STAT 443 - Forecasting

Cameron Roopnarine

Last updated: April 18, 2021

Contents

Chapter 1

Week 1

1.1 What is a time series?

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

In particular,

(1) $X_1, \ldots, X_n$ 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 upon 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 $X_1, \ldots, X_T$ 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 $X_1, \ldots, X_T$ is a **real-valued** (**scalar**) **time series**.

DEFINITION 1.1.4: Multivariate time series

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

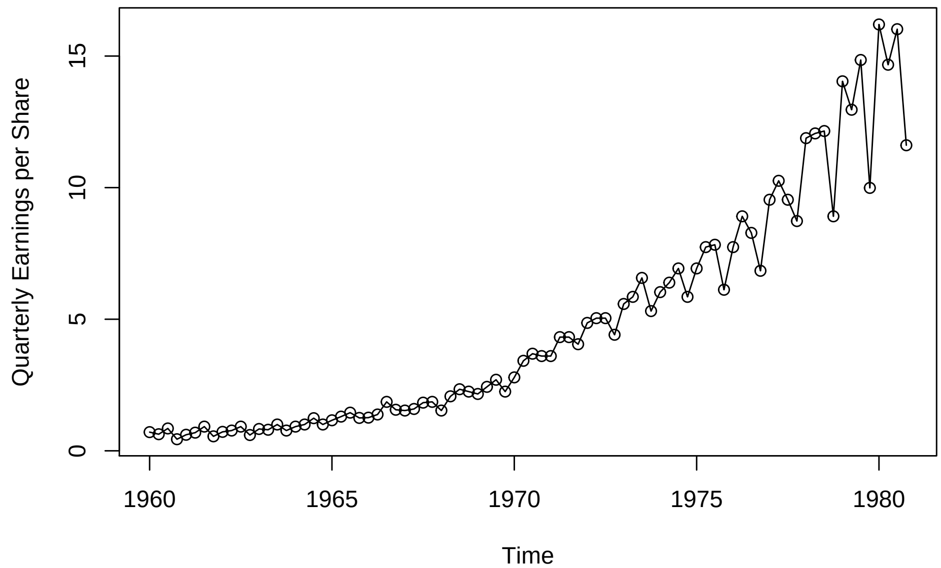


Figure 1.1: Quarterly Johnson and Johnson Earnings

```
# 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:

- The global temperature is steadily increasing over time.
- Heterogeneity exists within the mean over time.

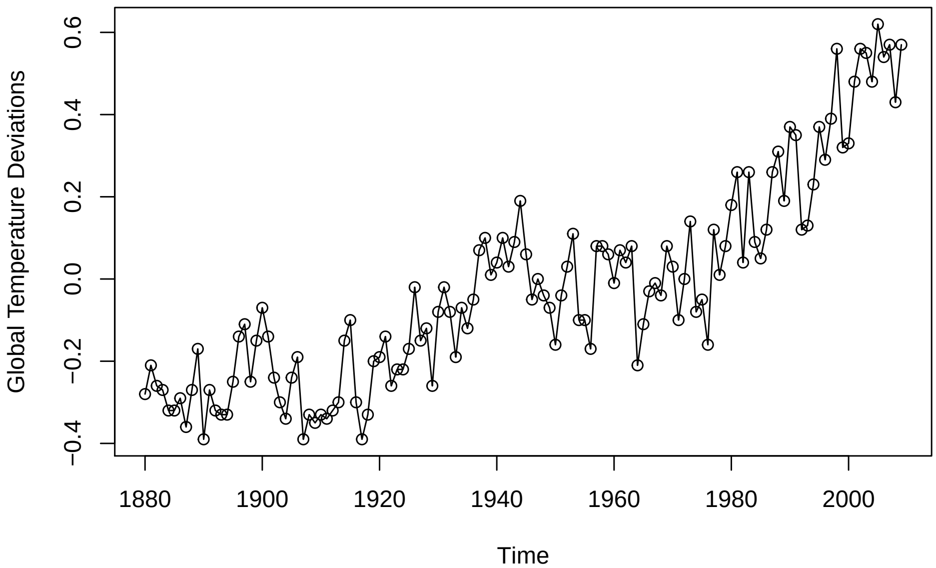


Figure 1.2: x_t is the deviation of global mean yearly temperature from the mean computed from 1951 to 1980

- 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, Observed stretch

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 $\mathbf{Z}$. 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 as can be seen in 1.3.

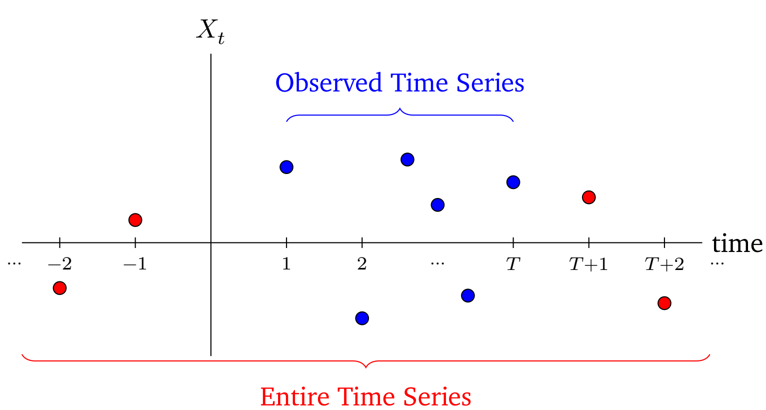


Figure 1.3: Time Series

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 by 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 $\{X_t\}_{t\in\mathbf{Z}}$

$$X_1, \dots, X_T \underset{\text{infer}}{\longrightarrow} \{X_t\}_{t\in\mathbf{Z}}$$

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

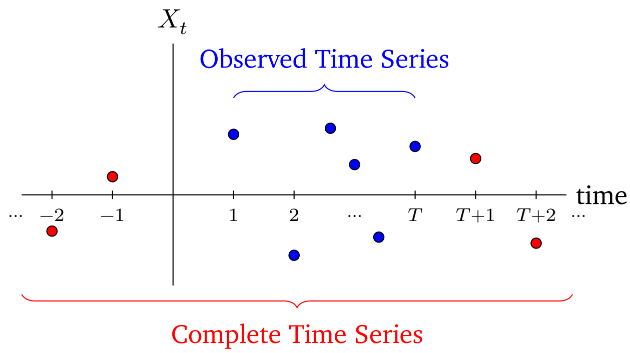
The worst case: $X_t \sim F_t$, where F_t is a *changing* function of t. If so, it is hard to pool the data $X_1, \dots, X_T$ to estimate F_t. If **serial dependence** occurs; that is, if the distribution of (X_t, X_{t+h}) depends strongly on t, then we have a similar problem in estimating e.g., $\mathrm{Cov}(X_t,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

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 $k \geq 1$, $i_1, \dots, i_k, h \in \mathbf{Z}$,

$$(X_{i_1}, \dots, X_{i_k}) \equiv (X_{i_1+h}, \dots, X_{i_k+h})$$

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 <u>not</u> 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

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

where 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 **first order stationary**.
- The property when $\gamma(t,s) = f(|t-s|)$ only depends on the lag $|t-s|$ is called **second order stationary**.
- For a second order stationary process,

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

Since $\gamma(h) = \gamma(-h)$, we normally only record $\gamma(h)$ for $h \geq 1$.

1.4 White Noise and Stationary Examples

DEFINITION 1.4.1: Strong white noise

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

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

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

DEFINITION 1.4.3: Gaussian white noise

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

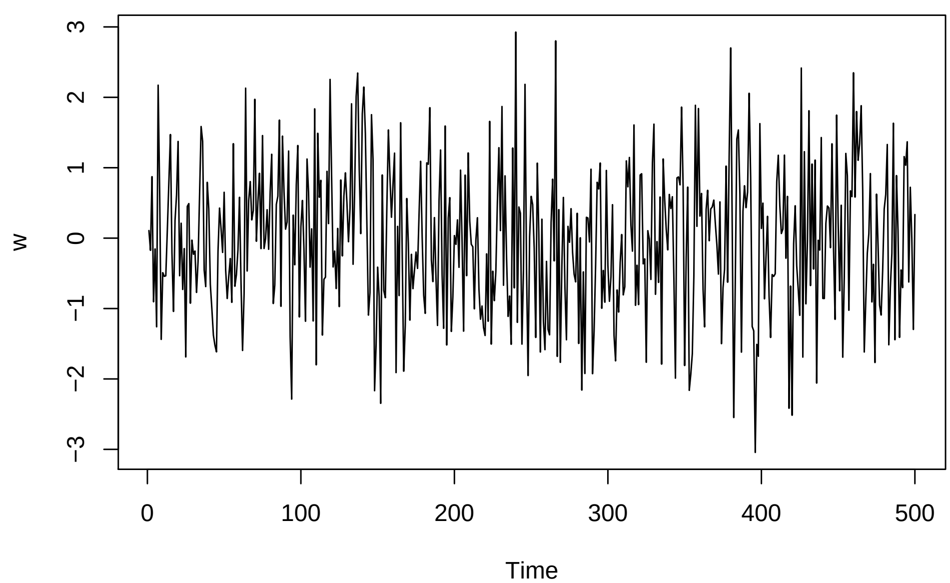


Figure 1.4: Gaussian White Noise of Length 500

```
# Figure 1.4
plot.ts(rnorm(500), main = "Gaussian White Noise", ylab = "w")
```

Figure 1.4 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 $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise, then $\mathbb{E}[W_t] = 0$; that is, the mean of W_t doesn't depend on t.

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

$\gamma(t,s)$ only depends on $|t-s|$. Therefore, $\{W_t\}_{t\in\mathbb{Z}}$ 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

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

where $\mathcal{B}$ 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**.

REMARK 1.4.7

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

THEOREM 1.4.8

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

REMARK 1.4.9

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

EXERCISE 1.4.10

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 <u>not</u> 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 <u>not</u> 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 <u>not</u> 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

$$(X_t) \equiv (X_0)$$

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

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

By the Cauchy-Schwarz inequality,

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

If $t < s$, then

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

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

$$(X_t, X_s) \equiv (X_{t-t}, X_{s-t}) \equiv (X_0, X_{s-t})$$

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 $k \in \mathbf{Z}_{\geq 1}$, $i_1 < i_2 < \cdots < i_k$ we have

$$(X_{i_1}, \dots, X_{i_k}) \sim \text{MVN}(\boldsymbol{\mu}_k(i_1,\dots,i_k), \varSigma_{k\times k}(i_1,\dots,i_k))$$

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

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.

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

Also,

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

Using the Gaussian assumption

$$(X_{i_1}, \dots, X_{i_k}) \equiv \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{k\times k}) \equiv (X_{i_1+h}, \dots, X_{i_k+h})$$

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

EXERCISE 1.5.5

Prove that if $\{X_t\}_{t\in\mathbf{Z}}$ is not 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 not strictly stationary.

1.6 † Theoretical L2 Framework for Time Series

- $X_t = \lim\limits_{h\to\infty} X_{h,t}$. 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; that is, $\mathbb{P}(X = 0) = 1$.

(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

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

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

The terms $\{W_t\}_{t\in\mathbf{Z}}$ are often called the "innovations" or "shocks" driving 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 $\varepsilon_t = g_t(W_t, W_{t-1}, \ldots)$ 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 1.5 is a small example of the global temperature data superimposed with the `lm()` estimate.

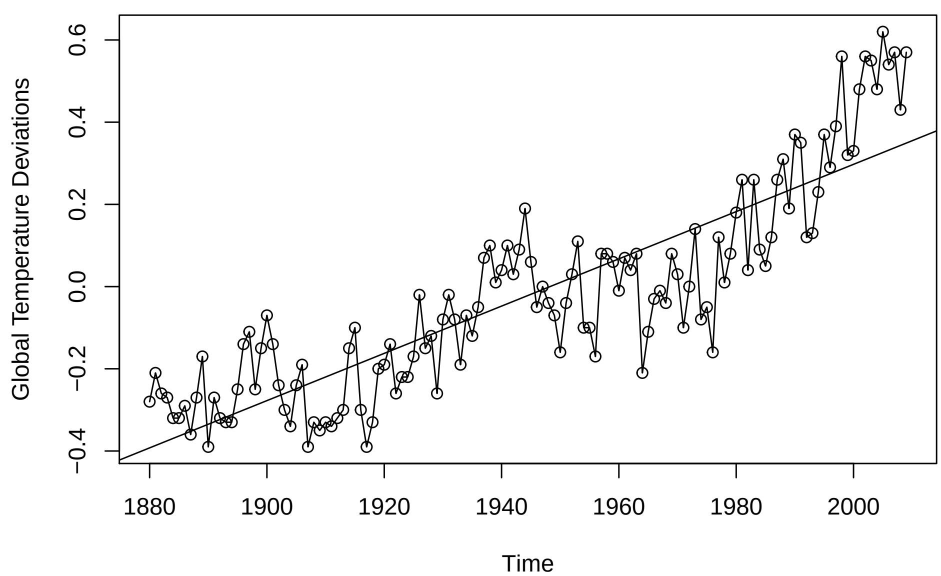


Figure 1.5: OLS estimate of linear trend

```
# Figure 1.5
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.

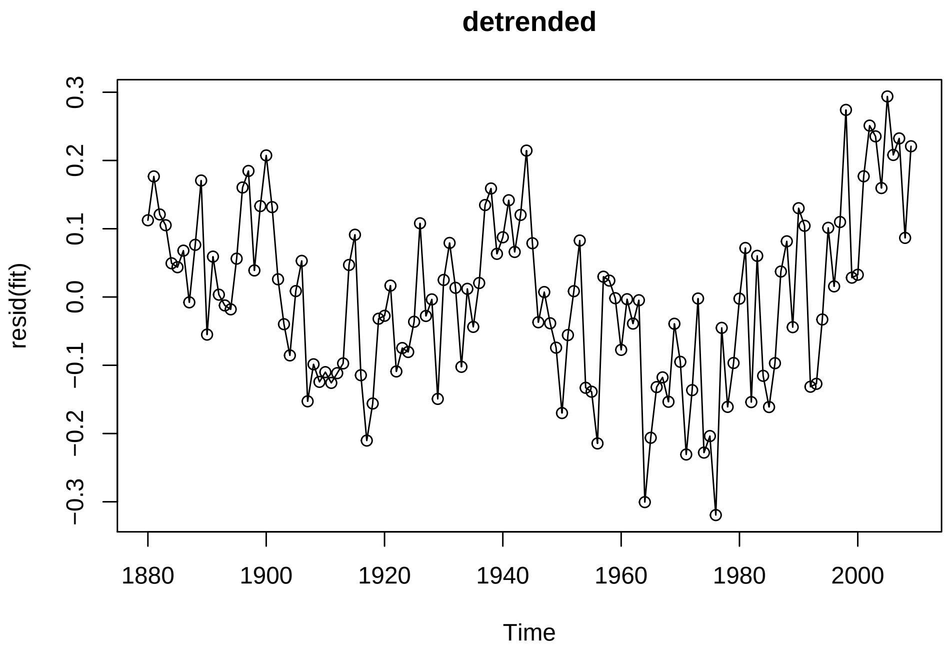


Figure 1.6: Residuals of OLS fit.

```
# Figure 1.6
plot(resid(fit), type = "o", main = "detrended")
```

In Figure 1.6: 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) which looks reasonably stationary. If the residuals were 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.

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

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.

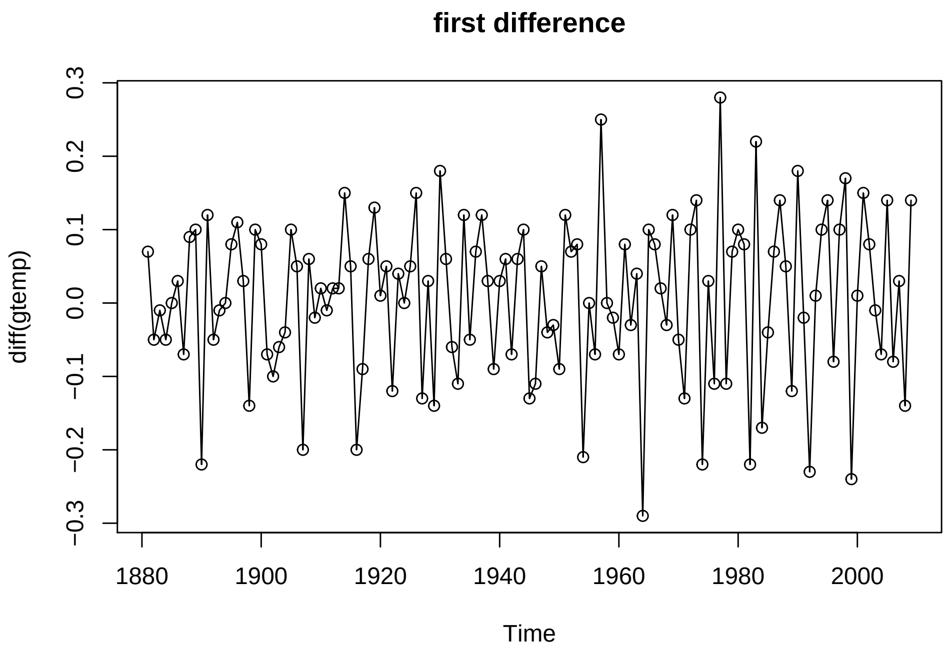


Figure 1.7: First differenced series. Average of first differenced series is $\hat{\delta} \approx 0.0066$

```
# Figure 1.7
plot(diff(gtemp), type = "o", main = "first difference")
```

In Figure 1.7: To see what this looks like in this temperature example, here is a plot of $\nabla X_t = X_t - X_{t-1}$ for Figure 1.2. 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:

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

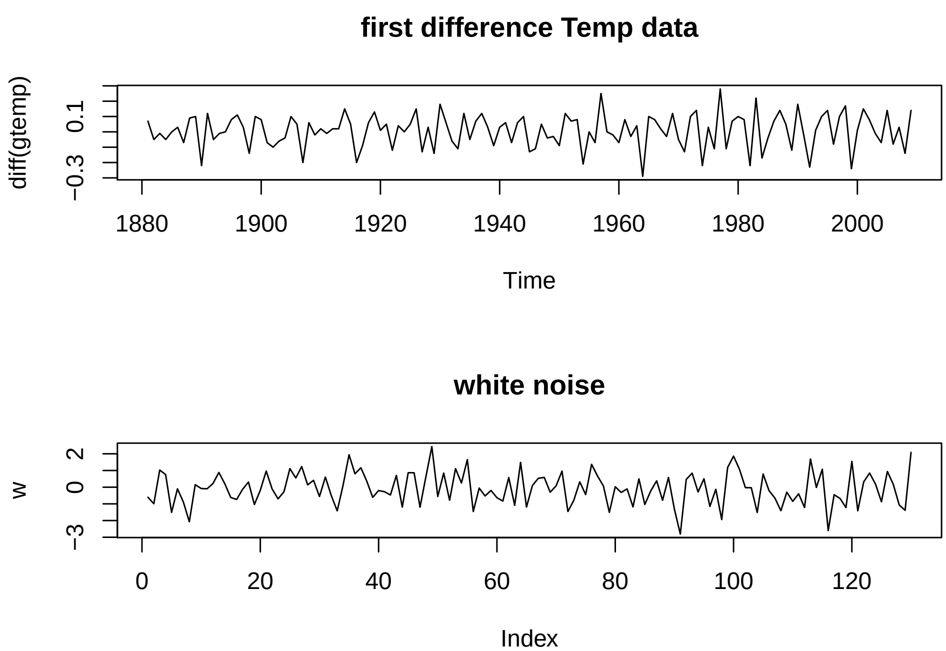


Figure 1.8: First Difference and White Noise

```
# Figure 1.8
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 1.8: If these two series behave in the same way, then it stands to reason that

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

Chapter 2

Week 2

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

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

$\{W_t\}_{t\in\mathbf{Z}}$ 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. 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

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

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

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

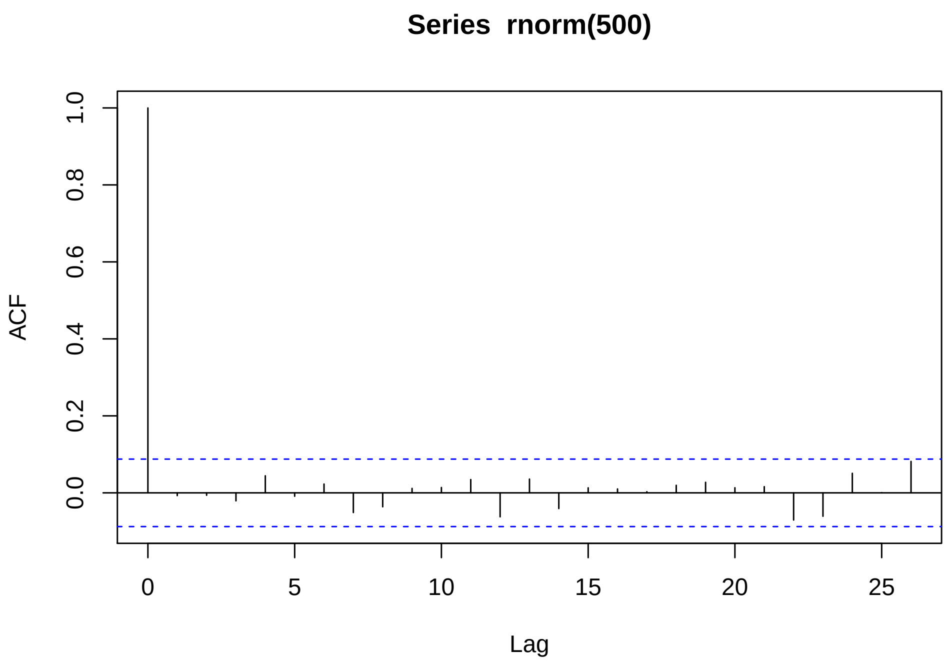


Figure 2.1: ACF of white noise, sample length 130

```
# 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 are 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.

Review/Introduce

(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 $X_n = o_p(1)$. Likewise, we also write $X_n \overset{p}{\to} X$ 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

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

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

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

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

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

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

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

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

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

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 $\theta \in \mathbf{R}$ where 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:

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

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 $\varepsilon > 0$,

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

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,

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

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

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.

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

due to strict stationarity.
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

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

Aside: For $\delta > 0$,

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

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

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

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

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 $\dagger\ 2+\delta$ 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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$,

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

Proof of Proposition 2.4.2

Since $\varphi(x) = |x|^{2+\delta}$ is convex where $a_1, \dots, a_n \in \mathbf{R}$, 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.

$$\sum_{t=1}^{a_T} X_t = \sum_{j=0}^{m} \sum_{\substack{t \equiv j \pmod{m+1} \\ 1 \le t \le a_T}} X_t$$

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

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

where C is the same constant as in Section 2.3.

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 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

$$\{Y_{m,n} : m, n \in \mathbf{Z}_{\ge 1}\}$$

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 $\varepsilon > 0$,

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

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,

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

by the Triangle Inequality. 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 an 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)

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

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 $X_1, \ldots, X_T$ is an observed time series in which we think was generated by a stationary process, then $\gamma(h) = \text{Cov}(X_t, X_{t+h})$ 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 without loss of generality that $\mathbb{E}[X_t] = 0$.

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

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

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,

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

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

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.

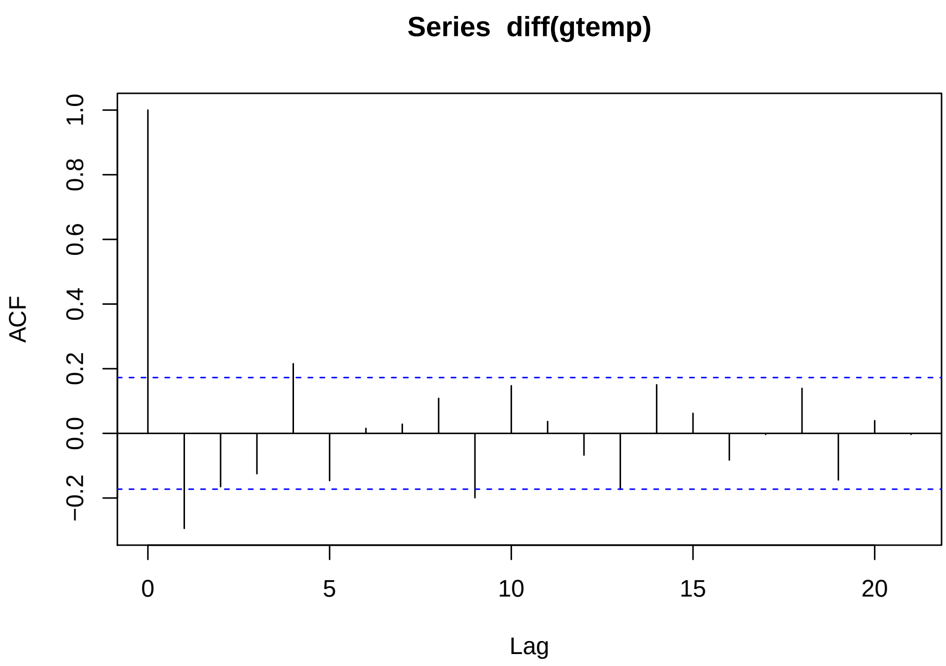


Figure 2.2: ACF of first differenced temperature data

```
# 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

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

Hence

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

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

```r
# 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."

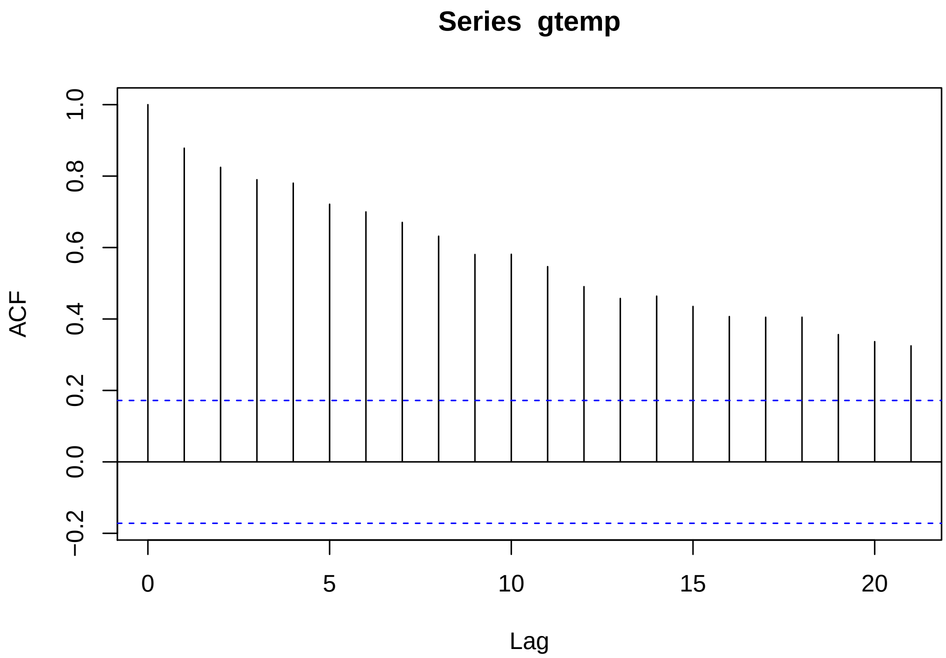


Figure 2.3: ACF of raw temperature data, sample length 130

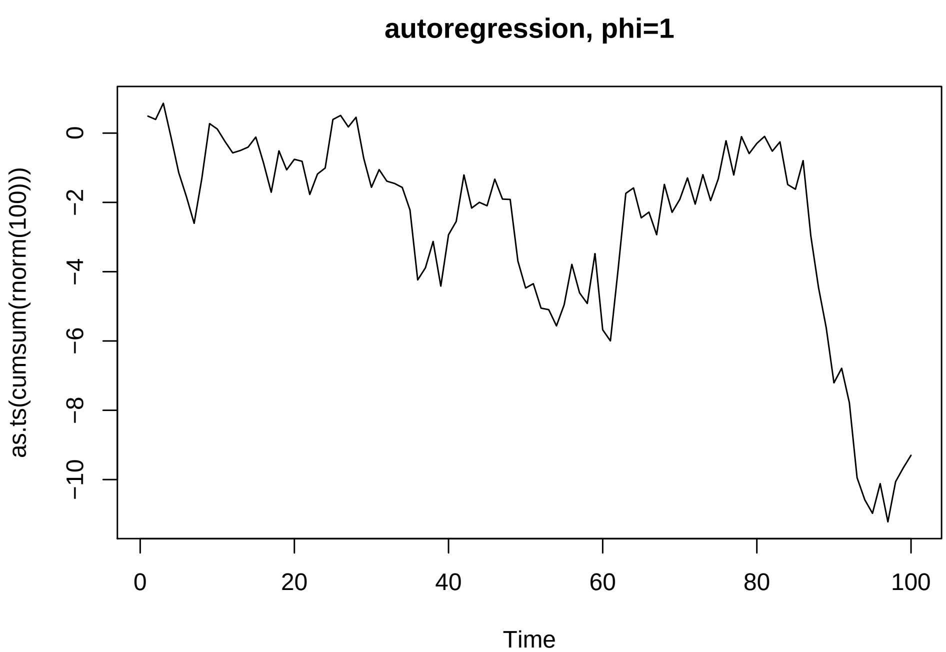


Figure 2.4: Realization of an AR(1) with $\phi = 1$ starting from $x_0 = 0$

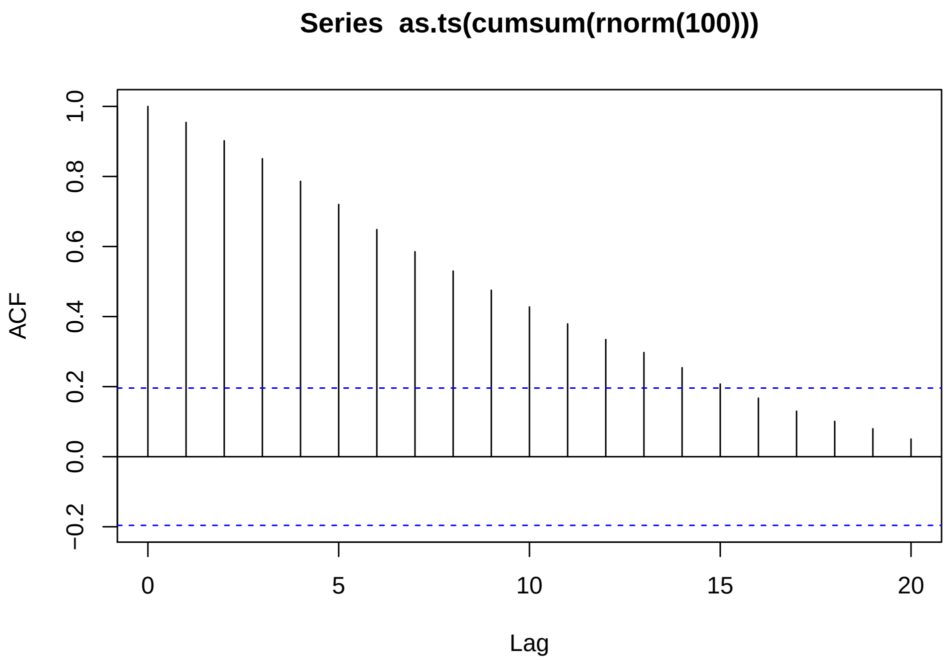


Figure 2.5: ACF of an AR(1) with $\phi = 1$ starting from $x_0 = 0$

- In Figure 2.5 there exists a similar pattern which is indicative of non-stationarity.

Chapter 3

Week 3

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

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

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

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 or MA(q), if there exists $\theta_1, \ldots, \theta_q \in \mathbf{R}$ 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 $\theta_0 = 1$. In other words, we've truncated the linear process representation at the level 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

- First difference of X_t:

$$\nabla X_t = (1 - B)X_t = X_t - BX_t = X_t - X_{t-1}$$

- Second difference of X_t:

$$\nabla^2 X_t = (1-B)^2 X_t = (1-2B+B^2)X_t = X_t - 2X_{t-1} + X_{t-2}$$

DEFINITION 3.1.4: Moving average operator

The **moving average operator** is defined by

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

DEFINITION 3.1.5: 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

$$\mathbb{E}[X_t] = \mathbb{E}\left[\sum_{\ell=0}^{q} \theta_\ell W_{t-\ell}\right] = 0 \quad (\theta_0 = 1)$$

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

$$\begin{aligned}
\gamma(h) &= \text{Cov}(X_t, X_{t+h}) \\
&= \mathbb{E}\left[\left(\sum_{\ell=0}^{q} \theta_\ell W_{t-\ell}\right)\left(\sum_{k=0}^{q} \theta_k W_{t+h-k}\right)\right] \qquad t-\ell = t+h-k \implies k = \ell + h \\
&= \begin{cases} \sigma_W^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & 1 \le h \le q \\ 0 & h > q \end{cases}
\end{aligned}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only display the values for $h \ge 0$. Note that $\gamma(q)$ cannot be zero because $\theta \neq 0$. The cutting off of $\gamma(h)$ after q lags is the signature of the $\text{MA}(q)$ model. Therefore,

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

REMARK 3.1.6

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.

In Figure 3.1, let's look an example now of what a moving average process would actually look like if we were to realize a moving average process. On the top of Figure 3.1, I've plotted a moving average process of order 0, which is just a strong white noise. Then, as we progress down to panel 2 and panel 3 I've calculated moving averages of orders 1 and 2 based on this strong white noise sequence. In the second panel, $X_t = W_t + W_{t-1}$, so this is a moving average process of order 1, in which $\theta_1 = 1$. In the third panel, we have a moving average process of order 2, in which $X_t = W_t + W_{t-1} + W_{t-2}$, which is a moving average process of order 2 where $\theta_1 = \theta_2 = 1$. One thing to observe when going from a moving average process of order 0 to 2 is that the time series is getting "smoother."

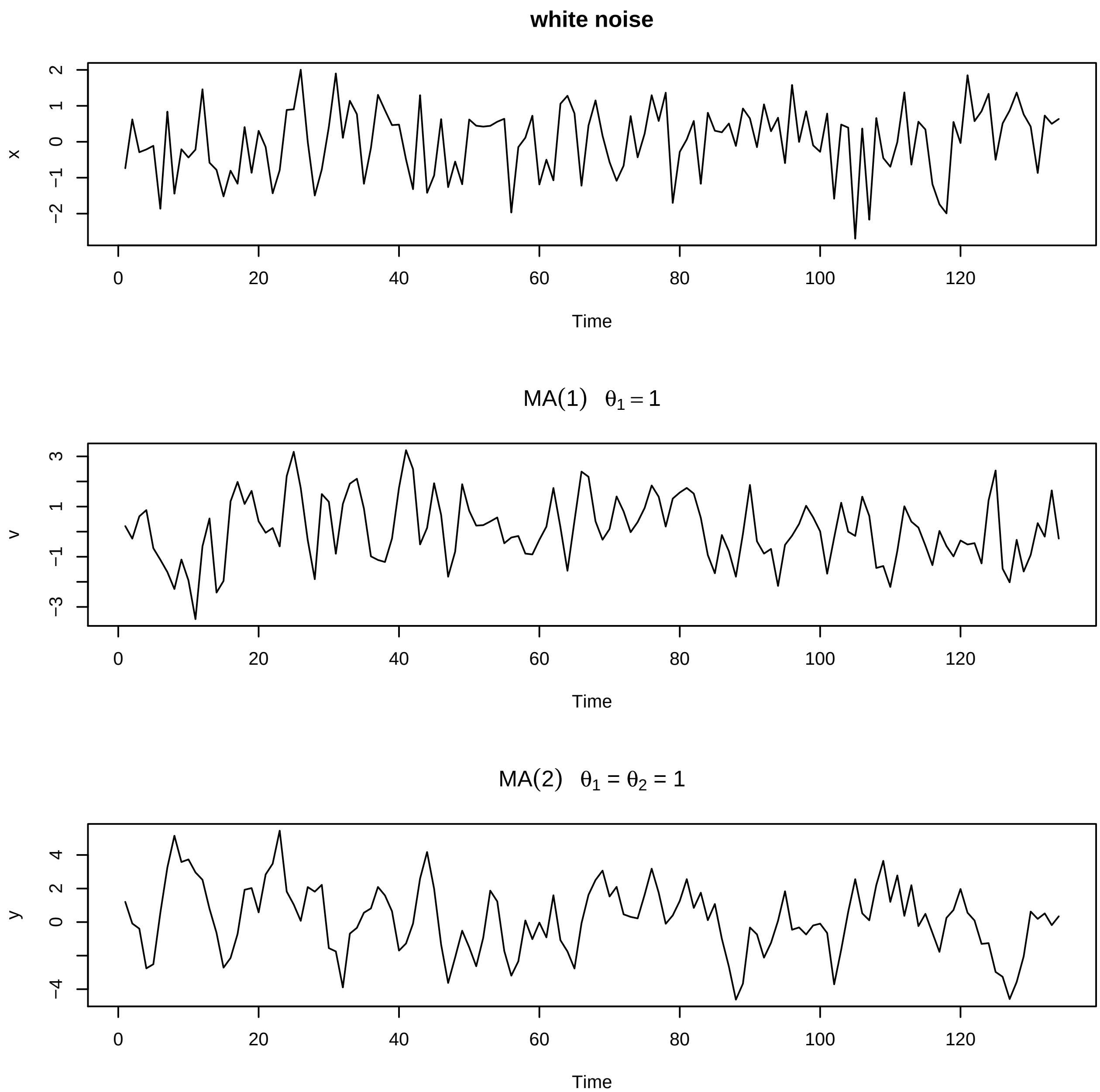


Figure 3.1: Realizations of MA processes with coefficients equal to 1

```
# Figure 3.1
par(mfrow = c(3, 1))
```

```
ma0.sim <- arima.sim(list(order = c(0, 0, 0), ma = c()), n = 134)
plot(ma0.sim, ylab = "x", main = "white noise")

ma1.sim <- arima.sim(list(order = c(0, 0, 1), ma = c(1)), n = 134)
plot(ma1.sim, ylab = "v", main = (expression(MA(1) ~  ~  ~ theta[1] == 1)))

ma2.sim <-
  arima.sim(list(order = c(0, 0, 2), ma = c(1, 1)), n = 134)
plot(ma2.sim, ylab = "y", main = (expression(paste(
  MA(2), ~  ~  ~ theta[1], " = ",  theta[2], " = ", 1
))))
```

In Figure 3.2, the difference is apparent since going from MA(0) to MA(1) shows that MA(1) has significant serial correlation at lag 1. Similarly, for MA(2) there is significant serial correlation at lag 2.

```
# Figure 3.2
acf(ma0.sim)
acf(ma1.sim)
acf(ma2.sim)
```

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, or 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.

Markov Property:

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

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

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

Let's see.

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

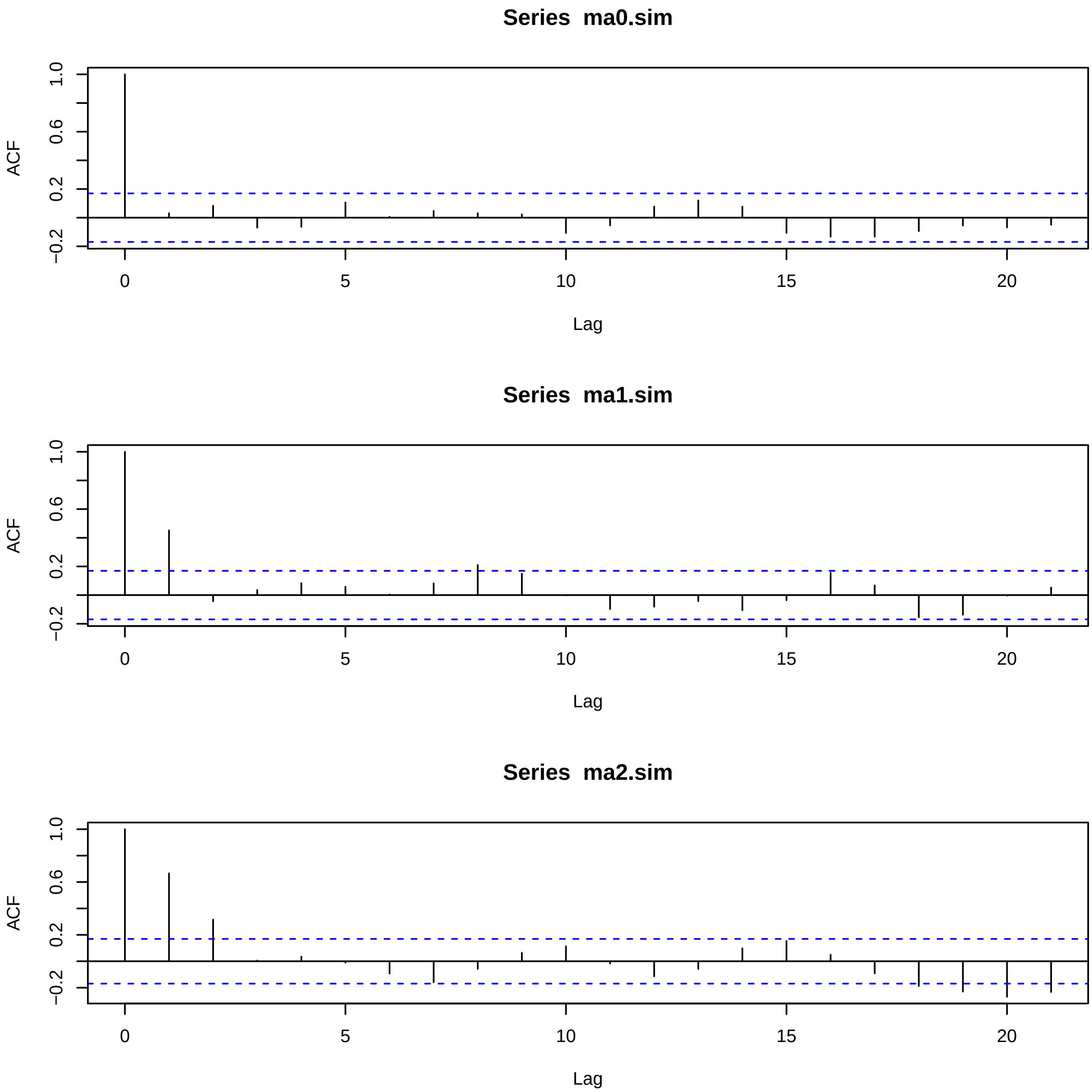


Figure 3.2: ACF plots of corresponding moving average series.

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

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

Therefore,

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

since $\displaystyle\sum_{j=1}^{\infty} \frac{1}{\phi^j} < \infty$. This sequence is strictly stationary since it is a Bernoulli shift. However, what we have derived is not desirable as this model is 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

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

Now, $\left|\text{Cov}(X_t, X_0)\right|^2 \le \mathbb{V}(X_t)\mathbb{V}(X_0) = (\mathbb{V}(X_0))^2$, so we get

$$\left|\text{Cov}(X_t, X_0)\right| \le \sqrt{\mathbb{V}(X_t)\mathbb{V}(X_0)} = \sqrt{(\mathbb{V}(X_0))^2} = \mathbb{V}(X_0)$$

Therefore, $-2\text{Cov}(X_t, X_0) \le 2\left|\text{Cov}(X_t, X_0)\right| \le 2\mathbb{V}(X_0)$. Finally,

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

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.

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

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

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.

```
# Figure 3.3
ar0.sim <- arima.sim(list(order = c(1, 0, 0), ar = c(0.5)), n = 134)
plot(ar0.sim, ylab = "x", main = (expression(AR(1) ~  ~  ~ phi[1] == 0.5)))

ar1.sim <- arima.sim(list(order = c(1, 0, 0), ar = c(0.9)), n = 134)
plot(ar1.sim, ylab = "y", main = (expression(AR(1) ~  ~  ~ phi[1] == 0.9)))

ar2.sim <-
  arima.sim(list(order = c(1, 0, 0), ar = c(-0.9)), n = 134)
plot(ar2.sim, ylab = "z", main = (expression(AR(1) ~  ~  ~ phi[1] == -0.9)))

# Figure 3.4
acf(ar0.sim)
acf(ar1.sim)
acf(ar2.sim)
```

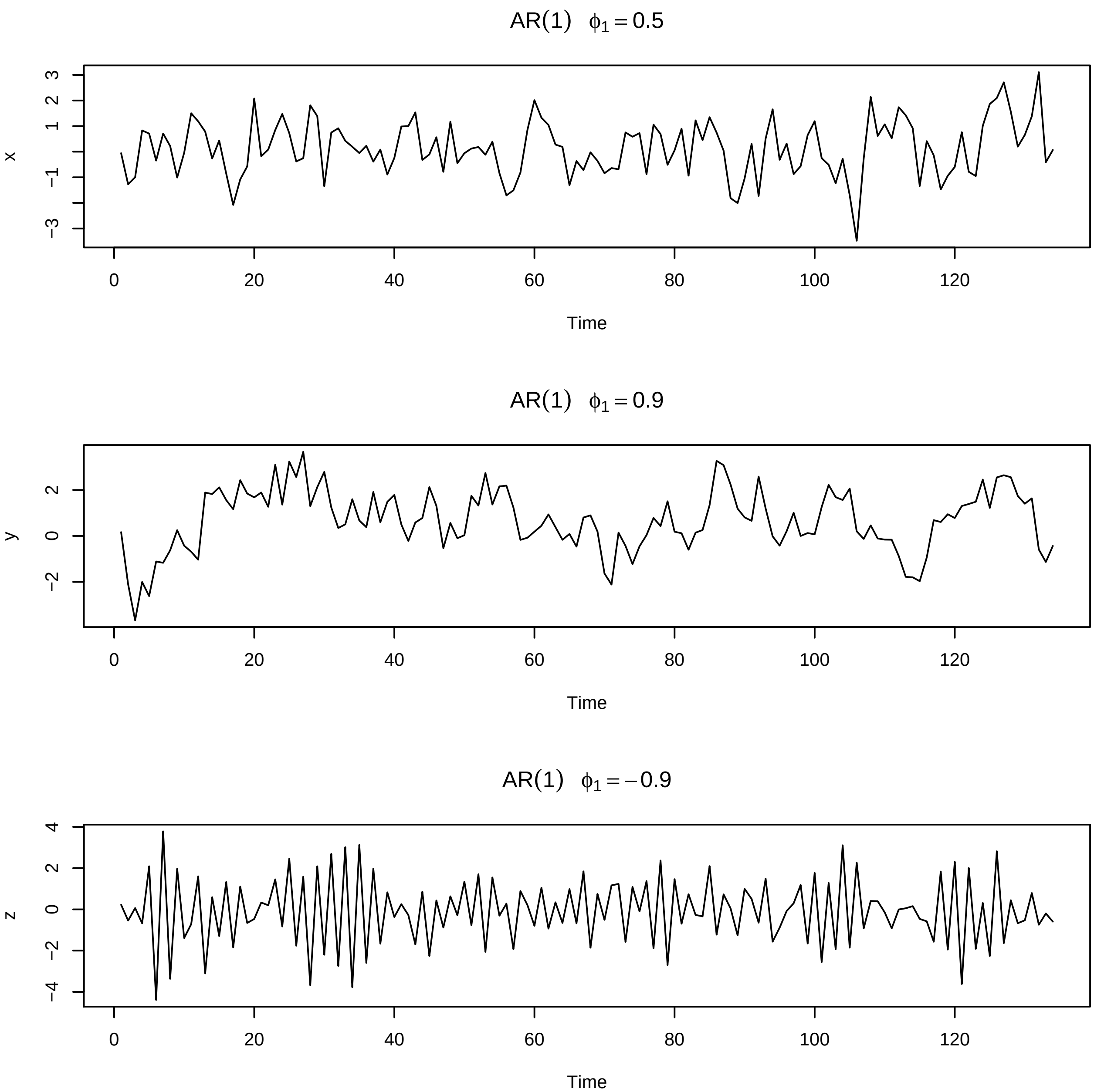


Figure 3.3: Realizations of AR(1) processes

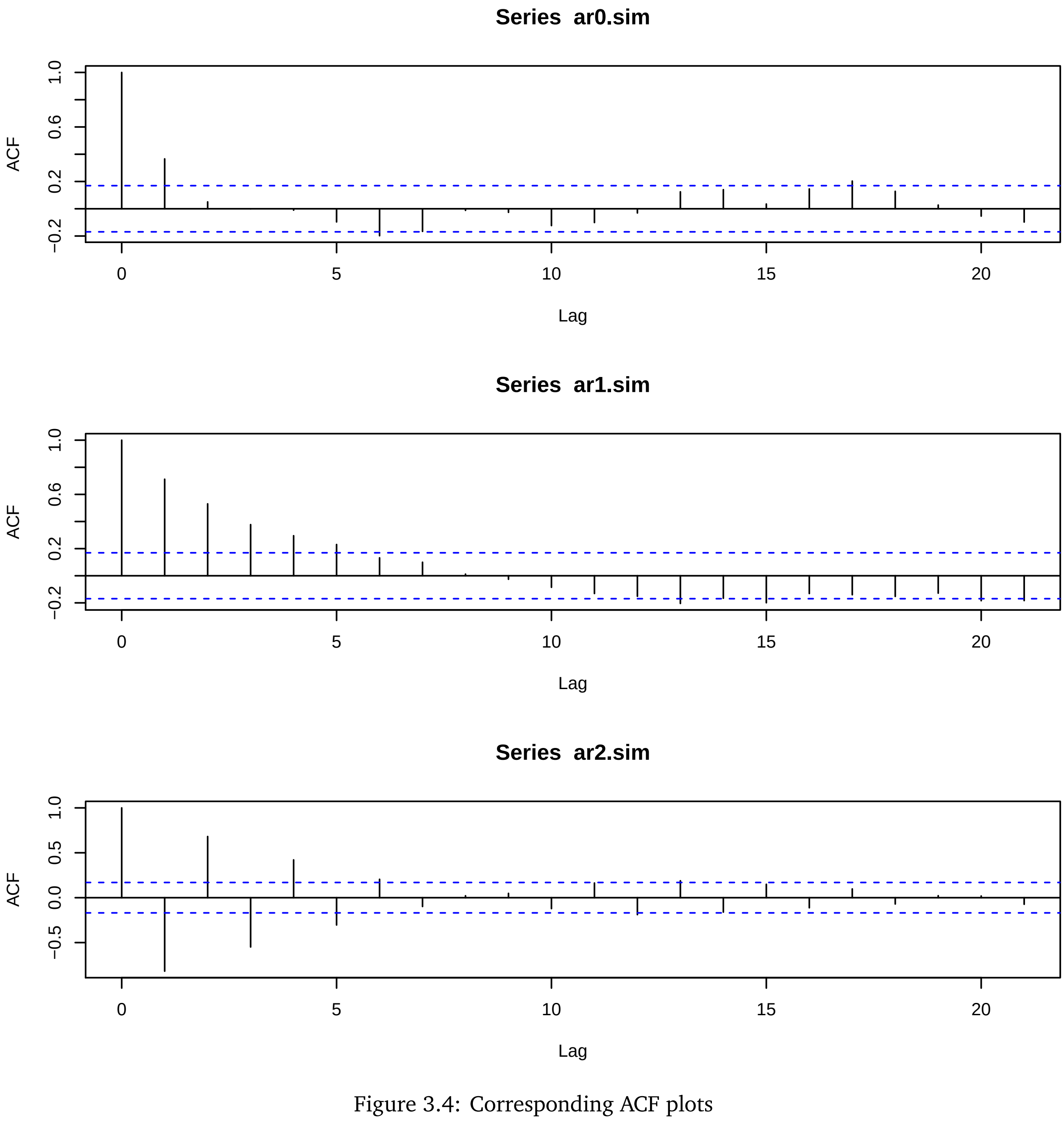


Figure 3.4: Corresponding ACF plots

DEFINITION 3.2.3: Autoregressive process, Autoregressive polynomial

We say X_t follows an **autoregressive process** of order p, or 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 $\phi(B)X_t = W_t$.

3.3 ARMA Processes

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:

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

If $W_t \sim$ strong white noise

$$X_t = \theta(B)W_t \quad (X_t \sim \text{MA}(q))$$
$$\phi(B)X_t = W_t \quad (X_t \sim \text{AR}(p))$$

Why not combine the two?

DEFINITION 3.3.1: Autoregressive moving average

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

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

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

This implies that the model is

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

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 3.3.2: Parameter redundancy

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

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

Therefore,

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

$$\phi(z) = 1 - 0.5z \implies \text{zero of } \phi \text{ is } z_0 = 2$$

$$\theta(z) = 1 - 0.5z \implies \text{zero of } \theta \text{ is } z_0 = 2$$

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 3.3.3: 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

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

where $\psi(B) = \sum_{\ell=0}^{\infty} \psi_\ell B^\ell$ and $\sum_{\ell=0}^{\infty}|\psi_\ell| < \infty$ with $\psi_0 = 1$.

DEFINITION 3.3.4: 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 3.3.5

Causality + Invertibility $\implies$ Information in $\{X_t\}_{t\le T}$ is the same as Information in $\{W_t\}_{t\le T}$ where $\{X_t\}_{t\le T}$ is an observed time series.

THEOREM 3.3.6: 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: $\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \phi_j z^j$ where $|z| \le 1$ so $\frac{1}{\phi}$ has a convergent power series representation for $|z| \le 1$.

THEOREM 3.3.7: 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

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

3.4 ARMA Process Examples and ACF

EXAMPLE 3.4.1

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 the unit circle in $\mathbb{C}$, X_t is stationary, causal, and invertible.

EXAMPLE 3.4.2

Suppose

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

where $X_t \sim$ 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

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

$$\implies \psi(z)\phi(z) = \theta(z)$$

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.

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

Simplifying,

$$
\psi_j = \begin{cases} 1 & j = 0 \\ \frac{7}{3}\left(-\frac{1}{4}\right)^j & j \geq 1 \end{cases}
$$

We can automate ψ_j in R with `ARMAtoMA()`.

```
library(astsa)
ARMAtoMA(ar=-1/4, ma=-1/3, 10)
```

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,

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

We can automate $\gamma_X(h)$ in R with `ARMAacf()`.
For $h \geq 1$, we have

$$
\begin{aligned}
\gamma_X(h) &= \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \\
&= \psi_0 \psi_h + \sum_{j=1}^{\infty} \psi_j \psi_{j+1} \\
&= \frac{7}{3}\left(-\frac{1}{4}\right)^h + \sum_{j=1}^{\infty}\left[\frac{7}{3}\left(-\frac{1}{4}\right)^j \frac{7}{3}\left(-\frac{1}{4}\right)^{j+1}\right] \\
&= \frac{91}{135}(-1)^h 4^{1-h}
\end{aligned}
$$

Then,

$$\begin{aligned}
\gamma_X(0) &= \sum_{j=0}^{\infty} \psi_j^2 \\
&= (1)^2 + \sum_{j=1}^{\infty} \psi_j^2 \\
&= 1 + \sum_{j=1}^{\infty} \frac{7}{3}\left(-\frac{1}{4}\right)^j \\
&= \frac{184}{135}
\end{aligned}$$

Therefore, the ACF for $h \geq 1$ is given by

$$\rho_X(h) = \begin{cases} 1 & h = 0 \\ \dfrac{\gamma_X(h)}{\gamma_X(0)} = \dfrac{\frac{91}{135}(-1)^h 4^{1-h}}{\frac{184}{135}} = \dfrac{91}{23}(-1)^h 2^{-2h-1} & h \geq 1 \end{cases}$$

Let's verify this in R.

```
round(ARMAacf(ar = -1 / 4, ma = -1 / 3, 5), 6)
h <- seq(1, 10, by = 1)
round((91 / 23) * (-1) ^ h * 2 ^ (-2 * h - 1), 6)
```

Output:

```
        0         1         2         3         4         5
 1.000000 -0.494565  0.123641 -0.030910  0.007728 -0.001932

          -0.494565  0.123641 -0.030910  0.007728 -0.001932
```

As we can see, this is correct.

Chapter 4

Week 4

4.1 Stationary Process Forecasting

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

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

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

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

where L is a loss function.

Frequently, the loss function is taken to be the *mean-squared error* (MSE)

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

When using MSE, it is natural to consider

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

L^2 is a Hilbert space when equipped with the inner product

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

Hilbert spaces are generalizations of Euclidean space ($\mathbf{R}^d$) in which the geometry and notation of projection are preserved.

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

DEFINITION 4.1.1: Closed Linear Subspace

We say $\mathcal{M} \subseteq L^2$ is a **closed linear subspace**, if

(i) Linearity: $X, Y \in \mathcal{M}$, $\alpha, \beta \in \mathbf{R}$ then $\alpha X + \beta Y \in \mathcal{M}$

(ii) Closed: If $X_n \to X$ (in the sense that $\mathbb{E}[(X_n - X)^2] \to 0$), and $X_n \in \mathcal{M}$, then $X \in \mathcal{M}$.

THEOREM 4.1.2: Projection Theoren

If $\mathcal{M}$ is a closed linear subspace in L^2 and $x \in L^2$, then there exists a unique $\hat{X} \in \mathcal{M}$ such that

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

Moreover, $\hat{X}$ satisfies the prediction equations/normal equations:

$$(X - \hat{X}) \in \mathcal{M}^{\perp} \implies \mathbb{E}[(X - \hat{X})Y] = 0 \quad (\forall Y \in \mathcal{M})$$

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

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

where $\mathcal{M}$ is a closed linear subspace based on the available data.

(1) $\mathcal{M}_1 = \{z : z = f(X_T, \dots, X_1), f$ is any Borel Measurable function$\}$. In this case

$$\hat{X}_{T+h} = \mathbb{E}[X_{T+h} \mid X_T, \dots, X_1]$$

Unfortunately $\mathcal{M}_1$ is enormous and complicated!

(2) $\mathcal{M}_2 = \overline{\text{Span}}(1, X_T, \dots, X_1) = \{Y : Y = \alpha_0 + \sum_{j=1}^{T} \alpha_j X_j,\ \alpha_0, \dots, \alpha_T \in \mathbf{R}\}$ which is the linear functions of $X_1, \dots, X_T$. $\hat{X}_{T+h}$ is called the **best linear predictor** (BLP).

4.2 Best Linear Prediction

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

$$\mathbb{E}[(X_{T+h} - \hat{X}_{T+h})^2] = \inf_{Y \in \mathcal{M}_2} \mathbb{E}[(X_{T+h} - Y)^2]$$

$\hat{X}_{T+h}$ is the best prediction among all linear functions of $X_T, \dots, X_1$.

DEFINITION 4.2.1: Projection

If $\hat{X}$ satisfies

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

we say that $\hat{X}$ is the **projection** of X onto $\mathcal{M}$, and we write $\hat{X} = \text{Proj}(X \mid \mathcal{M})$.

In particular, the BLP is

$$\hat{X}_{T+h} = \text{Proj}(X_{T+h} \mid \mathcal{M}_2)$$

Consider the case when $h = 1$. From the Projection Theorem, the BLP is of the form

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

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

$$\mathbb{E}[(X_{T+1} - \hat{X}_{T+1})Y] = 0 \quad (\forall Y \in \mathcal{M}_2)$$

In particular,

$$\mathbb{E}[(X_{T+1} - \hat{X}_{T+1})1] = 0 \quad (Y = 1)$$

$$\mathbb{E}[(X_{T+1} - \hat{X}_{T+1})X_j] = 0 \quad (1 \le j \le T,\, Y = X_j)$$

We have $T+1$ equations. Since $\mathbb{E}[X_j - \mu] = 0$,

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

Before proceeding, note that this implies

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

So we may assume without loss of generality that $\mu = 0$, therefore $\mathbb{E}[X_i X_j] = \gamma(j-i)$. Therefore,

$$0 = \mathbb{E}[(X_{T+1} - \hat{X}_{T+1})X_k] = \gamma(T+1-k) - \sum_{j=1}^{T} \phi_{T,j}\gamma(j-k) \quad (1 \le k \le T)$$

Therefore, we have linear system of equations for $\phi_{T,1}, \dots, \phi_{T,T}$:

$$\sum_{j=1}^{T} \phi_{T,j}\gamma(j-k) = \gamma(T+1-k)$$

Let

$$\boldsymbol{\gamma}_T = \begin{pmatrix} \gamma(T) \\ \vdots \\ \gamma(1) \end{pmatrix} \in \mathbf{R}^T$$

$$\mathit{\Gamma}_T = [\gamma(j-k), 1 \le j,k \le T] \in \mathbf{R}^{T\times T}$$

$$\boldsymbol{\phi}_T = (\phi_{T,1}, \dots, \phi_{T,T})^\top \in \mathbf{R}^T$$

this linear system may be expressed as

$$\mathit{\Gamma}_T \boldsymbol{\phi}_T = \boldsymbol{\gamma}_T \implies \boldsymbol{\phi}_T = \mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T$$

given that $\mathit{\Gamma}_T$ is invertible.

The BLP is of the form

$$\hat{X}_{T+1} = \boldsymbol{\phi}_T^\top \boldsymbol{X}_T = (\mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T)^\top \boldsymbol{X}_T$$

where $\boldsymbol{X}_T = (X_T, \dots, X_1)^\top \in \mathbf{R}^T$.

When is $\mathit{\Gamma}_T$ non-singular?

THEOREM 4.2.2

If $\gamma(0) > 0$, and $\gamma(h) \to 0$ as $h \to \infty$, then $\mathit{\Gamma}_T$ is non-singular.

Takeaway: Most stationary processes (those whose serial dependence decays over time) have non-singular $\mathit{\Gamma}_T$.

Note that

$$\hat{X}_{T+1}^2 = \boldsymbol{\gamma}_T^\top \mathit{\Gamma}_T^{-1} \boldsymbol{X}_T \boldsymbol{X}_T^\top \mathit{\Gamma}_T^{-1} \boldsymbol{\gamma}_T$$

Note that $\mathbb{E}[\boldsymbol{X}_T \boldsymbol{X}_T^\top] = \mathit{\Gamma}_T$. Therefore, $\mathbb{E}[\hat{X}_{T+1}^2] = \boldsymbol{\gamma}_T^\top \mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T$. Also, since

$$\mathbb{E}[X_{T+1}\boldsymbol{X}_T] = \boldsymbol{\gamma}_T \implies \mathbb{E}[X_{T+1}\hat{X}_{T+1}] = \boldsymbol{\gamma}_T^\top \mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T$$

It follows that the mean-squared prediction error is

$$\begin{aligned} P_{T+1}^T &= \mathbb{E}[(X_{T+1} - \hat{X}_{T+1})^2] \\ &= \mathbb{E}[X_{T+1}^2 - 2X_{T+1}\hat{X}_{T+1} + \hat{X}_{T+1}^2] \\ &= \gamma(0) - 2\boldsymbol{\gamma}_T^\top \mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T + \boldsymbol{\gamma}_T^\top \mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T \\ &= \gamma(0) - \boldsymbol{\gamma}_T^\top \mathit{\Gamma}_T^{-1}\boldsymbol{\gamma}_T \end{aligned}$$

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

4.3 Partial ACF

If $X_t \sim \text{ARMA}(p,q)$, then we might be able to identify p, q by looking at the ACF.

$$X_t \sim \text{AR}(p) \implies \text{ACF has a geometric decay}$$

$$X_t \sim \text{MA}(q) \implies \text{ACF is non-zero at the first } q \text{ lags, then zero beyond}$$

ACF of an ARMA(p,q) model can be calculated by calculating the linear process coefficients $\{\psi\}_{\ell=0}^{\infty}$. Automated in R using `ARMAacf()`.

In Figure 4.1, it looks like geometric decay. However, it is hard to tell the difference between the ARMA$(1,1)$ process and the AR(p) process via the ACF. Therefore, we want to define the *partial autocorrelation function*.

```
# Figure 4.1 (Omitted the PACF)
ACF = ARMAacf(ar = c(.8), ma = 1, 24)[-1]
PACF = ARMAacf(ar = c(.8),
               ma = 1,
               24,
               pacf = TRUE)
par(mfrow = c(1, 2))
plot(ACF,
     type = "h",
     xlab = "lag",
     ylim = c(-.8, 1))
abline(h = 0)
plot(PACF,
     type = "h",
     xlab = "lag",
     ylim = c(-.8, 1))
abline(h = 0)
```

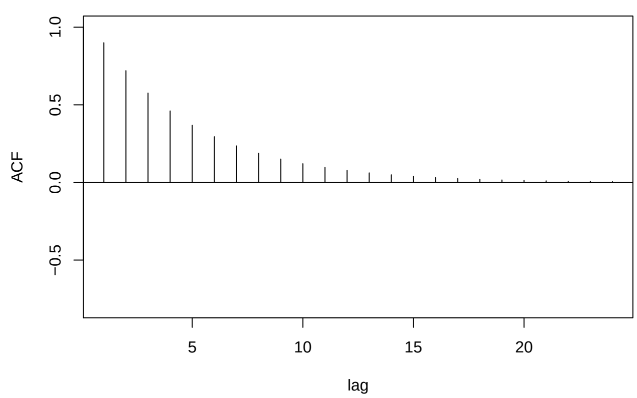


Figure 4.1: ARMA$(1,1)$: $X_t = 0.9X_{t-1} + W_t + 0.5W_{t-1}$

DEFINITION 4.3.1: Partial autocorrelation function

The **partial autocorrelation function** of a stationary process $\{X_t\}_{t\in\mathbb{Z}}$ is

$$\phi_{h,h} = \text{Corr}(X_{t+h} - \text{Proj}(X_{t+h} \mid X_{t+h-1}, \dots, X_{t+1}), X_t - \text{Proj}(X_t \mid X_{t+h-1}, \dots, X_{t+1}))$$

Interpretation: Autocorrelation between X_t and X_{t+h} after removing the linear dependence on the intervening variables $X_{t+h-1}, \dots, X_{t+1}$.

REMARK 4.3.2

If $X_t \sim \text{AR}(p)$, then $\phi_{h,h} = 0$ for $h \geq p+1$.

Proof of Remark 4.3.2

If $X_t \sim \text{AR}(p)$, then $X_{t+h} = \sum_{j=1}^{p} \phi_j X_{t+h-j} + W_{t+h}$.

$$\text{Proj}(X_{t+h} \mid X_{t+h-1}, \dots, X_{t+1}) = \sum_{k=1}^{h-1} \beta_k X_{t+h-k}$$

and minimizes

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

where the second term is minimized by setting $\beta_j = \phi_j$ for $1 \leq j \leq p$ and $\beta_j = 0$ for $j \geq p$. Note that W_{t+h} is independent of other terms. Hence,

$$X_{t+h} - \text{Proj}(X_{t+h} \mid X_{t+h-1}, \dots, X_{t+1}) = W_{t+h} \quad (h \geq p+1)$$

Therefore,

$$\phi_{h,h} = \text{Corr}(W_{t+h}, X_t - \text{Proj}(X_t \mid X_{t+h-1}, \dots, X_{t+1}))$$

which is independent by causality. Therefore, $\phi_{h,h} = 0$.

REMARK 4.3.3

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

$$\phi_{h,h} \neq 0$$

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

which is geometric decay.

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

Estimating the PACF

Using the BLP theory,

$$\hat{\phi}_{h,h} = (\hat{\Gamma}_h^{-1} \hat{\boldsymbol{\gamma}}_h)(h)$$

where

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

$$\hat{\boldsymbol{\gamma}}_h = [\hat{\gamma}(1), \dots, \hat{\gamma}(h)] \in \mathbf{R}^h$$

4.4 ARMA Forecasting

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

$$\hat{X}_{T+h} = \text{Proj}(X_{T+h} \mid \mathcal{M}_2) \approx \mathbb{E}[X_{T+h} \mid X_T, \dots, X_1]$$

by causality and invertibility $X_t \sim$ linear function of W_t.

Furthermore,

$$\hat{X}_{T+h} \approx \tilde{X}_{T+h} = \mathbb{E}[X_{T+h} \mid X_T, \dots, X_1, X_0, \dots]$$

which is geometric decay of the dependence on past values.

Since X_t is casual and invertible,

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

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

where $\psi_0 = \pi_0 = 1$. Note that ψ's and π's are computable by solving homogeneous linear difference equations.

These representations imply,

$$\text{Information in } (X_T, X_{T-1}, \dots) = \text{Information in } (W_T, W_{T-1}, \dots)$$

So

$$\tilde{X}_{T+h} = \mathbb{E}[X_{T+h} \mid X_T, X_{T-1}, \dots] = \mathbb{E}[X_{T+h} \mid W_T, W_{T-1}, \dots]$$

$$\begin{aligned}
\tilde{X}_{T+h} &= \mathbb{E}\left[\sum_{\ell=0}^{\infty} \psi_\ell W_{T+h-\ell} \,\middle|\, W_T, W_{T-1}, \dots\right] \\
&= \mathbb{E}\left[\sum_{\ell=0}^{h-1} \psi_\ell W_{T+h-\ell} \,\middle|\, W_T, \dots\right] + \mathbb{E}\left[\sum_{\ell=h}^{\infty} \psi_\ell W_{T+h-\ell} \,\middle|\, W_T, \dots\right] \\
&= \sum_{\ell=h}^{\infty} \psi_\ell W_{T+h-\ell} && \text{since } \psi_\ell W_{T+h-\ell} = 0 \text{ for } 0 \le \ell \le h-1
\end{aligned}$$

Also, using invertibility,

$$0 = \mathbb{E}[W_{T+h} \mid X_T, X_{T-1}, \dots] = \mathbb{E}\left[\sum_{\ell=0}^{\infty} \pi_\ell X_{T+h-\ell} \,\middle|\, X_T, \dots\right]$$

by independence, and furthermore, with $\pi_0 = 1$ we have

$$0 = \tilde{X}_{T+h} + \sum_{\ell=1}^{h-1} \pi_\ell \tilde{X}_{T+h-\ell} + \sum_{\ell=h}^{\infty} \pi_\ell X_{T+h-\ell}$$

Therefore,

$$\tilde{X}_{T+h} = -\sum_{\ell=1}^{h-1} \pi_\ell \tilde{X}_{T+h-\ell} - \sum_{\ell=h}^{\infty} \pi_\ell X_{T+h-\ell}$$

Truncated ARMA Prediction

$$\hat{X}_{T+h} = -\sum_{j=1}^{h-1} \pi_j \hat{X}_{T+h-j} - \sum_{j=h}^{T+h-1} \pi_j X_{T+h-j}$$

Residuals:

$$\hat{W}_t = \phi(B)\hat{X}_t - \theta_1 \hat{W}_{t-1} - \cdots - \theta_q \hat{W}_{t-q}$$

Mean initialization:

- $\hat{W}_t = 0$ for $t \leq 0$ and $t \geq T$.
- $\hat{X}_t = 0$ for $t \leq 0$ and $t \geq T+1$.
- $\hat{X}_t = X_t$ for $1 \leq t \leq T$.

Estimator for σ_W^2:

$$\hat{\sigma}_W^2 = \frac{1}{T}\sum_{t=1}^{T} \hat{W}_t^2$$

Mean Squared Prediction Error: Since $\hat{X}_{T+h} \approx \sum_{j=h}^{\infty} \psi_j W_{t+h-j}$,

$$P_{T+h}^T = \mathbb{E}\Big[(X_{T+h} - \hat{X}_{T+h})^2\Big] = \mathbb{E}\left[\left(\sum_{j=0}^{h-1} \psi_j W_{t-j}\right)^2\right] = \sigma_W^2 \sum_{j=0}^{h-1} \psi_j^2$$

Estimated Mean Squared Prediction Error:

$$\hat{P}_{T+h}^T = \hat{\sigma}_W^2 \sum_{j=0}^{h-1} \psi_j^2$$

Construction of Prediction Intervals: Since $\hat{X}_{T+h} \approx \mathbb{E}[X_{T+h} \mid X_T, \ldots]$,

$$\mathbb{E}\Big[\hat{X}_{T+h} - X_{T+h}\Big] = 0 \quad \text{(Tower Property)}$$

$$\mathbb{E}\Big[(\hat{X}_{T+h} - X_{T+h})^2\Big] = P_{T+h}^T$$

Hence

$$\frac{\hat{X}_{T+h} - X_{T+h}}{\sqrt{\hat{P}_{T+h}^T}}$$

is an approximately mean zero and unit variance random variable.

Suppose c_α is the α-critical value of this random variable, then

$$\hat{X}_{T+h} \pm c_{\alpha/2}\sqrt{\hat{P}_{T+h}^T}$$

is an approximate $(1-\alpha)$ prediction interval for X_{T+h}.

Choices for c_α:

(1) z_α (standard normal critical value).

Motivation: If W_t is Gaussian, then $X_t = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$ is Gaussian.

(2) Empirical critical value of residuals (standardized)

$$\frac{\hat{W}_t}{\sigma_W} \quad (1 \leq t \leq T)$$

(3) t-distribution, Pareto, or skewed distribution fit to standardized residuals.

Long Range Behaviour of ARMA Forecasts

Suppose $Y_t = s_t + X_t$ where $X_t \sim \text{ARMA}(p, q)$.

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

$\hat{Y}_{T+h}$ is converging fast to $\hat{s}_{T+h}$. Therefore, when we are doing ARMA forecasting in a trend + noise framework, we better get the trend correct for long range forecasts. Long range forecasts are only going to depend on the trend, and very little on the noise because we know that ARMA processes have a geometric decay to their dependent structure.

$$P_{T+h}^{T} = \sigma_W^2 \sum_{\ell=0}^{h-1} \psi_\ell^2 \xrightarrow{h\to\infty} \sigma_W^2 \sum_{\ell=0}^{\infty} \psi_\ell^2 = \gamma_X(0) = \sigma_W^2$$

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

4.5 ARMA Forecasting Example 1: Cardiovascular Mortality

[R Code] Cardiovascular Mortality

Slide 1

Let's give ARMA forecasting a try on real data.

Slide 2

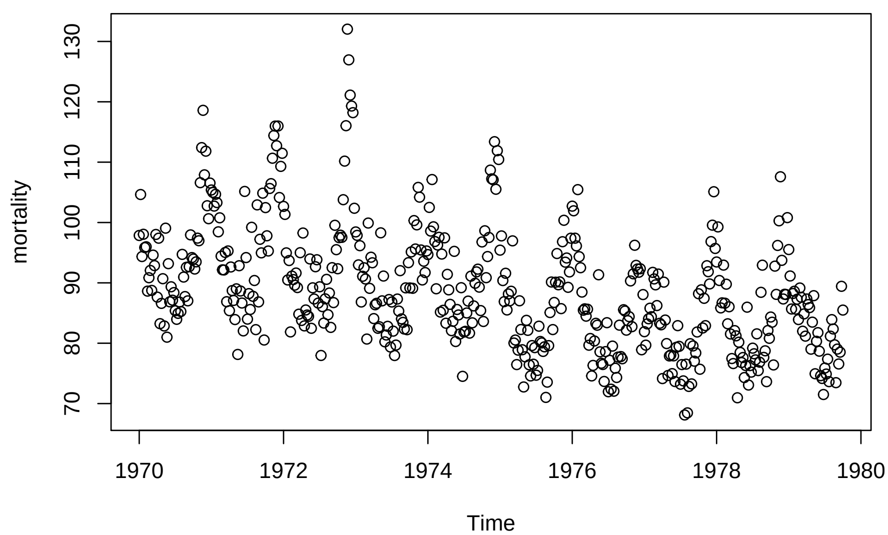


Figure 4.2: Weekly cardiovascular mortality, LA County.

Slide 3

Let X_t = cardiovascular mortality series. Our model is

$$X_t = s_t + Y_t$$

where $Y_t \sim \text{ARMA}(p, q)$.

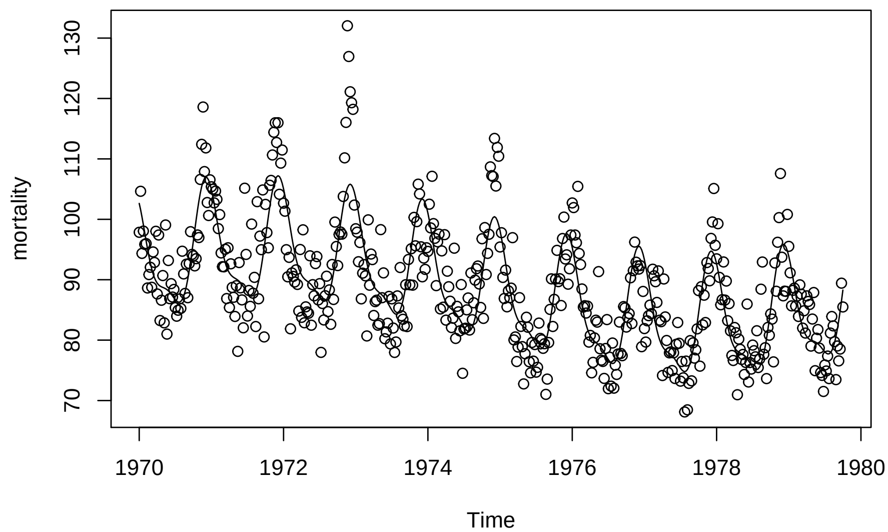
$$X_t = \underbrace{\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3}_{\text{polynomial}} + \underbrace{\beta_4 \sin\left(\frac{2\pi}{52}t\right) + \beta_5 \cos\left(\frac{2\pi}{52}t\right) + \beta_6 \sin\left(\frac{2\pi}{26}t\right) + \beta_6 \cos\left(\frac{2\pi}{26}t\right)}_{\text{seasonal}}$$

where the first four terms are the polynomial trends, the next two terms are the yearly cycle, and the last two are the half-yearly cycle.

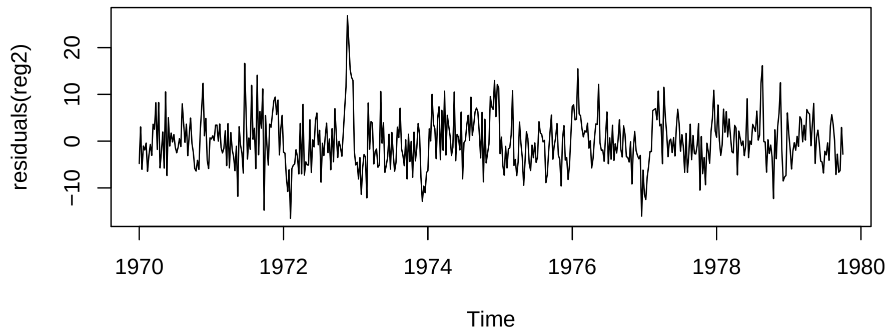
Decided on the trend using AIC, which will be discussed next week.

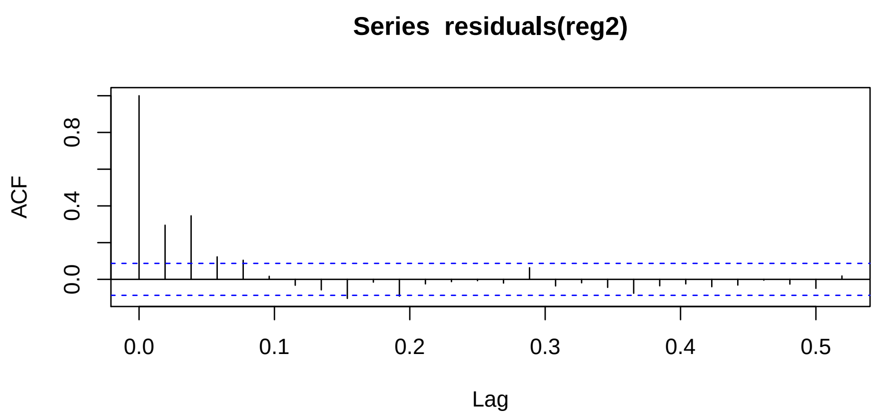
Slide 4

s_t estimated using ordinary least squares.



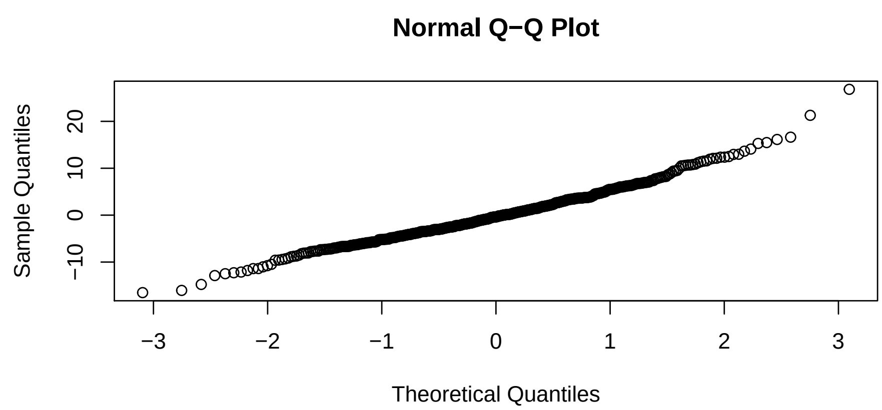
Slide 5





- $\hat{Y}_t = X_t - \hat{s}_t$ "seems reasonably stationary."
- Mild serial correlation in $\hat{Y}_t$ — Might be well modelled by $\text{MA}(2)$ or $\text{ARMA}(1,1)$.

Slide 6

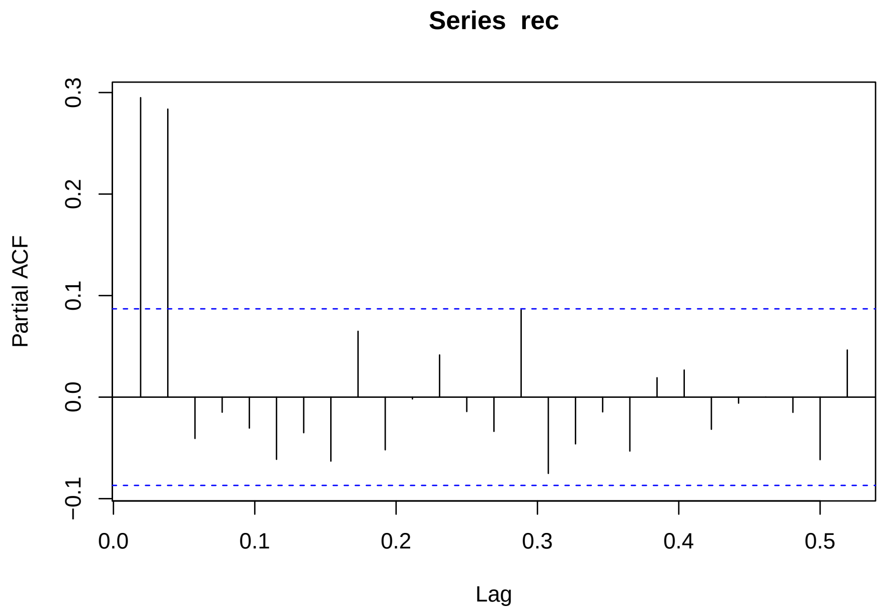


- $\hat{Y}_t$ seems reasonably normal, suggests using

$$\pm Z_{\alpha/2}\sqrt{P_{T+h}^{T}}$$

to construct prediction bounds.

Slide 7



Considering the PACF: On the first two lags these are large which is indicative of an autoregressive 2 structure, that is, AR(2) structure.

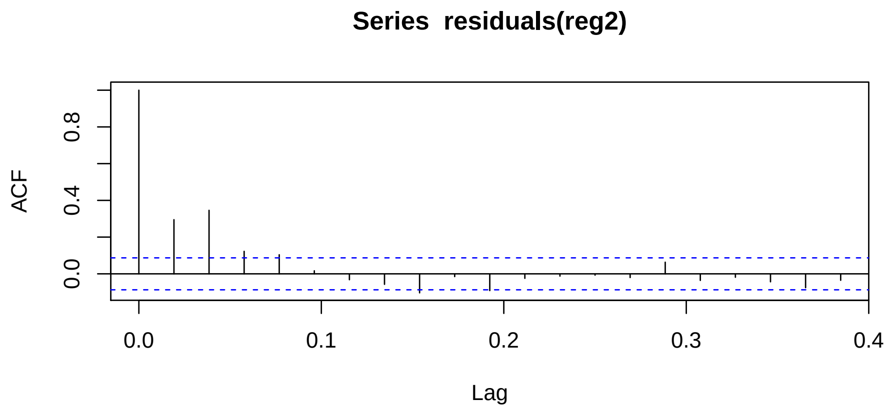
Slide 8

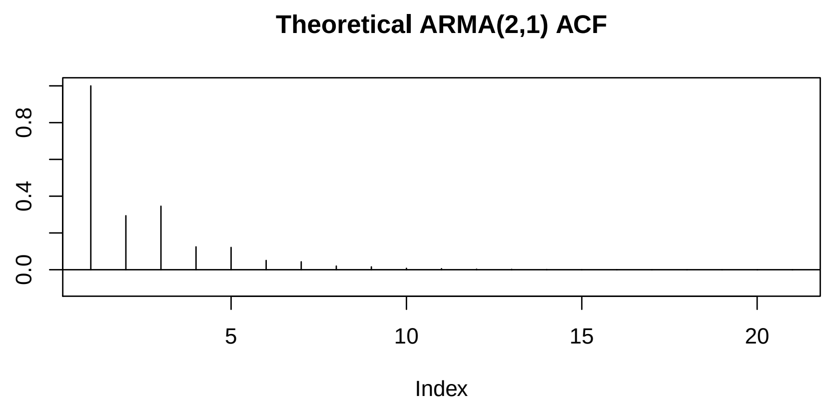
Model $\hat{Y}_t$ as ARMA$(2,1)$.

$$Y_t = 0.0885Y_{t-1} + 0.3195Y_{t-2} + W_t + 0.1328W_{t-1}$$

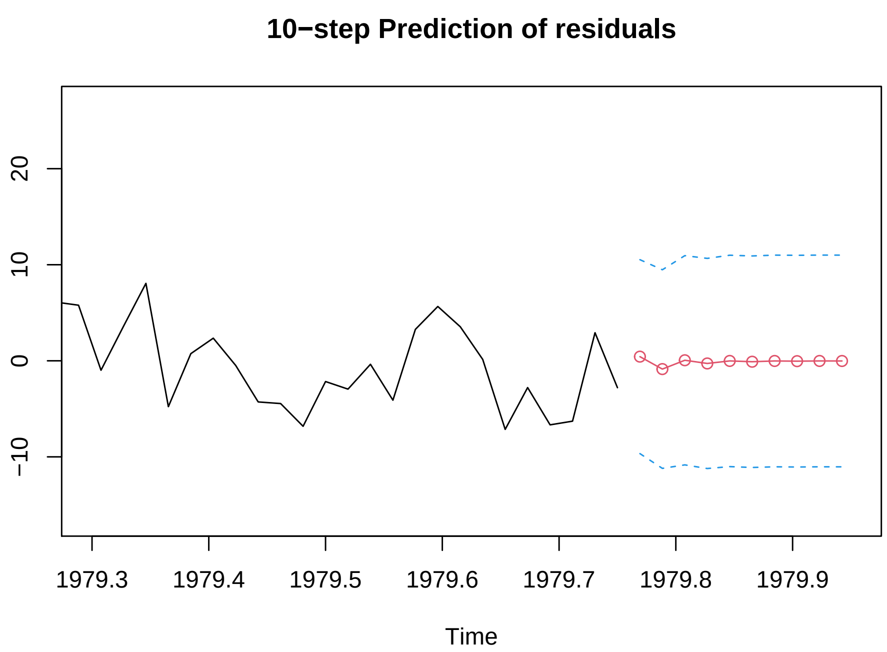
parameters estimated by MLE.

Slide 9





Slide 10



$\hat{Y}_{T+h|T}$, $h = 1, \ldots, 10$.

$$\hat{Y}_{T+h|T} \pm 1.96\sqrt{\hat{P}^T_{T+h}}$$

where 1.96 is the 97.5% critical value of $\mathcal{N}(0, 1)$.

Slide 11

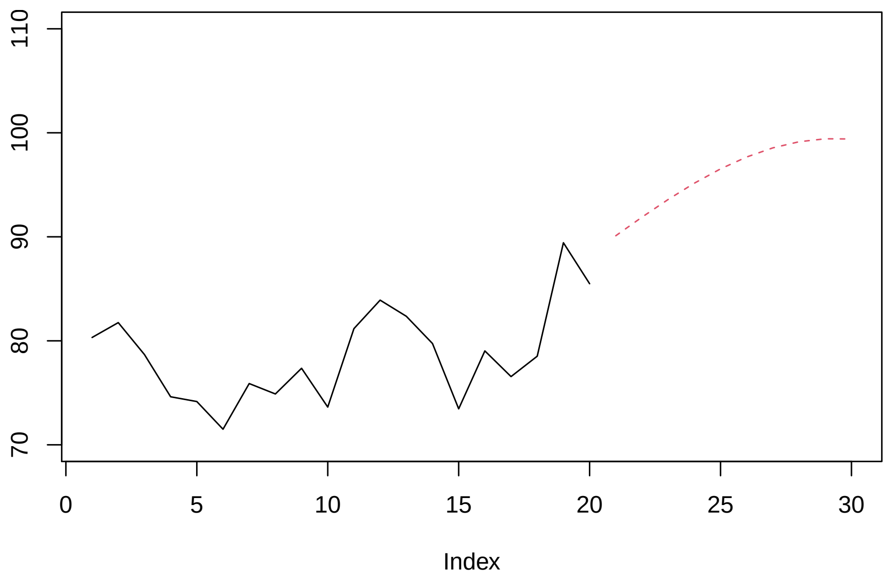


Figure 4.3: 30 weeks of data with predicted trend

Slide 12

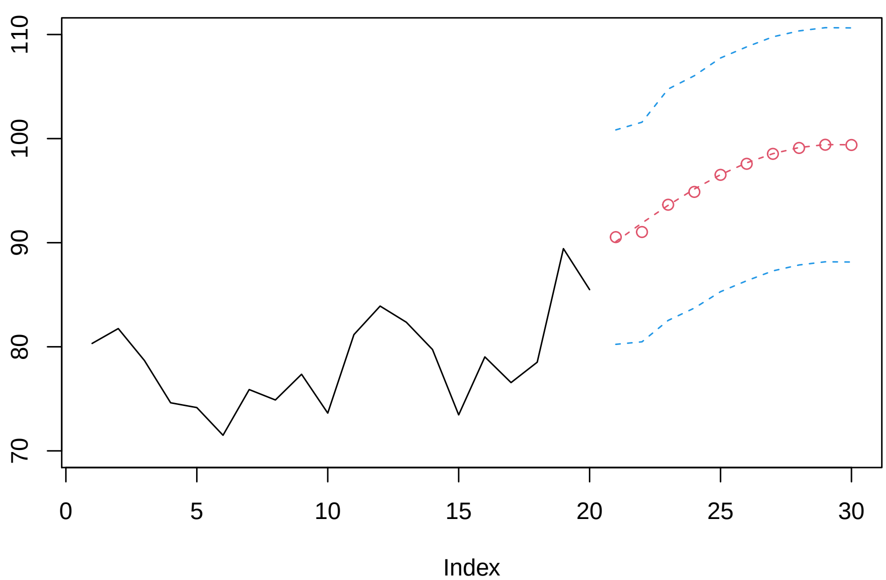


Figure 4.4: Forecasts with 95% prediction intervals

Fluctuations attribute to serial dependence. Red lines show that forecasts quickly converge to trend.

4.6 ARMA Forecasting Example 2: Johnson and Johnson

[R Code] Johnson and Johnson

X_t Johnson and Johnson Earnings.

$$X_t = e^{s_t + Y_t}$$

where Y_t is stationary. In this case,

$$\log(X_t) = s_t + Y_t$$

where $Y_t \sim \text{ARMA}(p, q)$.

Chapter 5

Week 5

5.1 ARMA Parameter Estimation: AR Case

Suppose we observe a time series $X_1, \ldots, X_T \sim \text{ARMA}(p, q)$

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

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

Our goal is to estimate

- $\phi_1, \ldots, \phi_p$ (AR parameters)
- $\theta_1, \ldots, \theta_q$ (MA parameters)
- σ_W^2 (white noise variance)

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

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

This leads to (upon some calculations):

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

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

where $X_t - \hat{\phi}X_{t-1}$ is estimated W_t and $\hat{\sigma}_W^2$ is the sample variance of residuals.

AR(p) case: $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t$. OLS: $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top \in \mathbf{R}^p$

$$\hat{\boldsymbol{\phi}} = \underset{\hat{\boldsymbol{\phi}}}{\arg\min} \sum_{t=p+1}^{T} (X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p})^2$$

$\hat{\boldsymbol{\phi}}$ admits a stationary and causal solution.

Solve using calculus (take first order partial derivatives set equal to zero), leads to a system of p linear equations of the form

$$\hat{\Gamma}_p \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\gamma}}_p$$

where

$$\hat{\Gamma}_p = (\hat{\gamma}(j-k),\ 1 \le j,k \le p) \in \mathbf{R}^{p\times p}$$
$$\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1). \dots, \hat{\gamma}(p))^\top$$

The resulting OLS estimator takes the form

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

Similar approach: use method of moments (set parameters so that empirical moments match theoretical causal moments induced by the model).

If $X_t \sim \mathrm{AR}(p)$, then for $1 \le h \le p$.

$$\begin{aligned}\gamma(h) &= \mathbb{E}[X_t X_{t+h}] \\ &= \mathbb{E}[X_t(\phi_1 X_{t+h-1} + \cdots + \phi_p X_{t+h-p} + W_{t+h})] \\ &= \phi_1\gamma(h-1) + \phi_2\gamma(h-2) + \cdots + \phi_p\gamma(h-p) + 0\end{aligned}$$

where the 0 occurs since $X_t \perp\!\!\!\perp W_{t+h}$.

This implies the linear system:

$$\boldsymbol{\gamma}_p = \Gamma_p \boldsymbol{\phi}$$
$$\boldsymbol{\gamma}_p = (\gamma(1), \dots, \gamma(p))^\top \in \mathbf{R}^p$$
$$\Gamma_p = [\gamma(j-k),\ 1 \le j,k \le p] \in \mathbf{R}^{p\times p}$$

Note that $X_t = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$ where $\psi_0 = 1$ and $W_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$ imply

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

which are **Yule-Walker Equations**.

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

Yule-Walker Estimators:

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

EXAMPLE 5.1.1

In the AR(1) case, 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 5.1.2

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

$$\frac{\hat{\phi}_{\text{OLS, } i}}{\hat{\phi}_{\text{YW, } i}} \xrightarrow[T\to\infty]{P} 1$$

OLS and YW estimates are asymptotically equivalent.

THEOREM 5.1.3

$$\sqrt{T}(\hat{\phi}_{YW} - \phi) \xrightarrow[T\to\infty]{D} MVN(0, \sigma_W^2 \Gamma_p^{-1})$$

$$\hat{\sigma}_W^2 \xrightarrow{P} \sigma_W^2$$

- *Optimal variance among all possible (asymptotically) unbiased estimators, hence* ***efficient****.*
- *Result can be used to obtain confidence intervals for ϕ.*

5.2 ARMA Parameter Estimation: MLE

Ordinary Least Squares and Yule-Walker equation 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.

Latent Variables (variables associated with W_t) $\implies$ MLE is best.

Suppose $X_t \sim$ AR(1) (causal)

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

where $W_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_W^2)$, then

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

is Gaussian. L^2 limits of Gaussian random variables are Gaussian. (MGF or characteristic function).

Moreover, $X_1, \dots, X_T$ are jointly Gaussian since

$$a_1 X_1 + \cdots + a_T X_T = \sum_{\ell=0}^{\infty} \phi^\ell \underbrace{(a_1 W_{1-\ell} + \cdots + a_T W_{t-\ell})}_{\text{Gaussian}}$$

MLE:

$$\mathcal{L}(\phi, \sigma_W^2) = f(X_T, X_{T-1}, \dots, X_1; \phi, \sigma_W^2)$$

where

- $\mathcal{L}(\phi, \sigma_W^2)$ is the likelihood of ϕ and σ_W^2.
- $f(X_T, X_{T-1}, \dots, X_1; \phi, \sigma_W^2)$ is the joint density of $X_T, \dots, X_1$ at the observed data. Gaussian Density.

Key idea in Time series: To evaluate the likelihood condition on the path/past!

$$\begin{aligned} f(X_T, \dots, X_1) &= f(X_T \mid X_{T-1}, \dots, X_1) f(X_{T-1}, \dots, X_1) \\ &\ \ \vdots \qquad\qquad \text{iterate} \\ &= f(X_T \mid X_{T-1}, \dots, X_1) f(X_{T-1} \mid X_{T-2}, \dots, X_1) \cdots f(X_2 \mid X_1) f(X_1) \\ &= \prod_{i=1}^{T} f(X_i \mid X_{i-1}, \dots, X_1) \end{aligned}$$

According to HW2:

$$X_i \mid (X_{i-1}, \dots, X_1) \sim \mathcal{N}(\phi X_{i-1}, \sigma_W^2)$$

Note that $X_i \mid (X_{i-1}, \dots, X_1) = X_i \mid X_{i-1}$, AR$(1)$.

Thus,

$$\begin{aligned}\mathcal{L}(\phi, \sigma_W^2) &= \prod_{i=2}^{T} \frac{1}{\sqrt{2\pi\sigma_W^2}} \exp\left\{-\frac{(X_i - \phi X_{i-1})^2}{2\sigma_W^2}\right\} f(X_1) \\ &= (2\pi\sigma_W^2)^{-\frac{T-1}{2}} \exp\left\{-\frac{\sum_{i=2}^{T}(X_i - \phi X_{i-1})^2}{2\sigma_W^2}\right\} f(X_1; \phi, \sigma_W^2)\end{aligned}$$

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

General ARMA(p, q) case: Again, $X_T, \dots, X_1$ are jointly Gaussian if $W_t \sim$ Gaussian.

$$\underbrace{L(\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_W^2)}_{\boldsymbol{\theta} \in \mathbf{R}^{p+q+1}} = \prod_{i=1}^{T} \underbrace{f(X_i \mid X_{i-1}, \dots, X_1)}_{\text{Gaussian}}$$

$$\begin{aligned}X_i \mid (X_{i-1}, \dots, X_1) &\sim \mathcal{N}(\mathbb{E}[X_i \mid X_{i-1}, \dots, X_1], \text{MSE}) \\ &\sim \mathcal{N}(\tilde{X}_{i|(i-1)}(\boldsymbol{\theta}), P_{i-1}^{i}(\boldsymbol{\theta}))\end{aligned}$$

where $P_{i-1}^{i}(\boldsymbol{\theta})$ is forecast MSE predicting X_i from $X_{i-1}, \dots, X_1$.

This likelihood can be maximized using numerical optimization. (Newton-Raphson Algorithm, Conjugate Gradient).

THEOREM 5.2.1: Chapter 8 of Brockwell and Davis, Hannan (1980)

The MLE's of $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma_W^2$ 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.

REMARK 5.2.2: Takeaway Message

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

5.3 Model Selection Diagnostic Tests

Using MLE, we can fit an ARMA(p, q) model to an observed series $X_1, \dots, X_T$.

Question: How do we select the orders p and q of the 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.

Model Diagnostics

If the ARMA(p, q) model fits the data well, then the estimated residuals should behave like white noise.

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

where

- $\tilde{X}_{t|(t-1)}$ is the truncated predicator of X_t based on $X_{t-1}, \dots, X_1$, and
- $\hat{P}_t^{t+1}$ is the estimated MSE.

This can be investigated by considering $\hat{\rho}_W(h)$ which is the empirical ACF of $\hat{W}_1, \dots, \hat{W}_T$.

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

We know that for each fixed h,

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

Also, for $j \neq h$,

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

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

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

Therefore,

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

Box-Ljung-Pierce Test [White Noise Test for ARMA(p, q) Models]

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

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

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

That is, we lose $p + q$ degrees of freedom for fitting the model.

The BLP test p-value is then computed as

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

REMARK 5.3.1

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

$$Q(T, H) \xrightarrow[T\to\infty]{P} \infty$$

Interpretation: If BLP p-values are small, the model is ill-fitting or under specified.

5.4 Model Selection Information Criteria

Suppose we are trying to select the orders p and q of an ARMA(p, q) model to fit $X_1, \dots, X_T$.

$$\boldsymbol{\phi} = \text{AR parameters}$$

$$\boldsymbol{\theta} = \text{MA parameters}$$

$$\sigma_W^2 = \text{white noise variance}$$

$$\mathcal{L}(X_1, \dots, X_T; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \sigma_W^2)$$

Natural idea: Maximize the likelihood of the data as a function of p and q.

Problem: The likelihood is (monotonically) increasing as a function of p and q. Maximizing would lead to overfitting.

Solution: Maximize the likelihood subject to a penalty term on the number of parameters (complexity) of the model. Let the number of parameters in the ARMA(p, q) model be denoted by $k = p + q + 1$.

$$-2\log(\mathcal{L}(X_1, \dots, X_T; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \sigma_W^2)) + P(T, k)$$

where $P(T, k)$ is an increasing function of k.

Optimal p and q balance model fit with the penalty for complexity.

Common Penalty Term Choices

- AIC$(p, q) = -2\log(\mathcal{L}(X_1, \dots, X_T; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \sigma_W^2)) + \dfrac{2k + T}{T}$.
  - Comes from estimating the Kullback–Leibler distance from the fitted model to the "true" model.
- BIC$(p, q) = -2\log(\mathcal{L}(X_1, \dots, X_T; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \sigma_W^2)) + \dfrac{k\log(T)}{T}$.
  - Comes from approximating and maximizing the posterior distribution of the model given the data.

Interpretation: Small AIC/BIC mean a better model.

Information criteria are also used in trend fitting. Suppose

$$X_t = s_t + Y_t = f_t(\boldsymbol{\beta}) + Y_t$$

where $\boldsymbol{\beta} \in \mathbf{R}^k$ and $f_t(\boldsymbol{\beta})$ is the trend we fit.

Estimate $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ using ordinary least squares.

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

Information criteria typically calculated assuming Y_t is a Gaussian white noise.

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

where for $P(T, k)$ we use AIC or BIC penalty.

REMARK 5.4.1

(1) In trend fitting, the assumption of Gaussian white noise residuals is often in doubt.
(2) AIC/BIC are not perfect! They are but one of many tools useful in model selection.
Strengths:
  (i) Easy to compute.
  (ii) Facilitates comparing many models quickly.
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.

5.5 ARIMA Models

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

DEFINITION 5.5.1: Integrated

We say a time series X_t is **integrated** to order d if $\nabla^d X_t$ is stationary, but $\nabla^j X_t$ for $1 \leq j \leq d$ is not stationary.

Motivation: If Y_t is stationary, and $X_t = \sum_{j=1}^{t} Y_j$, X_t is integrated to order 1. $Z_t = \sum_{j=1}^{t} X_j$ is integrated to order 2, and so on.

DEFINITION 5.5.2: ARIMA

We say X_t follows an **Autoregressive Integrated Moving Average Process** (ARIMA) of orders p, d, q if

$$\phi(B)(1-B)^d X_t = \theta(B) W_t$$

and write $X_t \sim$ ARIMA(p, d, q). Note that $\nabla^d X_t$ follows an ARMA(p, q) model.

Forecasting ARIMA(p, d, q) Processes

(1) $Y_t = \nabla^d X_t$ follows an ARMA(p, q) model and so can be forecast using truncated ARMA prediction.

(2) Forecasts $\hat{Y}_{T+h|T}$ can be used to forecast X_{T+h} by reversing the differencing.

EXAMPLE 5.5.3

For $d = 1$, $Y_{T+1} = X_{T+1} - X_T$ so $\hat{X}_{T+1|T} = X_T + \hat{Y}_{T+1|T}$. This can be iterated to produce longer Horizon forecasts.

Predicting MSE is approximately of the form

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

where $\psi_{j,*}^2$ is the coefficient of z^j in the power series expansion (centred at zero) of

$$\frac{\theta(z)}{\phi(z)(1-z)^d} \quad (|z| < 1)$$

Idea:

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

EXAMPLE 5.5.4

Let $X_t \sim \text{ARIMA}(0,1,0)$.

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

if we iterate t-times. If $Y_t = \nabla X_t$, then $\hat{Y}_{T+h|T} = 0$ (forecasting W_t's). Therefore,

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

Similarly, $\hat{X}_{T+h|T} = X_T$. The best predictor of random walk is last known location.
Prediction MSE:

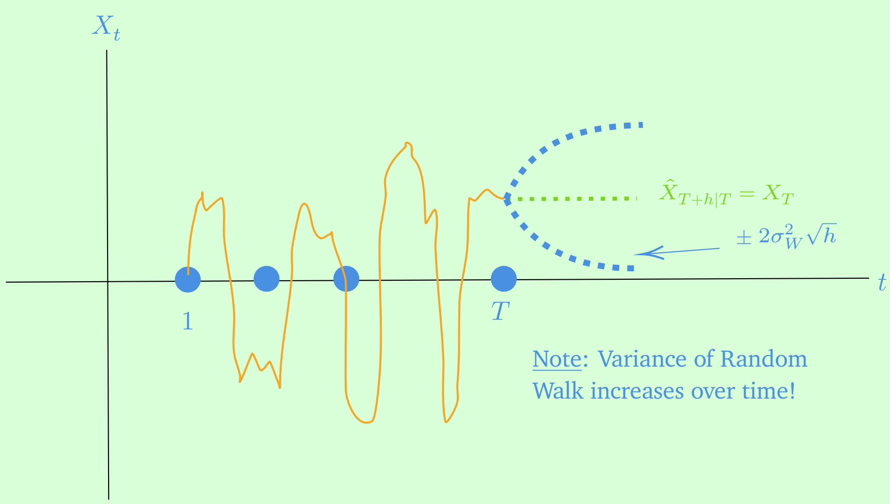
$$\frac{\theta(z)}{\phi(z)(1-z)^d} = \frac{1}{1-z} = \sum_{j=0}^{\infty} z^j \quad (|z| < 1)$$

$$\implies \psi_{j,*} = 1 \quad (\forall j)$$

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

Note that

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



If we forecast into the future, the forecast will be the last observed value. Also, if we plot prediction intervals, they would be of the form $\pm 2\sigma_W^2$MSE where MSE which is on the order of $\sqrt{h}$. In particular, these are increasing as a function of h. Therefore, the variance of a Random Walk will increase over time, and hence the prediction intervals will increase over time. This is a normal feature you see when you do ARIMA forecasts, and this is the basic reason why.

How to decide in practice on degree of differencing d:

(1) Eye-ball Test.

(2) Formal Stationary Tests [Dickey-Fuller, Kwiatkowski–Phillips–Schmidt–Shin (KPSS)].

(3) Cross-validation.

5.6 ARIMA Modelling Example

[R Code] ARIMA Modelling Example

Chapter 6

Week 6

6.1 SARIMA Models

Frequently, time series exhibit "seasonality."

Rough Definition of Seasonality

A time series X_t is said to be "seasonal" if it exhibits regular variation so that for some lag s, X_t is "similar" to X_{t-s}. Some sources of seasonality are weather or scheduled events. These typically lead to yearly, weekly, monthly, or quarterly cycles.

REMARK 6.1.1

ARIMA models are not ideal for modelling seasonality.

$$\text{ARIMA Models} \implies \text{Random Walk with Stationary Errors}$$

Random walks do not seasonality.

DEFINITION 6.1.2: Seasonal ARIMA

X_t is said to follow a **Seasonal ARIMA** model (SARIMA) of orders p, d, q and P, D, Q and seasonal period s if

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^d Y_t = \Theta_Q(B^s)\theta_q(B)W_t$$

We abbreviate the SARIMA p, d, q, P, D, Q model with seasonal period s as SARIMA$(p, d, q) \times (P, D, Q)_s$.

$$\begin{array}{rcccccccc}
\Phi_P(B) & = & 1 & - & \Phi_1 B & - & \cdots & - & \Phi_P B^P \\
\Phi_P(B^s) & = & 1 & - & \Phi_1 B^s & - & \cdots & - & \Phi_P B^{Ps} \\
\phi_p(B) & = & 1 & - & \phi_1 B & - & \cdots & - & \phi_p B^p \\
\Theta_Q(B) & = & 1 & + & \Theta_1 B & + & \cdots & + & \Theta_Q B^Q \\
\Theta_Q(B^s) & = & 1 & + & \Theta_1 B^s & + & \cdots & + & \Theta_Q B^{Qs} \\
\theta_q(B) & = & 1 & + & \theta_1 B & + & \cdots & + & \theta_q B^q
\end{array}$$

DEFINITION 6.1.3

The **seasonal** autoregressive and moving average polynomials are defined by

$$\Phi(z) = 1 - \Phi_1 z - \cdots \Phi_P z^P$$

$$\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$$

EXAMPLE 6.1.4

Let $X_t \sim \text{SARIMA}(1,1,1) \times (1,1,1)_{13}$.

$$\Phi(z) = 1 - \Phi_1 z$$

$$\phi(z) = 1 - \phi_1 z$$

$$\Theta(z) = 1 + \Theta_1 z$$

$$\theta(z) = 1 + \theta_1 z$$

Therefore,

$$(1 - \Phi_1 B^{13})(1 - \phi_1 B)\underbrace{(1 - B^{13})(1 - B)X_t}_{Y_t} = \Theta(B^{13})\theta(B)W_t$$

$$Y_t - \Phi_1 Y_{t-13} - \phi_1 Y_{t-1} - \phi_1 \Phi_1 Y_{t-14} = \text{MA term}$$

$$Y_t = f(Y_{t-13}, Y_{t-1}, \text{MA noise}, Y_{t-14})$$

where Y_{t-13} is the seasonal lag.

REMARK 6.1.5

(1) $Y_t = (1 - B^s)^D (1 - B)^d X_t$, a SARIMA model is just one big ARMA model for Y_t.
(2) Advantage over 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.

(2) Differencing and seasonal differencing can be decided upon by:

  (a) Eye-ball test and/or examining the ACF and PACF.

  (b) Stationarity tests.

  (c) Cross-validation.

  We will discuss (b) and (c).

(3) Choosing the order and estimating the components of $\Phi, \phi, \Theta, \theta$ can be done in the same was as with ARMA models.

6.2 SARIMA Cardiovascular Mortality Example

[R Code] SARIMA Cardiovascular Mortality Example

6.3 Time Series Cross-Validation

DEFINITION 6.3.1: Cross-validation

Cross-validation is a data driven model evaluation and selection tool for predictive models that entails the following.
(1) Splitting the available data into training and testing sets.
(2) Fitting models on the training sets.
(3) Evaluating predictions of the model on the tests sets as an overall evaluation of model quality.

Standard Cross-Validation

Suppose (Y_i, X_i) for $1 \le i \le n$ satisfy $Y_i = f(X_i) + \varepsilon_i$. Let M be a model used to estimate f using $\hat{f}$, with the goal of minimizing $L(Y_i, \hat{f}(X_i))$.

K-fold Cross-Validation

(1) Split (Y_i, X_i) for $1 \le i \le n$ randomly into K-groups $G_1, \dots, G_k$.

(2) For each $1 \le i \le K$, use M to estimate $\hat{f}^{(-j)}$ when the data G_i is left out.

(3) Evaluate error on G_i with

$$\text{CV}_j = \sum_{(Y_i, X_i) \in G_j} L(Y_i, \hat{f}^{-j}(X_i))$$

(4) The total cross-validation error of the model is:

$$\text{CV}(M) = \sum_{j=1}^{k} \text{CV}_j$$

REMARK 6.3.2

- K is often called the number of **folds**.
- If $K = n$, the procedure is often called the "leave-one-out" cross-validation.
- $K = 10$ is called "10-fold cross validation."

Problems with Time Series Cross-Validation

(1) Randomly splitting the data scrambles up any serial dependence relationships.

(2) In time series forecasting, it is often most natural to use the past (recent past) to predict future values.

Time Series Cross-Validation Algorithm

(1) Split the data into training and testing ranges $1 \le t_r \le T$ where $t_r \approx 0.75T$ is 75% of the training sample. The test sample is $X_{t_r+1}, \dots, X_T$.

(2) For each j in $t_r + 1, \dots, T$, use model to forecast $\hat{X}_{j+1|j}$ based on $X_1, \dots, X_j$. Calculate loss

$$L(\hat{X}_{j+1|j}; X_{j+1}) = L_j$$

(3) Cross-validation score of model

$$\text{CV}(M) = \sum_{j=t_r+1}^{T} L_j$$

REMARK 6.3.3

(1) If interested in longer horizon forecasting, you can compare

$$\hat{X}_{j+1|j}, \dots, \hat{X}_{j+h|j} \quad \text{to} \quad X_{j+1}, \dots, X_{j+h}$$

in the loss calculation step.

(2) Stationarity is *crucial* in time series cross validation since the model errors in the present must be similar to errors in the future.

(3) One normally cannot cross-validate everything as this is computationally infeasible.

6.4 Cross-Validation Example

[R Code] Cross-Validation Example

6.5 Simulated and Bootstrapped Prediction Intervals

Usually forecasts are of the form

$$\hat{X}_{T+1|T} = g(X_T, X_{T-1}, \dots, X_1, W_{T+1})$$

where W_{T+1} is a strong white noise innovation.

Often, even models are additive so that

$$\hat{X}_{T+1|T} = g(X_T, \dots, X_1) + W_{T+1}$$

Simple and powerful models to produce prediction intervals use simulation!

Simulated Prediction Intervals

(1) Choose a distribution for $\{W_t\}$. A common choice is $W_t \sim \mathcal{N}(0, \hat{\sigma}_W^2)$.

(2) For $b = 1, \dots, B$ where B is a large number, simulate $\left\{W_{T+1}^{(b)}\right\}$.

(3) Compute $\hat{X}_{T+1|T}^{(b)} = g(X_T, \dots, X_1) + W_{T+1}^{(b)}$ for $b = 1, \dots, B$.

(4) Denote the empirical q^{th} quantile of $\left\{\hat{X}_{T+1}^{(b)} : b = 1, \dots, B\right\}$ by $\hat{Q}_{T+1}(q)$. We set the $(1-\alpha)$ prediction interval as

$$\left(\hat{Q}_{T+1}\left(\frac{\alpha}{2}\right), \hat{Q}_{T+1}\left(1-\frac{\alpha}{2}\right)\right)$$

REMARK 6.5.1

For longer horizon forecasts, prediction intervals can be obtained by iteration:

$$\hat{X}_{T+h|T}^{(b)} = g(\hat{X}_{T+h-1|T}^{(b)}, \dots, \hat{X}_{T+1|T}^{(b)}, X_T, \dots, X_1) + W_{T+h}^{(b)}$$

The prediction interval is

$$\left(\hat{Q}_{T+h}\left(\frac{\alpha}{2}\right), \hat{Q}_{T+h}\left(1-\frac{\alpha}{2}\right)\right)$$

where $\hat{Q}_{T+h}(q)$ the empirical q^{th} quantile of $\hat{X}_{T+h}^{(b)}$.

Distributions to Choose for W_t

(1) $W_t \sim \mathcal{N}(0, \hat{\sigma}_W^2)$ where $\hat{\sigma}_W^2$ is estimated from residuals which leads to approximately the same "well known" prediction intervals.

(2) A distribution fit to the estimated residuals $\hat{W}_t$; e.g., a t-distribution, Pareto, etc.

(3) The empirical distribution of the residuals $\hat{W}_t$; that is, randomly drawing $\left\{\hat{W}_1, \dots, \hat{W}_T\right\}$ which is commonly known as **bootstrapping**.

Note: An important consideration of the bootstrap is that the residuals should be white! We can check the whiteness of the residuals using the ACF or a white noise test.

6.6 Bootstrap Prediction Intervals Example

[R Code] Bootstrap Prediction Intervals Example

Chapter 7

Week 7

7.1 Exponential Smoothing Models Introduction

- **ARIMA Models**: Model a time series, potentially after differencing towards stationarity, in terms of its autocorrelation (linear process).
- **Exponential Smoothing**: Flexibly model the trend and seasonality observed in a time series.

Simple Exponential Smoothing

Suppose we wish to forecast a time series $X_1, \dots, X_T$. Two extreme forecasts are

$$\hat{X}_{T+1|T} = X_T \quad \text{[Random Walk]}$$

$$\hat{X}_{T+1|T} = \bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t \quad \text{[IID Sequence]}$$

Compromise: *Exponential Smoothing*.

$$\hat{X}_{T+1|T} = \alpha X_T + \alpha(1-\alpha)X_{T-1} + \cdots + \alpha(1-\alpha)^{T-1}X_1$$

where $\alpha \in [0,1]$ is the **smoothing parameter**.

Weights applied to past observations decrease exponentially quickly.

Simple exponential smoothing can be stated as a recursive system of equations.

- Prediction Equation: $\hat{X}_{T+1} = \ell_T$.
- Smoothing/Level Equation: $\ell_T = \alpha X_T + (1-\alpha)\ell_{T-1} = \ell_T(\alpha, \ell_\sigma)$ which is a convex combination of last observed value and last prediction or "level."
- Initial Condition: ℓ_0.
- Parameters Defining Model are $\alpha \in [0,1]$ and ℓ_0.

Estimation may be conducted using MLE (later) or OLS. For OLS,

$$(\hat{\alpha}, \hat{\ell}_0) = \underset{0\le\alpha\le 1,\, \ell_0 \in \mathbf{R}}{\arg\min} \sum_{i=2}^{T} \left[X_i - \ell_i(\alpha, \ell_0)\right]^2$$

$$\hat{X}_{T+1} = \hat{\alpha} X_T + (1-\hat{\alpha})\ell_T(\hat{\alpha}.\hat{\ell}_0)$$

which can be calculated by iterating the level equation back to ℓ_0.

Linear Trend Exponential Smoothing

- Prediction Equation: $\hat{X}_{T+h} = \ell_T + hb_T$ where ℓ_T is the **level** and b_T is the **slope**.
- Level Equation: $\ell_T = \alpha X_T + (1-\alpha)(\ell_{T-1} + b_{T-1})$ which is the convex combination of last observation and last "level" or prediction.
- Trend/Slope Equation: $b_T = \beta(\ell_T - \ell_{T-1}) + (1-\beta)b_{T-1}$ where $\ell_T - \ell_{T-1}$ is the last "observed" slope or change in level.
- Initial Conditions: ℓ_0 and b_0.
- Parameters: $\alpha, \beta \in [0,1]$, $\ell_0, \beta_0 \in$ **R** which are estimated using MLE/OLS.

Trend + Seasonal Exponential Smoothing (Holt Winters ES, 1960s)

Suppose h is the forecast horizon of interest and time series has seasonal period p. Set $k = \lfloor (h-1)/p \rfloor$.

- Prediction Equation: $\hat{X}_{T+1} = \ell_T + hb_T + s_{T+h-p(k+1)}$ where ℓ_T is the level, b_T is the slope, and $s_{T+1-p(k+1)}$ is the seasonal effect.
- Level Equation: $\ell_T = \alpha(X_T - s_{T-p}) + (1-\alpha)(\ell_{T-1} + b_{T-1})$.
- Slope Equation: $b_T = \beta(\ell_T - \ell_{T-1}) + (1-\beta)b_{T-1}$.
- Seasonal Equation: $s_T = \gamma(X_T - \ell_{T-1} - b_{T-1}) + (1-\gamma)s_{T-p}$.
- Initial Conditions: $\ell_0, \beta_0, s_0, \dots, s_{-p+1}$.
- Parameters: $\alpha, \beta, \gamma \in [0,1]$, $\ell_0, \beta_0, s_0, \dots, s_{-p+1} \in$ **R**.

7.2 Exponential Smoothing as a State Space Model

Consider Simple Exponential Smoothing:

- Prediction Equation: $\hat{X}_{t|t-1} = \ell_{t-1}$.
- Level Equation: $\ell_t = \alpha X_t + (1-\alpha)\ell_{t-1}$.

Re-arranging the level equation gives

$$\ell_t = \ell_{t-1} + \alpha(\underbrace{X_t - \ell_{t-1}}_{\text{residual } \varepsilon_t}) = \ell_{t-1} + \alpha\varepsilon_t$$

Also, $X_t = \ell_{t-1} + \varepsilon_t$. Therefore, these equations can be reformulated as:

- Prediction Equation: $X_t = \ell_{t-1} + \varepsilon_t$.
- Level Equation: $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$.

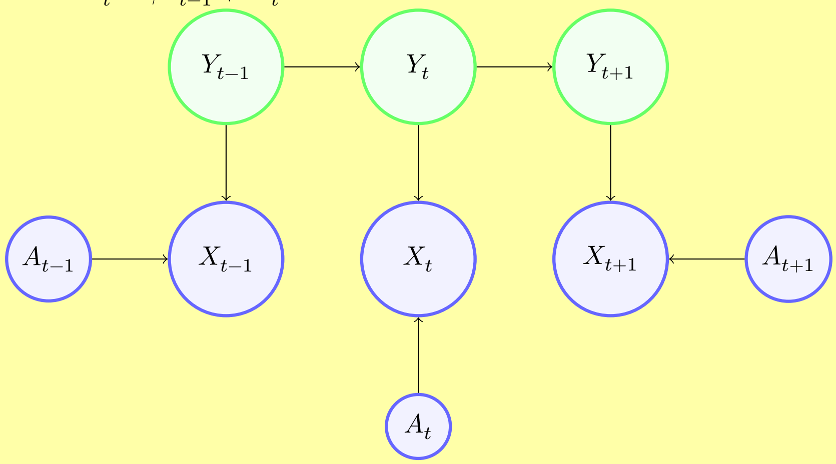
Why is this useful? If we make a parametric assumption on ε_t (e.g., $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$), then we can use Likelihood techniques (MLE, AIC, simulation based Prediction Intervals).

Such equations are examples of "State Space" Models:

DEFINITION 7.2.1: State space model

We say X_T follows a general **state space model** if:

- Observation Equation: $X_t = A_t Y_t + \varepsilon_t$ where A_t is the **measurement matrix**, Y_t is the **state vector** (unobserved), and ε_t is an **observation error**.
- State Equation: $Y_t = \phi Y_{t-1} + W_t$.



ε_t and W_t are white noise terms that may depend on each other.

EXAMPLE 7.2.2: State Space Models

- AR(1): $X_t = Y_t$ where $Y_t = \phi Y_{t-1} + W_t$ where $W_t \sim$ strong white noise.
- Simple Exponential Smoothing:

$$X_t = Y_{t-1} + \varepsilon_t$$

$$Y_t = Y_{t-1} + \alpha\varepsilon_t$$

  where $\varepsilon_t \sim$ strong white noise.

All ARMA and Exponential Smoothing models can be written in state-space form.

Parameter Estimation and Model Selection using State-Space Formulation

- $X_t = \ell_{t-1} + \varepsilon_t$.
- $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$.
- $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.
- Initial Condition: ℓ_0.

$$\mathcal{L}(X_1, \dots, X_T; \alpha, \ell_0, \sigma_\varepsilon^2) = \prod_{i=1}^{T} \underbrace{\mathcal{L}(X_i \mid X_{i-1}, \dots, X_1; \alpha, \ell_0, \sigma_\varepsilon^2)}_{\mathcal{N}(\ell_{i-1}(\alpha, \ell_0), \sigma_\varepsilon^2)}$$

Likelihood can be maximized numerically, and we use this to calculate AIC/BIC.

7.3 Multiplicative Exponential Smoothing Models

Standard Exponential Smoothing has "additive" errors, in the sense that

$$X_t = \ell_{t-1} + \varepsilon_t$$

$$\ell_t = \alpha X_t + (1-\alpha)\ell_{t-1}$$

Therefore, $\varepsilon_t = X_t - \ell_{t-1}$.

We can also formulate exponential smoothing in terms of "multiplicative" errors, in the sense that

$$\varepsilon_t = \frac{X_{t-1} - \ell_{t-1}}{\ell_{t-1}}$$

where we note that the error is relative to the previous level. Therefore,

$$X_t = \ell_{t-1}(1+\varepsilon_t)$$
$$\ell_t = \alpha X_t + (1-\alpha)\ell_{t-1} = \alpha\varepsilon_t\ell_{t-1} + \alpha\ell_{t-1} + (1-\alpha)\ell_{t-1} = \ell_{t-1}(1+\alpha\varepsilon_t)$$

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 propagates in the model.

- Additive: $\hat{X}_{T+1} = \ell_T + \sum_{j=T+1}^{T+h} \varepsilon_j$ where we note that the MSE scales like h.
- Multiplicative: $\hat{X}_{T+h} = \ell_T \prod_{j=T+1}^{T+h}(1+\varepsilon_j)$ where we note that the MSE (variance) is scaling like

$$\Big(\mathbb{E}[(1+\varepsilon_0)^2]\Big)^h$$

  which could grow very quickly as $h \to \infty$.

Multiplicative Linear + Trend and Holt Winters

Linear + Trend State Space Formulation:

$$\varepsilon_t = \frac{X_t - (\ell_{t-1} + b_{t-1})}{\ell_{t-1} + b_{t-1}}$$
$$X_t = (\ell_{t-1} + b_{t-1})(1+\varepsilon_t)$$
$$\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$$
$$b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Multiplicative Seasonal Exponential Smoothing

Let p be the seasonal period.

$$X_t = (\ell_{t-1} + b_{t-1})s_{t-p}(1+\varepsilon_t)$$
$$\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$$
$$b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$$
$$s_t = s_{t-p}(1+\gamma\varepsilon_t)$$

When to use Additive versus Multiplicative

Seasonal Exponential Smoothing Models:

(1) Multiplicative models imply that as the level increases (decreases) the seasonal fluctuations increase (decrease). Additive models suggest seasonal fluctuations remain constant as trend fluctuations.

$$\text{Seasonal Fluctuations} \uparrow \text{ as Level} \uparrow \implies \text{ Multiplicative.}$$

(2) Use AIC/BIC: The AIC can be evaluated for each state-space model and compared.

7.4 Exponential Smoothing Model Selection

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 (including additive versus multiplicative) models. Other options include:

- Cross-validation.
- Residual Analysis (white noise testing).

Prediction Intervals

Using the state-space formulation, valid prediction intervals may be computed using simulation.

EXAMPLE 7.4.1: Simple Exponential Smoothing

$$\hat{X}_{T+1|T} = \hat{\ell}_T$$

State-space formula:

$$\hat{X}_{T+1} \cong \hat{\ell}_T + \underbrace{\varepsilon_{T+1}}_{\mathcal{N}(0,\sigma_\varepsilon^2)}$$

(1) Estimate

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{T-1}\sum_{j=2}^{T}(X_j - \hat{\ell}_{T-1})^2$$

(2) Simulate

$$\hat{X}^{(b)}_{T+1|T} = \hat{\ell}_T + \underbrace{\varepsilon^{(b)}_{T+1}}_{\mathcal{N}(0,\hat{\sigma}_\varepsilon^2)}$$

(3) Use 5% and 95% sample quantiles of $X^{(b)}_{T+1|T}$, $b = 1, \dots, B$ as prediction intervals.

REMARK 7.4.2

In many cases, the prediction MSE assuming $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ can be computed explicitly. See § 7.7 of HA.

An important consideration in applying this approach is that ε_t should behave like Gaussian white noise. We can check this using a residual analysis.

- White noise tests, ACF plots.
- Quantile-Quantile plot for Normality.

7.5 J and J Exponential Smoothing Forecast

[R Code] J and J Exponential Smoothing Forecast

Chapter 8

Week 8

8.1 Neural Network Autoregression

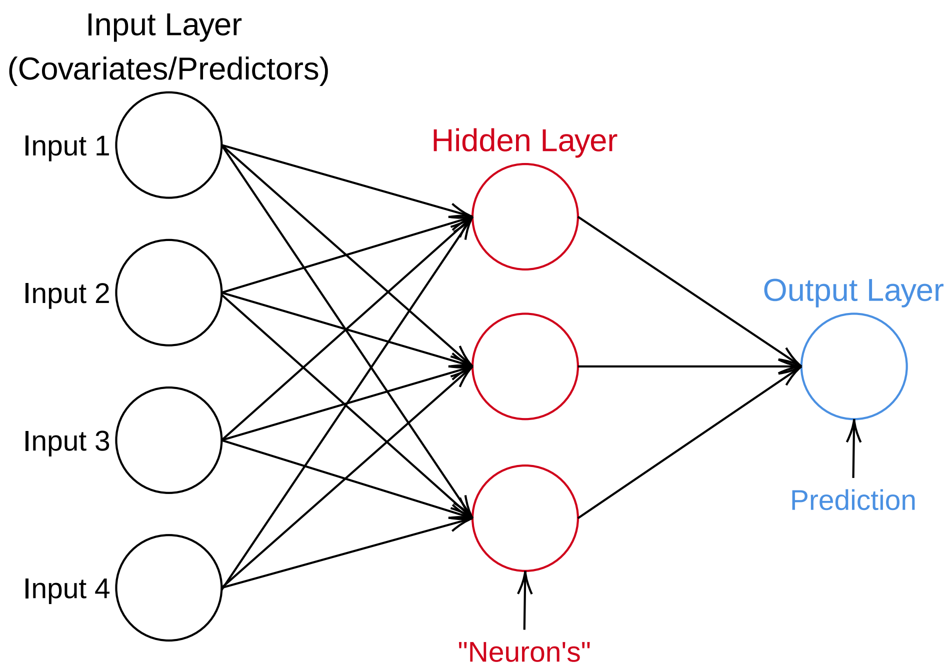


Figure 8.1: Simple Neural Network "Architecture"

It's possible to have several hidden layers and multiple neurons in each hidden layer.

Any particular layer in the neural network regression, the inputs are mapped to the neurons in the hidden layer using a simple linear transformation: inputs are mapped to the j^{th} neuron linearly. The value taken on the j^{th} neuron is

$$z_j = b_j + \sum_{i=1}^{4} w_{ij}x_i$$

where b_j is a function, x_i is the i^{th} input, and w_{ij} are the weights.

To calculate the inputs to the next layer, a non-linear transformation is applied. For example, using the sigmoid function:

$$S(z) = \frac{1}{1 + e^{-z}}$$

The final model is a complex non-linear function of the inputs.

Neural Network AR

- Input layer: $X_t, \ldots, X_{t-p}$.
- Output layer: X_{t+1}.

A neural network model with k hidden states (assuming one hidden layer) we call a NNAR(p, k) model.

REMARK 8.1.1

If $k = 0$, then NNAR(p) = AR(p). The inputs are mapped linearly to the outputs.

Seasonal Neural Network AR

- Input layer: $X_t, \ldots, X_{t-p}, X_{t-m}, X_{t-P_m}$ where m is the seasonal lag.
- Output layer: X_{t+1}.

We call this a NNSAR$(p, k, P)_m$ model.

The model selection of choosing k, p, and P can be carried out using cross-validation where the weights are estimated using ordinary least squares.

Prediction Intervals

If $\boldsymbol{X}_t = (X_t, \ldots, X_{t-p}, X_{t-m}, \ldots, X_{t-P_m})^\top$ denotes the vector of predictors, then we can posit an additive stochastic model for X_{t+1} as

$$X_{t+1} = f(\boldsymbol{X}_t) + \varepsilon_{t+1}$$

where f is the neural network.

By calculating the residuals $\hat{\varepsilon}_t = X_t - \hat{f}(\boldsymbol{X}_t)$, prediction intervals can be estimated using the bootstrap

$$X_{T+1}^{(b)} = \hat{f}(\boldsymbol{X}_T) + \hat{\varepsilon}_{T+1}^{(b)} \quad (b = 1, \ldots, B)$$

We can then construct a prediction interval by using the empirical quantiles from the simulated distribution of the forecast 1-step ahead. This process can be iterated multiple times to produce forecasts as well as prediction intervals for forecasts at longer time horizons.

[R Code] Neural Network Autoregression

8.2 Comparing Various Forecasting Methods

- [R Code] Comparing Various Forecasting Methods
- The M3-Competition: Results, Conclusions and Implications

8.3 Conditional Heteroscedasticity

$$\underbrace{\text{Hetero}}_{\text{different}}\text{-}\underbrace{\text{scedasticity}}_{\text{variance}}$$

$$\underbrace{\text{Hetero}}_{\text{same}}\text{-}\underbrace{\text{scedasticity}}_{\text{variance}}$$

EXAMPLE 8.3.1

If X_t is weakly stationary, then X_t is "homoscedastic" in the sense that $\mathbb{V}(X_t) = \sigma_X^2$ does not change over time.

DEFINITION 8.3.2: Heteroscedastic

We say a time series X_t is **heteroscedastic** if $\mathbb{V}(X_t) = \sigma_{X,t}^2$; that is, the variance depends on t and changes at some points.

REMARK 8.3.3

Heteroscedastic time series are <u>not</u> stationary.

Asset price data terminology: In the context of conditionally heteroscedastic time series, we often consider asset price or "financial" time series. Suppose $X_t =$ price of an asset at time t.

DEFINITION 8.3.4: Returns, Log-returns

If X_t is the value of an asset at time t, then the **return** (relative gain) Y_t of the asset at time t is

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

Furthermore, the **log-returns** of a positive asset price series X_t are

$$Y_t = \log\left(\frac{X_t}{X_{t-1}}\right) = \log(X_t) - \log(X_{t-1})$$

REMARK 8.3.5

"Volatility" $\iff$ "Variance".

[R Code] ARCH and GARCH Introduction

A common observation, especially prominent with financial and asset price data, is that periods of volatility or heteroscedastic tend to cluster.

Why? 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 as we will see in the next example.

EXAMPLE 8.3.6

Let $X_t \sim \text{AR}(1)$; that is, $X_t = \phi X_{t-1} + W_t$ where $|\phi| < 1$.

$$\mathbb{E}[X_t \mid X_{t-1}, X_{t-2}, \ldots] = \phi X_{t-1}$$

ARMA models "model" the conditional mean $X_{t-1}, X_{t-2}, \ldots$.

$$\mathbb{V}(X_t \mid X_{t-1}, X_{t-2}, \ldots) = \sigma_W^2$$

$X_{t-1}, X_{t-2}, \ldots$ leave the variance untouched.

DEFINITION 8.3.7: Conditionally heteroscedastic

We say a time series X_t is **conditionally heteroscedastic** if

$$\mathbb{V}(X_t \mid X_{t-1}, X_{t-2}, \dots) = \sigma^2_{X,t}$$

that is, the variance changes with t.

It's possible to have a time series X_t that's homoscedastic, but is also conditionally heteroscedastic.

8.4 ARCH and GARCH Models

DEFINITION 8.4.1: Autoregressive conditionally heteroscedastic (ARCH)

Let W_t be a unit variance strong white noise; that is, $\mathbb{E}[W_t] = 0$ and $\mathbb{V}(W_t) = 1$. We say X_t follows an **autoregressive conditionally heteroscedastic** (ARCH) model if there exists parameters $\omega > 0$, $\alpha_1 \geq 0$ such that $X_t = \sigma_t W_t$ where

$$\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2$$

where σ_t^2 is the conditional variance and W_t is a white noise.

REMARK 8.4.2

ARCH is from Robert Engle, 1982.

DEFINITION 8.4.3: Autoregressive conditionally heteroscedastic [ARCH(p)]

We say X_t follows an **autoregressive conditionally heteroscedastic** model of order p, if W_t is a strong white noise with $\mathbb{E}[W_t^2] = 1$ and

$$X_t = \sigma_t W_t$$

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2$$

where $p > 0$, $\omega > 0$, and $\alpha_j \geq 0$ for $j = 1, \dots, p$. We write $X_t \sim \text{ARCH}(p)$.

REMARK 8.4.4

(1) σ_t^2 is called the "conditional variance" or "volatility." Imagine that there exist a representation $X_t = g(W_t, \dots, W_{t-1})$ (stationary process satisfying the ARCH model). Then, for example, in the ARCH(1) case,

$$\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 = g_\sigma(W_{t-1} W_{t-2}, \dots)$$

Therefore,

$$\mathbb{V}(X_t \mid W_{t-1}, W_{t-2}, \dots) = \mathbb{V}(\sigma_t W_t \mid W_{t-1}, \dots) = \sigma_t^2 \mathbb{V}(W_t) = \sigma_t^2$$

$\mathbb{V}(W_t) = 1$ identifies σ_t^2 as conditional variance.

(2) Engle won the Nobel Prize in economics in part for "methods of analyzing economic time series with time varying volatility (ARCH)" in 2003.

(3) One problem noted early on was that ARCH(p) models required large orders of p to model asset returns which suggested *generalizing* the model.

DEFINITION 8.4.5: Generalized autoregressive conditional heteroskedasticity (GARCH)

We say X_t follows a **generalized autoregressive conditional heteroskedasticity** (GARCH) model if W_t is unit variance strong white noise and

$$X_t = \sigma_t W_t$$

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2$$

where $q \geq 0$, $p > 0$, $\omega > 0$, $\alpha_j \geq 0$ for $j = 1, \dots, p$, and $\beta_k \geq 0$ for $k = 1, \dots, q$. We write $X_t \sim$ GARCH(p, q).

REMARK 8.4.6

The GARCH(p, q) model was proposed by Bollerslev (1986).

REMARK 8.4.7

- GARCH$(p, 0) \equiv$ ARCH(p).
- GARCH$(0, 0)$ is a white noise.

PROPOSITION 8.4.8: Properties of GARCH

Suppose for the moment that there exists "a stationary and causal time series X_t satisfying the GARCH(p, q) model," $X_t = g(W_t, W_{t-1}, \dots) \implies \sigma_t^2 = g_\sigma(W_{t-1}, W_{t-2}, \dots)$, then

(1) $\mathbb{E}[X_t] = \mathbb{E}[\sigma_t]\mathbb{E}[W_t] = 0$ *since σ_t and W_t are independent.*

$$\gamma_X(h) = \mathbb{E}[X_{t+h}X_t] = \mathbb{E}[\sigma_{t+h}W_{t+h}\sigma_t W_t] = 0$$

since W_{t+h} is independent of the rest. Therefore, GARCH series have mean zero and are serially uncorrelated by construction.

(2) *Suppose $X_t \sim$ ARCH(1).*

$$\begin{aligned} X_t^2 &= \sigma_t^2 W_t^2 \\ &= \sigma_t^2 (W_t^2 + 1 - 1) \\ &= \sigma_t^2 + (W_t^2 - 1) \\ &= \omega + \alpha_1 X_{t-1}^2 + \sigma_t^2 (W_t^2 - 1) \end{aligned}$$

*Now, note that $\sigma_t^2 = g(W_{t-1}, W_{t-2}, \dots)$, and $W_t^2 - 1$ is a mean zero random variable. Hence, the **last** term is a weak white noise.*

Therefore, $X_t^2 \sim$ AR(1) process (weak white noise innovations).

(3) *In general, if $X_t \sim$ GARCH(p, q), then X_t^2 follows an ARMA model with weak white noise innovations.*

$$X_t \sim \textit{GARCH}(p, q) \implies X_t^2 \textit{ is serially correlated (ARMA).}$$

[R Code] ARCH and GARCH Models

8.5 Stationarity of GARCH Models

Suppose $X_t \sim$ GARCH(p, q) model.

Question: Under what conditions on $\omega, \alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_p$, does a stationary process $\{X_t\}_{t \in \mathbb{Z}}$ satisfy these questions?

REMARK 8.5.1

Suppose a stationary solution exists that is a causal Bernoulli shift; that is,

$$X_t = g(W_t, W_{t-1}, \dots) \implies \sigma_t^2 = g_\sigma(W_{t-1}, W_{t-2}, \dots)$$

If $\mathbb{V}(X_t) < \infty$, note

$$\mathbb{V}(X_\sigma) = \mathbb{V}(\sigma_t W_t) = \mathbb{E}[\sigma_t^2 W_t^2] = \mathbb{E}[\sigma_t^2] = \sigma_X^2$$

Using the GARCH recursion:

$$\mathbb{E}[\sigma_t^2] = \omega + \sum_{j=1}^{p} \alpha_j \mathbb{E}[X_{t-j}^2] + \sum_{k=1}^{q} \beta_k \mathbb{E}[\sigma_{t-k}^2]$$

$$\implies \sigma_X^2 = \omega + \sum_{j=1}^{p} \alpha_j \sigma_X^2 + \sum_{k=1}^{q} \beta_k \sigma_X^2$$

Solving gives

$$\sigma_X^2 = \frac{\omega}{1 - \sum_{j=1}^{p} \alpha_j - \sum_{k=1}^{q} \beta_k}$$

Suggests that in order for a solution to exist in L^2, we need at least

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$$

(Bollerslev, 1986)

Consider GARCH$(1,1)$ case; that is,

$$X_t = \sigma_t W_t$$
$$\sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2$$

In order to get a stationary solution for X_t that satisfies $X_t = \sigma_t W_t$, we need a stationary casual variance process.

Let $f(z) = \alpha z^2 + \beta$. Iterate GARCH recursion:

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2 \\
&= \omega + \alpha(\sigma_{t-1}^2 W_{t-1}^2) + \beta \sigma_{t-1}^2 \\
&= \omega + (\alpha W_{t-1}^2 + \beta)\sigma_{t-1}^2 \\
&= \omega + f(W_{t-1})(\omega + \alpha X_{t-2}^2 + \beta \sigma_{t-2}^2) \\
&= \omega + \omega f(W_{t-1}) + f(W_{t-1})(\alpha X_{t-2}^2 + \beta \sigma_{t-2}^2) \\
&= \omega + \omega f(W_{t-1}) + \omega f(W_{t-1}) f(W_{t-2}) + f(W_{t-1}) f(W_{t-2})(\alpha X_{t-3}^2 + \beta_{t-3} \sigma_{t-3}^2) \\
&\vdots \\
&= \omega\left(1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} f(W_{t-j})\right) \\
&= g_\sigma(W_{t-1}, W_{t-2}, \dots)
\end{aligned}$$

Posit solution

$$\sigma_t^2 = \omega\left(1 + \sum_{j=1}^{\infty} \prod_{i=1}^{j} f(W_{t-i})\right)$$

Question: When is this series well-defined?

$$\prod_{i=1}^{j} f(W_{t-i}) = \exp\left\{\sum_{i=1}^{j} \log[f(W_{t-j})]\right\}$$

Now, note that $\sum_{i=1}^{j} \log[f(W_{t-j})]$ is a random walk. Therefore,

$$\sum_{i=1}^{j} \log[f(W_{t-j})] \to \begin{cases} +\infty & \text{with probability 1 if } \mathbb{E}[\log[f(W_0)]] > 0 \\ -\infty & \text{with probability 1 if } \mathbb{E}[\log[f(W_0)]] < 0 \\ \text{oscillates between } -\infty \text{ and } +\infty & \text{if } \mathbb{E}[\log[f(W_0)]] = 0 \end{cases}$$

The good case is when $\mathbb{E}[\log[f(W_0)]] < 0$, and it causes the terms to tend to zero fast.

THEOREM 8.5.2

A stationary solution X_t exists to the GARCH$(1,1)$ equations if and only if

$$\gamma = \mathbb{E}[\log(\alpha W_0^2 + \beta)] < 0 \quad \textit{[Top Lyapunov Exponent]}$$

The solution is of the form

$$X_t = \sigma_t W_t$$

$$\sigma_t^2 = \omega\left(1 + \sum_{j=1}^{\infty} \prod_{i=1}^{j} (\alpha W_{t-j}^2 + \beta)\right) = g(W_{t-1}, W_{t-2}, \ldots)$$

where g is a function that is not linear; that is, we have a non-linear time series.

REMARK 8.5.3

(1) If $\gamma < 0$, $\omega = 0$ forces $X_t \equiv 0$. Therefore, we will normally assume $\omega > 0$.
(2) The condition $\gamma = \mathbb{E}[\log(\alpha W_0^2 + \beta)] < 0$ depends on the distribution of W_t.
(3) A sufficient condition is $\alpha_1 + \beta_1 < 1$.

Proof of Remark 8.5.3 (3)

Jensen's Inequality: If $f : \mathbf{R} \to \mathbf{R}$ is convex, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

and the opposite holds if f is concave. We note that $\log(x)$ is concave, hence

$$\mathbb{E}[\log(\alpha W_0^2 + \beta)] \leq \log(\mathbb{E}[\alpha W_0^2 + \beta]) = \log(\alpha + \beta) < 0$$

only when $\alpha + \beta < 1$.

REMARK 8.5.4: Second-order Stationarity of GARCH$(1,1)$ Equation

If $\alpha_1 + \beta_1 > 1$, we have seen that $\mathbb{V}(X_t)$ is not well-defined. If $\alpha_1 + \beta_1 < 1$, then

$$\mathbb{E}[\sigma_t^2] = \mathbb{E}\left[\frac{\omega}{1 - \alpha - \beta}\right] < \infty$$

Assuming $\alpha_1 + \beta_1 < 1$, then we know a stationary solution exists and in this case, X_t is weakly stationary and is a weak white noise.

$$\gamma_X(h) = \mathbb{E}[X_{t+h} X_t] = \mathbb{E}[\sigma_{t+h} W_{t+h} \sigma_t W_t] = 0$$

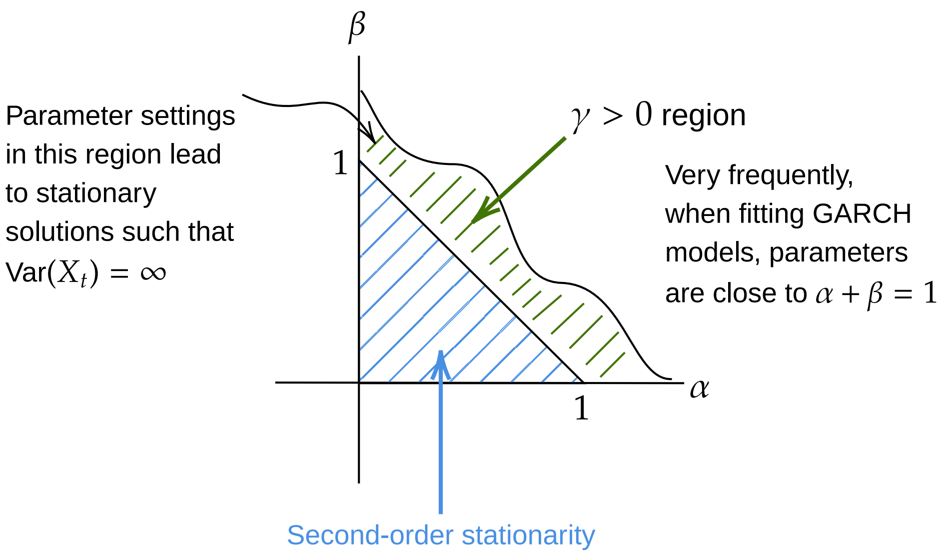


Figure 8.2: GARCH$(1, 1)$ "Region of Stationarity"

8.6 † Stationarity of General GARCH(p, q)

General conditions exist for when a GARCH(p, q) process has a strictly stationary solution: Let

$$\begin{aligned}
\tau_t &= (\beta_1 + \alpha_1 W_t^2, \beta_2, \dots, \beta_{q-1}) \in \mathbf{R}^{q-1} \\
\xi_t &= (X_t^2, 0, \dots, 0) \in \mathbf{R}^{q-1} \\
\alpha &= (\alpha_2, \dots, \alpha_{p-1}) \in \mathbf{R}^{p-2} \\
I_c &= c \times c \text{ identity matrix.} \\
N &= (\omega, 0, \dots, 0) \in \mathbf{R}^{p+q-1} \\
Y_t &= (\sigma_t^2, \dots, \sigma_{t-q+1}^2, X_t^2, \dots, X_{t-p+1}^2) \in \mathbf{R}^{p+q-1} \\
M_t &= \begin{bmatrix} \tau_t & \beta_q & \alpha & \alpha_p \\ I_{q-1} & 0 & 0 & 0 \\ \xi_t & 0 & 0 & 0 \\ 0 & 0 & I_{p-2} & 0 \end{bmatrix} \in \mathbf{R}^{(p+q-1)\times(p+q-1)}
\end{aligned}$$

THEOREM 8.6.1

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 autoregression 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.

$$r_t = \|A_t A_{t-1} \dots A_1\|$$

under some relatively mild conditions (ergodicity)

$$\gamma = \lim_{t\to\infty} \left[\frac{1}{t}\mathbb{E}[\log(r_t)]\right]$$

is well-defined and is called the top Lyapunov exponent of the sequence A_t for $t \in \mathbf{Z}$. This result is coming from Ergodic theory in the 1970s.

THEOREM 8.6.2

A stationary solution to the GARCH(p, q) equations exists if and only if

$$\gamma < 0$$

where γ is the top Lyapunov exponent of sequence M_t for $t \in \mathbb{Z}$ appearing in the Markov representation. When a stationary solution exists, it is causal and unique.

THEOREM 8.6.3: Theorem 1 of Bollerslev (1986)

A necessary and sufficient condition for there to exist a second order stationary solution to the GARCH(p, q) equations is that

$$\sum_{j=1}^{p} \alpha_j + \sum_{\ell=1}^{q} \beta_\ell < 1$$

8.7 Identifying GARCH Models

The decision to fit a volatility (GARCH) model to a time series often arises from

(1) Observing volatility (conditional heteroskedasticity) in a series.

(2) Conditional variance forecasting is of specific interest (e.g., risk analysis, financial TS 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 residuals.

Identifying Serial Correlation

Recall that the normal ACF bounds (blue lines) are constructed based on the assumption that the series is a *strong* white noise. A GARCH model is a *weak* white noise.

ACF Bounds for Weak White Noise

Suppose for example that $X_t \sim \text{GARCH}(1, 1)$, then

$$\gamma_X(h) = 0 \quad (h \geq 1)$$

$$\hat{\gamma}_X(h) \approx \frac{1}{T} \sum_{j=1}^{T-h} X_t X_{t+h} \implies \mathbb{E}[\hat{\gamma}_X(h)] = 0$$

$$\begin{aligned}
\mathbb{V}\left(\sqrt{T}\hat{\gamma}_X(h)\right) &= \frac{1}{T} \sum_{j=1}^{T-h} \sum_{k=1}^{T-h} \mathbb{E}[X_j X_{j+h} X_k X_{k+h}] \\
&= \frac{1}{T} \sum_{j=1}^{T-h} \sum_{k=1}^{T-h} \mathbb{E}[\sigma_j W_j \sigma_{j+h} W_{j+h} \sigma_{k+h} W_{k+h} \sigma_k W_{k+h}] \\
&= \frac{1}{T} \sum_{j=1}^{T-h} \mathbb{E}[X_{j+h}^2 X_j^2] \\
&\approx \mathbb{E}[X_0^2 X_{-h}^2]
\end{aligned}$$

- If $j > k$, then W_{j+h} is independent of the other terms.
- If $k > j$, then W_{k+h} is independent of the other terms.
- $\mathbb{E}[X_j + h^2 X_j^2]$ does not simplify to a product σ_X^4 since X_{j+h}^2 is correlated with X_j^2.

THEOREM 8.7.1

If X_t is a weak white noise (suitably weakly dependent), then

$$\sqrt{T}\hat{\gamma}_X(h) \xrightarrow[T\to\infty]{D} \mathcal{N}(0, \mathbb{E}[X_0 X_{-h}^2])$$

REMARK 8.7.2

(1) $\mathbb{E}[X_0^2 X_{-h}^2]$ can be consistently estimated from the sample:

$$\hat{\sigma}_h^2 = \frac{1}{T}\sum_{j=1}^{T-h} X_{j+h}^2 X_j^2$$

Therefore, an approximate $(1-\alpha)$ prediction interval for $\hat{\rho}(h)$ under the assumption of a weak white noise is

$$\pm \frac{1}{\sqrt{T}} z_{1-\alpha/2} \frac{\hat{\sigma}_h}{\hat{\gamma}(0)}$$

The blue line depends on h due to $\hat{\sigma}_h$.

(2) Note that

$$\mathbb{E}[X_0^2 X_{-h}^2] = \left(\mathbb{E}[X_0^2]\right)^2 + \underbrace{\text{Cov}(X_0^2, X_{-h}^2)}_{\text{GARCH} \implies \text{Cov}(\cdot)>0}$$

Hence, in a GARCH setting, the weak white noise intervals for ACF are (often) larger.

[R Code] Identifying GARCH Models

Chapter 9

Week 9

9.1 Tests for GARCH Models

Conditional heteroscedasticity is characterized by correlation in X_t^2. Formally, we can use a white noise test to X_t^2 to evaluate if X_t exhibits conditional heteroscedasticity.

THEOREM 9.1.1: Portmanteau (White Noise) Test of X_t^2

Let $\hat{\rho}_{X^2}(h)$ denote the empirical ACF of the series X_t^2 for $t = 1, \dots, T$. If X_t is a strong white noise with $\mathbb{E}[X^4] < \infty$, we define

$$Q(T, H) = T \sum_{h=1}^{H} \hat{\rho}_{X^2}(h) \xrightarrow[T\to\infty]{D} \chi^2(H)$$

where H is the number of lags we use. If $X_t \sim$ GARCH model, then

$$Q(T, H) \xrightarrow[T\to\infty]{P} \infty$$

The p-value of test for homoscedasticity versus conditional heteroscedasticity is given by

$$p = \mathbb{P}(\chi^2(H) \geq Q(T, H))$$

REMARK 9.1.2

(1) This test has several names in the literature, including "McLeod-Li Test."
(2) Often, it is applied to the GARCH models in order to evaluate goodness-of-fit of a GARCH model (and decide on p and q).

[R Code] Tests for GARCH Models

9.2 GARCH Parameter Estimation

Consider ARCH(1) case. We showed that if $X_t \sim$ ARCH(1), then $X_t^2 \sim$ AR(1); that is, $X_t^2 = \omega + \alpha X_{t-1}^2 + V_t$, where $V_t = \sigma_t^2(W_t^2 - 1)$ is a weak white noise.

Suggests estimating ω, α using least squares.

$$(\hat{\omega}, \alpha) = \underset{\omega \geq 0,\, 0<\alpha<1}{\arg\min} \sum_{t=2}^{T} \left[X_t^2 - (\omega + \alpha X_{t-1}^2)\right]^2$$

REMARK 9.2.1

This leads to consistent estimation for an ARCH(1) model.

For a general ARCH(p) model, we can also use least squares:

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{j=p+1}^{T} \left[X_j^2 - (\omega + \alpha_1 X_{j-1}^2 + \dots + \alpha_p X_{j-p}^2)\right]^2$$

where $\boldsymbol{\alpha} = (\omega, \alpha_1, \dots, \alpha_p)^\top$. Minimized by

$$\hat{\boldsymbol{\alpha}} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$$

$$X = \begin{bmatrix} 1 & X_p^2 & \cdots & X_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{T-1}^2 & \cdots & X_{T-p}^2 \end{bmatrix} \in \mathbf{R}^{(T-p)\times(p+1)}$$

$$\boldsymbol{Y} = (X_{p+1}^2, \dots, X_T^2)^\top \in \mathbf{R}^{T-p}$$

THEOREM 9.2.2: Chapter 7, Francq and Zakoïan

The OLS estimators of the ARCH(p) process 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 have a non-degenerate distribution.

Quasi-Maximum Likelihood Estimation

Let $X_t \sim$ ARCH(1); that is, $X_t = \sigma_t W_t$ and $\sigma_t^2 = \omega + \alpha X_{t-1}^2$.

We make a **parametric assumption** that $W_t \sim \mathcal{N}(0,1)$. Assuming the model admits a stationary and causal solution ($\omega > 0$ and $0 \le \alpha < 1$), then

$$\underbrace{X_t \mid X_{t-1}}_{\sigma_t^2 \text{ is known}} \sim \mathcal{N}(0, \omega + \alpha X_{t-1}^2)$$

$$\mathcal{L}(\omega, \alpha) = \prod_{t=2}^{T} \underbrace{\mathcal{L}(\omega, \alpha, X_t \mid X_{t-1}, \dots, X_1)}_{\mathcal{N}(0, \omega + \alpha X_{t-1}^2)}$$

which is maximized numerically.

General GARCH(p, q) Case

$$X_t \mid X_{t-1}, \dots, X_1 \overset{D}{\approx} X_t \mid \underbrace{X_{t-1}, X_{t-2}, \dots}_{\text{infinte past}} \sim \mathcal{N}(0, \sigma_t^2)$$

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} a_j X_{t-j}^2 + \sum_{\ell=1}^{q} \beta_\ell \sigma_{t-\ell}^2 = \sigma_t^2(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$\mathcal{L}(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=\max(p,q)+1}^{T} f_{\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}}(X_j \mid X_{j-1}, \dots, X_1)$$

where $f_{\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}}(X_j \mid X_{j-1}, \dots, X_1)$ is the conditional density of $\mathcal{N}(0, \sigma_j^2(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}))$.

REMARK 9.2.3

There is a catch to Quasi-Maximum Likelihood Estimation. As the equation

$$\sigma_t^2 = \sigma_t^2 = \omega + \sum_{j=1}^{p} a_j X_{t-j}^2 + \sum_{\ell=1}^{q} \beta_\ell \sigma_{t-\ell}^2 = \sigma_t^2(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

is iterated to calculate the conditional likelihood eventually things arise that are unknown:

$$\{X_j : j \leq 0\}$$

$$\{\sigma_j^2, j \leq 0\}$$

Therefore, we do some initializations:
- $\sigma_t^2 = \omega$ and $X_t^2 = \omega$ for $t \leq 0$.
- $\sigma_t^2 = \omega$ and $X_t^2 = 0$ for $t \leq 0$.

Note: if the series is "long," the initializations won't have much of an effect. However, we must be careful when fitting a GARCH model to short series.

Parameter Constraints:

$$(\hat{\omega}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \underset{\hat{\omega}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}}{\arg\max}\, \mathcal{L}(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

admits a stationary solution.

(1) "Hyper-Pyramid:"

$$(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta}) \in \left\{ \omega > 0, \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1, \alpha_i, \beta_j \geq 0 \right\}$$

solution is second-order stationary. Frequently, parameter estimates lie near the boundary (i.e., $\alpha + \beta = 1$) Most packages consider this region.

(2) $(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$: Top Lyapunov exponent < 0. Entire stationary region is searched. Some "better" packages implement this (e.g., SAS).

THEOREM 9.2.4: Chapter 6, Francq and Zakoïan

If $X_t \sim$ GARCH(p, q) admits a stationary and causal solution, then the Quasi-MLE (QMLE) estimators are consistent.
- *If $W_t \sim \mathcal{N}(0, 1)$ (actually, so that QMLE = MLE), then the estimators are efficient (achieve the smallest variance among consistent estimators).*
- *If $W_t \nsim \mathcal{N}(0, 1)$, the QMLE may not be efficient, but it is in several cases.*

Takeaway: QMLE estimation is the benchmark of GARCH model parameter estimation.

9.3 GARCH Residuals and Forecasting the Conditional Variance

If $X_t \sim \text{GARCH}(p, q)$, then $(\omega, \boldsymbol{\alpha}, \boldsymbol{\beta})$ can be estimated using QMLE to obtain $(\hat{\omega}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$,

Then, estimates of conditional variance can be computed by:

$$\hat{\sigma}_t^2 = \hat{\omega} + \sum_{j=1}^{p} \hat{\alpha}_j X_{t-j}^2 + \sum_{\ell=1}^{q} \hat{\beta}_\ell \hat{\sigma}_{t-\ell}^2 \qquad q + 1 \leq t \leq T$$

$$\hat{\sigma}_j^2 = \hat{\omega} + \sum_{\ell=1}^{\min(j,p)} \hat{\alpha}_\ell X_{j-\ell}^2 \qquad 1 \leq t \leq q$$

GARCH Residuals

$$X_t = \sigma_t W_t \implies W_t = \frac{X_t}{\sigma_t} \quad (\omega > 0)$$

Therefore, the residuals are given by

$$\hat{W}_t = \frac{X_t}{\hat{\sigma}_t}$$

Model diagnostics can be applied to $\hat{W}_t$ to check:

(1) "Whiteness" or "Squared Correlation."

(2) Normality.

(3) These also may be used in bootstrap procedures.

Forecasting the Conditional Variance

1-step ahead:

$$\hat{\sigma}^2_{T+1} = \hat{\omega} + \sum_{j=1}^{p} \hat{\alpha}_j X^2_{T-j} + \sum_{\ell=1}^{q} \hat{\beta}_\ell \hat{\sigma}^2_{T-\ell}$$

Initializations: $X_t^2 = \hat{\omega}$, $\hat{\sigma}_t^2 = \hat{\omega}$ for $t \leq 0$.

h-step ahead:

$$\hat{\sigma}^2_{T+h} = \hat{\omega} + \sum_{j=1}^{p} \hat{\alpha}_j \hat{X}^2_{T+h-j} + \sum_{\ell=1}^{q} \hat{\beta}_\ell \hat{\sigma}^2_{T+h-\ell}$$

$$\hat{X}_t^2 = \begin{cases} X_t^2 & t \leq T \\ \hat{\omega} \text{ or } \dfrac{\hat{\omega}}{1 - \sum_{j=1}^{p} \hat{\alpha}_j - \sum_{\ell=1}^{q} \hat{\beta}_\ell} & t > T \end{cases}$$

Chapter 10

Week 10

10.1 Choosing the Orders of a GARCH Model

(1) Use a GARCH$(1,1)$ model. "We do not find much evidence that the GARCH$(1,1)$ model is outperformed." Hansen, Peter R., and Asger Lunde (2001).

(2) Model Diagnostics: Consider the GARCH residuals

$$\hat{W}_t = \frac{X_t}{\hat{\sigma}_t}$$

   (a) Check for whiteness, BLP test applied for $\hat{W}_t$, and $\hat{W}_t^2$ (check for residual correlation in the squares).

   (b) Plot the ACF of $\hat{W}_t$ and $\hat{W}_t^2$.

(3) Use information criteria. If $\mathcal{L}(\hat{\omega}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is the maximized likelihood, then

$$\text{IC} = -2\log[\mathcal{L}(\hat{\omega}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})] + P(T,k)$$

where $k = 1 + p + q$ and $P(T,k)$ is the penalty term (AIC or BIC).

REMARK 10.1.1: Cross-validation

It is difficult to apply cross-validation in GARCH modelling since $\hat{\sigma}_t^2$ (object we are modelling) is unobserved.

Possible cross-validation criterion: Compare X_t^2 to $\hat{\sigma}_t^2$ (estimated from $X_{t-1}, \ldots, X_1$). It is not typical to do this (although maybe it should be).

[R Code] Choosing the Orders of a GARCH Model

10.2 Value at Risk Forecasting

One common application of GARCH modelling is to forecast the conditional quantile of the loss in price of financial assets.

DEFINITION 10.2.1: Horizon h loss

Suppose V_t is the price (value) of an asset at time t. The **horizon h loss** is denoted

$$L_{t,t+h} = -(\underbrace{V_{t+h} - V_t}_{\text{horizon } h \text{ return}})$$

DEFINITION 10.2.2: Value at risk

Let $\mathcal{F}_t$ denote all "information" available up to time t. For example, $\mathcal{F}_t = X_t, X_{t-1}, \dots, V_t, V_{t-1}, \dots$.
The horizon h **value at risk** at level α is denoted $\text{VaR}_{t,h}(\alpha)$, satisfies

$$\mathbb{P}(L_{t,h} > \text{VaR}_{t,h}(\alpha) \mid \mathcal{F}_t) \le \alpha$$

In practice, we take

$$\text{VaR}_{t,h}(\alpha) = \inf\{x : \mathbb{P}(L_{t,h} > x \mid \mathcal{F}_t) \le \alpha\}$$

That is, $\text{VaR}_{t,h}(\alpha)$ is the $(1-\alpha)$ conditional quantile of the loss distribution.

REMARK 10.2.3

If $L_{t,h} \mid \mathcal{F}_t$ is a continuous random variable, then $\text{VaR}_{t,h}(\alpha)$ satisfies

$$\mathbb{P}(L_{t,h} > \text{VaR}_{t,h}(\alpha)) = \alpha$$

EXAMPLE 10.2.4

If $L_{t,t+h} \mid \mathcal{F}_t \sim \mathcal{N}\left(m_{t,h}, \sigma^2_{t,h}\right)$, then

$$\text{VaR}_{t,h}(\alpha) = m_{t,h} + \sigma_{t,h}\Phi^{-1}(1-\alpha)$$

where

- $m_{t,h} = \mathbb{E}\left[L_{t,t+h} \,\middle|\, \mathcal{F}_t\right]$.
- $\sigma^2_{t,h} = \mathbb{V}\left(L_{t,t+h} \,\middle|\, \mathcal{F}_t\right)$.
- Φ^{-1} is the standard normal quantile function.

REMARK 10.2.5

Let $r_t = V_t - V_{t-1}$ be the simple returns, then

$$L_{t,t+h} = -\sum_{j=t+1}^{t+h} r_j \quad \text{[Telescoping Sum]}$$

Hence, if we can derive a model for $\{r_t\}_{t\in\mathbf{Z}}$ (e.g., a GARCH model), we can also obtain a model for $L_{t,t+h}$.
Similarly, if $r_t = \log(V_t/V_{t-1}) = \log(V_t) - \log(V_{t-1})$ denotes the log-returns, and $q_t(h,a)$ is the quantile of the conditional distribution of $r_{t+1} + \dots + r_{t+h}$, then

$$\text{VaR}_{t,h}(\alpha) = [1 - e^{q_t(h,a)}]V_t$$

"Model for returns/log-returns $\implies$ model for loss."

DEFINITION 10.2.6: RiskMetrics model

Let r_t denote the returns (or log-returns). The **RiskMetrics model** is defined by

$$\begin{aligned} r_t &= \sigma_t W_t & & W_t \sim \mathcal{N}(0,1) \\ \sigma_t^2 &= \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2 & & \text{[ETS Model for Conditional Variance]} \\ \text{VaR}_{t,1}(\alpha) &= \begin{cases} \sigma_{t+1}\Phi^{-1}(\alpha) & \text{if returns} \\ [1 - e^{q_t(h,a)}]V_t & \text{if log-returns} \end{cases} \end{aligned}$$

The h-step ahead VaR is approximated by

$$\mathrm{VaR}_{t,h}(\alpha) = \sqrt{h}\mathrm{VaR}_{t,1}(\alpha) \quad [\sqrt{h}\text{-scaling}]$$

REMARK 10.2.7

(1) $\sqrt{h}$-scaling derives from the assumption that $r_t = V_t - V_{t-1} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Therefore,

$$L_{t,t+h} = -\sum_{j=t+1}^{t+h} r_j \sim \mathcal{N}(0, \sigma^2 h) \quad \text{[Somewhat Dubious!]}$$

(2) The RiskMetrics model leads to a degenerate GARCH model ($\omega = 0$). It tends to underestimate σ_t^2.

A General Approach Using GARCH Models

- Step 1: Fit a GARCH model to the returns r_t.
- Step 2: Use the GARCH model to forecast $\hat{\sigma}_{t+1}^2$.
- Step 3: Set $q_t(1, \alpha) =$ a quantile of $r_{t+1} = \hat{\sigma}_{t+1}\hat{F}^{-1}(\alpha)$ where $\hat{F}$ is the distribution estimated from the GARCH residuals:
  (a) $\hat{F} \sim \mathcal{N}(0, 1)$ CDF.
  (b) $\hat{F} \sim t$ distribution, Pareto, etc.
  (c) $\hat{F} \sim$ Empirical CDF (Bootstrap).

For h-step ahead VaR forecasting:

- Option 1: Apply $\sqrt{h}$-scaling.
- Option 2: Use the GARCH model to simulate $r_{T+h}^{(b)}, \dots, r_{T+h}^{(b)}$, where the errors W_t are drawn from $\hat{F}$. Set $q_t(h, \alpha) =$ a quantile of $\sum_{j=t+1}^{t+h} r_j$ to be the empirical quantile of

$$\sum_{j=T+1}^{T+h} r_{T+j}^{(b)} \quad (b = 1, \dots, B)$$

  where B is large, (e.g., $B = 10^6$).

10.3 Backtesting and VaR Forecasts

DEFINITION 10.3.1: Backtesting

Backtesting returns to the practice of testing a predictive models' accuracy by applying it to historic data.

REMARK 10.3.2

Backtesting is a fancy finance term for cross-validation.

When backtesting VaR forecasts, we would be looking for:

- Correct Coverage: $\mathbb{P}\big(L_{t,t+h} > \mathrm{VaR}_{t,h}(\alpha)\big) \approx \alpha$.

- "Tightness/Sharpness to the Data:" If

$$\mathbb{P}\Big(L_{t,t+h} > \text{VaR}^1_{t,h}(\alpha)\Big) = \mathbb{P}\Big(L_{t,t+h} > \text{VaR}^2_{t,h}(\alpha)\Big)$$

  then whichever is larger is better.

1-step VaR Backtesting

Let $\mathbb{I}_{t+1}(\alpha) = \mathbb{I}\big\{L_{t,t+1} > \text{VaR}_{t,1}(\alpha)\big\}$. We should have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{t+1}(\alpha) \approx \alpha$$

- Historical Data Approach: $\hat{q}_t(1,\alpha)$ is the α empirical quantile of the last 250 returns.
- RiskMetrics: $\hat{q}_t(1,\alpha) = \hat{\sigma}_{t+1}\Phi^{-1}(\alpha)$, $\hat{\sigma}_{t+1}$ coming from the RiskMetrics "recursion" with $\lambda = 0.94$ and initialized by variance estimate from previous 250 observations.
- GARCH$(1,1)$-Gaussian: $\hat{q}_t(1,\alpha) = \hat{\sigma}_{t+1}\Phi^{-1}(\alpha)$, $\hat{\sigma}_{t+1}$ coming from GARCH$(1,1)$ fit.
- Non-parametric GARCH Bootstrap: $\hat{q}_t(1,\alpha)$ set to be a α quantile of simulated 1-step return from GARCH$(1,1)$ with errors drawn from GARCH$(1,1)$ residuals.

[R Code] Backtesting and VaR Forecasts

10.4 Asymptotics of Partial Sums of Stationary Random Variables

Suppose $\{X_t\}_{t\in\mathbf{Z}}$ is a strictly stationary time series; that is, $\mathbb{E}[X_t] = \mu$, and $\gamma_X(h) = \mathbb{E}[(X_t-\mu)(X_{t+h}-\mu)]$. We denote the estimator for μ by:

$$\bar{X} = \frac{1}{T}\sum_{i=1}^{T}X_i$$

Note that $\mathbb{E}[\bar{X}] = \mu$ and

$$\begin{aligned}\mathbb{V}(\bar{X}) &= \frac{1}{T^2}\sum_{j=1}^{T}\sum_{i=1}^{T}\mathbb{E}[(X_i-\mu)(X_j-\mu)] \\ &= \frac{1}{T^2}\sum_{h=1-T}^{T-1}(T-|h|)\gamma_X(h) \\ &\approx \frac{1}{T}\sum_{h=-\infty}^{\infty}\gamma_X(h) \quad \text{as } T\to\infty\end{aligned}$$

where $\gamma_X(h)$ is called the "long-run" variance of $\{X_t\}_{t\in\mathbf{Z}}$.

THEOREM 10.4.1

Under weak dependence conditions on $\{X_t\}_{t\in\mathbf{Z}}$ (e.g., if X_t is a linear process with $\sum_{\ell=0}^{\infty}\psi_\ell^2 < \infty$), then

$$\sqrt{T}(\bar{X}-\mu) \xrightarrow[T\to\infty]{D} \mathcal{N}\left(0, \sum_{h=-\infty}^{\infty}\gamma(h)\right)$$

Application: Inference for the mean of a stationary time series. Suppose $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary, $\mathbb{E}[X_t] = \mu$.

- H_0: $\mu = \mu_0$

- H_A: $\mu \neq \mu_0$

Test statistic:

$$Z_T = \frac{\sqrt{T}(\bar{X} - \mu_0)}{\sqrt{\sum_{h=-\infty}^{\infty} \gamma_X(h)}} \overset{D}{\approx} \mathcal{N}(0,1) \implies p = \mathbb{P}(|Z| > |Z_T|)$$

where $Z \sim \mathcal{N}(0,1)$.

Estimating the Long-Run Variance (LRV)

$\sigma^2_{\text{LRV}} = \sum_{h=-\infty}^{\infty} \gamma_X(h)$, a natural estimator is $\sum_{h=1-T}^{T-1} \hat{\gamma}_X(h)$. A problem here is that $\hat{\gamma}(T-1)$ is only based on a pair of observations.

Truncated Long-Run Variance Estimator

$$\hat{\sigma}^2_{\text{LRV}} = \sum_{h=-H}^{H} \hat{\gamma}_X(h)$$

H is the "bandwidth" or "truncation parameter." Normally, in order that $\hat{\sigma}^2_{\text{LRV}}$ would be consistent, we take $H = H(T) \xrightarrow{T\to\infty} \infty$. So,

$$\frac{H(T)}{T} \xrightarrow{T\to\infty} 0$$

Standard Choices of H

Default in most R functions that use truncated LRV estimators:

$$H(T) = \left\lfloor 4\left(\frac{T}{100}\right)^{1/4} \right\rfloor$$

Another one:

$$H(T) = \left\lfloor 12\left(\frac{T}{100}\right)^{1/4} \right\rfloor$$

Dependent Z-test or t-test

$$Z_T = \frac{\sqrt{T}(\bar{X} - \mu_0)}{\hat{\sigma}_{\text{LRV}}}$$

More conservative:

$$p = \mathbb{P}(|t_{T-1}| > |Z_T|)$$

Another one:

$$p = \mathbb{P}(|Z| > |Z_T|)$$

Partial Sum Process

Suppose $X_1, \dots, X_T$ are i.i.d. with $\mathbb{E}[X_i] = 0$ and $\mathbb{V}(X_i) = \sigma^2$. Define

$$S_T(x) = \frac{1}{\sqrt{T}} \sum_{i=1}^{\lfloor T_x \rfloor} X_i \quad \text{[Partial Sum Process]}$$

By CLT, $S_T(1) \overset{D}{\approx} \sigma\mathcal{N}(0,1)$ as $T \to \infty$. Also,

$$S_T(x) \xrightarrow[T\to\infty]{D} \sigma W(x) \quad \text{[Standard Wiener Process or Brownian-Motion]}$$

THEOREM 10.4.2

If $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary and suitably weakly dependent, then

$$S_T(x) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor T_x \rfloor}(X_t - \mu) \xrightarrow[T\to\infty]{D} \sigma_{LRV} W(x)$$

where

$$\sigma^2_{LRV} = \sum_{h=-\infty}^{\infty} \gamma_X(h)$$

10.5 KPSS Test

We are often interested in evaluating:

- H_0: $\{X_t\}_{t\in\mathbf{Z}}$ is stationary.
- H_A: $\{X_t\}_{t\in\mathbf{Z}}$ is non-stationary.

Other possible alternatives are:

- $H_{A,1}$: Change in level:
$$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_{k^*}] \neq \mathbb{E}[X_{k^*+1}] = \cdots = \mathbb{E}[X_T]$$
- $H_{A,2}$: Trend: $X_t = f(t) + \varepsilon_t$ where ε_t is stationary.
- $H_{A,3}$: Random-Walk [Unit Root]: $X_t = X_{t-1} + \varepsilon_t$.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

Consider

$$Z_T(x) = \frac{1}{\sqrt{T}}\sum_{i=1}^{\lfloor T_x \rfloor}(X_i - \bar{X}) = S_T - \frac{\lfloor T_x \rfloor}{T} S_T(1)$$

where $S_T(x) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor T_x \rfloor}(X_t - \mu)$. As we mentioned, fluctuations in $Z_T(x)$ as a function of x that are "large" indicate change in the level or random variable.

$$\text{KPPS}_T = \text{Measure of Fluctuations} = \frac{1}{T\hat{\sigma}^2_{\text{LRV}}}\sum_{k=1}^{T} Z_T^2(k/T)$$

Under H_0: $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary and weakly dependent with $\mathbb{V}(X_t) < \infty$.

$$\text{KPSS}_T = \frac{1}{T\hat{\sigma}^2_{\text{LRV}}}\sum_{t=1}^{T} Z_T^2(t/T) \approx \int_0^1 \left[\frac{Z_T(x)}{\sigma_{\text{LRV}}}\right]^2 dx$$

$$Z_T(x) = S_T(x) - \frac{\lfloor T_x \rfloor}{T} S_T(1) \xrightarrow[T\to\infty]{D} \sigma_{\text{LRV}}[W(x) - xW(1)]$$

Define $W(x) - xW(1)$ as the **Brownian Bridge** $B(x)$. Therefore,

$$\text{KPSS}_T \xrightarrow[T\to\infty]{D} \int_0^1 B^2(x)\,dx \quad \text{[Cramér-Von Mises Distribution]}$$

Under $H_{A,1}$ to $H_{A,3}$, $\text{KPSS}_T \xrightarrow[T\to\infty]{p} \infty$. If CVM $:= \int_0^1 B^2(x)\,dx$, then $p = \mathbb{P}(\text{CVM} > \text{KPSS}_T)$. Small p suggest non-stationarity.

REMARK 10.5.1

(1) Note that the null hypothesis of the KPSS test is stationarity, and so we only reject if there is *strong* evidence against stationarity.
"KPSS test is unlikely to identify series that only have mild non-stationarity."

(2) Test is powerful against:
- Changes in level.
- Trends.
- Random walk.

(3) Test is not powerful against:
- Heteroscedasticity (change in variance).

[R Code] KPSS Test

10.6 Diebold-Mariano Test

Notice that if we have two models

$$\begin{aligned} M_1 &\xrightarrow{\text{Forecasts}} \hat{X}_{t,1} \xrightarrow{\text{CV Errors}} \hat{e}_{t,1} = X_t - X_{t,1} \xrightarrow{\text{Loss}} \hat{L}_{t,1} = L(\hat{e}_{t,1}) \\ M_2 &\xrightarrow[\text{Forecasts}]{} \hat{X}_{t,2} \xrightarrow[\text{CV Errors}]{} \hat{e}_{t,2} = X_t - X_{t,2} \xrightarrow[\text{Loss}]{} \hat{L}_{t,2} = L(\hat{e}_{t,2}) \end{aligned}$$

where $L(x) = x^2 \implies$ MSE for example.

$$\text{CV Error} = \sum_{t \in \text{test sample}} \hat{L}_{t,i}$$

REMARK 10.6.1

Even if the models have the same predictive power, one of them will have "better" cross-validation error.
Question: Is the model "really" better?

Diebold-Mariano (1995) suggested testing

$$H_0\colon \mathbb{E}\left[\hat{L}_{t,1} - \hat{L}_{t,2}\right] = 0$$

Statistic: $D = \hat{L}_{t,1} - \hat{L}_{t,2}$ (average loss difference between models).

$$\bar{D} = \frac{1}{T}\sum_{t=1}^{T} D_t \quad [T\text{-length of test sample}]$$

Under the assumption that D_t is weakly dependent and stationarity, and if H_0 holds, then

$$\text{DM}_T = \frac{\sqrt{T}\bar{D}}{\hat{\sigma}_{\text{LRV}}(D)} \xrightarrow{D} \mathcal{N}(0,1)$$

Test of Equivalent Mean Loss:

$$p = \mathbb{P}(|Z| > |\text{DM}_T|)$$

[R Code] Diebold-Mariano Test

Chapter 11

Week 11

11.1 Multivariate Time Series Introduction

So far we have considered the case where $\{X_t\}_{t\in\mathbf{Z}}$, or an observed stretch $X_1, \dots, X_T$ are real numbers (take values in $\mathbf{R}$).

Frequently, we observe multiple time series at the same time. Suppose we observe d time series of length T.

$$\begin{matrix} X_{1,1} & \cdots & X_{1,T} \\ X_{2,1} & \cdots & X_{2,T} \\ \vdots & \ddots & \vdots \\ X_{d,1} & \cdots & X_{d,T} \end{matrix}$$

Conceptually, we might imagine that what we observe is a vector $\boldsymbol{X}_t = (X_{1,t}, \dots, X_{d,t})^\top \in \mathbf{R}^d$ for $1 \le t \le T$.

DEFINITION 11.1.1: Multivariate time series

Consider a vector-valued stochastic process $\boldsymbol{X}_t = (X_{1,t}, \dots, X_{d,t})^\top \in \mathbf{R}^d$, $t \in \mathbf{Z}$. We call such a process indexed by the integers, or an observed stretch $\boldsymbol{X}_1, \dots, \boldsymbol{X}_T$, a **multivariate** (**vector-valued**, **d-variate**) time series.

EXAMPLE 11.1.2

- $(X_{1,t}, \dots, X_{d,t})^\top$ could denote the log-returns of d-stocks.
- $(X_{1,t}, X_{2,t}, X_{3,t})^\top$ could denote the measurements of the position of an object at time t.

DEFINITION 11.1.3: Mean, Autocovariance matrix (Multivariate)

Consider a multivariate time series $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ of dimension $\boldsymbol{d}$. The **mean** of the process is

$$\mu_t = \mathbb{E}[\boldsymbol{X}_t] = \begin{pmatrix} \mathbb{E}[X_{1,t}] \\ \vdots \\ \mathbb{E}[X_{d,t}] \end{pmatrix}$$

The **autocovariance matrix** is

$$\Gamma(t,s) = \mathbb{E}\big[(\boldsymbol{X}_t - \mu_t)(\boldsymbol{X}_s - \mu_s)^\top\big] \in \mathbf{R}^{d\times d}$$

where $\Gamma(t,s)$ encodes the variances/covariances between all coordinates of the time series at times t and s.

DEFINITION 11.1.4: Weakly stationary, Strictly stationary (Multivariate)

We say a vector-valued time series $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ is **weakly stationary** if

$$\mu_t = \mathbb{E}[\boldsymbol{X}_t] = \boldsymbol{\mu} \quad \text{[does not depend on } t\text{]}$$

$$\Gamma(t+h,t) = \Gamma(h) \quad \text{[autocovariance only depends on the lag]}$$

We say $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ is **strictly stationary** if for all $h \in \mathbf{Z}$, $m \in \mathbf{N}$, $i_1, \dots, i_m \in \mathbf{Z}$, $\mathcal{B}_1, \dots, \mathcal{B}_m \subseteq \mathbf{R}^d$ ("measurable subsets") we have

$$\mathbb{P}(\boldsymbol{X}_{i_1} \in \mathcal{B}_1, \dots, \boldsymbol{X}_{i_m} \in \mathcal{B}_m) = \mathbb{P}(\boldsymbol{X}_{i_1+h} \in \mathcal{B}_1, \dots, \boldsymbol{X}_{i_m+h} \in \mathcal{B}_m)$$

"Finite dimensional distributions are shift-invariant."

PROPOSITION 11.1.5: Properties of Multivariate Stationary Processes

- $\Gamma(h) = \Gamma(-h)^\top$.

$$\begin{aligned} \Gamma(-h)^\top &= \Big\{\mathbb{E}\big[(X_{t-h} - \mu)(X_t - \mu)^\top\big]\Big\}^\top \\ &= \mathbb{E}\big[(X_t - \mu)(X_{t-h} - \mu)^\top\big] \\ &= \mathbb{E}\big[(\boldsymbol{X}_{t+h} - \mu)(\boldsymbol{X}_t - \mu)^\top\big] && \textit{by weak stationarity} \\ &= \Gamma(h) \end{aligned}$$

- *By the Cauchy-Schwarz inequality,*

$$|\Gamma(h)[i,j]| \le \Big\{\Gamma(0)[i,i]\Gamma(0)[j,j]\Big\}^{1/2}$$

  - $\Gamma(h)[i,j]$ *is the covariance between* $X_{i,t+h}$ *and* $X_{j,t}$.
  - $\Gamma(0)[i,i]$ *is the variance of* $X_{i,0}$.
  - $\Gamma(0)[j,j]$ *is the variance of* $X_{j,0}$.

DEFINITION 11.1.6: Autocorrelation matrix

The **autocorrelation matrix** is defined as

$$R(h)[i,j] = \frac{\Gamma(h)[i,j]}{\Big\{\Gamma(0)[i,i]\Gamma(0)[j,j]\Big\}^{1/2}}$$

REMARK 11.1.7

- $\Gamma(h)[i,i] = \gamma_i(h)$ is the autocovariance of the component series $X_{i,t}$.
- $R(h)[i,i]$ is the ACF of the time series $X_{i,t}$.

DEFINITION 11.1.8: Cross-covariance, Cross-correlation function

The **cross-covariance** between series $X_{1,t}$ and $X_{2,t}$ assumed to be stationary is

$$\gamma_{1,2}(h) = \mathbb{E}\big[(X_{1,t+h} - \mu_1)(X_{2,t} - \mu_2)\big] = \Gamma(h)[1,2]$$

The **cross-correlation function** is

$$\rho_{1,2}(h) = \frac{\gamma_{1,2}(h)}{\big[\gamma_1(0)\gamma_2(0)\big]^{1/2}} = R(h)[1,2]$$

DEFINITION 11.1.9: Empirical autocovariance matrix

If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ is an observed series of length T (assumed to arise from a weakly stationary series), then the **empirical autocovariance matrix** is

$$\hat{\Gamma}_h = \frac{1}{T}\sum_{t=1}^{T-h}(\boldsymbol{X}_{t+h} - \bar{X})(\boldsymbol{X}_t - \bar{X})^\top$$

where $\bar{X} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_t$.

$$\hat{R}_h = \text{diag}\big[\hat{\Gamma}(0)\big]^{-1/2}\hat{\Gamma}(h)\,\text{diag}\big[\hat{\Gamma}(0)\big]^{-1/2}$$

THEOREM 11.1.10

If $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ is weakly stationary and suitably weakly dependent, then

$$\left\|\hat{\Gamma}(h) - \Gamma(h)\right\| = \mathcal{O}_p\left(\frac{1}{\sqrt{T}}\right)$$

where $\|\cdot\|$ is any norm on matrices.
If $\{X_{1,t}\}$ and $\{X_{2,t}\}$ are each strong white noises with finite variance, then

$$\sqrt{T}\hat{R}(h)[1,2] \xrightarrow[T\to\infty]{D} \mathcal{N}(0,1)$$

Takeaway: The usual "blue lines" $[\pm 1.96/\sqrt{T}]$ can be used to measure for "strong cross correlation."

[R Code] Multivariate Time Series

11.2 Vector Autoregressive and Vector ARMA Models

Suppose $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ is a strictly stationary vector-valued process in $\mathbf{R}^d$.

DEFINITION 11.2.1: Vector autoregressive process

We say $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ follows a **vector autoregressive process** of order 1, denoted VAR(1), if there exists a matrix $A \in \mathbf{R}^{d\times d}$ so that

$$\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{W}_t$$

where $\{\boldsymbol{W}_t\}_{t\in\mathbf{Z}}$ is a strong white noise in $\mathbf{R}^d$; that is, $\{\boldsymbol{W}_t\}_{t\in\mathbf{Z}}$ is i.i.d., $\mathbb{E}[\boldsymbol{W}_t] = \boldsymbol{0}$, and $\mathbb{V}(\boldsymbol{W}_t) = \Sigma_W$, where Σ_W is the covariance matrix of $\boldsymbol{W}_t$.

Stationary Solution to VAR(1)

Suppose $A \in \mathbf{R}^{d\times d}$ satisfies $\|A\|_{op} = \sup_{\|\boldsymbol{x}\|=1}\|A\boldsymbol{x}\| < 1$ where $\boldsymbol{x} \in \mathbf{R}^d$ and $\|\cdot\|$ is the Euclidean Norm. Then, the VAR recursion is:

$$\begin{aligned}
\boldsymbol{X}_t &= A\boldsymbol{X}_{t-1} + \boldsymbol{W}_t \\
&= A[A\boldsymbol{X}_{t-2} + \boldsymbol{W}_{t-1}] + \boldsymbol{W}_t \\
&= A^2\boldsymbol{X}_{t-2} + A\boldsymbol{W}_{t-1} + \boldsymbol{W}_t \\
&\vdots \\
&= \sum_{j=0}^{M} A^j\boldsymbol{W}_{t-j} + A^{M+1}\boldsymbol{X}_{t-(M+1)}
\end{aligned}$$

REMARK 11.2.2

For any $\boldsymbol{y} \in \mathbf{R}^d$,

(1) $\|A\boldsymbol{y}\| = \left\| A\dfrac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\right\| \|\boldsymbol{y}\| \le \|A\|_{op}\|\boldsymbol{y}\|$

(2) $\left\|A^M\boldsymbol{y}\right\| = \|AA^{M-1}\boldsymbol{y}\| \le \|A\|_{op}\left\|A^{M-1}\boldsymbol{y}\right\| \le \cdots \le \|A\|_{op}^M\|\boldsymbol{y}\|$. Therefore,

$$\left\|A^{M+1}\boldsymbol{X}_{t-(M+1)}\right\| \le \|A\|_{op}^{M+1}\|\boldsymbol{X}_{t-(M+1)}\| \xrightarrow{M\to\infty} 0$$

THEOREM 11.2.3

If $\|A\|_{op} < 1$, there exists a stationary process $\boldsymbol{X}_t \in \mathbf{R}^d$ so that

$$\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{W}_t$$

$$\boldsymbol{X}_t = \sum_{\ell=0}^{\infty} A^{\ell}\boldsymbol{W}_{t-\ell} \quad \textit{[vector-valued linear process]}$$

- *A^ℓ is well-defined since A is a contraction.*

DEFINITION 11.2.4: Vector ARMA

We say $\{\boldsymbol{X}_t\}_{t\in\mathbf{Z}}$ follows a **vector ARMA** model of orders p and q if there exists coefficient matrices $A_1, \dots, A_p, B_1, \dots, B_q \in \mathbf{R}^{d\times d}$ so that

$$\boldsymbol{X}_t = \underbrace{A_1\boldsymbol{X}_{t-1} + \cdots + A_p\boldsymbol{X}_{t-p}}_{\text{VAR}} + \boldsymbol{W}_t + \underbrace{B_1\boldsymbol{W}_{t-1} + \cdots + B_p\boldsymbol{W}_{t-q}}_{\text{VMA}}$$

THEOREM 11.2.5

There exist a stationary and causal solution to the vector ARMA recursion if and only if

$$\det(I - \boldsymbol{A}(z)) \neq 0 \quad (|z| \leq 1, z \in \mathbf{C})$$

where $\boldsymbol{A}(z) = A_1 \boldsymbol{z} + \cdots A_p \boldsymbol{z}^p$ is a matrix-valued polynomial.

REMARK 11.2.6

(1) Due to the difficulties of estimating the MA components in even moderate dimensions, it is common to use pure VAR models.
(2) Parameter estimation is simple using least squares.

$$\hat{A}_1, \ldots, \hat{A}_p = \underset{A_1, \ldots, A_p}{\arg\min} \sum_{t=p+1}^{T} \left\| \boldsymbol{X}_t - A_1 \boldsymbol{X}_{t-1} - \cdots - A_p \boldsymbol{X}_{t-p} \right\|^2$$

where $\| \cdot \|$ is the Euclidean Norm.
(3) Model selection can be conducted using AIC/BIC, cross-validation.

11.3 Other Multivariate Time Series Odds and Ends

As with the VARMA models, many other similar results and models from scalar time series have counterparts for multivariate time series.

THEOREM 11.3.1: Vector M-dependent CLT

If $\{\boldsymbol{X}_t\}_{t \in \mathbf{Z}}$ is a strictly stationary M-dependent time series in $\mathbf{R}^d$ with $\mathbb{E}[\|\boldsymbol{X}_t\|^2] < \infty$, then

$$\underbrace{\sqrt{T}(\bar{X} - \mu)}_{\text{Random Variable in } \mathbf{R}^d} \xrightarrow{D} \boldsymbol{G}$$

where $\boldsymbol{G}$ is a Gaussian vector in $\mathbf{R}^d$ with $\mathbb{E}[\boldsymbol{G}] = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{G}) = \sum_{h=-M}^{M} \Gamma_h$.

Results like this can be extended to suitably weakly dependent processes, e.g.,

$$\boldsymbol{X}_t = \sum_{\ell=0}^{\infty} A_\ell \boldsymbol{W}_{t-\ell}$$

Such results can be used to establish CLT's for $\hat{\gamma}_h$, the empirical autocovariance matrix:

$$\sqrt{T}(\hat{\Gamma}_h - \Gamma_h) \xrightarrow[T \to \infty]{D} G$$

where G is a mean-zero Gaussian matrix.

Application: Multivariate White Noise/Portmanteau Tests (Hosking, Li and Mcleod, 1980s)

If $X_1, \ldots, X_T$ is a d-dimensional time series sampled from a strong white noise process, then

$$P_{T,H} = T \sum_{h=1}^{H} \operatorname{trace}\left(\hat{\Gamma}_h^\top \hat{\Gamma}_0^{-1} \hat{\Gamma}_h \hat{\Gamma}_0^{-1} \right) \xrightarrow[T \to \infty]{D} \chi^2(d^2 H)$$

Approximate p-value of white noise test:

$$p = \mathbb{P}\big(\chi^2(d^2 H) > P_{T,H}\big)$$

11.4 VaR Example

[R Code] VaR Example

Chapter 12

Week 12

12.1 Multiple Time Series Regression and Transfer Function Models

Problem

Suppose that we observe a bivariate time series $(Y_t, X_t)_{1 \le t \le T}$, and we are interested solely in forecasting Y_{T+h}. X_t can be thought of as an *exogenous* or *covariate* series that we would like to use to improve the forecast of Y_t.

Wrinkles on this theme include:

- Y_t is vector-valued.
- X_t is vector-valued.
- Both X_t and Y_t are vector-valued.

DEFINITION 12.1.1: ARMAX

Y_t is said to follow an **ARMAX** model (ARMA model with eXogenous variables) if there exists a (strong) white noise $\{Z_t\}_{t \in \mathbb{Z}}$ such that

$$Y_t = \beta X_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + Z_t + \theta Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

where βX_t is the regression on X_t (contemporaneous). Using the Backshift operator, we may write this model as:

$$\phi(B) Y_t = \beta X_t + \theta(B) Z_t \implies Y_t = \frac{\beta}{\phi(B)} X_t + \frac{\theta(B)}{\phi(B)} Z_t$$

DEFINITION 12.1.2: Simple linear regression model

Y_t is said to follow a **simple linear regression model** with ARMA errors if there exists a white noise sequence $\{Z_t\}_{t \in \mathbb{Z}}$ so that

$$Y_t = \beta X_t + V_t \implies Y_t = \beta X_t + \frac{\theta(B)}{\phi(B)} Z_t$$

$$\phi(B) V_t = \theta(B) Z_t \implies V_t = \frac{\theta(B)}{\phi(B)} Z_t$$

where $\phi(B), \theta(B)$ are p, q-degree polynomials respectively.

DEFINITION 12.1.3: Transfer function model

Y_t is said to follow a **transfer function model** with X_t if there exist finite degree polynomials β, ν, ϕ, θ, and a strong white noise sequence $\{Z_t\}_{t\in\mathbf{Z}}$ such that

$$Y_t = \frac{\beta(B)}{\nu(B)} + \frac{\theta(B)}{\phi(B)} Z_t$$

EXAMPLE 12.1.4: Full Transfer Function Models

ARMAX and Simple contemporaneous regression models are special examples of *full transfer function models*.

REMARK 12.1.5: Non-Stationarity

If a certain degree of differencing is required to make Y_t, X_t stationary, then we write the transfer function model as:

$$(1-B)^d Y_t = \frac{\beta(B)}{\nu(B)}(1-B)^d X_t + \frac{\theta(B)}{\phi(B)} Z_t$$

- When $d \geq 1$, $\beta(z) = \beta$, $\nu(z) = \phi(z)$, this is called an **ARIMAX** model.
- When $d = 0$, $\beta(z) = \beta$, $\nu(z) = 1$, $\phi(z) = (1-z)^q \phi^\star(z)$, this is called a **regression model with ARIMA errors**.
- Seasonality can be incorporated by using seasonal lags in the differencing and transfer function polynomials.

12.2 Fitting and Forecasting Transfer Function Models

Transfer function models:

$$Y_t = \frac{\beta(B)}{\nu(B)} X_t + \frac{\theta(B)}{\phi(B)} Z_t$$

- Regression model with ARIMA errors:
  - $\beta(B) = \beta$ where β is a constant.
  - $\nu(B) = 1$.
  - $\phi(B) = (1-B)^d \phi^\star(B)$.

Two-step estimation:

(1) Estimate $\hat{\beta}$ using ordinary least squares:

$$\arg\min_\beta \sum_{t=1}^{T} \left(Y_t - \beta X_t\right)^2$$

(2) Calculate residuals:

$$\hat{V}_t = Y_t - \hat{\beta} X_t$$

and then fit an ARIMA model to $\hat{V}_t$. This is what most packages do!

For general transfer function models, the parameters can be estimated by positing a likelihood (usually Gaussian) for the innovations Z_t, or "pre-whitening" the input and output series to identify and estimate the transfer function, and then fitting an ARIMA model to the residuals.

$$Y_t = \frac{\beta(B)}{\nu(B)} X_t + N_t = \sum_{j=0}^{\infty} v_j B^j \cong \sum_{j=0}^{k} v_j \beta_j \quad \text{where } N_t \text{ is ARIMA}$$

Suppose there exists θ_x and ϕ_x so that

$$\frac{\theta_x(B)}{\phi_x(B)}X_t = \alpha_t \leftarrow \text{ white noise (i.e., } X_t \sim \text{ARIMA)}$$

By then defining

$$\beta_t = \frac{\theta_x(B)}{\phi_x(B)}y_t$$

$$N_t^\star = \frac{\theta_x(B)}{\phi_x(B)}N_t \quad \text{(still follows ARIMA model)}$$

we get the transfer function equation that

$$\beta_t = V(B)\alpha_t + N_t^\star \cong \sum_{j=0}^{\infty} v_k \beta^j + N_t^\star$$

REMARK 12.2.1

If X_t and N_t are independent, then α_t and $W_t^\star$ are independent. Multiply LHS and RHS by α_{t-j}, take expectation.

$$\mathbb{E}[\beta_t \alpha_{t-j}] = v_j \sigma_\alpha^2 \implies v_j = \frac{\mathbb{E}[\beta_t \alpha_{t-j}]}{\sigma_\alpha^2} \implies \hat{v}_j = \frac{\mathbb{E}[\widehat{\beta_t \alpha_{t-j}}]}{\hat{\sigma}_\alpha^2}$$

where $\mathbb{E}[\beta_t \alpha_{t-j}]$ is the CCF of α_t with β_t at lag j.

We may then estimate an ARIMA model for the noise:

$$\hat{N}_t^\star = \beta_t - \hat{V}(B)\alpha_t$$

Can be reverse-engineered by applying $\dfrac{\phi_x(B)}{\theta_x(B)}$ to estimate the original transfer function model from Y to X (Box-Jenkins, 1970s).

Forecasting Transfer Function Models

Having estimated the parameters, a forecast for Y_{T+h} can be obtained by:

(1) Forecasting covariate series $\hat{X}_{T+h}$ for $j = 1, \dots, h$.

(2) Inputting forecast covariate series and forecasted noise series (ARIMA forecast) into the transfer function model.

REMARK 12.2.2

In many cases, the covariate series X_t does not need to be forecast since it is known in advance.

EXAMPLE 12.2.3

- X_t is a trend.
- X_t is a dummy (indicator) variable coding calendar effects:

$$X_t = \begin{cases} 1 & \text{day } t \text{ is a holiday} \\ 0 & \text{otherwise} \end{cases}$$

12.3 Regression with ARIMA Errors Example

[R Code] Regression with ARIMA Errors Example

12.4 State Space Models and Kalman Filtering and Smoothing

Suppose $Y_t \in \mathbf{R}^d$, a very good class of models for Y_t are state space models or Dynamic Linear Models.

- Observation Equation:
$$Y_t = A_t X_t + \Gamma u_t + V_t$$
- State Equation:
$$X_t = \Phi X_{t-1} + \xi u_t + W_t \quad (X_t \in \mathbf{R}^p)$$
  - A_t is a known design matrix.
  - X_t is a state variable.
  - u_t are exogenous variables.
  - V_t and W_t are noise.
  - $V_t \sim \mathcal{N}_d(0, R)$.
  - $W_t \sim \mathcal{N}_p(0, Q)$.

State space models originated in Aerospace and Signal processing research:

EXAMPLE 12.4.1

We are interested in the position $X_t \in \mathbf{R}^3$ of a spacecraft. We cannot measure the position exactly, but we can measure:
$$Y_t = \begin{pmatrix} \text{velocity}_t \\ \text{azimuth}_t \\ \text{altitude}_t \end{pmatrix}$$
We assume X_t is related to Y_t through a state space model: Y_t is obtained after linearly transforming X_t and adding noise.

Every model that we have discussed so far has a state space formulation:

EXAMPLE 12.4.2: ARMA(p, q) State Space Formulation

- $\phi(B)Y_t = \theta(B)W_t$.
- Let $r = \max(p, q + 1)$.
- $\phi_j = 0$ for $j > p$ and $\theta_j = 0$ for $j > q$ where $\theta_0 = 1$.

Then, one can check that
$$Y_t = [\theta_{r-1}, \theta_{r-2}, \dots, \theta_0]\boldsymbol{X}_t \quad \text{Observation Equation}$$
$$\boldsymbol{X}_t = \begin{pmatrix} X_{t-r+1} \\ \vdots \\ X_t \end{pmatrix} \in \mathbf{R}^r$$
$$\boldsymbol{X}_{t+1} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & & \vdots \\ \vdots & \vdots & 0 & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \\ \phi_r & \phi_{r-1} & \phi_{r-2} & \cdots & \phi_1 \end{pmatrix} \boldsymbol{X}_t + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_{t-1}$$
which is our State Equation.

REMARK 12.4.3

ETS, ARIMA, and GARCH models all have state space representations.

Why are state space models nice?

(1) Unifying Framework.

(2) Extra Flexibility/Generality. By specifying design matrices A_t and exogenous variables u_t, we can handle:

(a) Missing data.

(b) Full transfer function models.

Big Problem with State-Space Representation: Having observed Y_t, what can we say about X_t?

Kalman Filter (Rudolf Kalman, 1960s)

- A method for estimating X_t based in $\{Y_s : s \leq t\}$ which is an online estimation of X_t.

Kalman Smoothing

- A method to estimate X_t based on $\{Y_s : 1 \leq s \leq T\}$ which is a retrospective estimation of X_t.

REMARK 12.4.4

If (Y_t, X_t) follow the state space model with Gaussian innovations, they are jointly Gaussian. Therefore, the best guess of

$$(X_t \mid Y_s)_{s\leq t} = \mathbb{E}[X_t \mid Y_s : s \leq t]$$

This would be the best in mean-square sense even if (X_t, Y_t) are not jointly Gaussian.

State Space Model:

$$Y_t = A_t X_t + \Gamma u_t + V_t \quad (V_t \sim \mathcal{N}_d(0, R))$$
$$X_t = \Phi X_{t-1} + \xi u_t + W_t \quad (W_t \sim \mathcal{N}(0, Q))$$

Initial conditions: X_0 and P_0 (initial variance of X_0).

Let $X_t^s = \mathbb{E}[X_t \mid Y_k : k \leq s]$ and $P_t^s = \mathbb{E}[(X_t - X_t^s)(X_t - X_t^s)^\top]$ where P_t^s is the covariance matrix of forecast error of X_t based on X_t^s.

Kalman Filter

$$X_t^{t-1} = \Phi X_{t-1}^{t-1} + \xi u_t$$
$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi^\top + Q$$
$$X_t^t = X_t^{t-1} + K_t(y_t - A_t X_t^{t-1} - \Gamma u_t)$$
$$P_t^t = [I - K_t A - t] P_t^{t-1}$$

where $K_t = P_t^{t-1} A_t^\top \left[A_t P_t^{t-1} A_t^\top + R\right]^{-1}$ is the **Kalman Gain** which defines how much we alter X_t^t based on observing the deviation Y_t from $A_t X_t^{t-1} + \Gamma u_t$.

REMARK 12.4.5

(1) $(X_t^t, P_t^t) = f(X_{t-1}^{t-1}, P_{t-1}^{t-1})$ where f is linear. The term $(X_{t-1}^{t-1}, P_{t-1}^{t-1})$ says we only have to store and do the linear algebra with $X_{t-1}^{t-1}, P_{t-1}^{t-1}$ and Y_t to update state prediction. Can be done in real time.

(2) Formulas look complicated, but they are quite simple! Just came from calculating

$$\underbrace{(X_t Y_s)_{s\le t}}_{\text{Jointly Gaussian}}$$

Kalman Smoother

Infer X_t based on $\{Y_s : 1 \le s \le T\}$ with initial conditions X_0 and P_0 for $t = T, T-1, \dots,$ (we start from the end of the series).

$$X_{t-1}^{\top} = X_{t-1}^{t-1} + J_{t-1}(X_t^{\top} - X_t^{t-1})$$
$$P_{t-1}^{\top} = P_{t-1}^{t-1} + J_{t-1}(P_t^{\top} - P_t^{t-1})J_{t-1}^{\top}$$
$$J_{t-1} = P_{t-1}^{t-1} + \phi^{\top}[P_t^{t-1}]^{-1}$$

REMARK 12.4.6

Estimating of model parameters of state space model can be obtained using MLE.

$$\varepsilon_t = y_t - A_t X_t^{t-1} - \Gamma u_t \sim \mathcal{N}(0, R)$$

where X_t^{t-1} is our best guess of X_t based on $\{y_s : s \le t-1\}$ implicitly a function of parameters.

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{j=1}^{T} f_{\varepsilon_j}(\boldsymbol{\theta})$$

which is maximizing as a function of $\boldsymbol{\theta} = (R, Q, \xi, \Gamma, \Phi)^{\top}$.

- Very difficult optimization problem (Newton-Raphson, EM, MCMC)

Application to Missing Data

Suppose we observe a time series Y_t with missing values, we would like to infer the time series

$$X_t = \begin{cases} Y_t & Y_t \text{ known} \\ Y_t^{\star} & \text{unknown values of } Y_t \text{ when missing} \end{cases}$$

$$Y_t = A_t X_t$$

$X_t \sim$ ARIMA (or other) specification thought to model Y_t well.

$$A_t = \begin{cases} 1 & Y_t \text{ is observed} \\ 0 & Y_t \text{ is missing} \end{cases}$$

Infer X_t using Kalman Smoothing.

12.5 Kalman Smoothing Time Series Imputation Example

[R Code] Kalman Smoothing Time Series Imputation Example