STA457S 2023 Summer

Anton S.

- **Professor:** Esam Mahdi, `e.mahdi@mail.utoronto.ca`
- **Lectures:** MW 6-9 BA1160

Contents

1. Introduction

Time series can be defined as a collection of random variables indexed according to the order they are obtained in time. We can consider time series as a sequence of random variables

$$x_1, x_2, \dots, x_t, \dots$$

where x_t is obtained at t-th time point. In this course, the indexing variable t will typically be discrete and not continuous. I.e. $t \in \mathbb{N}$ or $t \in \mathbb{Z}$. A *time series* is a series of observed values (x_t), we call the unrealized model a *process* in this course.

Definition 1.0.1. A series is *stationary* if it remains around a mean value over time.

Examples: Daily temperature, stock prices, generally measurements

Box-Jenkins Methodology

1. **Identification:** Examine graphs and identify patterns and dependency in an observed time series. We look for: *trend, periodic trend, outliers, irregular change*
2. **Estimation:** Select a suitable fitted model for predicting future values.
3. **Diagnostic checking:** Goodness of fit tests and residual scores to estimate adequacy of the model, determine unaccounted for patterns.
4. **Forecasting:** Use model to forecast the future values.

We say forecasting instead of prediction to indicate foretelling closely into the future.

Financial Time Series

We motivate a lot of this course with financial data, so we define terminology for financial time series.

Definition 1.0.2. The *net return* from the holding period $t-1$ to t is

$$R_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

i.e. relative percent increase of (x_k) from $t-1$ to t.

Definition 1.0.3. The *simple gross return* from the holding period $t-1$ to t is

$$\frac{x_t}{x_{t-1}} = 1 + R_t$$

Definition 1.0.4. The *gross return over the most recent k periods* is defined as

$$1 + R_t(k) = \frac{x_t}{x_{t-k}} = \prod_{i=0}^{i=k} \frac{x_{t-i}}{x_{t-i-1}} = (1 + R_t) \dots (1 + R_{t-k})$$

Definition 1.0.5. The *log returns* or *continuously compounded returns* are denoted r_t and defined as

$$r_t = \log(1 + R_t) = \log(x_t) - \log(x_{t-1})$$

Returns are scale-free but not unitless since they depend on t.

Definition 1.0.6. The *volatility* is the conditional standard deviation of underlying asset return.

In most financial time series data, the scale of the volatility appears to be the same. Highly volatile periods tend to be clustered together.

We may decompose a financial time series as

$$x_t = \underbrace{T_t}_{\text{trend}} + \underbrace{s_t}_{\text{season}} + \underbrace{c_t}_{\text{cycle}} + \underbrace{I_t}_{\text{irregularity}}$$

If these components are corelated, use a multiplicative decomposition $x_t = T_t s_t c_t I_t$. If only some are corelated, use a mixed model, i.e. $x_t = s_t T_t + c_t + I_t$.

Time Series Models

Definition 1.0.7 (Moving average)**.** The k-th (odd) moving average of a time series (x_t) is defined as the sum of the k values of the time series around x_t. For example, the third moving average series for (x_t) is

$$y_t = \frac{1}{3}(x_{t-1} + x_t + x_{t+1})$$

If k is even, we reindex and define the time of the moving average to be at the middle of the times we evaluate. For example the 4-th moving average of (x_t) is

$$y_t = \frac{1}{4}(x_{t-2} + x_{t-1} + x_{t+1} + x_{t+2})$$

Moving averages allow us to 'smooth' a time series by reducing the noise while maintaining the trend in the series.

Definition 1.0.8 (White noise)**.** A *white noise process* is a collection of uncorrelated and identically distributed random variables (w_t), each with 0 mean and finite variance σ_w^2 for every t. If the white noise follows a normal distribution, i.e.

$$w_t \sim N(0, \sigma_w^2)$$

then it is *Gaussian white noise*. In the Gaussian case, independent and uncorrelated are the same, so w_t are i.i.d.

Definition 1.0.9 (Random walk)**.** A *random walk with drift* (x_t) is a series

$$x_t = \delta + x_{t-1} + w_t$$

where $w_t \sim \text{wn}(0, \sigma^2)$. For $t \geqslant 1$, δ is the *drift*. When $\delta = 0$, the series is simply a random walk:

$$x_t = x_{t-1} + w_t$$

The series is the same as in the previous time step plus a white noise shock. Therefore we may write

$$x_t = \delta t + \sum_{j=1}^{t} w_j, \quad t \geqslant 1$$

If $\delta \neq 0$, the series is not stationary.

Definition 1.0.10 (Signal in noise)**.** Many realistic models for generating time series assume an underlying sinusoidal signal:

$$x_t = A\sin(\omega t + \phi) + \omega_t$$

As a general note, the goal of time series analysis is to apply a series of transformations in order to reduce the remaining model to a white noise series. Through these transformations we address trends in the series, aiming to be left with only a noise series.

2. Characteristics of Time Series

A complete description of time series is provided by the joint distribution function.

Definition 2.0.1. The *mean function* is defined as

$$\mu_t = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx$$

μ_t is the expectation of the process at the given t, f_t is probability density of x_t.

Definition 2.0.2. The *autocovariance function* is defined as the second moment product

$$\gamma_x(s,t) = \mathrm{Cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

for all s,t. Note $\gamma_x(t,t) = \mathrm{Var}(x_t)$.

Covariance measures the 'linear relationship' of random variables (it is an inner product on the space). The following examples can be computed with the bilinearity properties of covariance.

Example 1. Consider white noise $w_t \sim \mathrm{wn}(0, \sigma^2)$. Then we have

$$\gamma_w(s,t) = \mathrm{Cov}(w_s, w_t) = \begin{cases} \sigma^2, & s = t \\ 0, & s \neq t \end{cases}$$

Example 2. Consider moving average $v_t = \frac{1}{3}(w_{t+1} + w_t + w_{t-1})$ with $w_t \sim \mathrm{wn}(0, \sigma^2)$. Then we can verify that

$$\gamma_v(s,t) = \begin{cases} \frac{1}{3}\sigma^2, & s = t \\ \frac{2}{9}\sigma^2, & |s-t| = 1 \\ \frac{1}{9}\sigma^2, & |s-t| = 2 \\ 0, & |s-t| > 2 \end{cases}$$

Note: Prof said this is a great exam question!

Example 3. For a random walk without drift, $x_t = \sum_{j=1}^{t} w_j$ and $w_t \sim \mathrm{wn}(0, \sigma^2)$, we have

$$\gamma_x(s,t) = \min\{s,t\}\,\sigma^2$$

since the w_t are uncorrelated random variables. Note $\mathrm{Var}(x_t) = t\sigma^2$.

Definition 2.0.3. The **autocorrelation function** is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

The autocorrelation function gives a profile of the linear correlation of the series at time t. Cauchy Schwarz implies $|\gamma(s,t)|^2 \leqslant \gamma(s,s)\gamma(t,t)$.

Definition 2.0.4. For multivariate time series we have the **cross-variance** function

$$\gamma_{xy}(s,t) = \mathrm{Cov}(x_s, y_t)$$

and **cross-correlation** function

$$\rho_{xy}(s,t) = \frac{\sqrt{\gamma_{xy}(s,t)}}{\sqrt{\gamma_x(s,s)}\sqrt{\gamma_y(t,t)}}$$

This can be extended to time series with arbitrary components.

Stationary Models

Definition 2.0.5. A **stationary process** x_t has constant mean and variance for all t.

Stationarity is defined uniquely, so there is only one way for a series to be stationary. It is preferred that estimators of parameters do not changed over time. In many cases, stationary data can be approximated with stationary ARMA models which we discuss later. They also avoid the problem of *spurious regression*.

Definition 2.0.6. A series x_t is **strong stationary** if for any $t_1, t_2, \ldots, t_n \in \mathbb{Z}$ where $n \geqslant 1$ and any scalar shift $h \in \mathbb{Z}$, the joint distribution of both series is the same:

$$P(x_{t_1} \leqslant c_1, \cdots, x_{t_n} \leqslant c_n) = P(x_{t_1+h} \leqslant c_1, \cdots, x_{t_n+h} \leqslant c_n)$$

We never actually know the joint distribution, but this definition allows us to make some theoretical observations about time series. The above implies

1. $p(x_t \leqslant c) = p(x_{t+h} \leqslant c)$
2. $\mu_t = \mu_s$ for all s, t
3. $\gamma(s,t) = \gamma(s+h, t+h)$

It cannot be checked whether any observed time series is strong stationary. This motivates *weak stationary*.

Definition 2.0.7. A process is **time invariant** if it does not depend on time.

Definition 2.0.8. A time series is **weak stationary invariant, covariance stationary, second-order stationary** if

1. μ_t is constant
2. $\gamma(s,t) = \mathrm{Cov}(x_s, x_t)$ depends on s, t only by the difference $|s-t|$: $\gamma(t+h, t) = \gamma(h, 0)$.

Proposition 1. A strong stationary series is weakly stationary. The converse is not true.

Definition 2.0.9. The **autocovariance function of a stationary time series** will be written as

$$\gamma(h,0) = \gamma(h) = \mathrm{Cov}(x_{t+h}, x_t)$$

Note $\gamma(h) = \gamma(-h)$.

Definition 2.0.10. The **autocorrelation function of a stationary time series** will be written as

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)}\sqrt{\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}$$

Definition 2.0.11. The stochastic process w_t is a **strong white noise process** with mean zero and variance σ_w^2 and written $w_t \sim \text{wn}(0, \sigma_w^2)$ if and only if it is i.i.d. with zero mean and covariance

$$\gamma_w(h) = E(w_t w_{t+h}) = \begin{cases} \sigma_w^2, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

A weak stationary *Gaussian* white noise process is strongly stationary, due to uncorrelated implying independent in this case.

Example 4. Consider moving average $v_t = \frac{1}{3}(w_{t+1} + w_t + w_{t-1})$ with $w_t \sim \text{wn}(0, \sigma^2)$. It is stationary since $\mu_{v,t} = 0$.

$$\gamma_v(h) = \begin{cases} \frac{1}{3}\sigma^2, & h = 0 \\ \frac{2}{9}\sigma^2, & h = 1 \\ \frac{1}{9}\sigma^2, & h = 2 \\ 0, & h > 2 \end{cases} \qquad \rho_v(h) = \begin{cases} 1 & h = 0 \\ \frac{2}{3} & h = 1 \\ \frac{1}{3} & h = 2 \\ 0, & h > 2 \end{cases}$$

Example 5. $x_t = \varepsilon_t$ where $\varepsilon_t \sim \text{i.i.d}(0,1)$ is weakly stationary.

Example 6. $x_t = t + \varepsilon_t$ where $\varepsilon_t \sim \text{i.i.d}(0,1)$ is not weakly stationary since μ_t depends on t.

Example 7. Suppose $X_t = A\sin(t+B)$ where $A \sim r.v.(0,1)$, $B \sim U([-\pi,\pi])$. This process is stationary.

$$\begin{aligned} E(X_t) &= E(A\sin(t+B)) = E(A)E(\sin(t+B)) = 0 \\ \gamma(h) &= \frac{1}{2}\cos(h) \end{aligned}$$

$\gamma(h)$ can be verified by integrating.

Transforming Nonstationary Series

The random walk process $x_t = \delta t + \sum_{j=1}^{t} w_j$ is not stationary if it has drift, since $E(x_t) = \delta t$ depends on time. Suppose $\delta = 0$ so the mean function is constant. In this case

$$\gamma(h) = \text{Cov}(x_t, x_{t+h}) = t\sigma^2 \text{ and } \rho(h) = \frac{\text{Cov}(x_t, x_{t+h})}{\sqrt{\text{Var}(x_t)\text{Var}(x_{t+h})}} = \frac{1}{\sqrt{1 + h/t}}$$

For large t and h much smaller than t, get $\gamma(h)$ is very close to 1. We can eliminate the stationarity in a random walk process by taking the difference of the x_t:

$$\nabla x_t = x_t - x_{t-1} = \varepsilon_t \sim \text{wn}(0, \sigma_w^2)$$

In the presence of d unit rots, we apply d differences to x_t:

$$\nabla^d x_t = (1-B)^d x_t = \varepsilon_t$$

Where B is the backwards shift $Bx_t = x_{t-1}$. For example, consider $x_t = a + bt + ct^2$. Then we may take second order differences:

$$z_t = \nabla^2 x_t = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = 2c$$

for $t \geqslant 3$. The R function `diff(x, lag, differences)` can be used for this. The series ∇x_t can be used to transform the time series into stationarity.

Definition 2.0.12. A series is **jointly stationary** if they are each stationary and

$$\gamma_{xy}(h) = \text{Cov}(x_{t+h}, y_t) = E(x_{t+h} - \mu_x)(y_t - \mu_y)$$

is only a function of the lag. The **cross correlation function** of two jointly stationary series is

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$$

We again have $-1 \leqslant \rho_{xy}(h) \leqslant 1$.

Example 8. Consider two series $x_t = w_t + w_{t-1}$ and $y_t = w_t - w_{t-1}$. We find the cross correlation

Definition 2.0.13. A **linear process** x_t is defined to be a linear combination

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_j, \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

We may verify $\gamma_x(h) = \sum_{j=-\infty}^{\infty} \psi_{t+h}\psi_t$. Only need $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ for process to have finite variance. Note that the moving average is an example of a linear process.

If a time series is stationary, we may estimate the mean with $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$. In this case,

$$\text{Var}(\overline{x}) = \frac{\sigma_x^2}{n}\left(1 + \sum_{h=1}^{n-1}(1 - h/n)\rho(h)\right)$$

Estimators

Definition 2.0.14. The **sample autocovariance** is defined as

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x})$$

The sum is restricted since x_{t+h} is not available for $t + h > n$. This estimator is preferred than the one dividing by $n - h$ since it is a non-negative definite function. The **sample autocorrelation** is defined as

$$\hat{\rho}(0) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}$$

This allows us to test whether the autocorrelation is statistically significant at some lags: for n sufficiently large, approximately we have $\hat{\rho}(h) \sim N(0, \frac{1}{n})$. I.e. the estimator is normally distributed with

$$\mu_{\hat{\rho}(h)} = 0 \text{ and } \sigma_{\hat{\rho}(h)} = \frac{1}{\sqrt{n}}$$

We can test $H_0 : \rho(h) = 0$, $H_a : \rho(h) \neq 0$. For $\alpha = 0.05$, have $|\hat{\rho}(h)| \geqslant \frac{2}{\sqrt{n}}$.

- The ACF **cuts off at lag** h if there no spikes at lags $> h$ in the ACF plot.
- The ACF **dies down** if it decreases in a steady fashion.
- If ACF dies down quickly, then the data is stationary. If it dies down very slowly, it is not stationary.

Vector Valued Time Series

Same as regular time series, except

$$\mathbf{x}_t = (x_{t,1}, \dots, x_{t,p}) \in \mathbb{R}^p$$

The transpose is denoted $\mathbf{x}_t$. The mean is $\mu_t = E(x_t) = (\mu_{t,1}, \dots, \mu_{t,p})$. If the process is stationary, $E(x_t) = \mu$, and has autocovariance matrix

$$\Gamma(h) = E(x_{t+h} - \mu)(x_t - \mu)'$$

with cross covariance functions $\gamma_{ij}(h) = E(x_{t+h,i} - \mu_i)(x_{t,j} - \mu_j)$. Note $\gamma_{ij}(h) = \gamma_{ji}(-h)$.

Definition 2.0.15. The sample autocovariance matrix

$$\hat{\Gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \overline{x})(x_t - \overline{x})'$$

where $\overline{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$. The symmetry property holds: $\hat{\Gamma}(h) = \hat{\Gamma}(-h)'$.

3. Time Series Regression and Exploratory Data Analysis

We develop regression models in univariate and multiple time series analyis. We calculate least squares estimators of regression parameters, do ANOVA, and assess our parameters. Then we perform lagged regression, and do transformations of time series to stationarity.

The multiple linear regression model relates the response x to independent variables z_i with the relationship

$$x = \beta_0 + \beta_1 z_1 + \ldots + \beta_q z_q + \varepsilon$$

where ε is some error term. We model

$$E(x \mid z_1, \dots, z_q) = \beta_0 + \beta_1 z_1 + \ldots + \beta_q z_q$$

The linear model is *linear in the coefficients* β_1, not in z_i.

Definition 3.0.1. The multiple linear regression model in time series is modelled with

$$x_t = \beta_0 + \beta_{t,1} + \ldots + \beta_q z_{t,q} + w_t$$

1. x_t is the **dependent time series**

2. $z_{t,1}, \dots, z_{t,q}$ are **independent series**.

3. w_t for different t are iid, $\text{wn}(0, \sigma_w^2)$. Note this is stronger than the usual assumption.

We collect $n > q$ observations of the time series, at various time points and predict $\hat{x}_t = \hat{\beta}_0 + \hat{\beta}_1 z_{t1} + \ldots \hat{\beta}_q z_{tq}$. We describe x_t as a linear combination of the other time series. We minimize the error via least squares:

$$Q(\beta_0, \dots, \beta_q) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - \hat{x}_t)^2$$

Then differentiate and minimize by setting

$$\frac{\partial Q}{\partial \beta_i}\Big|_{\beta_0,\dots,\beta_q} = 0$$

When $q = 1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_t - \overline{x})(z_t - \overline{z})}{\sum_{i=1}^n (z_t - \overline{z})^2}, \quad \hat{\beta}_0 = \overline{x} - \hat{\beta}_z \overline{z}$$

Exam: Should be on reference sheet.

Matrix Form

We can write the multiple linear regression model in terms of vector valued time series/matrix form. Consider $z_t \in \mathbb{R}^q$ with component-wise independent time series $z_{t,i}$. Each z_t can be seen as a column vector of z. Then for the model

$$x_t = \beta' z_t + w_t \quad w_t \sim \text{iid}(0, \sigma_w^2)$$

the least squares estimate is given by

$$\hat{\beta} = (z'z)^{-1} z'x = \left(\sum_{t=1}^n z_t z_t' \right)^{-1} \sum_{t=1}^n z_t x_t$$

The minimized **sum squared errors** can be written

$$\text{SSE} = \sum_{t=1}^n (x_t - \hat{\beta}' z_t)^2$$

The covariance matrix is given by

$$\text{Cov}(\hat{\beta}) = \sigma_w^2 C, \quad C = (zz')^{-1}$$

i.e. the exterior product. The **mean squared error** is

$$\text{MSE} = s_w^2 = \frac{\text{SSE}}{n - (q+1)}$$

which is an unbiased estimator for σ_w^2.

Hypothesis Testing and Model Selection

We may test the hypothesis $\beta_i = 0$ for $i > 0$ with the test statistic

$$t = \frac{\hat{\beta}_i \beta_i}{s_w \sqrt{c_{i,i}}} \sim t_{n-(q+1)}$$

where $c_{i,i}$ is the i-th diagonal element of the covariance matrix C. We can also test whether a subset of z_i influences x_t. The **reduced model** is

$$x_t = \beta_0 + \beta_1 z_{t1} + \ldots + \beta_r + z_{tr} + w_t$$

where $\beta_0, \dots, \beta_r$ are a subset of the original coefficients. Our null hypothesis is $\beta_{r+1} = \cdots = \beta_q = 0$. We are testing whether the SSE deviates statistically significantly once we reduce the model, since it will always reduce somewhat. Our null is that the subset model is correct, since we prefer more parsimonious models.

$$F = \frac{(SSE_R - SSE)/(q-r)}{SSE/(n-q-1)} = \frac{MSR}{MSE} \sim F_{q-r, n-q-1}$$

Note: $n - q - 1 - (n - r - 1) = q - r$ which gives the above degrees of freedom. Reject the more parsimonious model at level α in favor of H_a if $F \geqslant F_\alpha$.

Sources of Variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$z_{t;r+1;q}$	$q-r$	SSR	$MSR = \frac{SSR}{q-r}$	$F_0 = \frac{MSR}{MSE}$	etc
Error	$n-q-1$	SSE	$MSE = \frac{SSE}{n-q-1}$		
Total	$n-r-1$	SSE_0			

- The **sum of squares contributed by regression** (explained variation): $SSR = \sum_{t=1}^{n}(\hat{x}_t - \overline{x})^2$.
- The **sum of squares contributed by error** (unexplained variation): $SSE = \sum_{t=1}^{n}(\hat{x}_t - \hat{x}_t)^2$.
- The **total** sum squared is $SSE_0 = SSR + SSE$.
- The **coefficient of determination** is the proportion of total explained variation is

$$R^2 = SSR/SSE_0 = 1 - SSE/SSE_0$$

.

In order to compare models we also consider the adjusted R^2, which account for the number of predictors.

$$R_a^2 = \left(R - \frac{q}{n-1}\right)\left(\frac{n-1}{n-q-1}\right)$$

For a model with $k+1$ parameters, the least squares estimator of the variance is $\hat{\sigma}_k^2 = SSE(k)/n$, where $SSE(k)$ comes from the model without intercept. Frequently we use *information criteria* to select the best model with k predictors

- $AIC = n \log \hat{\sigma}_k^2 + 2(k+2)$
- $AIC_C = AIC + \frac{2(k+2)(k+3)}{n-k-3}$
- $BIC = \log \hat{\sigma}_k^2 + (k+2)\log n$

We prefer models with minimal information criteria.

Example 9. Consider a time series M_t which is modelled as depending on other series T_t, P_t.

- **Trend-only**: $M_t = \beta_0 + \beta_1 t + w_t$
- **Linear**: $M_t = \beta_0 + \beta_1 t + +\beta_1 T_t + w_t$ or $M_t = \beta_0 + \beta_1 t + +\beta_1 P_t + w_t$ etc.
- **Curvilinear**: $M_t = \beta_0 + \beta_1 t + +\beta_1 (T_t - \overline{T})^2 + \beta_2 P_t + w_t$

The model simultaneously minimizing AIC and BIC is best. Note that the quadratic term in the curvilinear model is centered, probably to account for average temperature in β_0. Given observations for these models, we perform F-test to see whether we can drop some predictors.

When dealing with temporal data, we also need to consider **lagged variables**. This predicts values of x_t from possible lags in z_t. Lagged regression can be done using `dynlm` in R.

Transformations to Stationarity

In order to satisfy many of our assumptions, it is necessary for a series to be stationary. This is often not the case and we often want to transform our data. To remove any change in the mean function μ_t, we detrend the model by decomposing it into

$$x_t = \mu_t + y_t$$

where μ_t is a fitted mean function, y_t is the residual series. Our assumption about errors is that they follows $iid(0, \sigma^2)$, which makes y_t stationary.

The backshift, forward, and difference operators act on time series by

- **Backshift:** $B^h(x_t) = x_{t-h}$
- **Forward:** $B^{-h}(x_t) = x_{t+h}$
- **Difference:** $\nabla^h(x_t) = (1-B)^h(x_t)$

Often taking the first difference is more effective than detrending in order to make the series stationary. ACF plots end up much better.

Example 10. Suppose that after differencing, the ACF plot had a significant value at $h = 4$. Then we model

$$X_t = \theta X_{t-4} + w_t$$

We see later that this is a "$MA(4) = ARMA(0,4)$" model.

When a model has drift, for example $X_t = \delta + X_{t-1} + w_t$, then differencing makes complete sense in order to get a stationary series.

Fractional differencing extends the notion of the difference operator $\nabla^d = (1-B)^d$ to fractional powers of $d \in \left(-\frac{1}{2}, \frac{1}{2}\right)$ which still define stationary processes, especially for **long memory time series**.

A method to suppress large fluctuations of x_t is through the **Box-Cox** transformations:

$$y_t = \begin{cases} \frac{x_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log x_t, & \lambda = 0 \end{cases}$$

which is a method of selecting the best non-linear transformation of x_t in order to minimize the variance of the errors.

Harmonic regression is used when a model contrains a **periodic** trend, allowing us to use trigonometric functions of t to do detrending.

$$x_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{L}\right) + \beta_3 \cos\left(\frac{2\pi t}{L}\right) + w_t$$

Adding trigonometric terms with different frequencies can help with complex seasonality patterns.

Filtering and Smoothing

Filtering/smoothing helps discover useful trends and seasonal components.

Definition 3.0.2. The **moving average smoother** is

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}$$

where $\sum_{j=-k}^{j} a_j = 1, a_j > 0$ makes a symmetric weighted moving average.

Definition 3.0.3. The **kernel smoothing** is

$$m_t = \sum_{i=1}^{n} w_i(t) x_i$$

where $w_i = K(\frac{t-i}{b}) / \sum_{j=1}^{n} K(\frac{j-i}{b})$ are weights, K is some kernel function. The wider the **bandwidth** b, the smoother the model.

4. ARIMA Models

We move into the core of time series analysis. ARMA models are defined, autocorrelation functions are derived, and stationarity, causality, and invertibility of series are evaluated. The Box-Jenkins methodology requires that the model used in describing and forecasting a series is stationary and invertible

Definition 4.0.1. x_t is **stationary** if it remains in statistical equilibrium with properties that do not change over time. x_t is **invertible** if its weights do not depend on time, and x_t can be expressed as a function of previous observations $x_{t-1}, \ldots$.

Definition 4.0.2. The **partial correlation** at lag k of x_t is

$$\text{Corr}(x_{t+k} - \hat{x}_{t+k}, x_t - \hat{x}_t)$$

where $\hat{x}_{t+k} = \beta_1 x_{t+k-1} + \beta_{k-1} z_{t+1}$ and $\hat{x}_t = \beta_1 x_{t+1} + \beta_{k-1} z_{t+k-1}$. Note coefficients are same but reversed. The partial autocorrelation allows us to detect whether a dependence at lag k is appropriate, and is part of the Box-Jenkins methodology.

Auto-Regressive Models

Once trends and seasonal effects are removed from a model, we might construct a linear model for a series with autocorrelation.

Definition 4.0.3. A time series x_t with zero mean is **autoregressive process of order** p, denoted $AR(p)$ if it can be written

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots \phi_p x_{t-p} + w_t$$

for $\phi_p \neq 0$, $w_t \sim \text{wn}(0, \sigma_w^2)$. With backshift operator, we can write this as a polynomial of order p in B,

$$\Phi_p(B) x_t = w_t$$

and $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$, $\phi_p \neq 0$. This is the **characteristic polynomial** of order p.

The second expresssion in terms of characteristic polynomial is preferred, we will see it simplifies our understanding later. If the mean μ of x_t, we may replace x_t by $x_t - \mu$, and rewrite as

$$x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots \phi_p x_{t-p} + w_t$$

with $\delta = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$.

Example 11. The AR(2) model $x_t = 1.5 + 1.2x_{t-1} - 0.5x_{t-2} + w_t$ is

$$x_t - \mu = 1.2(x_t - \mu) - 0.5(x_t - \mu) + w_t$$

Solving for μ with $1.5 = \mu(1 - 1.2 - 0.5)$, we see $\mu = 5$.

Suppose we fit an AR(h) model. In order to decide whether the fit model is a good fit, we check:

- The plot of the time series does not show any increase in variance or trend.
- The ACF plot must decay exponentially, have a wavelet form, or be oscillating (i.e. sign alternates) about 0.
- The PACF plot can be used to detect the correct order for the autoregressive model.

Causal Conditions

We study whether a process can be completely described by its previous values.

Definition 4.0.4 (Causal conditions for AR(1))**.** The autoregressive process of order 1, AR(1), $x_t = \phi x_{t-1} + w_t$ is a **causal process** if it is stationary with values that are not depending on the future. In this case, the absolute value of the root of $1 - \phi z = 0$ must lie outside the unit circle. AR(1) process is causal if

$$|z| = \left|\frac{1}{\phi}\right| > 1 \iff |\phi| < 1$$

A causal process is stationary, but a stationary process is not necessarily causal.

Example 12.

1. $(1 - 0.4B)x_t = w_t$ is causal since the root of $(1 - 0.4z) = 0$ satisfies $|z| = |1/0.4| > 1$.
2. $(1 + 1.8B)x_t = w_t$ is not causal since $|1/\phi| < 1$.

Definition 4.0.5 (Causal conditions for AR(2))**.** The AR(2) model

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

is **causal** when the roots of the characteristic polynomial

$$\Phi_2(z) = 1 - \phi_1 z - \phi_2 z^2$$

lie outside the unit circle

$$\left|\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}\right| > 1$$

Necessary and sufficient conditions for this are

$$|\phi_2| < 1 \quad \phi_1 + \phi_2 < 1 \quad \phi_2 - \phi_1 < 1$$

Example 13. 1. $x_t = 1.1x_{t-1} - 0.4x_{t-2}$ is causal.

2. $x + t = 0.6x_{t-1} - 1.3x_{t-2} + w_t$ is not stationary (necessary and sufficient conditons).

Definition 4.0.6 (Causal conditions for $\text{AR}(p)$)**.** The autoregressive process of order p, $\text{AR}(p)$,

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots \phi_p x_{t-p} + w_t$$

is a **causal process** if *all* roots of the characteristic polynomial

$$\Phi_p(z) = 1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p,\ \phi_p \neq 0$$

lie outside the unit circle.

The function `polyroot(a)`, where `a` is a vector with polynomial coefficients, can be used to find the roots.

Moving Average Models

These are analogous to autoregressive models, except moving average models depend on white noise terms instead of terms of the series itself. There is an analogous characteristic polynomial $\Theta_q(B)$, with the same root condition on *invertibility* instead of causality.

Definition 4.0.7. A time series x_t with zero mean is a **moving average process** of order q, denoted $\text{MA}(q)$, if it can be written

$$x_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots \theta_q w_{t-q}$$

where $w_t \sim \text{wn}(0, \sigma_w^2)$ and $\theta_q \neq 0$. This process has characteristic polynomial $x_t = \Theta_q(B)w_t$ where

$$\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots \theta_q B^q,\ \theta_q \neq 0$$

If the roots z_i of the polynomial $\Theta_q(z)$ satisfy $|z_i| > 1$ for all i, then the process $\text{MA}(q)$ is **invertible**.

Consider the $\text{MA}(1)$ process. The autocorrelation function $\rho(h) = \dfrac{\theta}{1+\theta^2}$ does not change after replacing θ by $1/\theta$. That is

$$x_t = w_t + \theta w_{t-1} \quad \text{and} \quad x_t = w_t + \frac{1}{\theta} w_{t-1}$$

have the exact same autocorrelation function is $\rho(h)$ (show later). This is why invertibility matters: if the polynomial $\Theta_q(z)$ has all roots lying outside the unit circle, then the noise coefficients $\theta_1, \dots, \theta_q$ are uniquely identified.
Compare the two models:

- $\text{AR}(p)$: $\Phi_p(B)x_t = w_t$
- Autoregressive process is always invertible, but not always causal.
- $\text{MA}(q)$: $x_t = \Theta_q(B)w_t$
- Moving average process is always causal, but not always invertible.

We check partial autocorrelation, autocorrelation plots. Out of a set of candidate models, we use AIC and BIC in order to perform model selection for AR and MA.

Auto-Regressive Moving Average Models

Definition 4.0.8. A time series x_t is an **auto-regressive moving average (ARMA)** of order (p, q) if it can be written

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{j=1}^{q} \theta_q x_{t-q}$$

also written as

$$\Phi_p(B)x_t = \Theta_q(B)w_t$$

If x_t has non-zero mean, we can rewrite the above with $\Phi_p(B)(x_t - \mu) = \Theta_q(B)w_t$. Can also be written in summation form with a constant term $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$. This is the `intercept` from `arima()`.

The ARMA satifies stationarity, invertibility, identifiability conditions if

- **Stationary:** Same condition as for AR(p) on $\Phi_p(z)$.
- **Invertible:** Same condition as for MA(q) on $\Theta_q(z)$.
- **Identifiable:** The model is not redundant. $\Phi_p(z)$ and $\Theta_q(z)$ have no common roots.

Example 14. The ARMA$(1, 2)$ model $x_t = 0.2x_{t-1} + w_t - 1.1w_t + 0.18w_{t-2}$ can be written as

$$(1 - 0.2B)x_t = (1 - 1.1B + 0.18B^2)w_t \implies x_t = (1 - 0.9B)w_t$$

which is really an ARMA$(0, 1)$ = MA(1) model. That is we can find the non-redundant expression by removing common roots of the characteristic polynomials.

Definition 4.0.9. The MA(q) process $x_t = \Theta_q(B)w_t$ where

$$\Theta_q(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$$

and $w_t \sim \text{wn}(0, \sigma_w^2)$ is **invertible** if it can be represented as a convergent infinite AR form: AR(∞). Multiply both sides of above by $\Theta_q(B)^{-1}$ to get

$$w_t = \Theta_q(B)^{-1}x_t$$

Recall combinatorics and writing the above as a product of geometric series (factor the polynomial). We denote

$$w_t\Theta_q(B)^{-1}x_t = \Pi_\infty(B)x_t = 1 - \sum_{i=1}^{\infty} \pi_i B^i = -\sum_{i=0}^{\infty} \pi_i B^i$$

Note we are ensured that $\sum_{i=0}^{\infty} |\pi_i| < \infty$ with $\pi_0 = -1$.

Recall the definition of a *linear process* as defined in Section 2. Above we have shown that x_t can be written as an infinite sum of white noise series, and is therefore a linear process.

Example 15. Consider $x_t = (1 + \theta B)w_t$. Then we have the geometric series

$$w_t = \frac{1}{1 - (-\theta B)}x_t = \sum_{k=0}^{\infty} (-1)^k \theta^k B^k x_t = \Pi_\infty(B)x_t$$

This gives the expression

$$\pi_i = (-1)^{i+1}\theta^i$$

and particularly

$$x_t = \sum_{k=1}^{\infty}(-1)^{i+1}\theta^i B^i x_t + w_t$$

Note *why* we need the condition for all the roots of Θ_p to be within the unit circle: we want each geometric series in the product to converge absolutely.

Example 16. Suppose $x = w_t + 0.4w_{t-1}$. This is invertible since $|\theta| = 0.4 < 1$. We can then write

$$x_t = w_t + 0.4x_{t-1} - 0.4^2 x_{t-2} + \cdots$$

In general we know $\Pi_\infty(B) = \Theta_q(B)^{-1}$, so the coefficients π_i can be obtained by equating

$$\begin{aligned}1 &= \Pi_\infty(B)\Theta_q(B)\\ &= 1 - (\pi_1 - \theta_1)B - (\pi_2 + \theta_1\pi_1 - \theta_2)B^2 \cdots\\ &- (\pi_j + \theta_1\pi_{j-1} + \cdots + \theta_{q-1}\pi_{j-q+1} + \theta_q\pi_{j-q})B^j\end{aligned}$$

All non-constant coefficients are 0,

$$\pi_j = -\theta_1\pi_{j-1} - \cdots - \theta_q\pi_{j-q}$$

Now what if we reverse this and do the same for a causal process?

Definition 4.0.10. The AR(p) process

$$\Phi_p(B)x_t = w_t$$

where $\Phi_p(B) = 1 - \sum_{j=1}^{p}\phi_j B^j$, $w_t \sim \text{wn}(0, \sigma_w^2)$ is **causal** if it can be represent as a convergent infinite MA(∞) form:

$$x_t = \Phi_p(B)^{-1}w_t = \Psi_\infty(B)w_t$$

where $\Phi_p(B)^{-1} = \Psi_\infty(B) = 1 + \sum_{k=1}^{\infty}\psi_k B^k$.

Using the same condition as before,

$$1 = \Psi_\infty(B)\Phi_p(B)$$

gives us $\psi_j = \phi_1\psi_{j-1} + \ldots + \phi_p\psi_{j-p}$. This Ψ is known as the **impulse response sequence**.