear function.

DEFINITION 2.1.1: Linear process

A time series $\{X_t\}_{t\in\mathbf{Z}}$ is said to be a **linear process** if there exists a strong white noise $\{W_t\}_{t\in\mathbf{Z}}$ and coefficient $\{\psi_\ell\}_{\ell\in\mathbf{Z}}$ where $\psi_\ell \in \mathbf{R}$, so that

$$\sum_{\ell=-\infty}^{\infty} |\psi_\ell| < \infty$$

and

$$X_t = \sum_{\ell=-\infty}^{\infty} \psi_\ell W_{t-\ell}$$

Note that the sum defining X_t is well-defined as a limit in L^2. Also, we must require that $\mathbb{V}(W_{t-\ell}) < \infty$.

DEFINITION 2.1.2: Causal linear process

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **causal linear process** if

$$X_t = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$$

Note that X_t only depends on W's in the "past."

EXAMPLE 2.1.3

$X_t = W_t$ is a linear process, so all ψ's are 0, except for $\psi_0 = 1$ which is a strong white noise sequence.

REMARK 2.1.4

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

EXAMPLE 2.1.5

$X_t = W_t + \theta W_{t-1}$ where $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise with finite variance, and X_t is a linear process.

$$\gamma_X = \begin{cases} (1+\theta^2)\sigma_W^2 & h = 0 \text{always non-zero} \\ \theta\sigma_W^2 & h = 1 \\ 0 & h \geq 2 \end{cases}$$

$\gamma_X(h)$ non-zero for $h \geq 1$ only where "lagged" terms in the linear process are non-zero. Suggests a way of sleuthing out what

$$g(W_t, W_{t-1}, ...) = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$$

must look like.

DEFINITION 2.1.6: Autocorrelation function

Suppose $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary. The **autocorrelation function** (ACF) of $\{X_t\}_{t\in\mathbf{Z}}$ is

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

Note since $\gamma(0) = \mathbb{V}(X_t) = \mathbb{V}(X_0)$ (since the process is stationary),

$$|\gamma(h)| = |\text{Cov}(X_t, X_{t+h})| \underset{\text{CS}}{\leq} \sqrt{\underbrace{\mathbb{V}(X_t)\mathbb{V}(X_{t+h})}_{\text{Same \# by stationarity}}} = \mathbb{V}(X_0)$$

Hence, $|\rho(h)| \leq 1 \implies -1 \leq \rho(h) \leq 1$.

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

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

where $\mu = \mathbb{E}[X_t]$. Hence, a sensible estimator is

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

which is the **sample mean** (**time series average**).

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

where $(X_t - \bar{X})(X_{t+h} - \bar{X})$ is the averaging over centred terms h-time steps apart.

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

EXAMPLE 2.1.7

$X_t = W_t$ where $\{W_t\}_{t \in \mathbf{Z}}$ is a strong white noise with $\mathbb{V}(W_t) = \sigma_W^2 < \infty$.

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

Therefore,

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

Note that it's always the case that

$$\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$$

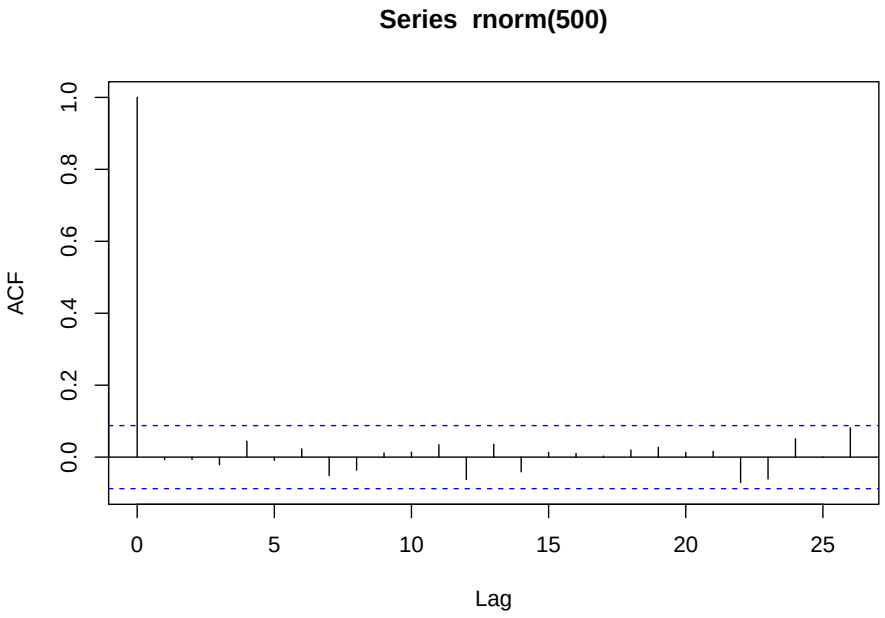


Figure 2.1: ACF of white noise, sample length 130

```
# Figure 2.1
acf(rnorm(500))
```

In Figure 2.1: Let's then have a look at what the empirical autocorrelation function looks like when we apply it to a strong white noise sample. In this case, we are considering a strong Gaussian white noise with variance 1. This is what the sample ACF looks like. What we're plotting here is on the x-axis we have the lags h, and on the y-axis we have the magnitudes of the autocorrelation $\hat{\rho}(h)$. What we're seeing here is $\hat{\rho}(0) = 1$ (by definition). However, for lags other than zero, for the other autocorrelations plotted, we can see that they are relatively small compared to $\hat{\rho}(0) = 1$, which is the point of the blue lines (explained in the next lecture). The basic interpretation of blue lines is that if an autocorrelation would go inside the blue lines then you could imagine that it would be consistent with the series being a strong white noise, which is what we observe here. There's small violations that can occur by sheer chance.

2.2 Modes of Convergence of Random Variables

$\hat{\gamma}(h)$ is an estimator of $\gamma(h)$ when the data is stationary, and we want to discuss the asymptotic properties of this estimator.

Introduce (Review)

1. Stochastic Boundedness (convergence of random variables): $\mathcal{O}(p)$ and $o(p)$
2. Convergence in Probability
3. Convergence in Distribution

DEFINITION 2.2.1: Bounded in probability

Suppose $\{X_n\}_{n\geq 1}$ is a sequence of random variables. We say that X_n is **bounded in probability** by Y_n if for all $\varepsilon > 0$, there exists real numbers M, N, so that for all $n \geq N$,

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

Notation: $X_n = \mathcal{O}_p(Y_n)$, and in English, we say "X_n is on the order of Y_n."

DEFINITION 2.2.2: Converges in probability

We say X_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 is a sequence of scalars, we abbreviate $\dfrac{X_n}{a_n}$ converges in probability to zero as

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

Hence, X_n converges in probability to zero is denoted $X_n = o_p(1)$. We also write $X_n \xrightarrow{p} X$ to denote X_n converges in probability to X.

DEFINITION 2.2.3: Converges in distribution

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

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

REMARK 2.2.4

When $F(x)$ is the CDF of a continuous random variable (e.g. a normal CDF), then

$$\lim_{n\to\infty} |F_n(y) - F(y)| = 0 \quad (\forall y \in \mathbf{R})$$

THEOREM 2.2.5: Markov's Inequality

If $\mathbb{E}[Y^2] < \infty$, *then*

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

Proof of: Theorem 2.2.5

$$\begin{aligned}
\mathbb{E}[Y^2] &= \mathbb{E}[Y^2\mathbb{I}\{|Y| \geq m\} + Y^2\mathbb{I}\{|Y| < m\}] \\
&= \mathbb{E}[Y^2\mathbb{I}\{|Y| \geq m\}] + \mathbb{E}[Y^2\mathbb{I}\{|Y| < m\}] \\
&\geq \mathbb{E}[Y^2\mathbb{I}\{|Y| \geq m\}] \\
&\geq m^2\mathbb{E}[\mathbb{I}\{|Y| \geq m\}] && \text{since } Y^2 \geq m^2 \\
&= m^2\mathbb{P}(|Y| \geq m)
\end{aligned}$$

REMARK 2.2.6: Generalization of Markov's Inequality

If $\mathbb{E}[Y^k] < \infty$, then

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

EXAMPLE 2.2.7

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

$$\begin{aligned}
\mathbb{V}(\bar{X}_T) &= \mathbb{E}[\bar{X}_T^2] \\
&= \frac{1}{T^2}\mathbb{E}\left[\left(\sum_{t=1}^{T} X_t\right)^2\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\underbrace{\mathbb{E}[X_tX_s]}_{\neq 0 \iff t=s} \\
&= \frac{T}{T^2}\sum_{t=1}^{T}\mathbb{E}[X_t^2] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[X_0^2] \\
&= \frac{\sigma^2}{T} && \text{since } \sigma^2 = \mathbb{E}[X_0^2]
\end{aligned}$$

Therefore, for $\varepsilon > 0$, by Markov's Inequality we have

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

Hence, $|\bar{X}_T| \xrightarrow{p} 0$

(2) $\bar{X}_T = \mathcal{O}_p(1/\sqrt{T})$, as before

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

So by Markov's Inequality, for $M > 0$

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

Hence $\sqrt{T}\bar{X}_T = \mathcal{O}_p(1) \implies \bar{X}_T = \mathcal{O}_p(1/\sqrt{T})$.

REMARK 2.2.8

Alternatively, we can show this using the CLT. By the CLT,

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

Therefore, if $F_T \sim$ CDF of $\sqrt{T}\bar{X}_T$ and $\Phi \sim$ CDF of $\mathcal{N}(0,1)$ random variable we have

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

For $\varepsilon > 0$, choose M such that

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

For this M, choose T_0 such that if $T \geq T_0$, then

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

and

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

Then,

$$\begin{aligned}
\mathbb{P}(|\sqrt{T}\bar{X}_T| \geq M) &= F_T(-M) + (1 - F_T(M)) \\
&= \Phi\left(-\frac{M}{\sigma}\right) + \left[1 - \Phi\left(\frac{M}{\sigma}\right)\right] + F_T(-M) - \Phi\left(-\frac{M}{\sigma}\right) + \Phi\left(\frac{M}{\sigma}\right) - F_T(M) \\
&\leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} \\
&= \varepsilon
\end{aligned}$$

REMARK 2.2.9

In general,

$$\frac{X_n}{a_n} \xrightarrow{D} \text{non-degenerate random variable} \implies X_n = \mathcal{O}_p(a_n)$$

REMARK 2.2.10: 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
$$X_n + Y_n = \mathcal{O}_p(\max(a_n, b_n))$$
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
$$X_n Y_n = o_p(1)$$

EXAMPLE 2.2.11

Suppose W_t is a strong white noise in L^2 with $\mathbb{E}[W_t^4] < \infty$. Let $X_t = W_t + \theta W_{t-1}$ for $\theta \in \mathbf{R}$. Show that
$$\hat{\gamma}(1) \xrightarrow{p} \theta\sigma_W^2$$

Solution.

$$\begin{aligned}
\bar{X}_T &= \frac{1}{T}\sum_{t=1}^{T} X_t \\
&= \frac{1}{T}\sum_{t=1}^{T}(W_t + \theta W_{t-1}) \\
&= \frac{1}{T}\sum_{t=1}^{T} W_t + \frac{\theta}{T}\sum_{t=}^{T} W_{t-1} \\
&= o_p(1) && \text{by WLLN}
\end{aligned}$$

$$\begin{aligned}
\hat{\gamma}(1) &= \frac{1}{T}\sum_{t=1}^{T-1}(X_t - \bar{X}_T)(X_{t+1} - \bar{X}_T) \\
&= \frac{1}{T}\sum_{t=1}^{T-1}[X_t X_{t+1} - X_t\bar{X}_T - \bar{X}_T X_{t+1} + (\bar{X}_T)^2] \\
&= \frac{1}{T}\sum_{t=1}^{T-1} X_t X_{t+1} - \frac{\bar{X}_T}{T}\sum_{t=1}^{T-1} X_t - \frac{\bar{X}_T}{T}\sum_{t=1}^{T-1} X_{t+1} + \frac{T-1}{T}(\bar{X}_T)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T-1} X_t X_{t+1} + R_1 + R_2 + R_3
\end{aligned}$$

Notice that $R_i = o_p(1)$ for $i = 1, 2, 3$ since, for example, $\bar{X}_T = o_p(1)$ and $\sum_{t=1}^{T} X_t = o_p(1)$ so their product is $o_p(1)$; so we only need to focus on the first term.

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T-1} X_t X_{t+1} &= \frac{1}{T}\sum_{t=1}^{T-1}(W_t + \theta W_{t-1})(W_{t+1} + \theta W_t) \\
&= \frac{1}{T}\sum_{t=1}^{T-1}\theta W_t^2 + G_1 + G_2 + G_3
\end{aligned}$$

Now,
$$\frac{1}{T}\sum_{t=1}^{T-1}\theta W_t^2 \xrightarrow{\text{a.s.}} \theta\mathbb{E}[W_t^2] = \theta\sigma_W^2$$

by strong law of large numbers. We now wish to calculate the variance of

$$G_1 = \frac{1}{T}\sum_{t=1}^{T-1} W_t W_{t+1}.$$

$$\mathbb{E}[G_1] = \frac{1}{T}\sum_{t=1}^{T-1} \mathbb{E}[W_t W_{t+1}] = 0$$

$$\begin{aligned}\mathbb{V}(G_1) &= \mathbb{E}[G_1^2] \\ &= \frac{1}{T^2}\sum_{t=1}^{T-1}\sum_{s=1}^{T-1} \underbrace{\mathbb{E}[W_t W_{t+1} W_s W_{s+1}]}_{\neq 0 \iff s=t} \\ &= \frac{T-1}{T^2}\sum_{t=1}^{T-1} \mathbb{E}[W_t^2 W_{t+1}^2] = \frac{T-1}{T^2}\sigma_W^4 \xrightarrow{T\to\infty} 0\end{aligned}$$

By Markov's Inequality,

$$G_1 = o_p(1)$$

and similarly, for G_2 and G_3.

2.3 † M-dependent CLT

Suppose X_t is a mean zero 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

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t = \sqrt{T}\bar{X}_T \overset{D}{\approx} \mathcal{N}(0, \sigma_X^2)$$

(2) If X_t is a strong white noise, what the approximate distribution of

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h} \underbrace{X_t X_{t+h}}_{\text{not iid}} + o_p(1)$$

$X_t X_{t+h} = Y_t$ is strictly stationary.

- Only way to understand how $\{X_t\}_{t\in\mathbf{Z}}$ behaves, we have to observe replicates of the process.
- If process is suitably "weakly dependent," then we can observe replicates of the process by viewing in on overlapping windows.

DEFINITION 2.3.1: m-dependent

We say a time series $\{X_t\}_{t\in\mathbf{Z}}$ is **m-dependent** for a positive integer m, if for all

$$t_1 < t_2 < \cdots < t_{d_1} < s_1 < s_2 < \cdots < s_{d_2} \in \mathbf{Z}$$

so that $t_{d_1} + m \leq s_1$, then

$$(X_{t_1}, \ldots, X_{t_{d_1}})$$

is **independent** of

$$(X_{s_1}, \ldots, X_{s_{d_2}})$$

EXAMPLE 2.3.2

$X_t = W_t + \theta W_{t-1}$ for $\theta \in \mathbf{R}$ and W_t is a strong white noise is 2-dependent.

THEOREM 2.3.3: Generalization of the standard CLT to m-dependent

Suppose X_t is a strictly stationary and m-dependent time series for $m \in \mathbf{Z}_{>0}$ with $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[X_t^2] < \infty$, then if

$$S_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t = \sqrt{T}\bar{X}_T \xrightarrow{T\to\infty}{}^{D} \mathcal{N}(0, \sigma_m^2)$$

where

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

Note that σ_m^2 is just the variance of S_T and can be easily calculated.

DEFINITION 2.3.4: Triangular array

We say $\{X_{i,j}, 1 \le j \le n_i, 1 \le i < \infty\}$ forms a **triangular array** of mean zero L^2 random variables, if $\mathbb{E}[X_{i,j}] = 0$, $\mathbb{E}[X_{i,j}^2] < \infty$, and for each i-fixed we have $X_{i,1}, \dots, X_{i,n_i}$ are independent with $n_i < n_i + 1$.

Visually, row-wise random variables are independent:

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

THEOREM 2.3.5: Linderberg-Feller CLT for Triangular Arrays

Let $\{X_{i,j}, 1 \le j \le n_i, 1 \le i < \infty\}$ be a triangular array of mean zero L^2 random variables. Define

$$\sigma_i^2 = \sum_{j=1}^{n} \mathbb{V}(X_{i,j})$$

and

$$S_i = \frac{1}{\sigma_i} \sum_{j=1}^{n_i} X_{i,j}$$

If for $\varepsilon > 0$,

$$\frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} \mathbb{E}[X_{i,j}^2 \mathbb{I}\{|X_{i,j}| > \varepsilon\sigma_i\}] \xrightarrow{i\to\infty} 0$$

then

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

Proof of: Theorem 2.3.3

Bernstein Blocking Argument: we take a given time series of length T.
Let a_T = big block size and m = little block size. We assume $a_T \to \infty$ as $T \to \infty$, but $\frac{a_T}{T} \to 0$. Then,

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

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

Since a_T is increasing up to infinity, for T sufficiently large, $a_T > m$, and so by m-dependence,

$$\sum_{t\in B_j} X_t$$

is independent of

$$\sum_{t\in B_k} X_t \quad (j \neq k)$$

similarly for $B_j, B_k \to b_j, b_k$.

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t = \frac{1}{\sqrt{T}}\sum_{j=1}^{N}\sum_{t\in B_j} X_t + \frac{1}{\sqrt{T}}\sum_{j=1}^{N}\underbrace{\sum_{t\in b_j} X_t}_{\text{iid}} + \text{Remainder} = G_1 + G_2 + G_3$$

We want to show the big blocks dominate.

$$\mathbb{V}(G_2) = \frac{1}{T}\sum_{j=1}^{N}\mathbb{E}\left[\left(\sum_{t\in b_j} X_t\right)^2\right] = \frac{N}{T}\mathbb{E}\left[\left(\sum_{t=1}^{m} X_t\right)^2\right]$$

Also,

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

Let $h = t - s$, set of possible values for h is $m - |h|$, so

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

noting that $\gamma(h) = \gamma(-h)$, therefore for C as a constant, we have

$$\mathbb{V}(G_2) = \frac{N}{T}C = \frac{\left\lfloor \frac{T}{a_T+m} \right\rfloor}{T}(C) \xrightarrow{a_T\to\infty} 0$$

and hence $G_2 = o_p(1)$.
Let's deal with the big block terms. Notice

$$G_1 = \frac{1}{\sqrt{T}} = \sum_{j=1}^{N}\sum_{t\in B_j} X_t = \sum_{j=1}^{N}\frac{\sum_{t\in B_j} X_t}{\sqrt{T}} = \sum_{j=1}^{N} Y_j$$

where Y_j is a triangular array. So, $\mathbb{V}(G_1) = \sum_{j=1}^{N}\mathbb{V}(Y_j)$.

$$\begin{aligned}
\mathbb{V}(Y_j) &= \mathbb{V}(Y_1) \\
&= \frac{1}{T}\mathbb{E}\left[\left(\sum_{t=1}^{a_T} X_t\right)^2\right] \\
&= \frac{1}{T}\sum_{t=1}^{a_T}\sum_{s=1}^{a_T}\mathbb{E}[X_t X_s] \\
&= \frac{1}{T}\sum_{h=1-a_T}^{a_T-1}(a_T-|h|)\gamma(h)
\end{aligned}$$

Note that since the process is m-dependent, $\gamma(h) = 0$ if $|h| \geq m$. Continuing,

$$\frac{1}{T}\sum_{h=1-a_T}^{a_T-1}(a_T - |h|)\gamma(h) = \sum_{h=-m}^{m}(a_T - |h|)\gamma(h)$$

Therefore,

$$\mathbb{V}(G_1) = \underbrace{\frac{N}{T}}_{\approx 1/a_T}\sum_{h=-m}^{m}(a_T - |h|)\gamma(h) \xrightarrow{T\to\infty} \sum_{h=-m}^{m}\gamma(h)$$

Therefore, the variance of G_1 is bounded. We showed $\sigma_N^2 = \mathbb{V}(G_1) \approx$ constant. So, we must show

$$\sum_{j=1}^{N}\mathbb{E}[\underbrace{Y_j^2}_{\text{iid}}\mathbb{I}\{|Y_j| > \varepsilon\sigma_N\}] = N\mathbb{E}[Y_1^2\mathbb{I}\{|Y_1| > \varepsilon\sigma_N\}] \xrightarrow{T\to\infty} 0$$

Aside:

$$\begin{aligned}\mathbb{E}[|Y|^{2+\delta}] &\geq \mathbb{E}[|Y|^{2+\delta}\mathbb{I}\{|Y| > \varepsilon\}] \\ &\geq \varepsilon^{\delta}\mathbb{E}[|Y|^2\mathbb{I}\{|Y| > \varepsilon\}]\end{aligned}$$

$$\implies \mathbb{E}[|Y|^2\mathbb{I}\{|Y| > \varepsilon\}] \leq \frac{\mathbb{E}[|Y|^{2+\delta}]}{\varepsilon^{\delta}}$$

It may be shown that for $C > 0$

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

So

$$\begin{aligned}N\mathbb{E}[Y_1^2\mathbb{I}\{|Y_1| > \varepsilon\sigma_N\}] &\leq \frac{N}{(\varepsilon\sigma_N)^{\delta}}C\left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}} \\ &= \frac{C}{(\varepsilon\sigma_N)^{\delta}}\frac{Na_T}{T}\left(\frac{a_T}{T}\right)^{\delta/2} \xrightarrow{T\to\infty} 0\end{aligned}$$

Therefore, by Theorem 2.3.3

$$\frac{G_1}{\sigma_N} \xrightarrow{T\to\infty}{}^{D} \mathcal{N}(0,1)$$

and since

$$\sigma_N^2 \to \sum_{j=-m}^{m}\gamma(j)$$

we have

$$G_1 \xrightarrow{D} \mathcal{N}\left(0, \sum_{h=-m}^{m}\gamma(h)\right)$$

Since $G_2 = o_p(1)$ we have

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

2.4 † Two Plus Delta Moment Calculation

We want to show

$$\mathbb{E}[|Y_1|^{2+\delta}] \leq C\left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}}$$

where

$$Y_1 = \frac{1}{\sqrt{T}}\sum_{t=1}^{a_T} X_t$$

a_T = big block size $\to \infty$ as $T \to \infty$

$$\frac{a_T}{T} \to 0$$

X_t are m-dependent random variables.

$$\mathbb{E}[|X_i|^{2+\delta}] < \infty \quad (\delta > 0) \iff \mathbb{E}\left[\left|\sum_{t=1}^{a_T} X_t\right|^{2+\delta}\right] \leq C a_T^{\frac{2+\delta}{2}}$$

THEOREM 2.4.1: Rosenthal's Inequality

If $X_1, \ldots, X_n$ are independent random variables with $\mathbb{E}[|X_i|^{2+\delta}] < \infty$ for $\delta > 0$, then

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

In particular, if $X_1, \ldots, X_n$ are i.i.d., then

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

Proof of: Theorem 2.4.1

See Petrov, Limit theorems of Probability Theory, p.g. 59.

PROPOSITION 2.4.2

For arbitrary random variables $X_1, \ldots, X_n$,

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

Proof of: Proposition 2.4.2

Since $\varphi(x) = |x|^{2+\delta}$ is convex where $a_1, \ldots, a_n \in \mathbf{R}$, by Jensen's Inequality,

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

Rearranging yields

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

Replace $a_i \sim X_i$, take expectation.

$$\sum_{t=1}^{a_T} X_t = \sum_{j=0}^{m} \sum_{\substack{t=j \\ \text{mod } (m+1) \\ 1\le t\le a_T}} X_t$$

Variables in the second sum are separated by at least m-time steps, and hence i.i.d. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{t=1}^{a_T} X_t\right|^{2+\delta}\right] &\le (m+1)^{(2+\delta)-1}\mathbb{E}\left[\left|\sum_{\substack{t=j \\ \text{mod } (m+1) \\ 1\le t\le a_T}} X_t\right|^{2+\delta}\right] && \text{by Proposition 2.4.2} \\
&\le (m+1)^{(2+\delta)-1}\left(\frac{a_T}{m+1}\right)^{\frac{2+\delta}{2}}\mathbb{E}[|X_1|^{2+\delta}] && \text{by Theorem 2.4.1} \\
&= C a_T^{\frac{2+\delta}{2}}
\end{aligned}$$

where C is the same constant as in the proof of Theorem 2.3.3.

2.5 † Linear Process CLT

EXAMPLE 2.5.1

$X_t = \sum_{\ell=0}^{m} \psi_\ell W_{t-\ell}$ where $\{W_t\}_{t\in\mathbf{Z}}$ is is a strong white noise in L^2.

A general linear process $X_t = \sum_{\ell=0}^{m} \psi_\ell W_{t-\ell}$ is not m-dependent.

THEOREM 2.5.2: Basic Approximation Theorem (BAT)

Suppose X_n is a sequence of random variables so that there exists an array

$$\{Y_{n,m}, m, n \ge 1\}$$

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) For all $\varepsilon > 0$,

$$\lim_{m\to\infty}\left[\limsup_{n\to\infty} \mathbb{P}(|X_n - Y_{n,m} > \varepsilon|)\right] = 0$$

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

REMARK 2.5.3

$Y_{m,n}$ is often an "m-dependent" approximation to X_n

Proof of: Theorem 2.5.2

Shumway and Stoffer using characteristic functions.

THEOREM 2.5.4: Linear Process CLT

Suppose $X_t = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$ *is a causal linear process with* $\sum_{\ell=0}^{\infty} |\psi_\ell| < \infty$ *with* $\{W_t\}_{t\in\mathbf{Z}}$ *is a strong white noise in* L^2*. If*

$$S_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t$$

then

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

Proof of: Theorem 2.5.4

X_t is strictly (and weakly) stationary.

$$\begin{aligned}
\gamma(h) &= \mathbb{E}[X_t X_{t+h}] \\
&= \mathbb{E}\left[\left(\sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}\right)\left(\sum_{j=0}^{\infty} \psi_j W_{t+h-j}\right)\right] \\
&= \sum_{\ell=0}^{\infty}\sum_{j=0}^{\infty} \psi_\ell \psi_j \mathbb{E}[W_{t-\ell} W_{t+h-j}] && \text{Fubini's Theorem} \\
&= \sum_{\ell=0}^{\infty} \psi_\ell \psi_{\ell+h} \sigma_W^2
\end{aligned}$$

Then,

$$\sum_{h=-\infty}^{\infty} \gamma(h) = \sum_{h=-\infty}^{\infty} \left|\sum_{\ell=0}^{\infty} \psi_\ell \psi_{\ell+h} \sigma_W^2\right| \le \sum_{\ell=0}^{\infty} |\psi_\ell| \sum_{h=-\infty}^{\infty} |\psi_h| \sigma_W^2 < \infty$$

So $\sum_{h=-\infty}^{\infty} \gamma(h)$ is well-defined. Note that $\mathbb{E}[S_T] = 0$ since $\mathbb{E}[X_t] = 0$. Also,

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

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

Define

$$X_{t,m} = \sum_{\ell=0}^{m} \psi_\ell W_{t-\ell}$$

$$S_{T,m} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_{t,m}$$

is a m-dependent approximation to S_T.

(1) By the m-dependent CLT,

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

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

$$S'_m \xrightarrow{D} \mathcal{N}\left(0, \sum_{h=-\infty}^{\infty} \gamma(h)\right)$$

(3)

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

So condition (3) of the BAT is satisfied using Markov's Inequality. Therefore,

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

2.6 Asymptotic Properties of Empirical ACF

If $X_1, \dots, X_T$ is an observed time series which we think was generated by a stationary process, then $\text{Cov}(X_t, X_{t+h})$ does not depend on t. Recall that

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

$$\rho(h) = \text{Corr}(X_t, X_{t+h}) = \frac{\gamma(h)}{\gamma(0)}$$

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

Questions:

(1) Are $\hat{\gamma}$ and $\hat{\rho}$ consistent?

(2) What is the approximate distribution of $\hat{\gamma}(h)$ and $\hat{\rho}(h)$.

Consistency: By adding and subtracting μ in the definition of $\hat{\gamma}(h)$, we may assume WLOG that $\mathbb{E}[X_t] = 0$.

Suppose $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary, and

$$X_t = g(W_t, W_{t-1}, \dots)$$

We first need to establish the consistency of

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

where X_t's are not i.i.d. so Law of Large numbers does not work. Instead, we would use the Ergodic Theorem, but we will not cover it here. Therefore,

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

Furthermore,

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

where we note that the last three terms converge in probability to 0 by the Ergodic Theorem.

Also, note that $\mathbb{E}[X_t X_{t+h}] = \gamma(h)$ and

$$X_t X_{t+h} = g_h(W_{t+h}, W_{t+h-1}, ...)$$

Again, by the Ergodic Theorem,

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

Therefore, $\hat{\gamma}(h) \xrightarrow{P} \gamma(h)$ and $\hat{\rho}(h) = \dfrac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \xrightarrow{P} \rho(h)$ under strict stationarity and $\mathbb{E}[X_t^2] < \infty$.

Distribution of $\hat{\gamma}(h)$: Consider simple (but most important case) when $\{X_t\}_{t\in\mathbf{Z}}$ is a strong white noise with $\mathbb{E}[X_t^4] < \infty$. The finite 4th moment assumption is not really assumed here, but this will be explained why it's classically assumed.

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

Similarly,

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

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

$$\mathbb{V}(\tilde{\gamma}(h)) = \mathbb{E}[\tilde{\gamma}^2(h)] = \frac{1}{T^2}\sum_{t=1}^{T-h}\sum_{s=1}^{T-h} \mathbb{E}[X_t X_{t+h} X_s X_{s+h}]$$

is non-zero only when $t = s$, so

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

where $\mathbb{E}[X_t^2] = \sigma_X^2$. Therefore,

$$\mathbb{V}(\sqrt{T}\tilde{\gamma}(h)) \xrightarrow{T\to\infty} \sigma_X^4$$

THEOREM 2.6.1

If $\{X_t\}_{t\in\mathbf{Z}}$ is a strong white noise with $\mathbb{E}[X_t^4] < \infty$, then

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

Proof of: Theorem 2.6.1

Using Martingale CLT which is derived from m-dependent CLT.

COROLLARY 2.6.2

It follows that if

$$\sqrt{T}\hat{\gamma} \xrightarrow{D} \mathcal{N}(0, \sigma_X^4)$$

and $\hat{\gamma}(0) \xrightarrow{P} \sigma_X^2$ (SLLN), then by Slutsky's Theorem,

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

If $\{X_t\}_{t\in\mathbf{Z}}$ 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 $\Phi(z_{\alpha/2}) = 1-\alpha$. Hence,

$$\left(\frac{-1.96}{\sqrt{T}}, \frac{1.96}{\sqrt{T}}\right)$$

is an approximate 95% prediction interval for $\hat{\rho}(h)$ assuming the data is generated by a strong white noise process.

Now, we know that the blue boundaries are $\pm\frac{1.96}{\sqrt{T}}$ in Figure 2.1. Also, we might be able to say that exists mild serial correlation at lag 1 of the ACF for Figure 2.2 since there are lines that go outside the blue boundaries.

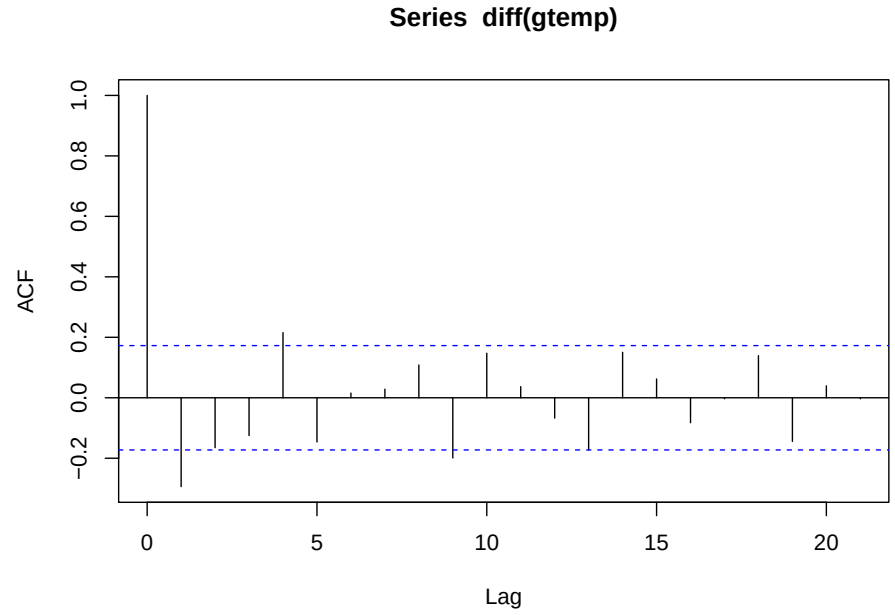


Figure 2.2: ACF of first differenced temperature data

```
# Figure 2.2
plot(acf(diff(gtemp)))
```

2.7 Interpreting the Autocorrelation Function (Non-stationary)

We have an excellent understanding of how $\hat{\rho}(h)$ behaves when $X_1, \ldots, X_T$ is a strong white noise.

- Consistency:

$$\hat{\rho}(h) \xrightarrow{P} 0 \quad (h \geq 1)$$

- Distribution:

$$\hat{\rho}(h) \overset{D}{\approx} \mathcal{N}\left(0, \frac{1}{T}\right) \quad (T \text{ is large})$$

What happens when we calculate the empirical ACF for a non-stationary time series?

EXAMPLE 2.7.1

$X_t = t + W_t$ where W_t is a strong white noise. Note that X_t has a linear trend, and hence not stationary. First,

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

Also,

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

where R is the remainder with the white noise terms. Note that the dominant term; that is, the $\mathcal{O}(T^2)$ doesn't depend on h.
It follows that in this case that

$$\frac{\hat{\rho}(h)}{T^2} \xrightarrow{T\to\infty} C \quad (\forall h)$$

Hence

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

Moral: If X_t has a trend that is not properly removed, $\hat{\rho}(h)$ is likely to be large.

```
# Figure 2.3
acf(gtemp)
# Figure 2.4
plot(as.ts(cumsum(rnorm(100))), main = "autoregression, phi=1")
# Figure 2.5
acf(as.ts(cumsum(rnorm(100))))
```

- Looking back at Figure 1.2, we see that this time series has an upwards trend. Therefore, based on what we just did, we expect that the ACF should be very large (close to 1) at each lag for this time series. Clearly, Figure 2.3 is indicative of a strong trend or non-stationarity.
- In Figure 2.4, we are plotting

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

  with $X_0 = 0$ and $X_t = \sum_{j=1}^{t} W_j$ which is non-stationary. Some people say it has a "stochastic trend."
- In Figure 2.5 there exists a similar pattern which is indicative of non-stationarity.

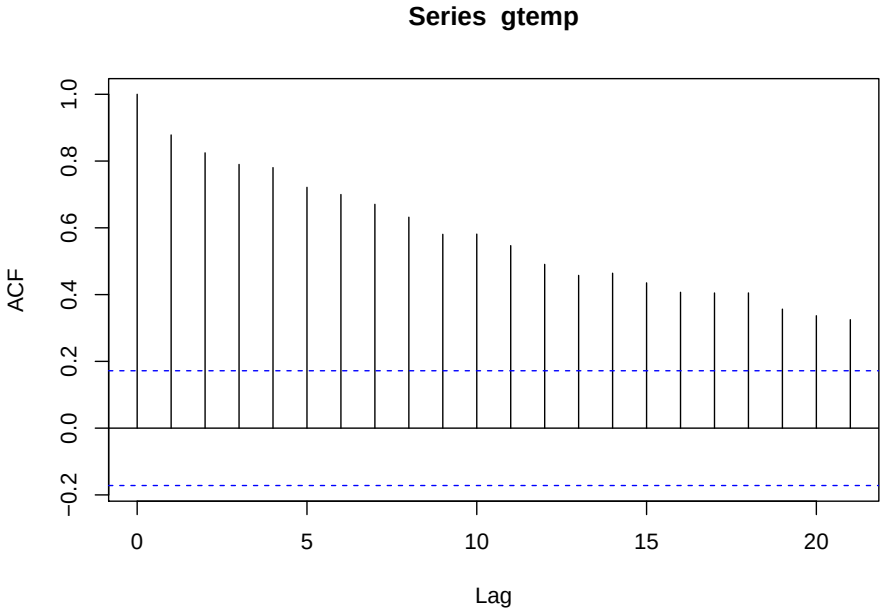


Figure 2.3: ACF of raw temperature data, sample length 130

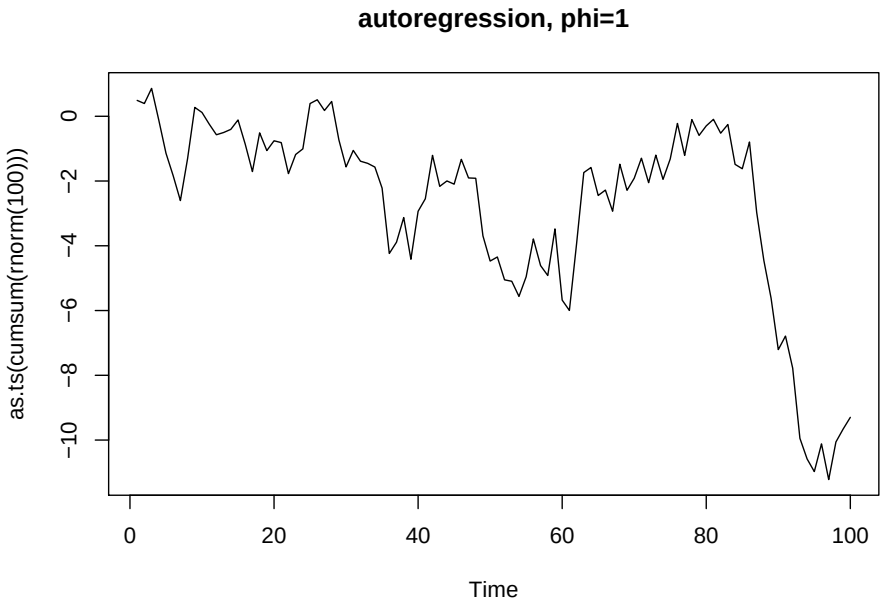


Figure 2.4: Realization of an AR(1) with $\phi = 1$ starting from $x_0 = 0$

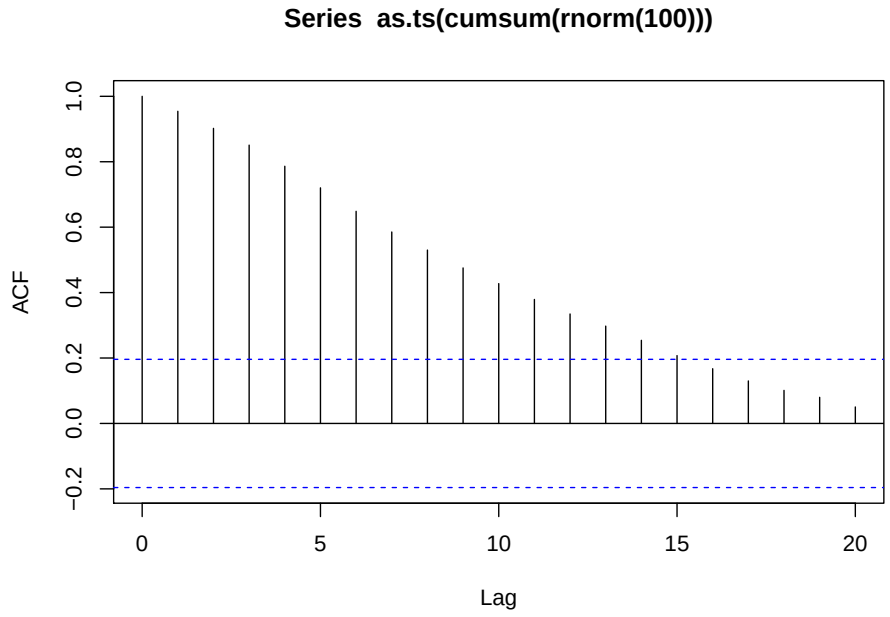


Figure 2.5: ACF of an AR(1) with $\phi = 1$ starting from $x_0 = 0$

Chapter 3

ARIMA Models

3.1 Moving Average Processes

Suppose X_t is stationary. Identify serial dependence using ACF $\hat{\rho}(h)$. If the lines go out of the dotted blue boundaries, namely $\pm\frac{1.96}{\sqrt{T}}$, within the ACF plot of $\hat{\rho}(h)$, then we suspect serial dependence.

Posit

$$X_t = g(W_t, W_{t-1}, ...) = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell} \quad \text{[Linear Process]}$$

Not feasible to estimate infinitely many parameters $\{\psi\}_{\ell=0}^{\infty}$. Assume coefficients arise from a parsimonious linear model f.

DEFINITION 3.1.1: Moving average process

Suppose $\{W_t\}_{\in \mathbf{Z}}$ is a strong white noise with $\mathbb{V}(W_t) = \sigma_W^2 < \infty$. We say X_t is a **moving average process** of order q if there exists $\theta_1, ..., \theta_q \in \mathbf{R}$ with $\theta_q \neq 0$ such that

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q} = \sum_{\ell=0}^{q} \theta_\ell W_{t-\ell}$$

where $\theta_0 = 1$. We abbreviate this definition as MA(q).

DEFINITION 3.1.2: 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 $a, b \in \mathbf{R}$

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

EXAMPLE 3.1.3

$\nabla X_t =$ first difference of $X_t = (1 - B)X_t$

DEFINITION 3.1.4: Moving average polynomial

The **moving average polynomial** is defined as

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

If $X_t \sim \text{MA}(q)$, then

$$X_t = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q} = \theta(B) W_t$$

which is a succinct expression defining $\text{MA}(q)$.

Properties of MA(q) Processes

1. $\text{MA}(0)$ process is a strong white noise.
2. If $X_t \sim \text{MA}(q)$, then

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

$$\mathbb{V}(X_t) = \mathbb{E}\left[\left(\sum_{\ell=0}^{q} \theta_\ell W_{t-\ell}\right)^2\right] = \sum_{\ell=0}^{q} \theta_\ell^2 \sigma_W^2$$

$$\begin{aligned}\gamma(h) &= \text{Cov}(X_t, X_{t+h}) \\ &= \mathbb{E}\left[\left(\sum_{\ell=0}^{q} \theta_\ell W_{t-\ell}\right)\left(\sum_{k=0}^{q} \theta_k W_{t+h-k}\right)\right] \\ &= \begin{cases} \sum_{j=0}^{q-|h|} \theta_j \theta_{j+h} \sigma_W^2 & 0 \le h \le q \\ 0 & h > q \end{cases}\end{aligned}$$

Therefore,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{\sum_{j=0}^{q} \theta_j^2} & 0 \le h \le q \\ 0 & h \ge q+1 \end{cases}$$

REMARK 3.1.5

By choosing $\theta_1, \ldots, \theta_q$ appropriately, we can get any ACF we want $\rho(h)$ where $1 \le h \le q$.

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

3.2 Autoregressive Processes

DEFINITION 3.2.1: Autoregressive process

Suppose $\{W_t\}_{t \in \mathbf{Z}}$ is a strong white noise with $\mathbb{V}(W_t) = \sigma_W^2 < \infty$. We say X_t is an **autoregressive process** of order 1, abbreviated $\text{AR}(1)$ if there exists a constant ϕ such that

$$X_t = \phi X_{t-1} + W_t \quad (t \in \mathbf{Z})$$

Using the backshift operator, this may also be expressed as

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

Interpretation

Prediction: Form a linear model (regression) predicting X_t as

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

where X_t is the dependent variable and X_{t-1} is the covariant/independent variable.

Markovian Property:

$$X_t \mid (X_{t-1}, X_{t-2}, ...) = X_t \mid X_{t-1}$$

Question: Does there exist a stationary process X_t satisfying the following?

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

Let's see.

$$\begin{aligned}
X_t &= \phi X_{t-1} + W_t \\
&= \phi(X_{t-z} + W_t) + W_t && (z \in \mathbf{Z}) \\
&= \phi^2(X_{t-2}) + \phi W_{t-1} + W_t \\
&\vdots && k \text{ times} \\
&= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j W_{t-j} && \text{if } |\phi| > 1, \text{ the sum diverges}
\end{aligned}$$

Suppose $|\phi| < 1$, then

$$\xrightarrow[k\to\infty]{L^2\text{-sense}} 0 + \sum_{j=0}^{\infty} \phi^j W_{t-j}$$

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

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

Therefore, X_t satisfies AR(1) equation.

THEOREM 3.2.2

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

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

What if $|\phi| > 1$? If $X_t = \phi X_{t-1} + W_t$ for $t \in \mathbf{Z}$, then that implies $X_t = X_{t+1}/\phi - W_{t+1}/\phi$. Iterating k-times similarly as before, we get

$$X_t = \frac{X_{t+k}}{\phi^k} - \sum_{j=1}^{k} \frac{W_{t+j}}{\phi^j} \xrightarrow[k\to\infty]{L^2\text{-sense}} -\sum_{j=1}^{\infty} \frac{W_{t+j}}{\phi^j}$$

given that $\sum_{j=1}^{\infty} \frac{1}{\phi^j} < \infty$. This sequence is strictly stationary since it is a Bernoulli shift. Future dependent, normally we try to avoid this.

What if $|\phi| = 1$? In this case we claim that there is no stationary process such that $X_t = \phi X_{t-1} + W_t$. Let's prove this. Suppose $|\phi| = 1$. If $X_t = X_{t-1} + W_t$, then

$$X_t = \sum_{j=1}^{t} W_j + X_0 \quad \text{(by iterating)} \implies X_t - X_0 = \sum_{j=1}^{t} W_j$$

Now,

$$\mathbb{V}(X_t - X_0) = \mathbb{V}(X_t) + \mathbb{V}(X_0) - 2\text{Cov}(X_t, X_0) \leq 4\mathbb{V}(X_0)$$

where in the last inequality we used the fact that X_t is stationary. Furthermore,

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

Properties of Causal AR(1) for $|\phi| < 1$

(1) The span of dependence of X_t is "infinite"

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

(2) ACF.

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

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

where in the first sum we let $t - \ell = t + h - k$ and in the second sum we let $k = \ell + h$ for $\ell = 0, 1, 2, \ldots$. Hence,

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

Note: this decays geometrically in the lag parameter.

DEFINITION 3.2.3: Autoregressive process, Autoregressive polynomial

We say X_t follows an **autoregressive process** of order p, denoted AR(p) if there exists coefficients $\phi_1, \dots, \phi_p \in \mathbf{R}$ with $\phi_p \neq 0$ such that

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

We also define the **autoregressive polynomial** to be

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

$X_t \sim$ AR(p) if $\phi(B)X_t = W_t$.

We've seen the moving average polynomial:

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

and the autoregressive polynomial:

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

Why not combine the two?

DEFINITION 3.2.4: Autoregressive moving average (ARMA)

Given a strong white noise sequence W_t, we say that X_t is an autoregressive moving average process of orders p and q, denoted ARMA(p, q) if

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

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad (\phi_p \neq 0)$$
$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \quad (\theta_q \neq 0)$$

This implies that the model

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t + W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q}$$

Using ARMA models to model autocorrelation: ARMA combines the following two points.

- MA(q): ACF may be specified at lags $1, \dots, q$
- AR(p): ACF has geometric decay/oscillations

REMARK 3.2.5: Parameter redundancy

Consider $X_t = W_t$ where $X_t \sim$ MA(0), then

$$0.5X_{t-1} = 0.5W_{t-1}$$

Therefore,

$$X_t - 0.5X_{t-1} = W_t - 0.5W_{t-1} \implies X_t \sim \text{ARMA}(1,1)$$
$$\phi(z) = 1 - 0.5z \implies \text{zero of } \phi \text{ is } z_0 = z$$
$$\theta(z) = 1 - 0.5z \implies \text{zero of } \theta \text{ is } z_0 = z$$

Parameter redundancy manifests as shared zeros in ϕ and θ. We always assume the models are "reduced" by factoring and diving away common zeros in ϕ.

DEFINITION 3.2.6: Causal ARMA

We say an ARMA(p, q) is causal if there exists $\{X_t\}_{t\in\mathbf{Z}}$ satisfying $\phi(B)X_t = \theta(B)W_t$ and

$$X_t = \sum_{\ell=0}^{\infty} \phi_\ell W_{t-\ell} = \phi(B)W_t \quad \text{[Causal Linear Process Solution]}$$

where $\phi(B) = \sum_{\ell=0}^{\infty} \phi_\ell B^\ell$ and $\sum_{\ell=0}^{\infty} |\phi_\ell| < \infty$ with $\phi_0 = 1$.

DEFINITION 3.2.7: Invertible ARMA

An ARMA(p, q) is invertible if there exists $\{X_t\}_{t\in\mathbf{Z}}$ satisfying $\phi(B)X_t = \theta(B)W_t$ and

$$W_t = \sum_{\ell=0}^{\infty} \pi_\ell X_{t-\ell} = \pi(B)X_t$$

where $\pi(B) = \sum_{\ell=0}^{\infty} \pi_\ell B^\ell$ and $\sum_{\ell=0}^{\infty} |\pi_\ell| < \infty$ with $\pi_0 = 1$.

REMARK 3.2.8

Causal + Invertibility $\implies$ Information in $\{X_t\}_{t\le T}$ is the same as Information in $\{W_t\}_{t\le T}$. Also, note that $\{X_t\}_{t\le T}$ is an observed time series.

THEOREM 3.2.9: Causality

By the fundamental theorem of algebra, the autoregressive polynomial $\phi(z)$ has p roots, say $z_1, \dots, z_p \in \mathbf{C}$. If

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

then there exists a stationary and causal X_t to the ARMA equations: $\phi(B)X_t = \theta(B)W_t$.

$$X_t = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$$

The coefficients $\{\psi_\ell\}_{\ell=0}^{\infty}$ satisfy

$$\sum_{\ell=0}^{\infty} |\psi_\ell| < \infty$$

in fact,

$$|\psi_\ell| \le \frac{1}{\rho^\ell}$$

which is the geometric decay. Also,

$$\psi(z) = \sum_{\ell=0}^{\infty} \psi_\ell z^\ell = \frac{\theta(z)}{\phi(z)} \quad (|z| \le 1)$$

In essence,

$$X_t = \frac{\theta(B)}{\phi(B)} W_t = \sum_{j=0}^{\infty} \psi_j B^j W_t$$

Key: $\dfrac{1}{\phi(z)} = \sum_{j=0}^{\infty} \phi_j z^j \quad |z| \le 1$. *$(1/\phi)$ has a convergent power series rep.* $|z| \le 1$

THEOREM 3.2.10: Invertibility

If $z_1, \dots, z_q$ are the zeros of $\theta(z)$ and $\min_{1\le i\le q}|z_i| > 1$, then X_t is invertible,

$$W_t = \sum_{\ell=0}^{\infty} \pi_\ell X_{t-\ell}$$

Coefficients $\{\pi_\ell\}_{\ell=0}^{\infty}$ satisfy

$$\pi(z) = \sum_{\ell=0}^{\infty} \pi_\ell z^\ell = \frac{\phi(z)}{\theta(z)} \quad (|z| \le 1)$$

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

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

3.3 ARMA Process Examples and ACF

EXAMPLE 3.3.1

Consider the ARMA$(2,2)$ model

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

Questions:

- Is there a stationary and causal solution to X_t?
- Is it invertible?
- Is there parameter redundancy?

AR polynomial:

$$\phi(z) = 1 - \frac{1}{4}z - \frac{1}{8}z^2$$

MA polynomial:

$$\theta(z) = 1 - \frac{5}{6}z + \frac{1}{6}z^2$$

Roots for ϕ:

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

Roots for θ: $2, 3$

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

Note that $\phi(z)$ and $\theta(z)$ share common $(z-2)$ which indicates that the parameters are redundant. Therefore, X_t satisfies an ARMA$(1,1)$ with

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

Since the roots of ϕ and θ are outside of the unit circle in $\mathbb{C}$, X_t is stationary, causal, and invertible.

EXAMPLE 3.3.2

Suppose

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

where $X_t \sim \text{ARMA}(1,1)$.

$$\phi(z) = 1 + \frac{1}{4}(z) \implies \text{Root is } -4.$$

So X_t is stationary and causal, and can be represented as a linear process:

$$X_t = \sum_{\ell=0}^{\infty} \psi_\ell w_{t-\ell}$$

We need to calculate the coefficients ψ_ℓ.
We know

$$\psi(z) = \sum_{\ell=0}^{\infty} \psi_\ell z^\ell = \frac{\theta(z)}{\phi(z)} \quad (|z| \leq 1)$$
$$\implies \psi(z)\phi(z) = \theta(z)$$

Note that both $\psi(z)\phi(z)$ and $\theta(z)$ are power series, therefore we can calculate ψ_ℓ by matching coefficients.

- $\phi(z) = 1 + \frac{1}{4}z$
- $\theta(z) = 1 - \frac{1}{3}z$
- $\psi(z)\phi(z) = \theta(z)$

Let's compute it.

$$\begin{aligned}
z^0: &\quad \psi_0 = 1 \\
z^1: &\quad \frac{\psi_0}{4} + \psi_1 = -\frac{1}{3} && \implies \psi_1 = \frac{-7}{12} \\
z^2: &\quad \frac{\psi_1}{4} + \psi_2 = 0 && \implies \psi_2 = -\frac{7}{12}\left(\frac{1}{4}\right) \\
&\vdots \\
z^\ell: &\quad \frac{\psi^{\ell-1}}{4} + \psi^\ell = 0 && \implies \psi^\ell = -\frac{7}{12}\left(\frac{1}{4}\right)^{\ell-1}
\end{aligned}$$

Finite linear difference equation must be solved. (Automated in R with `ARMAtoMA()`). If X_t is a stationary and causal solution to the ARMA(p,q) model.

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

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

Note that

$$t - j = t + h - k, \implies k = h + j, j = 0, 1, 2, \ldots \quad \mathbb{E}[X_{t-j}^2] = {\sigma_w}^2$$

Therefore,

$$\gamma_X(h) = \sigma_W^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$$

Coefficients can be solved for as in the previous examples by solving a finite difference equation. (Automated in R with `ARMAacf`).