Time Series:
Section 3. Modeling SARIMA processes

Master in Mathematics and Applications

Docente Responsável:
Maria da Conceição Lopes Costa,
lopescosta@ua.pt

Universidade de Aveiro,
2024/2025

Section 3: Modeling SARIMA processes

1. Exploratory Analysis
2. Model Identification
   - Exploratory Analysis
   - Stationarization
   - Model Order Selection
3. Parameter Estimation
4. Diagnostic Evaluation
   - Model Statistical Quality Assessment
   - Assessment of the suitability of the model-residual analysis

First step of general procedure: Exploratory Analysis

- Plot the data through the cronogram
- See if there are discontinuities (level changes), outliers, variance changes, seasonal effects (annual change), trend, ...

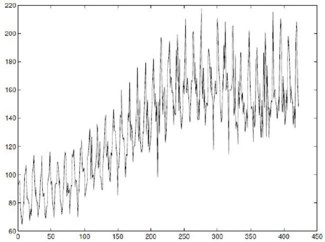


Figura 1: Monthly beer production (megalitres) in Australia since January 1956

(a) (b)

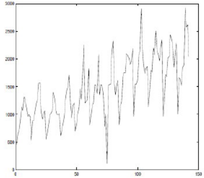


Figura 2: Monthly sales of red wine in Australia (January 1980 to October 1991);(a) transcription error in observation 75; (b)time series corrected.

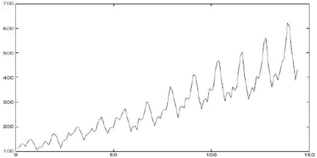


Figura 3: Number of air passengers (thousands) from Jan 1949 to Dec 1960; Exist trend, heteroscedasticity and seasonality.

Some useful transf. and adjustments besides the Box-Cox:

- Adjustment to the length of the month: the fact that the months vary in number of days can cause problems in interpretating seasonality. The adjustment is made as follows:

$$W_t = X_t \times \frac{365.25/12}{\text{number of days in month t}}$$

- Adjustment to the number of working days: this adjustment is necessary because the number of working days in a month varies over the years. After adjusting X_t to the month length, the adjustment is:

$$W_t = X_t \times \frac{\text{number of working in an average month}t}{\text{number of working days in month t}}$$

- Adjustments to mobile holidays and interventions (eg new legislation)

Box and Jenkins (1970) suggested the following methodology:

Model Identification $\longrightarrow$ Parameter Estimation $\longrightarrow$ Diagnostic Evaluation

2. Model Identification

- Plot the sample ACF and PACF to confirm the presence of trend and/or seasonality, possibly non-stationary
- If the sample ACF tends very slowly to zero, it shows non-stationarity
- If the sample ACF presents a periodic behavior slowly tending to zero, it is evidence of non-stationarity in seasonality

2. Model Identification

Example: Monthly production of cow's milk

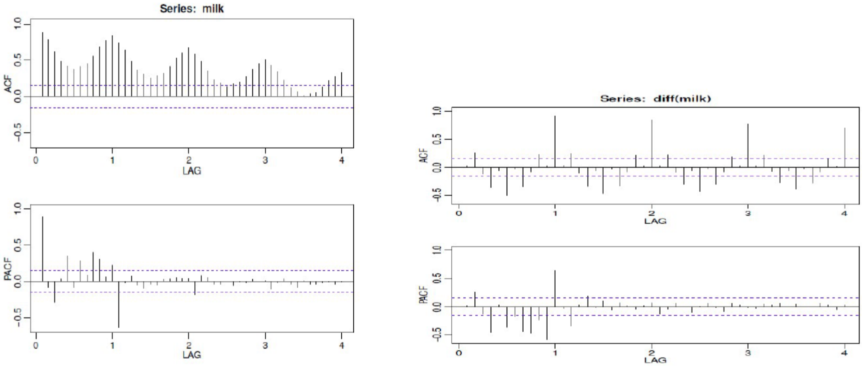


Figura 4: left: series with first difference in season (S=12); right:series of differences at 1 and 12.

2. Model Identification

Example: Monthly production of cow's milk

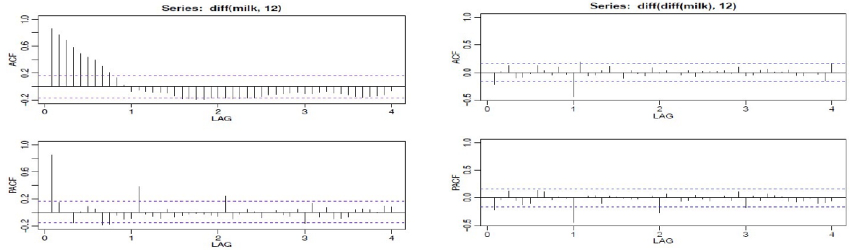


Figura 5: left: Time series of first difference in period S=12; right: TS of the differences in 1 and 12.

To identify the model, you have to analyse the sample acf and sample pacf of the stationary serie to identify the orders p and q of a possible ARMA model. The TS should have more than 100 of observations.

2. Model identification

Order selection

Furthermore, to choose the order of the models automatically or to select one of several models, the following information criteria are used:

- AIC (Akaike Information Criteria)

$$AIC = -2log(L) + 2(p + q + k + 1)$$

- AIC_c (Corrected Akaike Information Criteria)

$$AIC_c = AIC + \frac{2(p + q + k + 1)(p + q + k + 2)}{n - p - q - k - 2}$$

- BIC (Bayesian Information Criteria)

$$BIC = AIC + log(n)(p + q + k - 1)$$

Notation: L$\hookrightarrow$ likelihood; p, q $\hookrightarrow$ orders of ARMA process, c $\hookrightarrow$ drift; $k = 1$ if $c \neq 0$; $k = 0$ if $c = 0$

Remark: These information criteria can be seen as goodness of fit measures: balance between fit error and the number of model parameters; Choose the model that **minimizes** them.

3. Parameter estimation

Suppose we have $(x_1, \ldots, x_n)$ of stationary and invertible ARMA process with p and q fixed, with mean zero.
(you may transform the data, considering $\hookrightarrow x_k = y_k - \bar{y}$)

Goal: To estimate the vector parameter

$$\theta = a_1, \ldots, a_p, b_1, \ldots, b_q, \sigma_e^2)$$

- Moments method: Yule-Walker equations for AR process (linears); Non linear equations for MA models
- Least Squares method: minimizes the sum of the squared errors of one step (conditioned) predictors: $\sum_{t=1}^{n-1}(X_{t+1} - E(X_{t+1}|x_1, \ldots, x_t))^2$
- Maximum likelihood method: assuming $\{e_t\}$ is gaussian, maximizes the conditional likelihood function.

3. Parameter estimation

- In the general case of stationary and invertible ARMA models, the maximum likelihood and least squares methods (and Yule-Walker equations for AR models) lead to optimal estimators

  $\hookrightarrow$ Least Square estimators have the same asymptotic properties as maximum likelihood estimators.

- The parameter estimators have an asymptotically normal distribution, are centered and their variance is known.
- Estimates are obtained numerically (non-linear optimization methods).
- Different estimates can be obtained when using different software

Asymptotic properties: particular cases

- AR(1): $\hat{a} \sim \mathrm{AN}\left(a, \frac{1}{n}(1-a^2)\right)$
- AR(2): $\begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} \sim \mathrm{AN}\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \frac{1}{n}\begin{bmatrix} 1-a_1^2 & -a_1(1-a_2) \\ -a_1(1-a_2) & 1-a_2^2 \end{bmatrix}\right)$
- MA(1): $\hat{b} \sim \mathrm{AN}\left(b, \frac{1}{n}(1-b^2)\right)$
- ARMA(1, 1):
$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} \sim \mathrm{AN}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \frac{1}{n}\frac{1-ab}{(a-b)^2}\begin{bmatrix} (1-a^2)(1-ab) & (1-a^2)(1-b^2) \\ (1-a^2)(1-b^2) & (1-b^2)(1-ab) \end{bmatrix}\right)$

Parameter estimation of TS of Monthly production of cow's milk, modelled by a $SARIMA(1,1,0) \times (1,1,0)_{12}$

```
> milk.fit2=arima(milk,order=c(1,1,0),seasonal=list(order=c(1,1,0),period=12))
> milk.fit2

Call:
arima(x = milk, order = c(1, 1, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:
          ar1     sar1
      -0.2454  -0.4581
s.e.   0.0783   0.0714

sigma^2 estimated as 59.31:  log likelihood = -537.79,  aic = 1081.59
```

Model stationarity: must be checked by the software during the estimation

4. Diagnostic Evaluation: Model Statistical Quality Assessment

Model Statistical Quality Assessment:

- **Statistical significance of the model**

  $\hookrightarrow$ Estimates must be significantly different from zero

  With a 5% significance level, if θ is estimated by $\hat{\theta}$, with standard error (s.e.), then $\bar{\theta}$ is **significantly different from zero** if $0 \notin (\hat{\theta} - s.e., \hat{\theta} + s.e.)$

- **Model stability**

  The different parameters present in the model must not be correlated

  Empirical rule: correlation between any two parameter estimates must be less than 0.7.

4. Diagnostic Evaluation: Model Statistical Quality Assessment

Model SARIMA $(1, 1, 1) \times (1, 1, 0)_{12}$

> milk.fit3=sarima(milk, 1,1,1,1,1,0,12)

```
> milk.fit3 # to view the results
$fit

Call:
arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
    include.mean = !no.constant, optim.control = list(trace = trc, REPORT = 1,
        reltol = tol))

Coefficients:
          ar1      ma1     sar1
      -0.1936  -0.0556  -0.4583
s.e.   0.2298   0.2268   0.0714
```

4. Diagnostic Evaluation: Model Statistical Quality Assessment

Model whose parameter estimates **are significant and not correlated:**

Model SARIMA $(0, 1, 1) \times (1, 1, 0)_{12}$

$>$ milk.fit4=sarima(milk, 0,1,1,1,1,0,12)

```
> milk.fit4 # to view the results
$fit

Call:
arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
    include.mean = !no.constant, optim.control = list(trace = trc, REPORT = 1,
        reltol = tol))

Coefficients:
          ma1     sar1
      -0.2284  -0.4551
s.e.   0.0724   0.0713
```

```
> milk.fit4$fit$var.coef
              ma1          sar1
ma1  0.0052482652 0.0004850028
sar1 0.0004850028 0.0050866888
```

4. Diagnostic Evaluation: Residual Analysis

Residuals: $\hat{e}_t = x_t - \hat{x}_t$

- $x_t \hookrightarrow$ observed value
- $\hat{x}_t \hookrightarrow$ estimated value according to the model

The residuals obtained from the model estimation must be **uncorrelated**

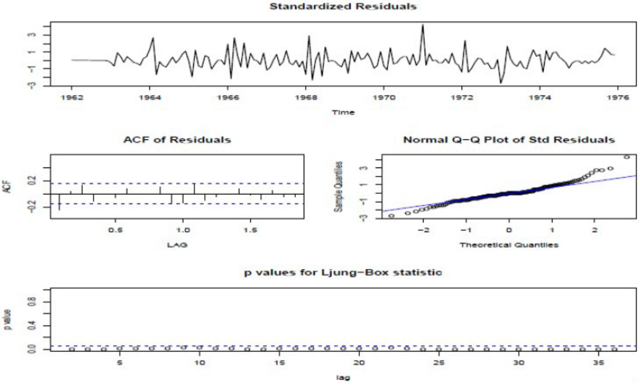
- **Bartlett test:** if the residuals are a realization of an approximately i.i.d process then the sample autocorrelations of the residuals have a $N(0, 1/n)$ distribution
- **Ljung-Box test:** under the assumption that the residuals are a realization of an i.i.d.model, then

$$Q_{LB} = n(n+2) \sum_{j=1}^{h} \hat{\rho}^2(j)/(n-j) \sim \chi_h^2$$

- **QQ-plot:** to assess the hypothesis that the residuals are normally distributed (or statistical tests: Kolmogorov-Smirnov test, Shapiro-Wilk or Anderson-Darling test)

Example of a no suitable model :

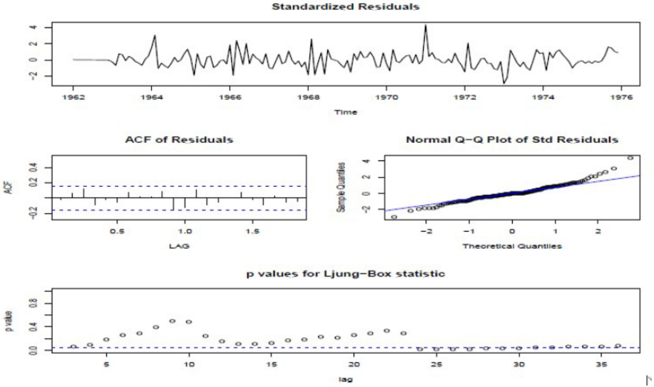
>milk.fit2=sarima(milk,0,1,0,1,1,0,12)



4. Diagnostic Evaluation: Residual Analysis

Example of a suitable model :

>milk.fit4=sarima(milk,0,1,1,1,1,0,12)



4. Diagnostic Evaluation

Model Choice: We have three suitable models

Modelo	Coeficientes		AIC	AICc	BIC
SARIMA$(1,1,0,1,1,0)_{12}$	AR1	SAR1			
	-0.24	-0.46	5.106	5.119	4.143
	(0.078)	(0.071)			
SARIMA$(0,1,1,1,1,0)_{12}$	MA1	SAR1			
	-0.23	-0.46	5.110	5.123	4.147
	(0.072)	(0.071)			
SARIMA$(0,1,1,0,1,1)_{12}$	MA1	SMA1			
	-0.22	-0.62	4.988	5.001	4.026
	(0.075)	(0.063)			