STAT 443: Forecasting

Rui Gong

February 19, 2021

Contents

1 Time Series

1.1 Introduction

Definition 1

We say $x_1, \dots, x_T$ is an (observed) time series of length T if x_t denotes an observation obtained at time t. In particular, the observations are ordered in time.

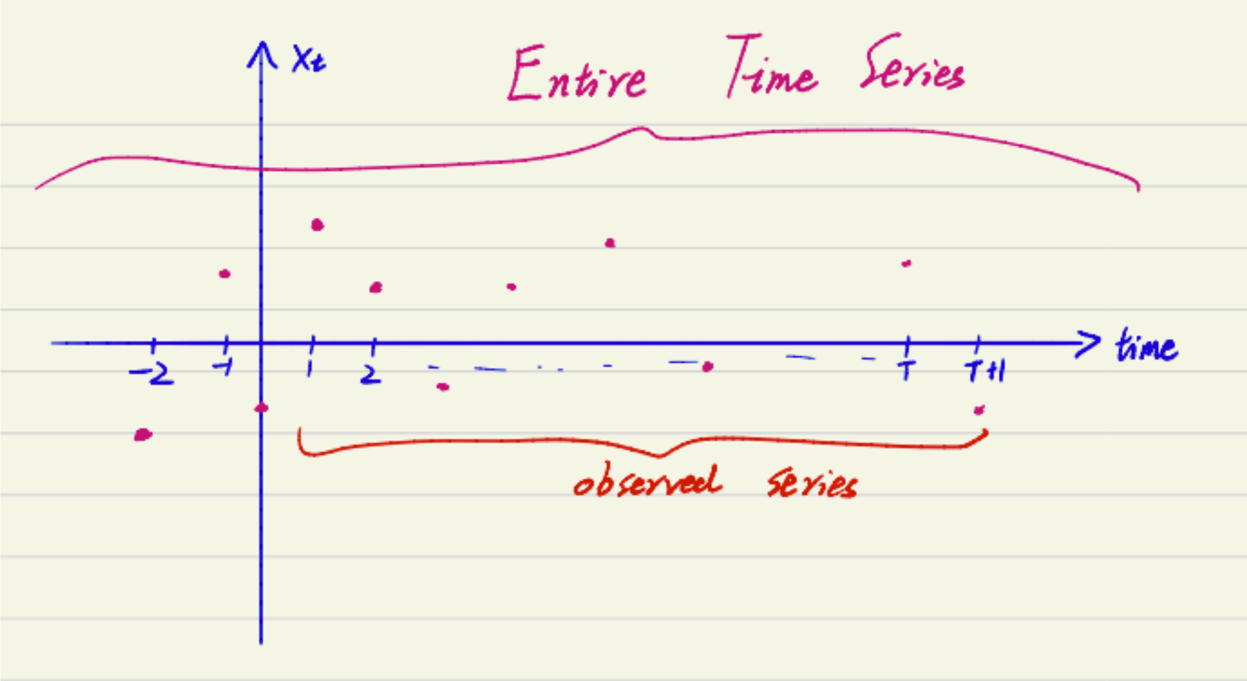
- If $X_t \in \mathbb{R}$, we say $x_1, \dots, x_T$ is a real-valued or scalar time series.
- If $X_t \in \mathbb{R}^p$, we say $x_1, \dots, x_T$ is a multivariate or vector valued time series.

With the time series data, comparing to classical statistics, we still care about prediction and inference.
However, in contrast, the data oftern exhibit:

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

Definition 2

Formally, we say $\{X_t\}_{t\in\mathbb{Z}}$ is a time series if $\{X_t : t \in \mathbb{Z}\}$ is a Stocahstic Proces indexed by $\mathbb{Z}$. This means that there is a common probability space $(\Omega, \mathcal{F}, P)$ so that $\forall t \in \mathbb{Z},\ X_t : \Omega \to \mathbb{R}$ is a random variable. In relation to the original definition, we say $x_1, \dots, x_T$ is an observed stretch or a realization or a sample path of length T from $\{X_t\}_{t\in\mathbb{Z}}$.



1.2 Forecasting

Consider a time series $x_1, \dots, x_T$. Based on $x_1, \dots, x_T$, we should 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 3

For $h \geqslant 1$, 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.

- $\hat{X}_{T+h} =$ forecast
- $h =$ horizon

There are two primary goals in forecasting:

1. Choose f_h "optimally".
   Normally, we or the practitioner have some measure, say $L(*, *)$, in mind for determining how "close" $\hat{X}_{T+h}$ is to X_{T+h}. We then wish to choose f_h such that
   $$L(X_{T+h}, f_h(X_T, \dots, X_1)) \text{ is minimized}$$
   Mose common measure $L(*, *)$ is Mean-Squared Error (MSE), where
   $$L(x, y) = E[(x - y)^2]$$
2. Quantify the uncertainty in the forecase
   This entails providing some description of how close we expect $\hat{X}_{T+h}$ to be to X_{T+h}

Example

Suppose every minute, we flip a coin such that

$$H \to 1$$
$$T \to -1$$

$X_t =$ outcome in minute$t, t = 1, \dots, T$. This produces a time series of length T, which is a random sequence of 1's and -1's.
Note $E[X_t] = 0$, for $h \geqslant 1$, consider $\hat{X}_{T+h} = f(X_T, \dots, X_1)$

$$\begin{aligned} L(X_{T+h}, \hat{X}_{T+h}) &= E[(X_{T+h} - \hat{X}_{T+h})^2] \\ &= \underbrace{E[X^2_{T+h}]}_{Var(X_t)} + E[\hat{X}^2_{T+h}] - 2\ \underbrace{E[X_{T+h}\hat{X}_{T+h}]}_{E[X_{T+h}]\hat{E}[X_{T+h}]=0} \\ &= E[X^2_{T+h}] + E[\hat{X}^2_{T+h}] \end{aligned}$$

which is minimzed by taking $\hat{X}_{T+h} = 0$
There is nothing "wrong" with the forecast, but ideally would also be able to say that the sequence appears to be random.

How can we quantify the uncertainty in forecasting?
The predictive distribution

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

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

$$P(X_{T+h} \in I_\alpha | X_T, \ldots, X_1) = \alpha (\alpha = 0.95, e.g.)$$

often such intervals take the form

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

Remark.

1. Estimiating predictive distributions leads one towards estimating the joint distribution of $X_{T+h}, X_T, \ldots, X_1$ (ARMA,ARIMA,etc).
2. It is important that we acknowledge that some things cannot be predicted!!!

1.3 Definition of Stationary

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

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

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

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

- Worst Case: $X_t \sim F_t$, where F_t is a changing function of t. If so, it's hard to pool the data $X_1, \ldots, X_T$ to estimate F_t
- Serial Dependence: If the distribution of (X_t, X_{t+h}) depends strongly on t, we have a similar problem in estimating. (e.g. $cov(X_t, X_{t+h})$)

Definition 4

We say that a time series $\{X_T\}_{t\in\mathbb{Z}}$ is strongly stationary or strictly stationary if $\forall k \geqslant 1, i_1, \ldots, i_k, h \in \mathbb{Z}$

$$(X_{i_1}, \ldots, X_{i_k}) \overset{D}{=} (X_{i_{1+h}}, \ldots, X_{i_{k+h}})$$

for all $k = 1, 2, \ldots$, all time points $i_1, \ldots, i_k$, and all $h \in \mathbb{Z}$ In other words, shifting the window on which you view the data does NOT change its distribution.

This implies that if $F_t =$ CDF of X_t, then

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

Definition 5

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

$$\mu_t = E[X_t]$$

and the autovariance function of the series is

$$\gamma(t, s) = E[(X_t - \mu_t)(X_s - \mu_s)] = cov(X_t, X_s)$$

Definition 6

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

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

i.e., $\gamma(t, s)$ is a function of $|t - s|$

In this case, we usually write

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

and we call the input h the **"lag"** parameter.
Additional terminology:

- The property when $E[X_t] = \mu$ does note depend on t is oftern called **"first order"** stationary.
- The property when $\gamma(t, s) = \gamma(|t - s|)$ only depends on the lag $|t - s|$ is called **"second order"** stationary.
- For a second order stationary process

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

Normally, we only record $\gamma(h), h \geqslant 0$

1.4 White Noise and Stationary Examples

Definition 7

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

Definition 8

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

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

Definition 9

We say $\{X_t\}_{t\in\mathbb{Z}}$ is a Gaussian white noise if

$$X_t \underset{iid}{\sim} N(0, \sigma^2)$$

Example

Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Then $E[W_t] = 0$ (doesn't depend on t).

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

$\{W_t\}_{t\in\mathbb{Z}}$ weakly stationary (γ only depends on $|t-s|$).
$\{W_t\}_{t\in\mathbb{Z}}$ is also strictly stationary. Let $k \geqslant 1$, $i_1 < i_2 < \ldots < i_k$, $k \in \mathbb{Z}$.

$$\begin{aligned} P(W_{i_1 \leqslant t_1, \ldots, W_{i_k} \leqslant t_k}) &= \prod_{j=1}^{k} P(W_{i_j} \leqslant t_j) \\ &= \prod_{j=1}^{k} P(W_{i_{j+h}} \leqslant t_j) \\ &= P(W_{i_{1+h}, \ldots, W_{i_{k+h}}} \leqslant t_k) \end{aligned}$$

Example

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

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

Then $E[X_t] = E[W_t + \theta W_{t-1}] = 0$,

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

When $|t-s|= 0$,

$$E[(W_t+\theta W_{t-1})^2] = E[W_t^2] + \theta^2 E[W_{t-1}^2] + 2E[\theta W_t W_{t-1}] = (1+\theta^2)\sigma_w^2 + 0$$

When $t = s+1$ (or $s = t+1$)

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

When $|t-s|> 1$, $W_t + \theta W_{t-1}$ is independent of $W_s + \theta W_{s-1}$
Continued: $\{X_t\}_{t\in\mathbb{Z}}$ is also strictly stationary. Suppose $k \geqslant 1, i_1, \ldots, i_k, h \in \mathbb{Z}, (i_1 < \ldots, i_k)$,

$$\begin{aligned} P(X_{i_1} \leqslant t_1, \ldots, X_{i_k} \leqslant t_k) &= P(W_{i_1} + \theta W_{i_1 - 1} \leqslant t_1, \ldots, W_{i_k} + \theta W_{i_k - 1} \leqslant t_k) \\ &= P\left[\begin{pmatrix} W_{i_1} \\ \vdots \\ W_{i_k} \end{pmatrix} \in B\right] \\ &= P\left[\begin{pmatrix} W_{i_1+h} \\ \vdots \\ W_{i_k+h} \end{pmatrix} \in B\right] \\ &= P(X_{i_1+h} \leqslant t_1, \ldots, X_{i_k+h} \leqslant t_k) \end{aligned}$$

where B is a subset of $\mathbb{R}^{i_k - i_1 + 1}$

Definition 10

Suppose $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a strong white noise. Then if $X_t = g(\varepsilon_t, \varepsilon_{t-1}, \ldots,)$ for some function:

$$g : \mathbb{R}^\infty \to \mathbb{R}$$

, we say that $\{X_t\}_{t\in\mathbb{Z}}$ is a Bernoulli shift

Theorem 1

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

Remark. Nobert Wiener conjectured that every stationary sequence is a Bernoulli shift (The TRUTH is almost every one is).

Example

Suppose W_t is strong white noise. Let

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

This is called a two-sided Random Walk. You can show that X_t is firt-order stationary, but not second-order stationary. (Consider the case when s, t have different signs and the same signs.)

1.5 Weak VS Strong Stationary

Sadly,

$$X_t \text{ strictly stationary} \nrightarrow X_t \text{ weakly stationary}$$

Ex: Suppose $X_t \underset{iid}{\sim}$ Cauchy Random Variables. i.e.

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

Then $E[X_t]$ doesn't exist, and hence not weakly stationary. But it's strongly stationary because it's a strong white noise.

If X_t strictly stationary and $E[X_0^2] < \infty \implies X_t$ is weakly stationary. Note that if X_t is strictly stationary, then

$$(X_t)\overset{D}{X_0} \implies E[X_t] = E[X_0](\text{Not depend on } t)$$

also,

$$Var(X_t) = Var(X_0)$$

By Cauchy-Schwarz inequality,

$$\gamma(t,s) = cov(X_t, X_s) \leqslant Var(X_t) < \infty$$

and suppose $t < s$,

$$\begin{aligned} cov(X_t, X_s) &= cov(X_0, X_{s-t}) = f(|t-s|) \\ (X_t, X_s) &\overset{D}{=} (X_{t-t}, X_{s-t}) \\ &\overset{D}{==} (X_0, X_{s-t}) \end{aligned}$$

Definition 11

$\{X_t\}_{t\in\mathbb{Z}}$ is said to be a Gaussian Process (or Gaussian times series) if for each $k \geqslant 1, i_1 < i_2 < i_k$,

$$(X_{i_1}, \dots, X_{i_k}) \sim MultiNormal(\underline{\mu}_k(i_1, \dots, i_k), \Sigma_{k\times k}(i_1, \dots, i_k)) = N_k(\underline{\mu}_k, \Sigma_{k\times k})$$

where

$$\underline{\mu}_k = \begin{bmatrix} E[X_{i_1}] \\ \vdots \\ E[X_{i_k}] \end{bmatrix}, \ \Sigma_{k\times k} = (cov(X_{i_j}, X_{i_r})_{1\leqslant j,r\leqslant k})$$

Proposition

If X_t is weakly stationary and Gaussian, then X_t is strictly stationary.

Proof. If X_t weakly stationary, $E[X_t] = \mu, \forall t$, and

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

$$\begin{aligned} Var(X_{i_1}, \ldots, X_{i_k}) &= [cov(X_{i_j}, X_{i_r})_{1 \leqslant j, r \leqslant k}] \\ &= [cov(X_0, X_{i_r - i_j})] \\ &= [cov(X_0, X_{i_r + h - (i_j + h)})] \\ &= [cov(X_{i_j+h}, X_{i_r+h})] \\ &= Var(X_{i_1+h}, \ldots, X_{i_k+h}) \end{aligned}$$

Using Gaussian assumption, we know

$$(X_{i_1}, \ldots, X_{i_k}) \stackrel{D}{=} N_k(\underline{\mu}, \Sigma_{k \times k}) \stackrel{D}{=} (X_{i_1+h, \ldots, X_{i_k+h}})$$

Hence, $\{X_t\}_{t \in \mathbb{Z}}$ is stricly stationary. $\square$

Exercise. Prove that if X_t is not weak;y stationary in this sense then X_t is not strictly stationary. (Hint:either $E[X_t]$ depends on t or $\gamma(X_t, X_s)$ is not a function of $|t - s|$)

1.6 Theoretiacl (L^2) framework for time series (optional)

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

Definition 12

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

Definition 13

We say that $\{X_t\}_{t\in\mathbb{Z}}$ is an L^2-time series if $X_t \in L^2, \forall t \in \mathbb{Z}$

Remark. L^2 is a Hilbert space when equipped with inner-product, $x, y \in L^2$

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

where $\langle *, *\rangle$ is an inner product.

1. Linear: $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$
2. $\langle X, X\rangle = E[X^2] = 0 \Leftrightarrow x = 0\, a.s.(i.e.\ P(X = 0) = 1)$
3. Symmetric: $\langle X, Y\rangle = \langle Y, X\rangle$

L^2 is also complete with this inner product i.e., whenever $X_n \in L^2$ so that $E[(X_n - X_m)^2] = 0$ as $n, m \implies \infty$, then $\exists X \in L^2$ such that $X_n \to X$ i.e. $E[(X_n - X)^2] \to \infty$
This follows from the "famous" Riesz-Fisher Theorem.

1.7 Useful tools for time series

1. Existence

$$X_{t,n} = \sum_{j=0}^{n} \psi_j \varepsilon_{t-j},\ \{\varepsilon_t\} \text{ is a strong WN}$$

Since $n > m$,

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

as $n, m \to 0$ if e.g. $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, then there must exist a Random Variable X_t such that $X_t = \lim_{n\to\infty} X_{t,m} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ and $X_t \in\ L^2$.

2. Projection Theorem and Forecasting
Forecasting can often be cast as finding a random variable y among a collection of possible forecast $\mathcal{M}$ (e.g. $\mathcal{M} = span\{X_T, \ldots, X_1\}$), such that

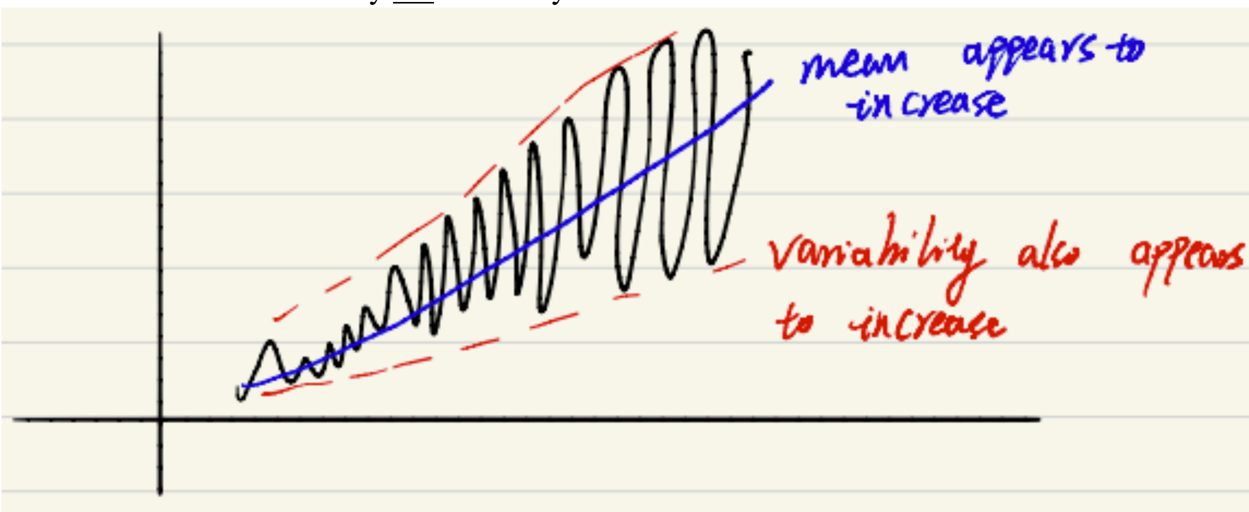
$$y = \arg\inf_{z\in\mathcal{M}E[(X_{T+h}-z)^2]}$$

when $\mathcal{M}$ is a closed linear subspace of L^2, the projection theorem gaurantees that such a y exists, and it must satisfy

$$\langle X_{T+h} - y, z\rangle = 0, \forall z \in \mathcal{M}$$

1.8 Signal+Noise Models

"Ideally", a time series that we are considering was generated from a stationary process. If so, we can pool data to estimate the process underlying structure (e.g. its marginal distribution, and serial dependence structure).
Most time series are evidently not stationary



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

- S_t is the deterministic "signal" or "trends of the series.
- ε_t is the "noise" added to the signal satisfying $E[\varepsilon_t] = 0$.
  There exists a (strong) white noise W_t such that

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

  The terms $\{W_t\}$ are often called the "innovation" or "shock" driving the random behaviour of X_t

Example 1

$\varepsilon_t = g_t(W_t, W_{t-1}, \ldots)$ might be $\varepsilon_t = \sum_{j=0}^{t} W_j$ (Random Walk), $\varepsilon_t = \sigma(t)W_t$ (changing variance models)
Goal: Estimate S_t, and infer the structure of $\varepsilon_t = g(W_t, W_{t-1}, \ldots)$

Goal: Estimate S_t, and infer the structure of $\varepsilon_t = g(W_t, W_{t-1}, \ldots)$
For the temperature data example, we may posit that

$$S_t = \beta_0 + \beta_t \text{[Linear Trend]}$$

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

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

, note $\beta_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$, $\beta_0 = \overline{y} - \beta_1 \overline{x}$

Definition 14

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

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

If the trend is now 0 (only noise left), there appears to be substantial serial dependence remaining in the series.

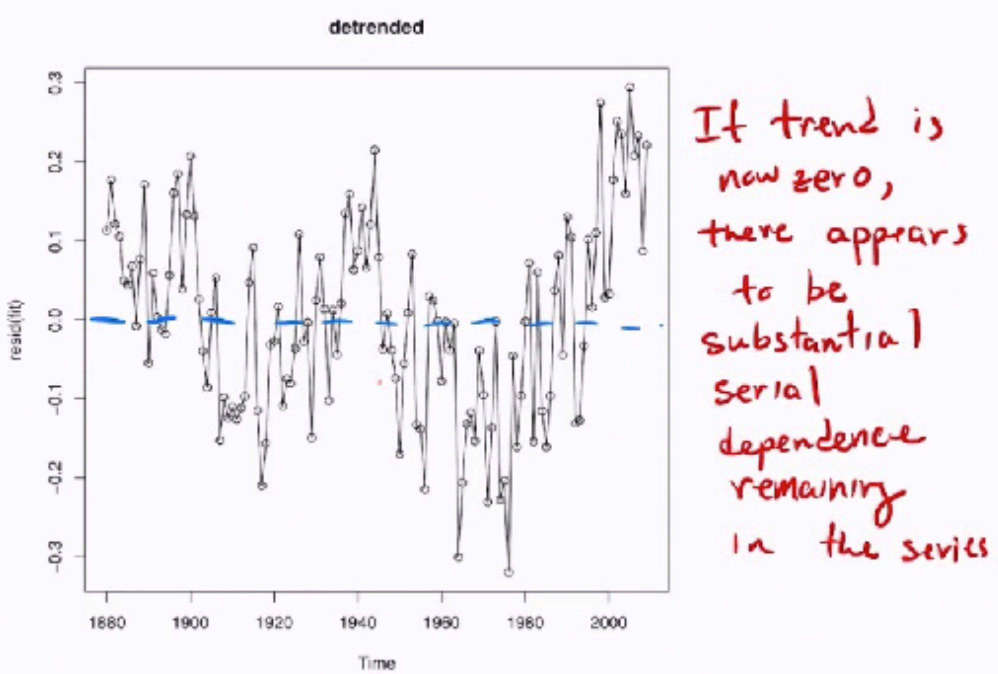


Figure: Residuals of OLS fit. A "Detrended" Time Series... Maybe not

1.9 Time Series Differencing

Signal+Noise Models: $X_t = S_t + \varepsilon_t$
Hopefully, upon estimating S_t with $\hat{S}_t$, we find $X_t - \hat{S}_t = \hat{\varepsilon}_t$ (Detrended Series) looks reasonably stationary.
If so, we might proceed in estimating the structure of $\{\hat{\varepsilon}_t\}_{t=1,\ldots,T}$ as if it were stationary.

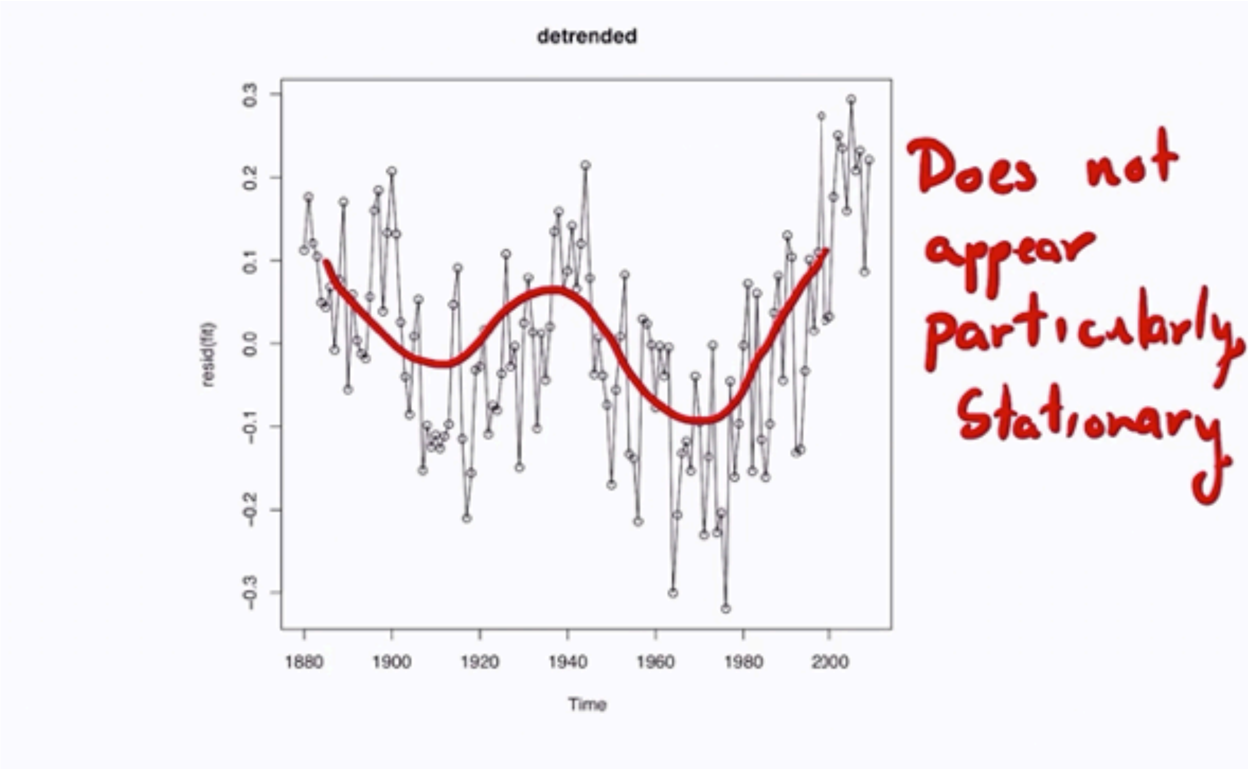


Figure: Residuals of OLS fit. A "Detrended" Time Series... Maybe not

Posit a random walk with drift model:

$$X_t = \sigma + X_{t-1} + \varepsilon_t,\ \varepsilon \sim \text{Strong White Noise}$$

Note here the σ is a drift term, constant

$$\begin{aligned}
X_t &= \sigma + X_{t-1} + \varepsilon_t \\
&= \sigma + \sigma + X_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\
&\vdots \\
&= \underbrace{t * \sigma + X_0}_{\text{linear}} + \underbrace{\sum_{j=1}^{t} \varepsilon_j}_{\text{Random Walk noise}}
\end{aligned}$$

Notice that under the Random Walk Model

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

so if X_t follows a random walk model, then the series $Y_t = \nabla X_t$ should have behave like a white noise shifted by σ.

Definition 15

Differencing a time series constitutes computing the difference between successive terms.
A diffrenced 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 $X_2 - X_1, X_3 - X_2, \ldots, X_T - X_{T-1}$ (length $T - 1$).
Higher order differences are calculated recursively, so

$$\underbrace{\nabla^d X_t}_{d^{th} \text{ order difference}} = \nabla^{d-1} \nabla X_t (\nabla^0 X_t = X_t)$$

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

- Pros
  - Differencing does not require parameter estimation (Don't estimate S_t)
  - Higher order differencing can reduce even very "trendy" series to look more like noise.
- Cons
  - Differencing can "wash away" features of time 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 2: Differencing Complicate Series

$X_t = W_t$, where $W_t \sim$ Strong White Noise :

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

$$\gamma_x(h) = cov(X_t, X_{t+h}) = \begin{cases} \sigma_w^2, & h = 0 \\ 0, & h \geqslant 1 \end{cases}$$

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

1.10 Autocorrelation and Empirical Autocorrelation:

Usually through either detrending or differencing, we arrive at a series X_t that we may consider as stationary.
Given such a series, we wish to estimate g, so that

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

where $\{W_t\}$ is an "innovation" sequence (strong white noise)

Definition 16

A time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be a linear process, if there exists a strong white noise $\{W_t\}_{t\in\mathbb{Z}}$, and coefficients $\{\psi_l\}_{l\in\mathbb{Z}}, \psi_l \in \mathbb{R}$, such that $\sum_{l=-\infty}^{\infty}|\psi_l| < \infty$, and $X_t = \sum_{-\infty}^{\infty}\psi_l W_{t-l}$[It's a well-defined as a limit in L^2, and it might depend on the future.]

Definition 17

$\{X_t\}_{t\in\mathbb{Z}}$ is a causal linear process, if

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

It only depends on W's in the "past".

Remark. Linear processes are strictly stationary (Bernoulli Shift)

Example 3

$X_t = W_t + \theta W_{t-1}$, $W_t \sim$ Strong White Noise. X_t is a linear process.

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

Note: When $h = 0$, $\gamma_X(h)$ is always non-zero. When $h = 1$, $\gamma_X(h)$ is non-zero if θ ("lagged" term coefficients) in the linear process are non-zero.
Suggests a way of slewthing out what $g(W_t, W_{t-1}, \ldots) = \sum_{l=0}^{\infty}\psi_l W_{t-l}$ must look like.

Definition 18

Suppose X_t is weakly stationary. The <u>autocorrelation function</u> of X_T (Abbrev: ACF) is

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

Note since $\gamma(0) = Var(X_t) = Var(X_0)$m

$$|\gamma(h)| = |cov(X_t, X_{t+h})| \leqslant \sqrt{Var(X_t)Var(X_{t+h})} = Var(X_0)$$

by stationary, $Var(X_t) = Var(X_{t+h}) = Var(X_0)$. Also,

$$|\rho(h)| \leqslant 1 \implies -1 \leqslant \rho(h) \leqslant 1$$

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

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

Hence a sensible estimator is

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t = \overline{X} \text{ (Sample mean/Time series avg.)}$$

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

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

Example 4

$X_t = W_t$, $W_t \sim$ Strong White Noise $Var(W_t) = \sigma_W^2 < \infty$

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

$$\implies \rho_X(h) = \begin{cases} 1, & h = 1 \longleftarrow \rho(0) = \gamma(0)/\gamma(0) = 0 \\ 0, & h \geqslant 1 \end{cases}$$

1.11 Modes of Convergence of Random Variables

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

1. Stochastic Boundedness(Op and op notation)
2. Convergence in Probability
3. Convergence in Distribution

Definition 19

Suppose $\{X_n\}_{n\geqslant 1}$ is a sequence of random variable, we say that X_n is bounded in probability by Y_n if $\forall \varepsilon > 0, \exists M, N \in \mathbb{R}$ such that $\forall n \geqslant \mathbb{N}$,

$$P\left(|X_n/Y_n|> M\right) \leqslant \varepsilon$$

Shorthand:$X_n = Op(Y_n) \implies$ "X_n is on the order of Y_n"/

Definition 20

We say X_n converges in probability to X if $\forall \varepsilon > 0$,

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

If a_n is a sequence of scalars, we abbriviate X_n/a_n converges in probability to zero as

$$X_n = op(a_n) \iff P(|X_n/a_n|> \varepsilon) \to 0, \text{ as } n \to 0, \forall \varepsilon > 0$$

Hence, X_n converges to zero in probability denoted as

$$X_n = op(1)$$

We also write $X_n \xrightarrow{P} X$ to denote X_n converges to X in probability.

Definition 21

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

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

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

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

Useful Tool: Chebyshev's Inequality: If $E[Y^2] < \infty$, then

$$\begin{aligned} E[Y^2] &= E[Y^2 \mathbb{1}_{|Y| \geqslant M} + Y^2 \mathbb{1}_{|Y| < M}] \\ &= E[Y^2 \mathbb{1}_{|Y| \geqslant M}] + E[Y^2 \mathbb{1}_{|Y| < M}] \\ &\geqslant E[Y^2 \mathbb{1}_{|Y| \geqslant M}] \\ &\geqslant M^2 E[\mathbb{1}_{|Y| \geqslant M}] \\ &= M^2 P(|Y| \geqslant M) \end{aligned}$$

which give us the Chebyshev's Inequality:

$$P(|Y| \geqslant M) \leqslant \frac{E[Y^2]}{M^2}$$

Generally when $E[|Y|^k] < \infty$, $P(|Y| \geqslant M) \leqslant \frac{E[|Y|^k]}{M^k}$

Example 5

Suppose X_n is a strong white noise in $L^2(E[X_0^2] < \infty)$, and let $\overline{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$, then

1. $|\overline{X}_T| = op(1)$
   For $\varepsilon > 0$,

$$\begin{aligned}
Var(\overline{X}_T) &= E[\overline{X}_T^2] \\
&= \frac{1}{T^2} E\left[\left(\sum_{t=1}^{T} X_t\right)^2\right] \\
&= \frac{1}{T^2}\left(\sum_{t=1}^{T}\sum_{t=1}^{T} E[X_t X_s]\right) \text{ the expectation is non-zero only then } t = s \\
&= \frac{1}{T^2}\sum_{t=1}^{T} E[X_t^2] \\
&= \frac{1}{T^2}\sum_{t=1}^{T} E[X_0^2] = \frac{\sigma^2}{T} \; (\sigma^2 = E[X_0^2])
\end{aligned}$$

   Hence we will have,

$$P(|\overline{X}_T| > \varepsilon) \leqslant \frac{E[\overline{X}_T^2]}{\varepsilon^2} = \frac{\sigma^2/T}{\varepsilon^2} \to 0$$

   as $T \to \infty$.
   Hence, $\overline{X}_T = op(1)$

2. $\overline{X}_T = Op(\frac{1}{\sqrt{T}})$,

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

   so by Chebyshev's, for $M > 0$,

$$P(|\sqrt{T}\overline{X}_T| > M) \leqslant \frac{Var(\sqrt{T}\overline{X}_T)}{M^2} = \frac{\sigma^2}{M^2} \to 0, \text{ as } M \to \infty$$

   Note: if we look at the definition, we should know the equation above shall work for any T large enough, so if we keep T in the equation, it cannot work.
   Hence, $\sqrt{T}\overline{X}_T = Op(1) \Rightarrow \overline{X}_T = Op(\frac{1}{\sqrt{T}})$.

Alternatively, we can show this using the Central Limit Theorem by the CLT $\sqrt{T}\overline{X}_T \xrightarrow{D} N(0, \sigma^2)$. Therefore, if $F_T \sim$CDF of $\sqrt{T}\overline{X}_T$, $\Phi \sim$CDF of $N(0,1)$ random variable.

$$|F_T(x) - \Phi(x/\sigma)| \to 0, \text{ as } T \to \infty, \forall x \in \mathbb{R}$$

For $\varepsilon > 0$, choose M such that $\Phi(-\frac{M}{\sigma}) = 1 - \Phi(M/\sigma) \leqslant \frac{\varepsilon}{4}$. For this M, choose T_0, so $T \geqslant T_0 \Rightarrow$ $|F_T(-M) - \Phi(-M/\sigma)| \leqslant \varepsilon/4$ and $|F_T(M) - \Phi(M/\sigma)| \leqslant \varepsilon/4$. Then,

$$\begin{aligned} P(|\sqrt{T}\overline{X}_T \geqslant M) &= F_T(-M) + (1 - F_T(M)) \\ &= \Phi(-M/\sigma) + (1 - \Phi(M/\sigma)) + F_T(-M) + -\Phi(-M/\sigma) + \Phi(M/\sigma) - F_T(M) \\ &\leqslant \varepsilon/4 + \varepsilon/4 + \varepsilon/4 + \varepsilon/4 \\ &= \varepsilon \end{aligned}$$

Remark. In general,

$$\frac{X_n}{a_n} \xrightarrow{D} \text{Non-degenerate R.V.} \Rightarrow X_n = Op(a_n)$$

Remark. Algebra of Op and op notation.

1. $X_n = Op(a_n), Y_n = Op(b_n) \Rightarrow X_n + Y_n = Op(\max\{a_n, b_n\})$
2. $X_n = op(1), Y_n = op(1), X_n + Y_n = op(1)$
3. $X_n = op(1), Y_n = op(1), X_n * Y_n = op(1)$

Example 6

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

Proof.

$$\overline{X}_T = \overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t = \frac{1}{T}\sum_{t=1}^{t}(W_t + \theta W_{t-1}) = \frac{1}{T}\sum_{t=1}^{T} W_t + \frac{\theta}{T}\sum_{t=1}^{T} W_{t-1} = op(1)$$

$$\begin{aligned} \hat{\gamma}(1) &= \frac{1}{T}\sum_{t=1}^{T-1}(X_t - \overline{X})(X_{t+1} - \overline{X}) \\ &= \frac{1}{T}\sum_{t=1}^{T-1} X_t X_{t+1} + \frac{T-1}{T}\overline{X}^2 - \overline{X}\frac{1}{T}\sum_{t=1}^{T-1} X_t - \overline{X}\frac{1}{T}\sum_{t=1}^{T-1} X_{t+1} \\ &= \frac{1}{T}\sum_{t=1}^{T-1} X_t X_{t+1} + R_{1,T} + R_{2,T} + R_{3,T} \end{aligned}$$

Notice that, $R_{i,T} = op(1), i = 1, 2, 3$

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

Now, $\frac{1}{T}\sum_{t=1}^{T}\theta W_t^2 \overset{SLLN}{\longrightarrow} \theta E[W_t^2] = \theta\sigma_W^2$ We take a look at $G_{1,T}$,

$$G_{1,T} = \frac{1}{T}\sum_{t=1}^{T} W_t W_{t+1},\ E[G_{1,T}] = \frac{1}{T}\sum_{t=1}^{T}\underbrace{E[W_t W_{t+1}]}_{=0}$$

$$\begin{aligned}
Var(G_{1,T}) = E[G_{1,T}^2] &= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\underbrace{E[W_t W_{t+1} W_s W_{s+1}]}_{<\infty;\neq 0 \text{ only if } s=t} \\
&= \frac{1}{T^2}\sum_{t=1}^{T} E[W_t^2 W_{t+1}^2] \\
&= \frac{T}{T^2}\sigma_W^2 \to 0 \text{ as } T \to \infty
\end{aligned}$$

By Chebyshev's Inequality, $G_{1,T} = op(1)$ (Similar steps for $G_{2,T}, G_{3,T}$). Then we can write

$$\hat{\gamma}(1) = \frac{1}{T}\sum_{t=1}^{T}\theta W_t^2 + \sum op(1)$$

Hence we have

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

$\square$

1.12 M-dependent CLT (Optional)

Suppose X_t is a mean zero, strictly stationary time series ($E[X_t^2 < \infty]$). Note we didn't assume X_t are iid. We frequently faces with the problem:

1. What is the approximate distribution of

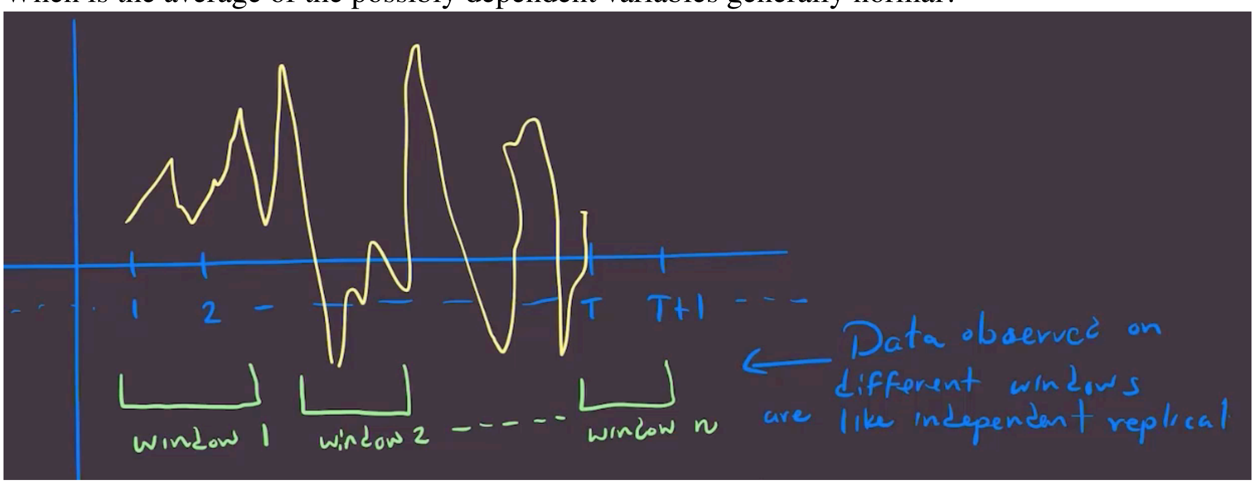
$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t = \sqrt{T}\overline{X}_T \overset{D}{\approx} N(0, \sigma_x^2)?$$

2. If X_t is a strong white noise. What's the approximately distribution of

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

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

When is the average of the possibly dependent variables generally normal?



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

Definition 22

We say a time series $\{X_t\}_{t\in\mathbb{Z}}$ is m-dependent for $m \in \mathbb{Z}_+$, if for all $t_1 < t_2 \ldots < t_{d_1} < s_1 < s_2 < \ldots < s_{d_2} \in \mathbb{Z}$ such that $t_{d_1} + m \leqslant s_1$ and

$$(X_{t_1}, \ldots, X_{t_{d_1}}) \text{ is independe of } (X_{s_1}, \ldots, X_{S_{d_s}})$$

it means two windows separated by (at least) m units are independent.

Example 7

$X_t = W_t + \theta W_{t-1}$ where W_t is a strong white noise is 2-dependent.

Theorem 2

Suppose X_t is a strictly stationary, and m-dependent time series with $E[X_t] = 0, E[X_t^2] < \infty$. Then

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

where

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

This is a generalization of the standard CLT to m-dependence.

Definition 23

Preliminaries: We say $\{X_{i,j}, 1 \leqslant j \leqslant n_i, 1 \leqslant i \leqslant \infty\}$ forms a triangular array of mean zero L^2 random variables, if $E[X_{i,j}] = 0, E[X_{i,j}^2] < \infty$, for each i-fixed $X_{i,1}, \ldots, X_{i,n_i}$ are independent, and $n_i < n_{i+1}$

$$\begin{aligned}
&X_{1,1}, \ldots, X_{1,n_1} \\
&X_{1,1}, \ldots, \ldots, X_{2,n_2} \longleftarrow \text{Row-wise random variables are independent} \\
&\vdots, \ldots, \ldots, \ldots, \ldots, \ddots
\end{aligned}$$

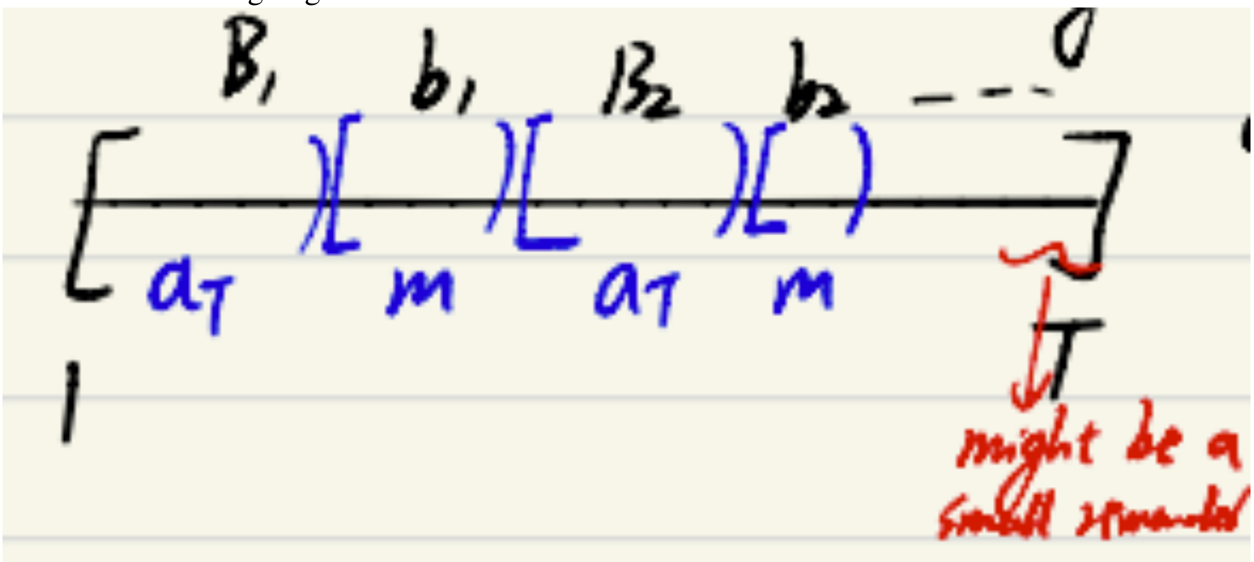
Theorem 3: Lindeberg-Feller CLT for triangular array

et $\{X_{i,j}, 1 \leqslant j \leqslant n_i, 1 \leqslant i \leqslant \infty\}$ be a triangular array of mean zero L^2-rvs. Define $\sigma_i^2 = \sum_{j=1}^{n_i} Var(X_{i,j})$ and $S_i = \frac{1}{\sigma_i} \sum_{j=1}^{n_i} X_{i,j}$ (Row-wis sum standardized).
(Lindeberg's Condition) If for $\varepsilon > 0$,

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

Then $S_i \xrightarrow{D} N(0, 1)$
The indicator in the condition is looking for the variable that contributes a non-negligible varaince. The whole summation is calculating the percentage of the variance that are contributed by those variables with significant variance. Sometimes it's called a uniform asymptotic negligible condition, it's saying that all of the random variable are negligible in the sense none of them contribute significantly to the variance.

Proof. of M-dependent CLT
"Bernstein Blocking Argument"



$$a_T = \text{Big Block Size},\ m = \text{little block size}$$

Assume $a_T \to \infty$ as $T \to \infty$, $\frac{a_T}{T} \to 0$.

$$N = \text{number of blocks} = \left\lfloor \frac{T}{m + a_T} \right\rfloor$$
$$B_j = \{i : (j-1)(a_T + m) + 1 \leqslant i \leqslant j a_T + (j-1)m\}$$
$$b_j = \{i : j a_T + (j-1)m + 1 \leqslant i \leqslant j(a_T + m)\}$$

Since $a_T \nwarrow \infty$, for T sufficiently large, $a_T > m$ and so by m-dependence, $\sum_{t\in B_j} X_t$ is independent of $\sum_{t\in B_k} X_t (j \neq k)$. Similar for $b_j, b_k, j \neq k$.

$$\frac{1}{\sqrt{T}} = \frac{1}{\sqrt{T}} \sum_{j=1}^{N} \sum_{t\in B_j} X_t = \frac{1}{\sqrt{T}} \sim_{j=1}^{N} \sum_{t\in b_j} + \text{Remainder}$$
$$= G_{1,T} + G_{2,T} + G_{3,T}$$
$$Var(G_{2,T}) = \frac{1}{T} \sum_{j=1}^{N} E\left[(\sum_{t\in b_j} X_t)^2 \right] \underbrace{=}_{\text{strict stationary}} \frac{N}{T} E\left[(\sum_{t=1}^{m} X_t)^2 \right]$$
$$E\left[(\sum_{t=1}^{m} X_t)^2 \right] = \sum_{t=1}^{m} \sum_{s=1}^{m} E\left[X_t X_s\right] = \sum_{t=1}^{m} \sum_{s=1}^{m} \gamma(|t-s|) = \sum_{h=1-m}^{m-1} (m - |h|)\gamma(h) < \infty$$

$$\implies Var(G_{2,T}) = \frac{N}{T} * \text{constant} = \left\lfloor \frac{T}{a_T + m} \right\rfloor / T * \text{constant} \to 0\ [a_T \to \infty]$$

Hence, as $T \to \infty$, $a_T \to \infty$, we will have $G_{2,T} = op(1)$ by Chebyshev's Inequality.
Notice $G_{1,T} = \frac{1}{\sqrt{T}} \sum_{j=1}^{N} \sum_{t\in B_j} X_t = \sum_{j=1}^{N} \frac{\sum_{t\in B_j} X_t}{\sqrt{T}}$, and we let $Y_{j,T} = \frac{\sum_{t\in B_j} X_t}{\sqrt{T}}$ (this variable

forms a triangular array, imagining each row shares the same T)

$$
\begin{aligned}
Var(G_{1,T}) &= \sum_{j=1}^{N} Var(Y_{j,T}) \\
Var(Y_{j,T}) &= Var(Y_{1,T}) \\
&= \frac{1}{T} E\left[\left(\sum_{t=1}^{a_T} X_i\right)^2\right] \\
&= \frac{1}{T} \sum_{t=1}^{a_T} \sum_{s=1}^{a_T} E[X_t X_s] \\
&= \frac{1}{T} \sum_{h=1-a_T}^{a_T-1} (a_T - |h|)\gamma(h) \\
&= \frac{1}{T} \sum_{h=-m}^{h=m} (a_T - |h|)\gamma(h) \text{ if } |h| \geqslant m \text{, then } \gamma(h) = 0 \text{ by m-independence}
\end{aligned}
$$

$$
\implies Var(G_{1,T}) = \frac{N}{T} \sum_{h=-m}^{m} (a_T - |h|)\gamma(h) \approx \frac{1}{a_T} \sum_{h=-m}^{m} (a_T - |h|)\gamma(h) \xrightarrow[T\to\infty]{} \sum_{h=-m}^{m} \gamma(h)
$$

Hence we know the variance of $G_{1,T}$ is bounded.
Check Lindeberg's Condition: $\sigma_N^2 = Var(G_{1,T}) \approx$ const, so we must show:

$$
\begin{aligned}
&\sum_{j=1}^{N} E\left[\underbrace{Y_{j,T}^2}_{iid} \mathbb{1}_{\{|Y_{j,T}|>\varepsilon\sigma_N\}}\right] \\
=&N * E\left[Y_{j,T}^2 \mathbb{1}_{\{|Y_{j,T}|>\varepsilon\sigma_N\}}\right] \to 0 \text{ as } T \to \infty
\end{aligned}
$$

Aside $E[|Y|^{2+\delta}] \underset{\delta>0}{\geqslant} E[|Y|^{2+\delta}\mathbb{1}_{\{|Y|>\varepsilon\}}] \geqslant \varepsilon^\delta E[|Y|^2 \mathbb{1}_{\{|Y|>\varepsilon\}}]$, so we have

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

It may be shown that $E[|Y_{j,T}^{2+\delta}] \leqslant \text{const}\left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}}$, so

$$
\begin{aligned}
N * E\left[Y_{j,T}^2 \mathbb{1}_{\{|Y_{j,T}|>\varepsilon\sigma_N\}}\right] &\leqslant \frac{N}{(\varepsilon\sigma_N)^\delta} \text{const} \left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}} \\
&= \frac{\text{const}}{(\varepsilon\sigma_N)^\delta} \frac{N a_T}{T} \left(\frac{a_T}{T}\right)^{\frac{\delta}{2}} \to 0 (T \to \infty)
\end{aligned}
$$

This implies $\frac{G_{1,T}}{\sigma_N} \xrightarrow{D} N(0,1)$, and since $\sigma_N^2 \to \sum_{h=-m}^{m} \gamma(j)$, we have

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

Since, at the beginning, we've shown that $G_{2,T} = op(1)$, so we have

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

as required. $\square$

1.13 $2+\delta$ Moment Calculation

We want to show that

$$E[|Y_{1,t}|^{2+\delta}] \leqslant \text{constant } \left(\frac{a_T}{T}\right)^{\frac{2+\delta}{2}}$$

, where $Y_{1,T} = \frac{1}{\sqrt{T}}\sum_{t=1}^{a_T} X_t$, and a_T = Big Block Size $\rightarrow \infty$, $(T \rightarrow \infty)$, $\frac{a_T}{T} \rightarrow 0$. X_t are m-denpendent random variables. Want

$$E[|X_i|^{2+\delta}] < \infty(\delta > 0) \Leftrightarrow \text{const} a_T^{\frac{2+\delta}{2}}$$

Tools: Rosenthal's Inequality. If $X_1, \ldots, X_n$ are independent RV's with $E[|X_i|^{2+\delta}] < \infty(\delta > 0)$, then

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

In particular, if $X_1, \ldots, X_n$ are iid, then

$$E[|\sum_{i=1}^n X_i|^{2+\delta}] \leqslant c_p n^{(2+\delta)/2} E[|X_1|]^{2+\delta}$$

For proof: see Petrov, Limit theorems of probability theory, P59.
Tool: For arbitrary RV's $X_1, \ldots, X_n$,

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

proof: By Jensen's Inequality, for all real numbers $a_1, \ldots, a_n$

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

$$\Longrightarrow |\sum_{i=1}^n a_i|^{2+\delta} \leqslant n^{(2+\delta)-1}\sum_{i=1}^n |a_i|^{2+\delta}$$

Replace a_i with X_i, take expectation.

Proof.

$$\sum_{t=1}^{a_T} X_t = \sum_{j=0}^{m} \sum_{\substack{\forall k \text{ mod } m=j, t=k+1 \\ 1\leqslant t \leqslant a_T}} X_t$$

so $\sum_{\substack{\forall k \text{ mod } m=j, t=k+1 \\ 1\leqslant t \leqslant a_T}} X_t$, variables in this sum separated by at least m-time steps, and are hence iid. So we got,

$$
\begin{aligned}
E\left[\left|\sum_{t=1}^{a_T} X_t\right|^{2+\delta}\right] &\leqslant (m+1)^{(2+\delta)-1} \sum_{j=0}^{m} E\left[\left|\sum_{\substack{\forall k \bmod m=j, t=k+1 \\ 1\leqslant t\leqslant a_T}} X_t\right|^{2+\delta}\right] \\
&\leqslant (m+1)^{(2+\delta)-1} \sum_{j=0}^{m} \left(\frac{a_T}{m+1}\right)^{\frac{2+\delta}{2}} c_p E[|X_1|]^{2+\delta} \\
&= (m+1)^{(2+\delta)-1} m \left(\frac{a_T}{m+1}\right)^{\frac{2+\delta}{2}} c_p E[|X_1|]^{2+\delta} \\
&= \text{const } * {a_T}^{\frac{2+\delta}{2}}
\end{aligned}
$$

□

1.14 Linear Process CLT

If $X_t \sim$ m-dependent, strictly stationary, $E[X_t] = 0, E[X_t^2] < \infty$, then

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

EX:$X_t = \sum_{t=0}^{m} \psi_l W_{t-l}$, whhere $\{w_t\}_{t\in\mathbb{Z}}$ is a strong White noise in L^2.
A general linear process

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

is <u>not</u> m-dependent, because it depends on the white noise arbitrarily back to the past.

Theorem 4: Basic Approximation Theorem BAT

Suppose X_n is a sequence of random variables so that there exists an array $\{Y_{m,n}, m, n \geqslant 1\}$,

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

Then $X_n \xrightarrow{D} Y$ as $n \to \infty$.
Normally, $Y_{m,n}$ is often an "m-dependent approximation to X_n. Proof is in Shumway and Stoffer.

Theorem 5: Linear Process CLT

Suppose

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

is a causal linear process with $\sum_{l=0}^{\infty} |\psi_l|< \infty$, $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise in L^2. Then if $S_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t$,

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

,where the variance of the S_T is the "long-run variance" of X_t

X_t is strictly (and weakly) stationary.

$$\begin{aligned}\gamma(h) = E[X_t X_{t+h}] &= E\left[\left(\sum_{l=0}^{\infty}\psi_l W_{t-l}\right)\left(\sum_{j=0}^{\infty}\psi_j W_{t+h-j}\right)\right] \\ \text{Fubin's Theorem} &= \sum_{l=0}^{\infty}\sum_{j=0}^{\infty}\psi_l\psi_j \underbrace{E[W_{t-l}W_{t+h-j}]}_{\neq 0,\text{ if } j=l+h} \\ &= \sum_{l=0}^{\infty}\psi_l\psi_{l+h}\sigma_W^2\end{aligned}$$

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

so $\sum_{h=-\infty}^{\infty}\gamma(h)$ is well-defined.

$$\begin{aligned}E[S_T] &= E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_t\right) = 0 \ (E[X_t] = 0) \\ Var(S_T) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}E[X_tX_s] = \frac{1}{T}\sum_{h=1-T}^{T-1}(T-|h|)\gamma(h) \\ &= \sum_{h=1-T}^{T-1}\left(1-\frac{|h|}{T}\right)\gamma(h) \\ &\xrightarrow{\text{by Dominated Convergence}} \sum_{h=-\infty}^{\infty}\gamma(h)\end{aligned}$$

Note: $\left(1-\frac{|h|}{T}\right)\gamma(h) \leqslant \underbrace{|\gamma(h)|}_{\text{summable}}$

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

1. By the m-dependent CLT

$$S_{T,m} \xrightarrow{D} N(0, \sum_{h=-m}^{m}\gamma_m(h)) =: S'_m,\ \gamma_m(h) = E[X_{t,m}X_{t+h,m}]$$

2. By Dominated Convergence $\sum_{h=-m}^{m}\gamma_m(h) \underset{m\to\infty}{\longrightarrow} \sum_{h=-\infty}^{\infty}\gamma(h)$, and hence

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

3.

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

so condition (3) of the BAT is satisfied using Chebyshev's Inequality. Hence

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

□

1.15 Aymptotic Properties of Empirical ACF

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

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}\left(X_t - \overline{X}\right)\left(X_{t+h} - \overline{X}\right)$$
$$\rho(h) = Corr(X_t, X_{t+H}) = \frac{\gamma(h)}{\gamma(0)}, \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

Questions:

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

Answer:

1. Consistency: By adding and subtracting μ in the difinition of $\hat{\gamma}(h)$, we may assume WLOG that $E[X_t] = 0$.
   Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is strictly stationary, and
   $$X_t = g(W_t, W_{t-1}, \dots,)$$
   which is a Bernoulli shift.
   Then
   $$\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{P} 0$$
   by the ergodic theorem (X_t is Ergodic).
   Further more
   $$\begin{aligned}\hat{\gamma}(h) &= \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})(X_{t+h} - \overline{X})\\ &= \underbrace{\frac{1}{T}\sum_{t=1}^{T-h} X_t X_{t+h}}_{\text{Dominant term}} - \underbrace{\frac{\overline{X}}{T}\sum_{t=1}^{T-h} X_t}_{\xrightarrow{P} 0} - \underbrace{\frac{\overline{X}}{T}\sum_{t=1}^{T} X_{t+h}}_{\xrightarrow{P} 0} + \underbrace{\frac{T-h}{T}\overline{X}^2}_{\xrightarrow{P} 0}\end{aligned}$$
   Note: $E[X_t X_{t+h}] = \gamma(h),\ X_t X_{t+h} = g_h(W_{t+h}, W_{t+h-1}, \dots,)$ (Still Ergodic). Again by the Ergodic Theorem:
   $$\frac{1}{T}\sum_{t=1}^{T-h} X_t X_{t+h} \xrightarrow{P} \gamma(h)$$
   which gives us
   $$\hat{\gamma}(h) \xrightarrow{P} \gamma(h),\ \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \xrightarrow{P} \rho(h)$$

under strict stationarity and $E[X_t^2] < \infty$.

2. Distribution of $\hat{\gamma}(h)$: Consider simple (but perhaps most important) case: X_t is a strong white noise. $E[X_t^4] < \infty$
   Finite 4^{th} moment assumption is not really needed here but I will explain why it is classically assumed.
   $$\hat{\gamma}(h) \xrightarrow{P} 0 \text{ in this case by strong white noise}$$
   Similarly as before
   $$\hat{\gamma}(h) = \underbrace{\frac{1}{T}\sum_{t=1}^{T-h} X_t X_{t+h}}_{\tilde{\gamma}(h)} + \text{smaller terms}$$
   Hence,
   $$\begin{aligned}
   E[\tilde{\gamma}(h)) &= \frac{1}{T}\sum_{t=1}^{T-h} E[X_t X_{t+h}] = 0 (h \geqslant 1) \\
   Var(\tilde{\gamma}(h)) &= E[\tilde{\gamma}^2(h)] \\
   &= \frac{1}{T^2}\sum_{t=1}^{T-h}\sum_{s=1}^{T-h} \underbrace{E[X_t X_{t+h} X_s X_{s+h}]}_{\neq 0 \leftrightarrow t=s} \\
   &= \frac{1}{T^2}\sum_{t=1}^{T-h} E[X_t^2 X_{t+h}^2] \\
   &= \frac{T-h}{T^2}\sigma_x^4 \; (E[X_t^2] = \sigma_X^2)
   \end{aligned}$$
   Therefore,
   $$var(\sqrt{T}\tilde{\gamma}(h)) \underset{T\to\infty}{\longrightarrow} \sigma_X^4$$

Theorem 6

If X_t is a strong white noise with $E[X_t^4] < \infty$,

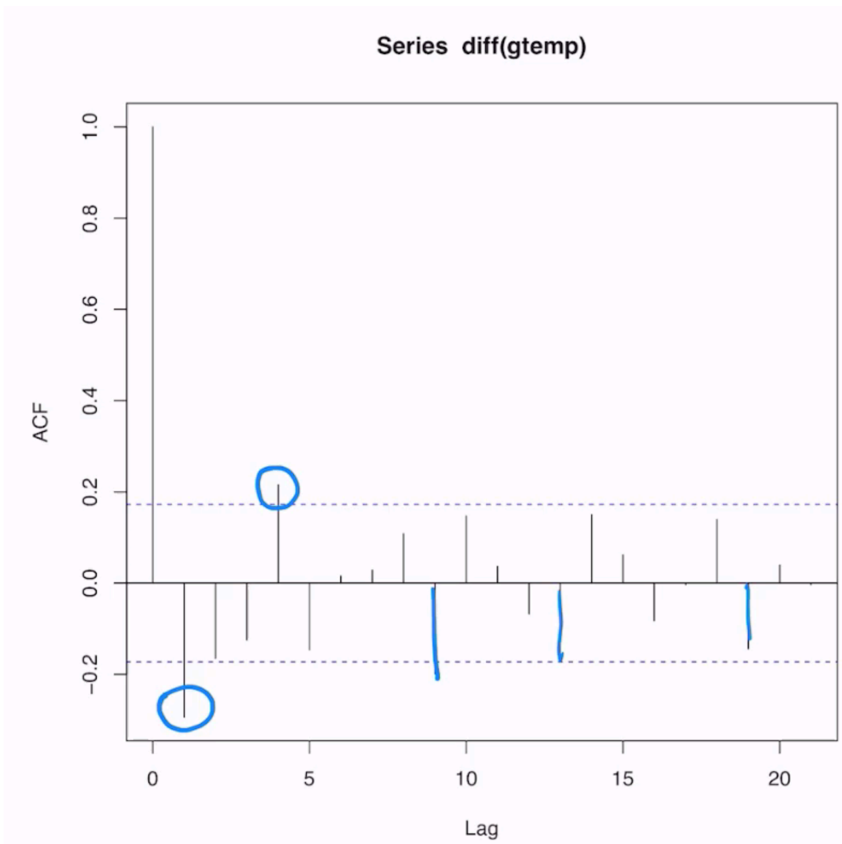
$$\sqrt{T}\tilde{\gamma}(h) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T-h} \underbrace{X_t X_{t+h}}_{\text{Not iid}} \xrightarrow{D} N(0, \sigma_X^4)$$

The convergence can be obtained by M(h+1)-dependent CLT and Martingale CLT.

It follows that
$$\sqrt{T}\hat{\gamma}(h) \xrightarrow{D} N(0, \sigma_X^4)$$
Since $\hat{\gamma}(0) \xrightarrow{P} \sigma_X^2$, by Slutsky's Theorem,
$$\sqrt{T}\frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \sqrt{T}\hat{\rho}(h) \xrightarrow{D} N(0,1)$$

Useful Tool: If X_t 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 (T large), where $\Phi(Z_\alpha) = 1-\alpha$. 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.
Hence, if the data is a strong white noise, for the most of time the ACF should lie in this interval. Also, since our empirical autocorrelation is consistent, we know if the true autocorrelation is non-zero, for T large enough, the empirical autocorrelation will be outside of this interval.

Example 8



Indicative of mild serial correlation at lag 1

1.16 Interpreting the ACF

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

$$\hat{\rho}(h) \xrightarrow{P} 0 \ (h \geqslant 1) \qquad\qquad \hat{\rho}(h) \overset{D}{\approx} N\left(0, \frac{1}{T}\right) \ (T \text{ is large})$$

What happens when we calculate the Empirical ACF for non-stationary data?

Example 9

$X_t = t + W_t$ ($W_t \sim \ S.W.N.$), we can see that X_t has a linear trend.

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

It follows in this case that

$$\frac{\hat{\gamma}(h)}{T^2} \to \text{Const for all } h \ (T \to \infty)$$

Hence,

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

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

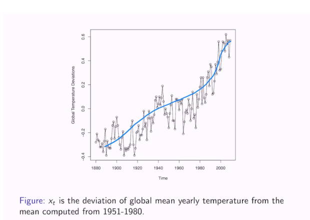
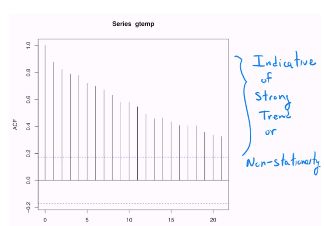
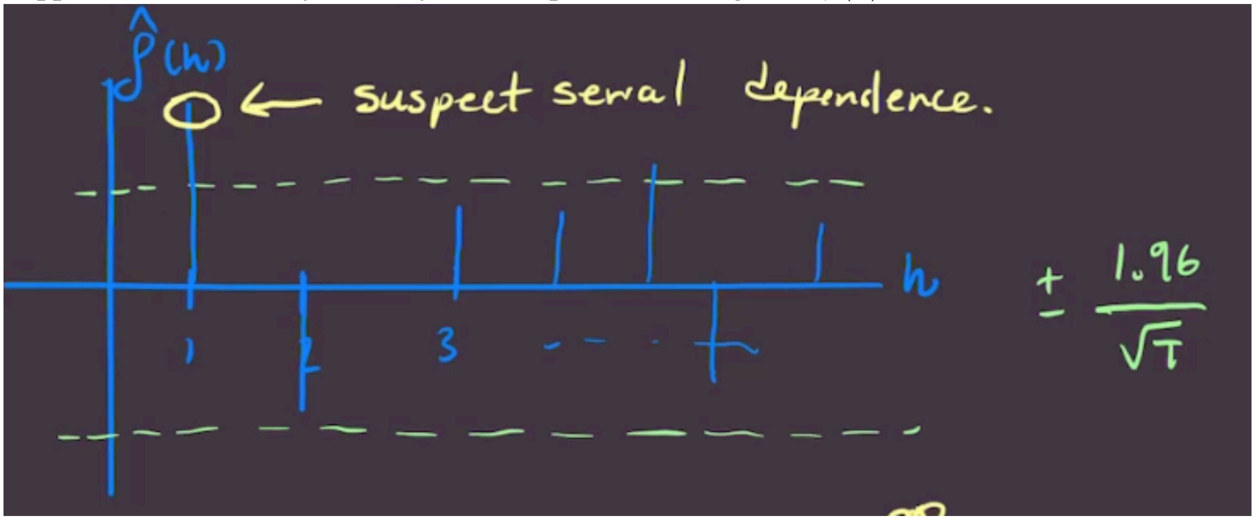


Figure: x_t is the deviation of global mean yearly temperature from the mean computed from 1951-1980.



1.17 Moving Average Processes

Suppose X_t is stationary. Identify serial dependence using ACF $\hat{\rho}(h)$



Posit $X_t = g(W_t, W_{t-1]}, \ldots) = \sum_{l=0}^{\infty} \psi_l W_{t-l}$ [Linear Process].
Not feasible to estimate infinitely many parameters $\{\psi_l\}_{l=0}^{\infty}$
Assume coefficients arise from a parsomonious linear model for X_t

Definition 24

Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise with $Var(W_t)\sigma_W^2 < \infty$. We say X_t is a **Moving Average Process of order** q (Abbrev. $MA(q)$) if there exists coefficient $\theta_1, \ldots, \theta_q \in \mathbb{R}, \theta_q \neq 0$, so that

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

which is a truncated linear process for order q

Definition 25

The Backshift operator, B, is defined by

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

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

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

Example

$$\nabla X_t = \text{first diff. of } X_t = (1 - B)X_t$$

Definition 26

We sat $\theta(x) = 1 + \theta_1 x + \ldots, \theta_q X^q$ is the Moving Average Polynomial. If $X_t \sim MA(q)$,

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

which is succinct expression defining $MA(q)$

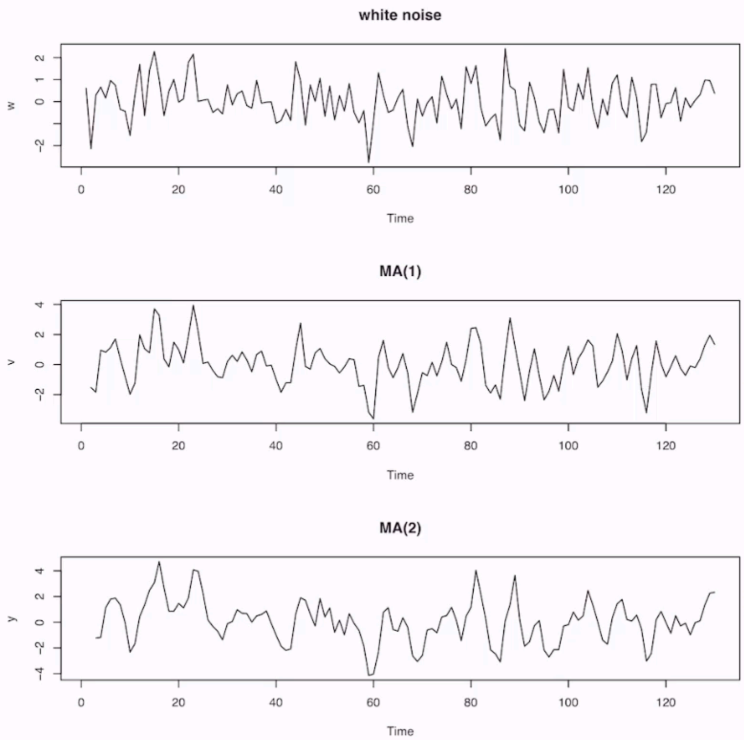
Properties of $MA(q)$ Processes:

1. $MA(q)$ process= Strong White Noise.
2. If $X_t \sim MA(q)$, then

$$\begin{aligned}
E[X_t &= E\left[\sum_{k=0}^{q} \theta_l W_{t-l}\right] = 0 \\
Var(X_t) &= E\left[\left(\sum_{l=0}^{q} \theta_l W_{t-l}\right)^2\right] = \sum_{l=0}^{q} \theta_l^2 \sigma_W^2 \\
\gamma(h) = Cov(X_t, X_{t+h}) &= E\left[\left(\sum_{l=0}^{1} \theta_l W_{t-l}\right)\left(\sum_{k=0}^{q} \theta_k W_{t+h-k}\right)\right] \\
&= \begin{cases} \sum_{j=0}^{q-|h|} \theta_j \theta_{j+h} \sigma_W^2, & 0 \leqslant h \leqslant q \\ 0, & h > q \end{cases} \\
\rho(h) = \frac{\gamma(h)}{\gamma(0)} &= \begin{cases} \frac{\sum_{j=0}^{q-|h|} \theta_j \theta_{j+h}}{\sum_{j=0}^{q} \theta_j^2}, & 0 \leqslant h \leqslant q \\ 0 & h \geqslant q+1 \end{cases}
\end{aligned}$$

Note: By choose $\theta_1, \ldots, \theta_q$ appropriately, we can get any ACF we want, $\rho(h), 1 \leqslant h \leqslant q$

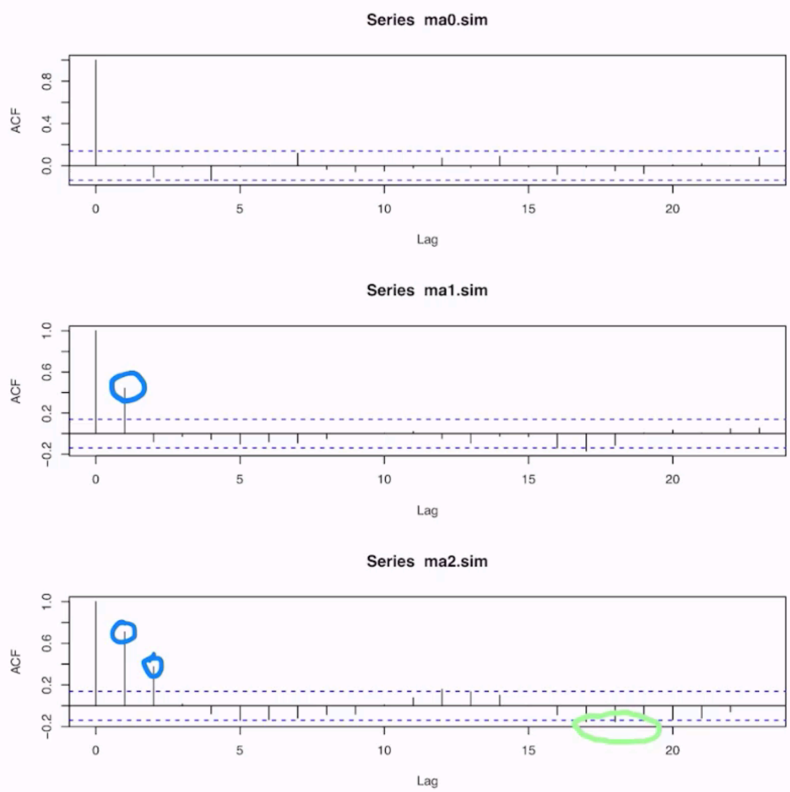
3. $X_t \sim MA(q) \implies X_t$ is q-dependent



$X_t = w_t$

$X_t = w_t + w_{t-1}$
$(\theta_1 = 1)$

$X_t = w_t + w_{t-1} + w_{t-2}$
$(\theta_1 = \theta_2 = 1)$



MA(1)

MA(2)

1.18 Autoregressive Processes

Definition 27

Suppose $\{W_t\}_{t\in\mathbb{Z}}$ is a strong white noise with $Var(W_t) < \infty$. We say X_t is an Autoregressive Process of order 1 (Abbrv. $AR(1)$) if there exists a constant ϕ so that

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

Using Backshift operator, this may also be expressed as

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

Interpretation:

- Prediction: Form a linear model (Regression) for 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 covariate/independent variable.
- Markovian Property:
$$X_t | X_{t-1}, X_{t-2}, \ldots = X_t | X_{t-1}$$

Question: Does there exist a stationary process X_t satisfying

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

$$\begin{aligned}
X_t &= \phi X_{t-1} + W_t, z \in \mathbb{Z} \\
&= \phi(\phi X_{t-2} + W_{t-1}) + W_t = \phi^2 X_{t-2} + \phi W_{t-1} + W_t \\
&\vdots \\
&= \phi^k X_{t-k} + \sum_{j=0}^{K-1} \phi^j W_{t-j}
\end{aligned}$$

So ,if $|\phi| > 1$, X_t blows-up. Suppose $|\phi| < 1$,, we have

$$\xrightarrow{L^2 \text{ sense}} 0 + \sum_{j=0}^{\infty} \phi^j W_{t-j} \leftarrow \text{ Causal Linear Process}$$

Moreover, if $X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}$, X_t is strictly stationary, and

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

X_t satisfies $AR(1)$ equation

Theorem 7

If $|\phi| < 1$, then there exists a strictly stationary and Causal Linear Process X_t so that

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

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

$$\begin{aligned} X_t &= X_{t+1}/\phi - W_{t+1}/\phi \\ &= \vdots \\ &= X_{t+k}/\phi^k - \sum_{j=1}^{k} \frac{W_{t+j}}{\phi^j} \\ &\xrightarrow{L^2-\text{sense}} -\sum_{j=1}^{\infty} \frac{W_{t+j}}{\phi^j} \end{aligned}$$

This sequence is strictly stationary! (Bernoulli-Shift). It depends on the future. Normally we try to avoid this.
What if $|\phi| = 1$?
In this case there is no stationary process X_t so that

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

Proof. $\phi = 1$. If $X_t = X_{t-1} + W_t$, then suppose it's stationary

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

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

$$Var(X_t - X_0) = Var(X_t) + Var(X_0) - 2cov(X_t, X_0) \leqslant 4Var(X_0)$$

$$Var(\sum_{j=1}^{t} W_j) = t\sigma_W^2 \to \infty, \text{as } t \to \infty$$

Contradiction. □

Properties of Causal $AR(1)$ [$|\phi| < 1$].

1. The span of dependence of X_t is "infinite"

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

2. ACF.

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

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

Hence

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

[Note: this decays geometrically in the lag parameter]

Definition 28

We say X_t follows an autoregressive process of order p (Abbrv. $AR(p)$) if there exists coefficients $\phi_1, \ldots, \phi_p \in \mathbb{R}$ ($\phi_p \neq 0$) so that

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

We define

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

to be the Autoregressive Polynomial. $X_t \sim AR(p)$, if

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

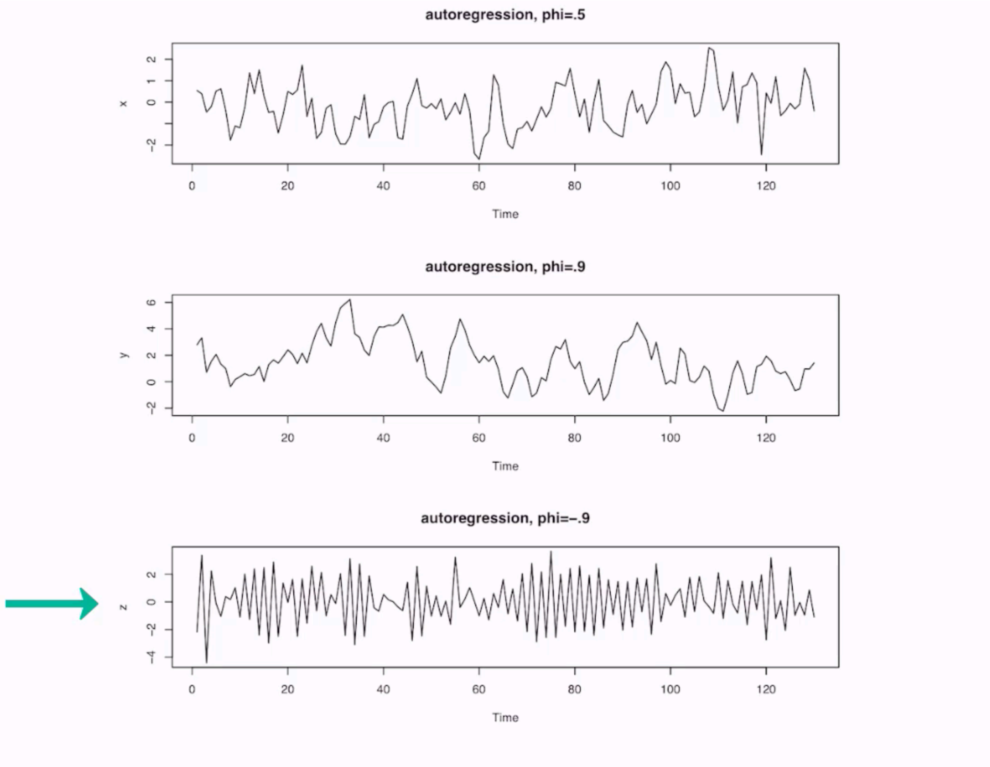


Figure: Realizations of AR(1) processes

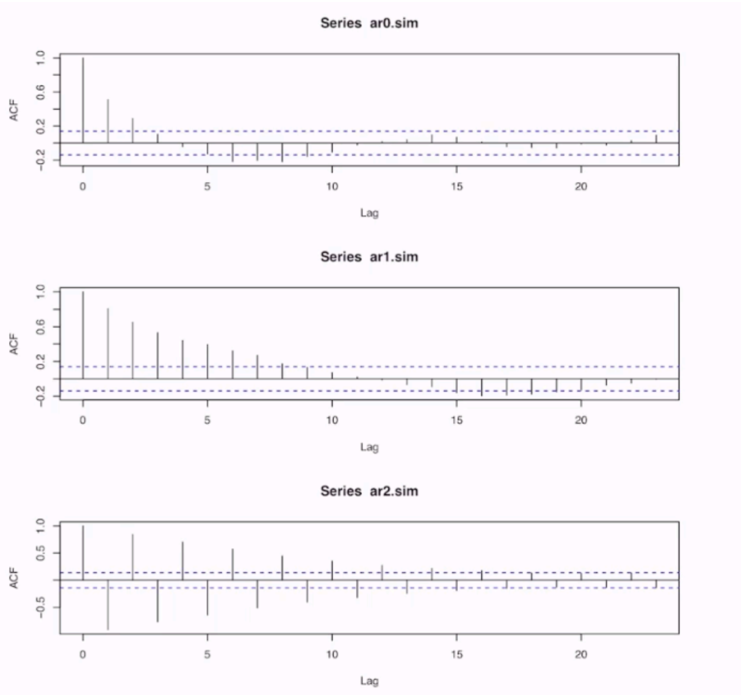


Figure: Corresponding ACF plots

1.19 Autoregressive Moving Average Processes

Moving Average Poly.

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

Autoregressive Poly.

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

If W_t ~Strong white noise,

$$X_t = \theta(B)W_t\ (X_t \sim MA(p))$$
$$\phi(B)X_t = W_t\ (X_t \sim AR(p))$$

Why not combine the two!!!

Definition 29

Given a strong white noise sequence W_t, we say that X_t is an Autoregressive Moving Average Process of orders $p\&q$ (Abbrv, $ARMA(p,q)$), if

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

where

$$\phi(x) = 1 - \phi_1 x - \ldots - \phi_p x^p\ (\phi_p \neq 0)$$
$$\theta(x) = 1 + \theta_1 x + \ldots + \theta_1 x^1,\ (\theta_q \neq 0)$$

This implies the model

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

Using ARMA models to model Autocorrelation:
$MA(q)$:ACF may be specified at lags $1, \ldots, q$
$AR(p)$: ACF has geometric decay/oscillations
$ARMA$ combine the two

Remark. Parameter Redundancy Consider $X_t = W_t\ (X_t \sim MA(0))$, then $0.5X_{t-1} = 0.5W_{t-1}$

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

where

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

Note if we observe the $ARMA$ above, we know we can degrade it to a $MA(0)$ model as above. Parameter redundancy manifests as shared zeros in the $\phi\&\theta$. We always assume models are "reduced" by factoring and dividing away common zeros in $\phi(z)$ and $\theta(z)$.

Definition 30

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

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

which is a Causal Linear Process Solution

Definition 31

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

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

W_t can be expressed as a linear function of X_t

Causality+Invertibility $\implies$ Information in $\{X_t\}_{t\leqslant T}$ is the same as Information in $\{W_t\}_{t\leqslant T}$

Theorem 8: Causality

By the fundamental theorem of algebra, the autoregressive polynomial $\phi(z)$ has p roots, say $z_1, \ldots, z_p \in \mathbb{C}$ (Complex Plane).
If $\rho = \min_{1\leqslant j\leqslant p}|z_j|> 1$, then there exists a stationary and causal X_t to the $ARMA$ equations: $\phi(B)X_t = \theta(B)W_t$, $X_t = \sum_{l=0}^{\infty} \psi_l W_{t-l}$.
The coefficients $\{\psi_l\}_{l=0}^{\infty}$ satisfy $\sum_{l=0}^{\infty}|\psi_l|< \infty$ [In fact: $|\psi_l|\leqslant \frac{1}{\rho^l} \leftarrow$ Geometric Decay]. And

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

In essence, $X_t = \frac{\theta(B)}{\phi(B)} W_t = \sum_{j=0}^{\infty} \psi_j B^j W_t$
Key: $\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j,\ |z|\leqslant 1$ ($\frac{1}{\phi}$ has a convergent power series representation $|z|\leqslant 1$.)

Theorem 9: Invertibility

If $Z_1, \dots, Z_q$ are the zeros of $\theta(z)$, and $\min_{1 \leqslant j \leqslant q} |z_i| > 1$, then X_t is invertible,

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

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

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

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

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

So the zeros of $\theta(z), \phi(z)$ are not in the unit circle.

1.20 Proof of Causality&Stationaryity condition for ARMA Processes

Suppose $\psi(z) = \sum_{l=0}^{\infty} \psi_l z^l$, where $\sum_{l=0}^{\infty} |\psi_l| < \infty$. Define $\psi(B)X_t = \sum_{l=0}^{\infty} \psi_l X_{t-l}$.

Lemma 10

If $\{X_t\}_{t\in\mathbb{Z}}$ is a stationary (in any sense) process in L^2, then

$$Y_t = \sum_{l=0}^{\infty} \psi_l X_{t-l} = \psi(B)X_t$$

is stationary (in the same sense).

Proof. If Y_t is well-defined, stationarity follows easily. Since if X_t is strictly stationary $\implies$ Y_t strictly stationary. (Bernoulli shift of X_t).
If X_t is weakly stationary. (Assume $E[X_t] = 0$,

$$E[Y_t Y_{t+h}] = E\left[\left(\sum_{l=0}^{\infty} \psi_l X_{t-l}\right)\left(\sum_{k=0}^{\infty} \psi_k X_{t+h-k}\right)\right] = \sum_{l=0}^{\infty}\sum_{k=0}^{\infty} \psi_l \psi_k \gamma_X(h-k+l)$$

which doesn't depend on t.
Y_t is well-defined as a limit on L^2; By Cauchy-Schwarz, $\gamma_X(h) \leqslant Var(X_0)$. So if $Y_{t,n} = \sum_{l=0}^{n} \psi_l X_{t-l}$, then for $n > m$,

$$\begin{aligned} E[(Y_{t,n} - Y_{t,m})^2] = E\left[\left(\sum_{l=m+1}^{n} \psi_l X_{t-l}\right)^2\right] &= \sum_{l=m+1}^{n}\sum_{k=m+1}^{n} \psi_l\psi_k\gamma_X(k-l) \quad \leqslant Var(X_0)\sum_{l=m+1}^{n}\sum_{k=m+1}^{n} |\psi_l||\psi_k| \\ &\leqslant Var(X_0)\left(\sum_{l=m+1}^{n} |\psi_L|\right)^2 \\ &\to 0 \text{ Since } \sum_{l=0}^{\infty} |\psi_l| < \infty \end{aligned}$$

Therefore, $Y_t = \lim_{n\to\infty} Y_{t,n}$ is well defined in L^2 □

Corollary 11

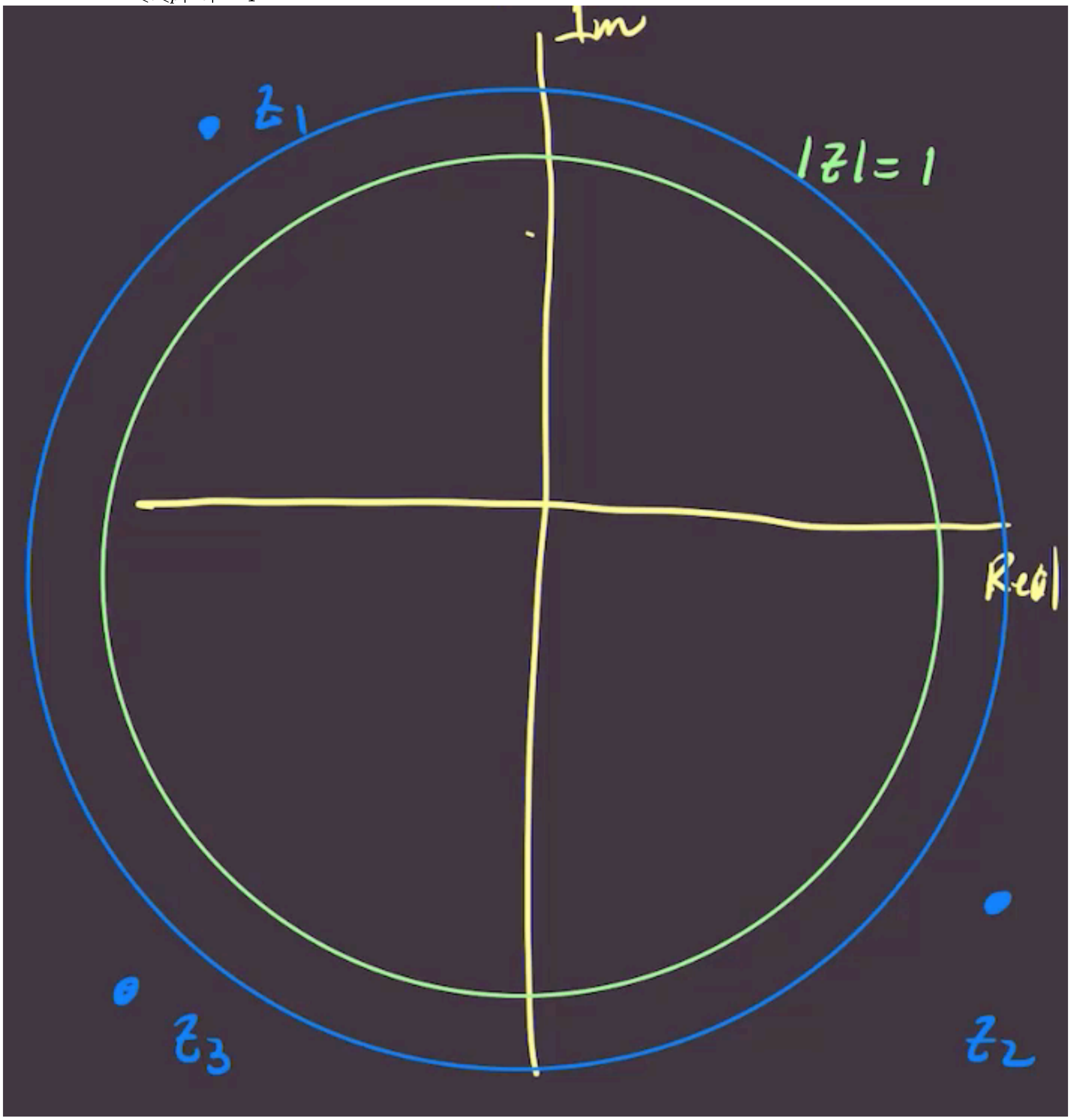
Notice then that if X_t is stationary, $\alpha(z) = \sum_{l=0}^{\infty} \alpha_l z^l, \beta(z) = \sum_{l=0}^{\infty} \beta_l z^l$, with $\sum |\alpha_l| < \infty, \sum |\beta_l| < \infty$. Then

$$Y_t = \alpha(B)\beta(B)X_t = \sum_{l=0}^{\infty}\left(\sum_{j=0}^{l} \alpha_j \beta_{l-j}\right) X_{t-l}$$

Where $\sum_{j=0}^{l} \alpha_j\beta_{l-j}$ is the coefficient of z^l in the power series $\alpha(z)\beta(z)$

Moral:Iteratively applying Backshift operations has the same ”Algebra” as power series multiplication.

Proof. Causality Theorem. Suppose $\phi(Z)$=Autoregressive Polynomial has zeros $z_1, \ldots, z_p \in \mathbb{C}$ so that $\min_{1\leqslant i\leqslant p}|z_i|> q$



Then there must exist $\epsilon > 0$ so that

$$\min_{1\leqslant i\leqslant p}|z_i|> 1 + \epsilon$$

Hence the function $\xi(z) = \frac{1}{\phi(z)}$ is Holomorphic (Analytic) on the set $\{z \in \mathbb{C} : |z|\leqslant 1+\frac{\epsilon}{2}\}$. Hence, $\xi(z)$ must have a power series representation converging on $|z|\leqslant 1 + \frac{\epsilon}{2}$

$$\xi(z) = \sum_{l=0}^{\infty} \xi_l z^l$$

Since $\sum_{l=0}^{\infty} \xi(1+\frac{\epsilon}{2})^l < \infty$, the sequence $|\xi_l|(1+\frac{\epsilon}{2})^l \leqslant k$ for some $k \in \mathbb{R}$. Hence $|\xi_l| \leqslant k(1+\frac{\epsilon}{2})^{-l}$, and hence $\sum_{l=0}^{\infty} |\xi_l| < \infty$.
Define $X_t = \xi(B)\theta(B)W_t$, then

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

Hence $X_t = \xi(B)\theta(B)W_t =: \frac{\theta(B)}{\phi(B)}W_t$ solves the ARMA equations. □

Remark. If $\phi(z) = 0, |z| < 1$ (zeros inside the unit circle), then

$$\frac{1}{\phi(z)} = \sum_{-\infty}^{\infty} \xi_l z^l, 1 - \epsilon < |Z| < 1 + \epsilon$$

In this case, $X_t = \xi(B)\theta(B)W_t = \sum_{l=-\infty}^{\infty} \psi_l W_{t-l}$ (Two sided Linear process, Not Causal, future dependent).
If $\phi(z) = 0$ for some $|z| = 1$m there is no stationary solution [Unit Root Time Series].

1.21 ARMA Processes: Example

Consider a $ARMA(2,2)$ model,

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

Is there a stationary and Causal Solution X_t? Is it invertible? Is there parameter redundancy?

$$\begin{aligned}
\text{AR poly: } \phi(z) &= 1 - \frac{1}{4}z - \frac{1}{8}z^2 \\
\text{MA poly: } \theta(z) &= 1 - \frac{5}{6}z + \frac{1}{6}z^2 \\
\text{Roots of } \phi : \frac{2 \pm \sqrt{4 + 4 * 8}}{-2} &= -1 \pm 3 = -4, 2 \\
\text{Roots of } \theta : 2, 3 &
\end{aligned}$$

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

and they share a common zero, shows parameters are redundant.
X_t satisfies an $ARMA(1,1)$ with

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

Since the roots of ϕ and θ are outside of the unit circle in . X_t is stationary causal and invertible.

Example 10

Suppose $X_t = -\frac{1}{4}X_{t-1} + W_t - \frac{1}{3}W_{t-1}$, then $X_t \sim\ ARMA(1,1)$. $\phi(z) = 1 + \frac{1}{4}z \implies$ Root is -4. So X_t is stationary and Causal, and can be represented as a linear process:

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

We know

$$\psi(z) = \sum_{l=0}^{\infty} \psi_l z^l = \frac{\theta(z)}{\phi(z)}, |z| \leqslant 1$$
$$\implies \psi(z)\phi(z) = \theta(z) \implies \text{Calculate } \psi_l \text{ by matching coefficients}$$

Note:

$$\begin{aligned}
&\phi(z) = 1 + \frac{1}{4}z,\ \theta(z) = 1 - \frac{1}{3}z \\
&\psi(z)\phi(z) = \theta(z) \\
&\ \implies z^0 : \psi_0 = 1 \\
&\ \implies z^1 : \frac{\psi_0}{4} + \psi_1 = -\frac{1}{3} \implies \psi_1 = -\frac{7}{12} \\
&\ \implies z^2 : \frac{\psi_1}{4} + \psi_2 = 0 \implies \psi_2 = -\frac{7}{48} \\
&\vdots \\
&\ \implies z^l : \frac{\psi_{l-1}}{4} + \psi_l = 0 \implies \psi_l = -\frac{7}{12}\left(\frac{1}{4}\right)^{l-1}
\end{aligned}$$

Where $\frac{\psi_{l-1}}{4} + \psi_l$ is called a finite linear difference equation and it must be solved. It is automated in the $ARMAtoMA$ function in R.

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

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

$$\begin{aligned}
\gamma_X(h) = E[X_t X_{t+h}] &= E\left[\left(\sum_{j=0}^{\infty} \psi_j W_{t-j}\right)\left(\sum_{k=0}^{\infty} \psi_k W_{t+h-k}\right)\right] \\
&= \sigma_W^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}
\end{aligned}$$

Coefficients ψ_j can be solved for as in the previous example by solving a finite difference equation. Automated in the $ARMAacf$ function in R.

1.22 L2 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

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

to forecast 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 Mean-Squared Error (MSE)

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

when using MSE, it is natural to consider

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

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

$$\langle x, y \rangle = E[xy]$$

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

$$proj(x \to y) = \langle x, y \rangle \, y$$

Theorem 12: Projection Theorem

We say $M \leqslant L^2$ is a closed linear subspace, if

- Linearity: $x, y \in M, \alpha, \beta \in \mathbb{R}, \alpha x + \beta y \in M$
- Closed: If $X_n \to X$ ($E[(X_n - X)^2] \to 0$), and $X_n \in M$, then $X \in M$

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

$$E\left[(x - \hat{x})^2\right] = \inf_{y \in M} E\left[(x - y)^2\right]$$

Moreover, $\hat{x}$ satisfies

- Prediction Equations/Normal Equations: $x - \hat{x} \in M^{\perp} \implies E[(x - \hat{x})h] = 0, \forall y \in M$

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

$$E\left[(x_{T+h} - \hat{x}_{T+h})^2\right] = \inf_{y \in M} E\left[(x_{T+h} - y)^2\right]$$

where M is a closed linear subspace based on the available data.

1. $M = M_1 = \{z : z = f(x_T, \ldots, X_1),\ f$ is any Borel Measurable function$\}$ In this case,

$$\hat{x}_{T+h} = E[x_{T+h}|x_T, \ldots, x_1]$$

   which is the ideal situation. Unfortunately, M_1 is enormous and complicated! (you have lots of functions to consider)

2. $M = M_2 = \overline{span}\{1, x_T, \ldots, x_1\} = \{y : y = \alpha_0 + \sum_{j=1}^{T} \alpha_j x_j\}$ where $\alpha_0, \ldots, \alpha_T \in \mathbb{R}$ so they are the linear functions of $x_1, \ldots, x_T$.
   $\hat{x}_{T+h}$ is called the Best Linear Predictor (BLP)

1.23 Best Linear Prediction

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

$$E[(x_{T+h} - \hat{x}_{T+h})^2] = \inf_{y \in M_2} E[(x_{T+h} - y)^2]$$

where

$$M_2 = \overline{span}\{1, x_T, \ldots, x_1\} = \{y : y = \alpha_0 + \sum_{j=1}^{T} \alpha_j x_j\}$$

$\hat{x}_{T+h}$ is the best predictor among all linear functions of $x_T, \ldots, x_1$.

Definition 32

If $\hat{x}$ satisfies

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

we say $\hat{x}$ is the projection of x onto M. Write

$$\hat{x} = proj(x|M)$$

BLP $\hat{x}_{T+h} = proj(x_{T+h}|\overline{span}\{1, x_T, \ldots, x_1\})$

Consider the case when $h = 1$. The BLP is of the form

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

where $\mu = E[x_t]$. $\hat{x}_{T+1}$ must satisfy the prediction equations, which is

$$E[(x_{T+1} - \hat{x}_{T+1})y] = 0, \forall y \in \overline{span}\{1, x_T, \ldots, x_1\}$$

In particular,

$$E[(x_{T+1} - \hat{x}_{T+1}) * 1] = 0, \; y = 1$$
$$E[(x_{T+1} - \hat{x}_{T+1}) * x_j] = 0, \; 1 \leqslant j \leqslant T, y = x_j$$

Since $E[x_j - \mu] = 0$, we have

$$0 = E[x_{T+1} - \hat{x}_{T+1}] = \mu - \phi_{T,0} + 0 \implies \phi_{T,0} = \mu$$

Before proceeding, note that this implies

$$E[(x_{T+1} - \hat{x}_{T+1})x_j] = E[(x_{T+1} - \mu - (\hat{x}_{T+1} - \mu))(x_j - \mu)]$$

so we may assume WLOG $\mu = 0 \implies E[x_i x_j] = \gamma(j - i)$
Therefore, (expand the last equation above and notice $\phi_{T,0} = 0$

$$0 = E[(x_{T+1} - \hat{x}_{T+1})x_k] = \gamma(T + 1 - k) - \sum_{j=1}^{T} \phi_{T,j}\gamma(j - k), 1 \leqslant k \leqslant T$$

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

which is a linear system of equations of $\phi_{T,1} \dots, \phi_{T,T}$
If

$$\underline{\gamma}_T = \begin{pmatrix} \gamma(T) \\ \vdots \\ \gamma(1) \end{pmatrix} \in \mathbb{R}^T, \underline{\Gamma}_T = [\gamma(j-k), 1 \leqslant j, k, \leqslant T] \in \mathbb{R}^{T \times T}$$

and $\phi_T = (\phi_{T,1}, \dots, \phi_{T,T})^T \in \mathbb{R}^T$, this linea system may be expressed as

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

The BLP is then of the form

$$\begin{aligned} \hat{x}_{T+1} &= \underline{\phi}_T^T \underline{X}_T = (\underline{\Gamma}_T^{-1} \underline{\gamma}_T)^T \underline{X}_T, \text{ where} \\ \underline{X}_T &= (x_1, \dots, x_T)^T \end{aligned}$$

Theorem 13

If $\gamma(0) > 0$, and $\gamma(h) \to 0$ as $h \to \infty$, then $\underline{\Gamma}_T$ is non-singular.
Takeaway: Most stationary processes (those whose serial dependence decays over time) have non-singular $\underline{\Gamma}_T$

Note that $\hat{x}_{T+1}^2 = \underline{\gamma}_T^T \Gamma_T^{-1} \underline{X}_T \underline{X}_T^T \Gamma_T^{-1} \underline{\gamma}_T$

$$\implies E[\hat{x}_{T+1}^2] = \underline{\gamma}_T^T \Gamma_T^{-1} \underline{\gamma}_T$$

also, since $E[x_{T+1} \underline{X}_T] = \underline{\gamma}_T \implies E[x_{T+1} \hat{x}_{T+1}] = \underline{\gamma}_T^T \Gamma_T^{-1} \underline{\gamma}_T$
It follows that the Mean-Squared prediction error is

$$\begin{aligned} P_{T+1}^t &= E[(x_{T+1} - \hat{x}_{T+1})^2] = E[x_{T+1}^2 - 2x_{T+1}\hat{x}_{T+1} + \hat{x}_{T+1}^2] \\ &= \gamma(0) - 2\underline{\gamma}_T^T \Gamma_T^{-1} \underline{\gamma}_T + \underline{\gamma}_T^T \Gamma_T^{-1} \underline{\gamma}_T = \gamma(0) - \underline{\gamma}_T^T \Gamma_T^{-1} \underline{\gamma}_T \end{aligned}$$

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

1.24 Partial Autocorrelation

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

$$\begin{aligned} X_t \sim AR(p) &\implies \text{ACF has geometric decay} \\ X_t \sim MA(p) &\implies \text{ACF is non-zero at first } q \text{ lags, then zero beyond.} \end{aligned}$$

ACF if an $ARMA(p,q)$ model can be calculated by calculating the linear process coefficients $\{\psi_l\}_{l=0}^{\infty}$
Automated in R using $ARMA_{acf}$ function.

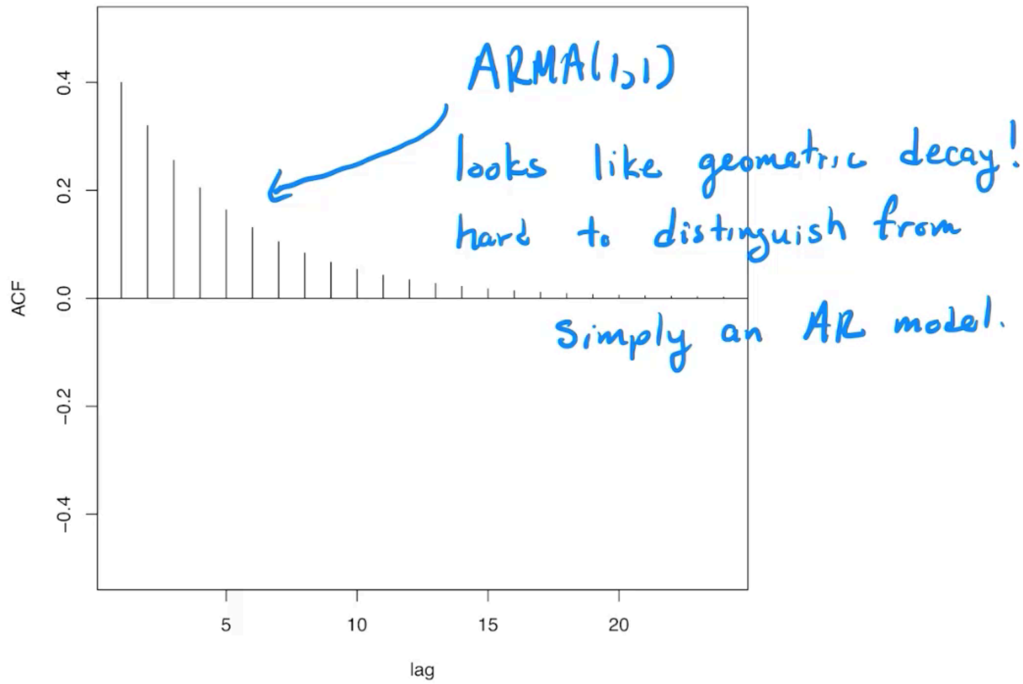


Figure: ARMA(1,1): $x_t = .9x_{t-1} + w_t + .5w_{t-1}$. It is hard to tell the difference between this and an AR(p) ACF

Definition 33

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

$$\phi_{h,h} = Corr\left(X_{t+h} - Proj(X_{t+h}|X_{t+h-1},\ldots,X_{t+1}), X_t - Proj(X_t|X_{t+h-1},\ldots,X_{t+1})\right)$$

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

Remark. If $X_t \sim AR(p)$, which is causal, then $\phi_{h,h} = 0$ for $h \geqslant p+1$

Proof.

$$X_t \sim AR(p) \implies X_{t+h} = \sum_{j=1}^{p} \phi_j X_{t+h-j} + W_{t+h}$$

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

and minimizes

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

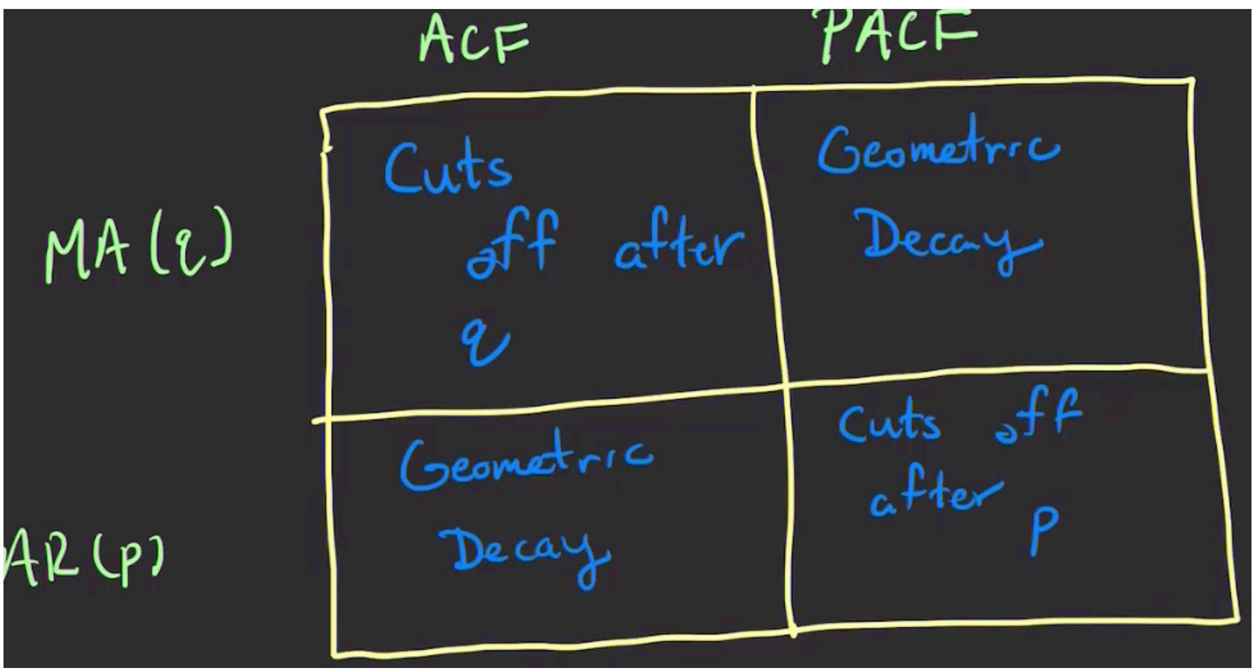
where the second term can be minimized by setting $\beta_j = \phi_j, 1 \leqslant j \leqslant p, \beta_j = 0, h \geqslant p+1$
Hence,

$$X_{t+h} - Proj(X_{t+h}|X_{t+h-1}, \ldots, X_{t+1}) = W_{t+h} \ (h \geqslant p+1)$$
$$\implies \phi_{h,h} = Corr(W_{t+h}, X_t - Proj(X_t|X_{t+h-1}, \ldots, X_{t+1})) = 0$$

we get it is 0 by causality, because $X_t - Proj(X_t|X_{t+h-1}, \ldots, X_{t+1})$ is a term that only depends on something before $t+h$ but not W_{t+h} itself. □

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

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



Estimating the PACF: Using the BLP theory

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

where

$$\begin{aligned}\hat{\Gamma}_h &= [\hat{\gamma}(j-k), 1 \leqslant j, k \leqslant h] \in \mathbb{R}^{h \times h} \\ \hat{\gamma}_h &= [\hat{\gamma}(1), \ldots, \hat{\gamma}(j)] \in \mathbb{R}^h\end{aligned}$$

1.25 Casual and Invertible ARMA Process Forecasting

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

$$\hat{X}_{T+h} = Proj(X_{T+h}|\overline{span}\{1, X_T, \dots, X_1\}) \approx E[X_{T+h}|X_T, \dots, X_1]$$

because by the Causality and Invertibility, $X_t \sim$ linear function of W_t
Further, $\hat{x}_{T+h} \approx \tilde{x}_{T+h} = E[x_{t+h}|X_T, \dots, x_1, x_0, \dots]$ because Geometric decay of the dependence on past values.
Since x_t is causal and invertible, then

$$x_t = \sum_{l=0}^{\infty} \psi_l w_{t-l},\ w_t = \sum_{l=0}^{\infty} \pi_l x_{t-l}\ (\pi_0 = \psi_0 = 1)$$

Note: ψ_l's and π_l'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 $\tilde{x}_{T+h} = E[x_{T+h}|x_T, x_{T-1}, \dots] = E[x_{T+h}|w_T, w_{T-1}, \dots]$

1.
$$\begin{aligned}\tilde{x}_{T+h} &= E[\sum_{l=0}^{\infty} \psi_l w_{T+h-l}|w_T, w_{T-1}, \dots] \\ &= E[\sum_{l=0}^{h-1} \psi_l w_{T+h-l}|w_T, \dots] + E[\sum_{l=h}^{\infty} \psi_l w_{T+h-l}|w_T, \dots]\end{aligned}$$

   Notice one term is independent of the given information, so it's just the mean which is 0, the second term is a function of the given information, so the equation is

$$\sum_{l=h}^{\infty} \psi_l w_{T+h-l}$$

   Also, using invertibility

$$\begin{aligned}0 = E[w_{T+h}|X_T, X_{T-1}, \dots] &= E[\sum_{l=0}^{\infty} \pi_l X_{T+h-l}|X_T, X_{T-1}, \dots] \\ &\underset{\pi_0=1}{=} \tilde{x}_{T+h} + \sum_{l=1}^{h-1} \pi_l \tilde{x}_{T+h-l} + \sum_{l=h}^{\infty} \pi_l x_{T+h-l}\end{aligned}$$

   so we have

$$\implies \tilde{x}_{T+h} = -\sum_{l=1}^{h-1} \pi_l \tilde{x}_{T+h-l} - \sum_{l=h}^{\infty} \pi_l x_{T+h-l}$$

Truncated ARAM Prediction:

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

notice that we truncated the last term to the observed information.
Residuals:

$$\hat{w}_t = \phi(B)\hat{x}_t - \theta_1 \hat{w}_{t-1} - \ldots - \theta_2 \hat{w}_{t-q}$$

Mean Initialization:

$$\hat{w}_t = 0, t \leqslant 0, t \geqslant T, \; \hat{x}_t = 0, t \leqslant 0, \hat{x}_t = x_t, 1 \leqslant t \leqslant T$$

Estimator for $\sigma_W^2 : \hat{\sigma}_W^2 = \frac{1}{T}\sum_{t=1}^T \hat{w}_t^2$
Mean Squared Prediction Error:
Since $\hat{x}_{T+h} \approx \sum_{j=h}^{\infty} \psi_j w_{t-j}$,

$$P_{T+h}^T = E[(x_{T+h} - \hat{x}_{T+h})^2] = E[(\sum_{j=0}^{h-1} \psi_j w_{t-j})^2] = \sigma_W^2 \sum_{j=0}^{h-1} \psi_j^2$$

Estimated Mean Square Prediction Error:

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

Construction of Prediction Intervals:
Since $\hat{x}_{T+h} \approx E[x_{T+h} | x_T, x_{T-1}, \ldots]$, then

$$E[\hat{x}_{T+h} - x_{T+h}] = 0, \text{ Tower Property}$$
$$E[(\hat{x}_{T+h} - x_{T+h})^2] = P_{T+h}^T$$

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 the Random Variable. Then

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

is an approximate $1-\alpha$ prediction interval for x_{T+h}.
Choices for c_α:

1. z_α which is the standard normal critical value
   Motivation: If w_t is Gaussian, then $x_t = \sum_{l=0}^{\infty} \psi_l w_{t-l}$ is Gaussian.

2. Empirical Critical Value of Residuals (standardized)

$$\frac{\hat{w}_t}{\sigma_W},\ 1 \leqslant t \leqslant T$$

3. t-distribution, Pareto, or skewed distribution fit to standardized Residuals.

Long Range Behaviour of ARAMA forecasts: Suppose $Y_t = S_t + X_t\ X_t \sim ARMA(p,q)$,

$$\hat{Y}_{T+h} = \hat{S}_{T+h} + \hat{X}_{T+h} = \hat{S}_{T+h} + \sum_{j=h}^{\infty} \psi_j W_{T+h-j}$$

The last term goes to 0 geometrically when h increases.
$\hat{Y}_{T+h}$ is converging fast to $\hat{S}_{T+h}$: Better get the trend for long Range Forecasts!

$$P_{T+h}^{T} = \sigma_W^2 \sum_{l=0}^{h-1} \psi_L^2 \to \sigma_W^2 \sum_{l=0}^{\infty} \psi_l^2 = \gamma_x(0)$$

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

1.26 ARMA Forecasting: Example

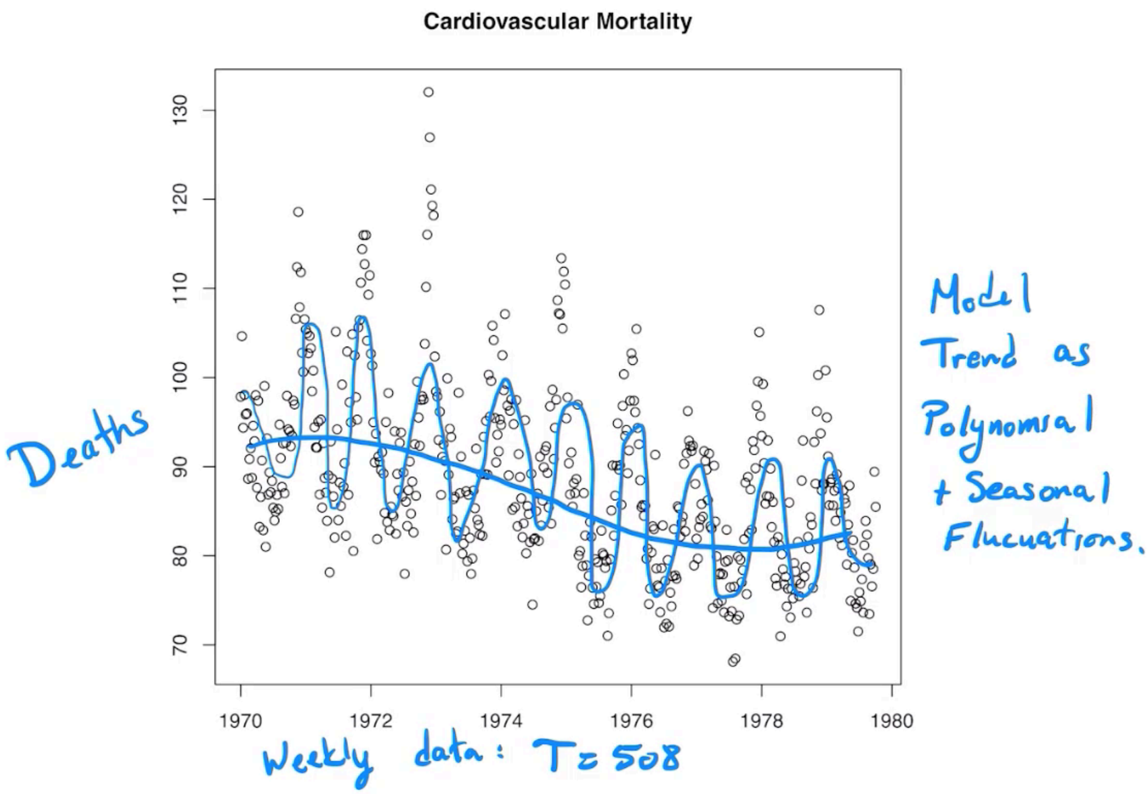


Figure: Weekly cardiovascular mortality, LA County.

$$X_T = \text{ Cardiovascular Mortality Series}$$

Model

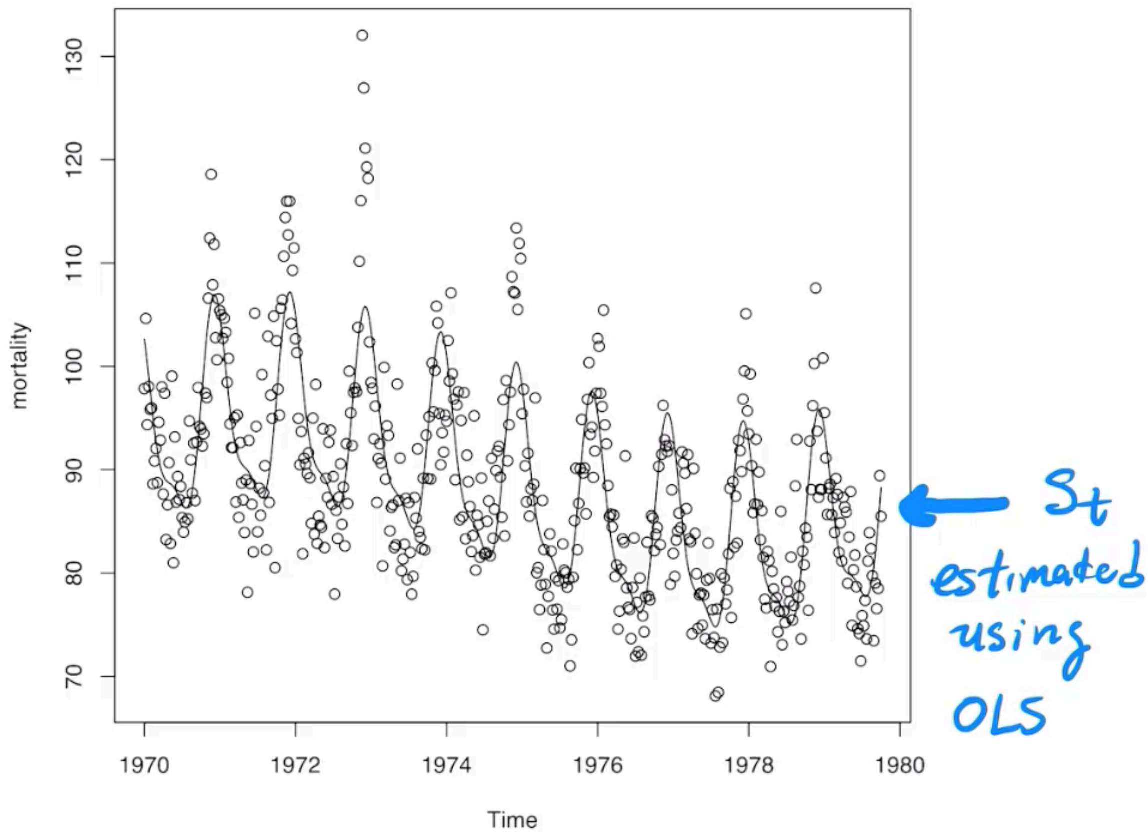
$$X_t = S_t + Y_t,\ Y_t \sim ARMA(p,q) \text{ process}$$

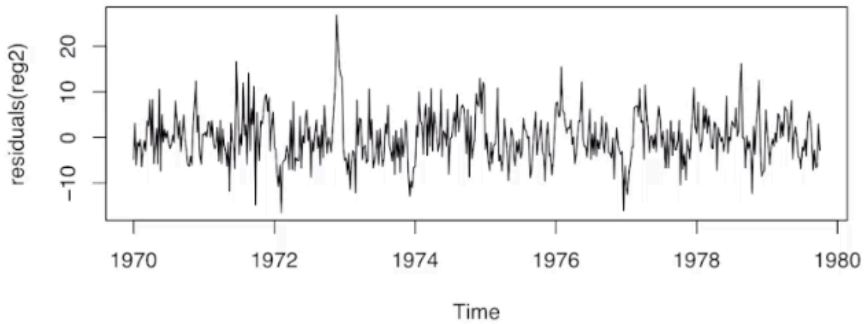
where

$$S_t = \text{ Seasonal + Polynomial trend}$$

$$= \underbrace{\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3}_{\text{Polynomial}} + \underbrace{\beta_4 \sin\left(\frac{2\pi}{52}t\right) + \beta_5 \cos\left(\frac{2\pi}{52}t\right)}_{\text{Yearly Cycle}} + \underbrace{\beta_6 \sin\left(\frac{2\pi}{26}t\right) + \beta_7 \cos\left(\frac{2\pi}{26}t\right)}_{\text{Half-Yearly Cycle}}$$

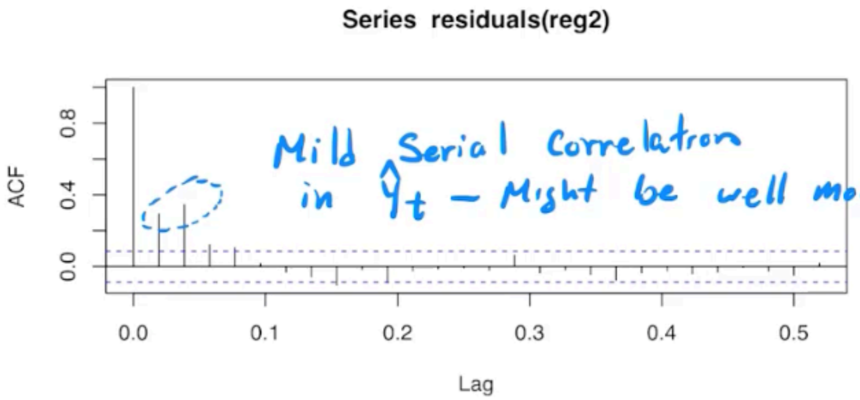
Decided on this trend using AIC (later)


130
120
110
100
90
80
70
mortality
1970
1972
1974
1976
1978
1980
Time
S_t estimated using OLS



$\hat{Y}_t = X_t - \hat{S}_t$

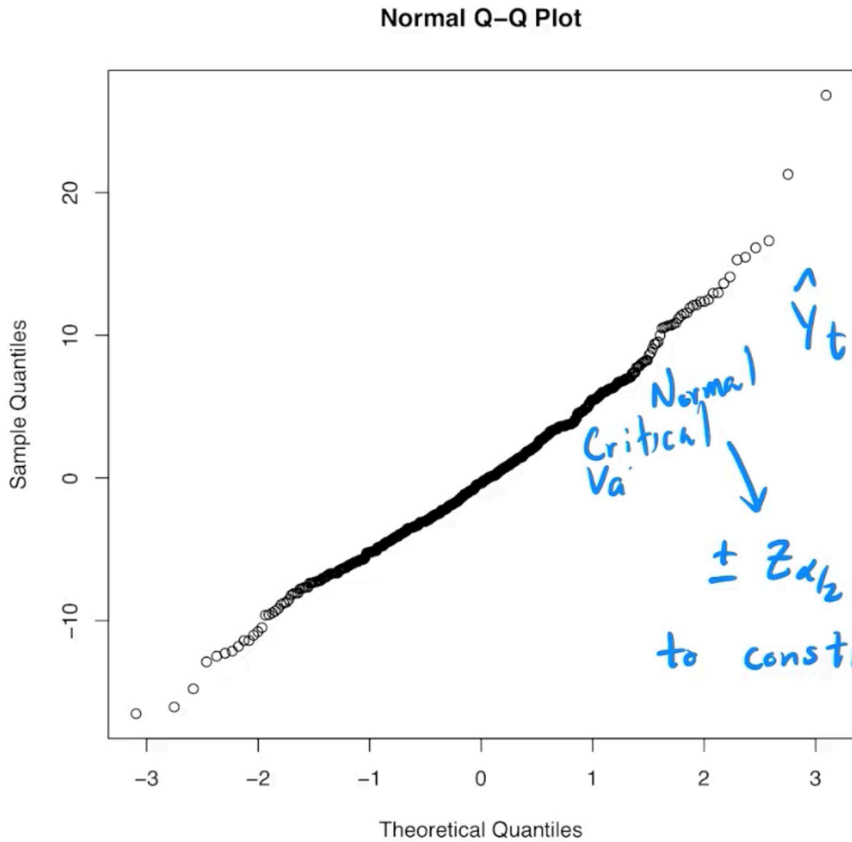
"Seems reasonably stationary"



Mild Serial Correlation in $\hat{Y}_t$ – Might be well modelled by

MA(2)

ARMA(1,1)



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

Normal Critical Va

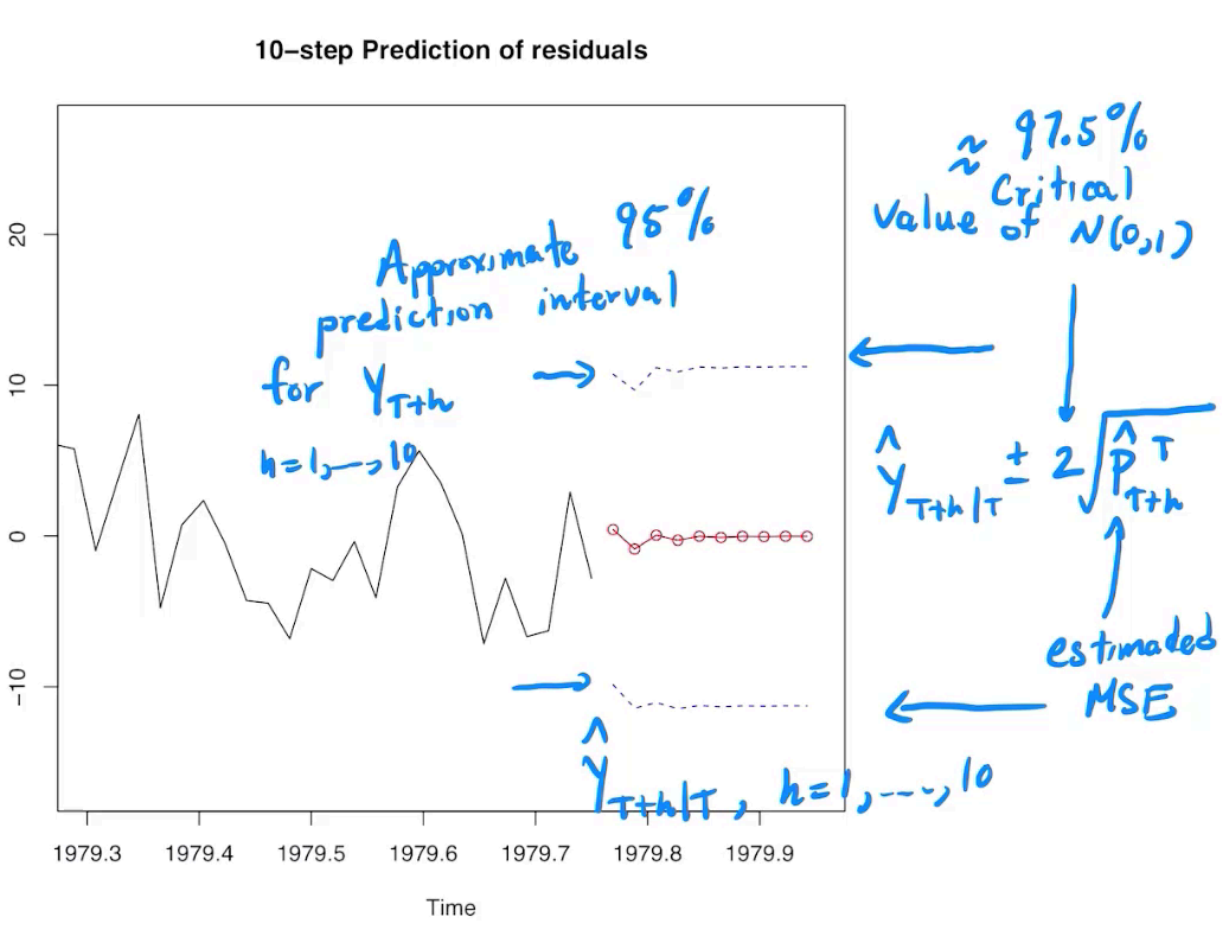
$$\pm z_{\alpha/2}\sqrt{P^T_{T+h}}$$

to construct prediction bounds.


Series rec
PACF indicative
of AR(2) structure
Partial ACF
Lag
0.3
0.2
0.1
0.0
-0.1
0.0
0.1
0.2
0.3
0.4
0.5

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

$$Y_t = \underbrace{0.0885Y_{t-1} + 0.3195Y_{t-2} + W_t + 0.1328W_{t-1}}_{\text{param. by MLE}}$$



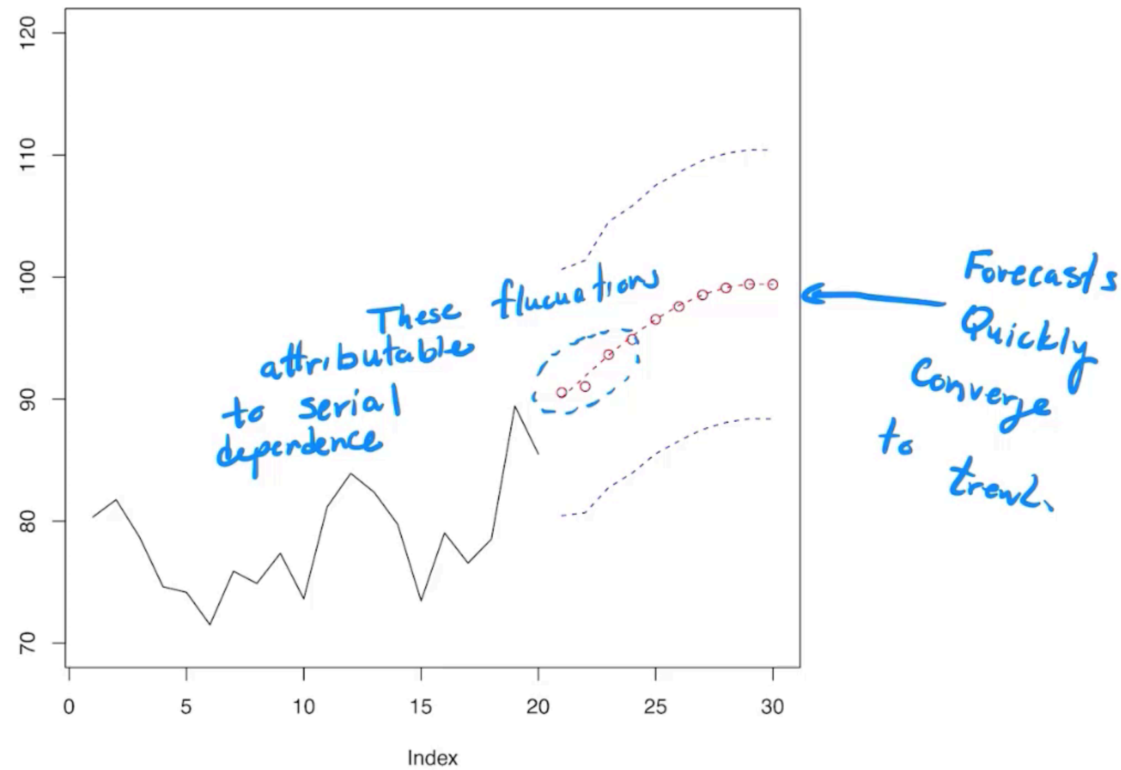


Figure: Forecasts with 95% prediction intervals

1.27 Estimating $ARMA(p, q)$ Parameters: AR Case

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

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

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

Goal: Estimate $\underbrace{\phi_1, ..., \phi_p}_{\text{AR parameters}}$; $\underbrace{\theta_1, ..., \theta_q}_{\text{MA parameters}}$; $\underbrace{\sigma_w^z}_{\text{white noise variance}}$

- AR(1) case: $X_t = \phi X_{t-1} + w_t, \quad Ew_t^2 = \sigma_w^2$

  Idea: use ordinary least squares(OLS).

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

  This leads to (upon some calculus):

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

  $\sigma_w^2 = \frac{1}{T-1}\sum_{t=2}^{T}(\underbrace{X_t - \phi X_{t-1}}_{\text{estimated } w_t})^2 \quad \longleftarrow$ Sample Variance of Residuals.

- AR(p) Case: $\quad X_t = \phi_1 X_{t-1} t - \cdots + \phi_p X_{t-p} + w_t$

  OLS: $\underline{\phi} = (\phi_1, ..., \phi_p)^T \in \mathbb{R}^p$

  $\hat{\underline{\phi}} = \underset{\substack{\underline{\phi}: X_t \text{admits a staionary} \\ \text{and Casual Solution}}}{argmin} \sum_{t=p+1}^{T} (X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p})^2$

  Solve using calculus (Take first order partial derivatives, set equal to zero).

  This leads to a system of p linear equations of the form

$$\hat{\Gamma}_p \hat{\underline{\phi}} = \hat{\underline{\gamma}}_p; \quad \hat{\Gamma}_p = (\hat{\gamma}(j-k), 1 \le j, k \le p) \in \mathbb{R}^{p \times p}$$

  $\hat{\gamma}_p = (\hat{\gamma}(1), ..., \hat{\gamma}(p))^T$

  The resulting OLS estimator takes the approximate form:

$$\hat{\underline{\phi}} = \underline{\hat{\Gamma}}_p^{-1} \hat{\underline{\gamma}}_p, \quad \hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\underline{\gamma}}_p^T \hat{\Gamma}_p^{-1} \hat{\gamma}_p.$$

- Similar approach: use Method of Moments (Set parameters so that empirical moments match theoretical moments induced by the model)
  If $X_t \sim AR(p)$, then for $1 \leq h \leq p$,

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

  This implies the linear system: $\underline{\gamma_p} = \underline{\Gamma_p \phi}$; $\gamma_p = (\gamma(1), \cdots, \gamma(p))^T \in \mathbb{R}^{p\times p}$
  $\underline{\Gamma}_p = [\gamma(j-k); 1 \leq j, k \leq p] \in \mathbb{R}^{p\times p}$

- Note that $X_t = \sum_{l=0}^{\infty} \psi_l w_{t-l}$, $\psi_0 = 1$ and $w_t = X_t - \phi_1 X_{t-1} - \cdots \phi_p X_{t-p}$.

$$\left.\begin{aligned}\Rightarrow \sigma_w^2 = E[X_t w_t] &= E[X_t(X_t - \phi_1 X_{t-1} - \cdots \phi_p X_{t-p})] \\ &= \gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p) \\ \underline{\gamma}_p = \Gamma_p \underline{\phi} &\end{aligned}\right\} \text{Yule-Walker Equations}$$

  $\Rightarrow$ Yule-Walker Estimators: $\hat{\phi} = \underline{\hat{\Gamma}}_p^{-1} \underline{\hat{\gamma}}_p, \quad \hat{\sigma}_w^2 = \hat{\gamma}(0) - \underline{\hat{\gamma}}_p^T \underline{\hat{\Gamma}}_p^{-1} \underline{\hat{\gamma}}_p$

  Example: In the $AR(1)$ case, the YW estimators are

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

Theorem 14

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

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

OLS and YW estimates are asymptotically equivalent. The i here means the i th autoregressive process coefficients.

Theorem 15

$$\sqrt{T}(\underline{\hat{\phi}}_{YW} - \underline{\phi}) \xrightarrow[T\to\infty]{D} N_P(\underbrace{0, \sigma_\omega^2 \Gamma_p^{-1}}_{\substack{\text{Optimal Variance among all possible (asymptotically)} \\ unbrasedestimators.[Efficient]}})$$

$\hat{\sigma}_w^2 \xrightarrow{p} \sigma_w^2$

Result can be used to obtain confidence interval for ϕ.

1.28 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 observable, and YW equations are not linear in the MA parameters.
Latent variables (e.g. variables associated with the noise w_t) $\Longrightarrow$ MLE is best.

- Suppose $X_t \sim AR(1)$

$$X_t = \phi X_{t-1} \quad , \quad w_t \underset{iid}{\sim} N(0, \sigma_w^2) \quad \text{(Gaussian Distributional Assumption on Noise)}$$

$$\text{Then } X_t = \sum_{l=0}^{\infty} \phi^l w_{t-l} \quad \text{is Gaussian}$$

  L^2-limits of Gaussian RV's are Gaussian (MGF or characteristic Function)

- Moreover, $X_1, ..., X_T$ are jointly Gaussian, since

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

  MLE: $L(\phi, \sigma_w^2) = f(X_T, X_{T-1}, ...X_1; \phi, \sigma_w^2)$
  and $L(\phi, \sigma_w^2)$ is likelihood of ϕ, σ_w^2, f is joint density of $X_T, ..., X_1$ evaluated at the observed data (Gaussian Density).

- Key idea in Time Series: To evaluate the likelihood, condition on the path/past!

$$\begin{aligned} f(X_T, ..., X_1) &= f(X_T | X_{T-1}, ...X_1) f(X_{T-1}, ..., X_1) \\ &= f(X_T | X_{T-1}, ..., X_1) f(X_{T-1} | X_{T-2}, ...X_1) ... f(X_2 | X_1) f(X_1) \\ &= \Pi_{i=1}^T f(X_i | X_{i-1}, ...X_1) \end{aligned}$$

  According to HWZ: $X_i | X_{i-1}, ...X_1 \sim N(\phi X_{i-1}, \sigma_w^2)$ by $X_t \sim AR(1)$

- Thus

$$\begin{aligned} L(\phi, \sigma_w^2) &= \Pi_{i=2}^T \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{(X_i - \phi X_{i-1})^2}{2\sigma_w^2}} \cdot f(X_1) \\ &= (w\pi\sigma_w^2)^{-\frac{T-1}{2}} e^{-\sum_{i=2}^T \frac{(X_i - \phi X_{i-1})^2}{2\sigma_w^2}} \cdot f(X_1; \phi, \sigma_w^2) \end{aligned}$$

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

- General ARMA(p,q) Case: Again $X_T, ..., X_1$ are jointly Gaussian if $w_t \sim Gaussian$

$$L(\phi_1, ..., \phi_p, \theta_1, ..., \theta_q, \sigma_w^2) = \Pi_{i=1}^T f(X_i | X_{i-1}, ..., X_1)$$
$$X_i | X_{i-1}, ..., X_1 \sim N(E(Xi | X_{i-1}, ..., X_1), MSE) \sim N(\tilde{X_{i|i-1}}(\underline{\theta}), P_{i-1}^i(\underline{\theta}))$$

  This likelihood can be maximized using numerical optimization.(Newton-Raphson Algorithm conjugate gradient). Note $\underline{\theta}$ is the vector $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_w^2)$

Theorem 16: chapter 8 of Brockwell and Davis, Hannan(1980)

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

Take away message:

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

1.29 Selecting the Orders of $ARMA(p, q)$ Model

Using Maximum Likelihood Estimation, we can fit an ARMA(p, q) model to an observed series $X_1, ..., X_T$.

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

$$\underline{\widehat{W}_t} = \frac{X_t - \tilde{X}_{t \mid t-1}}{\sqrt{\widehat{P}_t^{t-1}}}$$

$$should\ behave\ like\ white\ noise.$$

$\tilde{X}_{t \mid t-1} \sim$ truncated predictor of X_t based on $X_{t-1}, \dots, X_1$.
$\widehat{P}_t^{t-1} \sim$ estimated MSE.

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

As a measure of how "white" the residuals are, it is common to evaluate the cumulative significance of $\widehat{\rho}_W(h) \quad 1 \leq h \leq H$ by applying a "white noise test".

Suppose $W_1, \dots, W_T$ is a strong White Noise, and $\widehat{\rho}_W(h)$ is the empirical ACF of this series.

We know: $\sqrt{T}\widehat{\rho}_W(h) \underset{\rightarrow}{D} N(0, 1)$ for each fixed h. Also, for $j \neq h$,

$$\begin{aligned} Cov(\sqrt{T}\widehat{\gamma}_W(h), \sqrt{T}\widehat{\gamma}_W(j)) &= TE[\sum_{t=1}^{T} W_t W_{t+h}][\sum_{s=1}^{T} W_s W_{s+j}] \\ &= T \sum_{t=1}^{T} \sum_{s=1}^{T} \underbrace{EW_t W_{t+h} W_s W_{s+j}}_{\text{Always zero!}} = 0 \end{aligned}$$

Box-Ljung-Pierce Test (White Noise test for ARMA(p, q) models)

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

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

$$Q(T, H) = \overrightarrow{T \to \infty}^{D} \quad \chi^2(\qquad \underbrace{H - (p+q)}_{\text{Lose p + q degrees of freedom for fitting model}} \qquad)$$

The BLP test p-value is then computed as $P_{BLP} = P(\chi^2(H - (p+q)) > Q(T, H))$.

Remark. If $X_t \sim \ ARMA(p, q)$, and $\widehat{W}_t$ are calculated based on an ARMA(p', q') model where p' < p or q' < q **(Model is under specified)**, then

$$Q(T, H) \underset{\rightarrow}{P} \infty \; as \; T \to \infty.$$

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

1.30 Model Selection: Information Criteria

Model Selection: Information Criteria

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

$\underline{\phi} = AR\ parameters$ $\quad \sigma_w^2 =$ white noise variance.

$\underline{\theta} = MA\ parameters.$

$L(X_1, \ldots, X_T; \underbrace{\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}_w^2}_{\text{Maximum likelihood Estimators}})$ $\leftarrow$ **Natural idea: Maximize the likelihood of the data as a function of p,q.**

Problem: The likelihood is (monotonically) increasing as a function of p, 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 $k = p + q + 1$.

$$-2 \underbrace{log(L(X_1, \ldots, X_T;}_{\text{Minimize, decreasing function of k}} \widehat{\phi}, \widehat{\theta}, \widehat{\sigma}_w^2)) + \underbrace{p(T, k)}_{\text{Increasing function of k.}}$$

Optimal p and q Balance model fit with the penalty for complexity. Common Penalty Term Choices:

$AIC(p, q) = -2log(L(X_1, \ldots, X_T; \widehat{\phi}, \widehat{\theta}, \widehat{\sigma}_w^2) + \frac{2k+T}{T}$
comes from estimating the KullbackLeibler distance from the fitted model to the "true" model.

$BIC(p, q) = -2log(L(X_1, \ldots, X_T; \widehat{\phi}, \widehat{\theta}, \widehat{\sigma}_w^2) + \frac{klog(T)}{T}$
comes from approximating and maximizing the posterior distribution of the model given the data.

Interpretation: Smaller AIC/BIC = Better model. Information Criteria are also use in trend fitting:

Suppose

$$x_t = s_t + y_t = \overbrace{f_t(\underbrace{\beta}_{\text{vector of parameters in } \mathbb{R}^k.})}^{\text{trend we fit}} + y_t$$

Estimate β with $\widehat{\beta}$ using ordinary least squares.

$$RSS_T = \sum_{t=1}^{T} (x_t - f_t(\widehat{\beta}))^2$$

Information Criteria typically calculated assuming Y_t is Gaussian White Noise and are of the form

$$RSS_T + \underbrace{p(T,k)}_{\text{use AIC or BIC penalty.}}$$

Remarks:

1. In trend fitting, the assumption of Gaussian white noise residuals is often in doubt.
2. AIC/BIC are not perfect! They are lout one of many tools useful in model selection.
   - Strengths:
     (a) easy to compute
     (b) Facilitates comparing many models quickly.
   - Weakness:
     (a) Likelihood must be specified.
     (b) 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.

1.31 ARIMA Models:

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

Definition 34

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

Motivation:
If y_t is stationary, and $X_t = \sum_{j=1}^{t} y_j$, then X_t is integrated to order 1; $Z_t = \sum_{i=1}^{t} X_i$ is integrated to order 2, etc

Definition 35

We say X_t follows an Autoregressive Integrated Moving Average Process of orders p, d, q (Abbrv. $X_t \sim ARIMA(p, d, q)$), if

$$\phi(B) \underbrace{(1-B)^d X_t}_{\nabla^d X_t \text{ follows an } ARIMA(p,q)} = \theta(B) W_t$$

and X_t is integrated to order d.

Forecasting $ARIMA(p, d, q)$ processes:

1. $y_t = \nabla^d X_t$ follows and $ARMA(p, q)$ model, and so can be forecasted using truncated ARIMA prediction.
2. Forecasts $\hat{y}_{T+h|T}$ can be used to forecast X_{T+h} by reversing the differencing. For example, say $d = 1$, then $y_{T+1} = X_{T+1} - X_T$, so $\hat{X}_{T+1|T} = X_T + \hat{y}_{T+1|T}$. This can be iterated to produce longer Horizon forecasts.

Prediction MSE is approximately of the form

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

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

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

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

Example 11

$X_t \sim ARIMA(0,1,0)$, then

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

If $y_t = \nabla X_t$, $\hat{y}_{T+h|T} = 0$ (Forecasting W_t's), implies that

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

Similarly,

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

Best Predictor of Random Walk is the last know location.

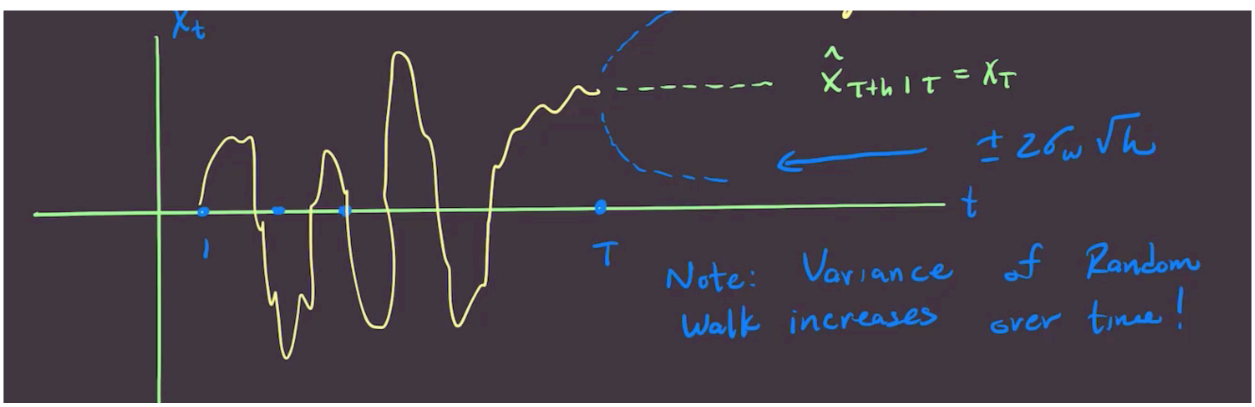
Prediction MSE:

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

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

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

Note: $E[(\hat{X}_{T+h|T} - X_{T+h})^2] = E[(\sum_{j=T+1}^{T+h} W_j)^2] = h\sigma_w^2$



How to decide in practice an degree of differencing d:

1. Eye-ball test (look when the differencing looks stationary)
2. Formal Stationary Tests (Dicky-Fuller, KPSS test)
3. Cross-Validation