# Identifying Models and Forecasting

(Tsay: Chapters 2.1 – 2.6)

Isabel Casas

# Previously

An ARMA(p,q) process $Y_t$:

$$\phi(L)\,Y_t = c + \theta(L)\epsilon_t$$

with $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ and $\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$

- The process is stable and stationary if the roots $\phi(z) = 0$ are all outside the unit circle
- In this case the ARMA process can be written as a MA($\infty$)
- The process is invertible if the roots of $\theta(z) = 0$ are all outside the unit circle
- In this case the ARMA process can be written as a AR($\infty$)

# Forecasting

- Box and Jenkins modelling philosophy
- Forecasting AR processes
- Forecasting MA processes
- Forecasting ARMA processes

# Box and Jenkins modelling philosophy

- Given a data set, how do we fit the proper ARIMA (p,d,q)?
- ARIMA (p,d,q) is the autoregressive integrated moving average process:

$$\phi^*(\mathsf{L}) = \phi(\mathsf{L})(1-\mathsf{L})^d Y_t = c + \theta(\mathsf{L})\varepsilon_t, \quad \{\varepsilon_t\} \sim \mathsf{IID}(0, \sigma^2). \quad (1)$$

- Box and Jenkins modelling strategy offers a coherent data-driven way of building ARMA or ARIMA models.
- Presented in Box and Jenkins (1970), it consists of three stages (econometritians terminology):
  1. Identification (Specification)
  2. Estimation
  3. Diagnostic checking (Evaluation)

# Box and Jenkins modelling philosophy

- ARIMA(0,1,0) = random walk:

$$Y_t = c + Y_{t-1} + \epsilon_t \;\Rightarrow\; \Delta Y_t = c + \epsilon_t$$

- ARIMA(0, 2, 0) $= \Delta^2 Y_t = c + \epsilon_t$
- ARIMA(0, d, 0) $= \Delta^d Y_t = c + \epsilon_t$
- ARIMA(p,d,q)$= \phi(L)\Delta^d Y_t = c + \theta(L)\epsilon_t$

Remember that $\Delta^d = (1 - L)^d$

Q: How is an ARIMA(1,1,1)?

# Box and Jenkins modelling philosophy

- **Indentification**
  1. Check $d$ in (1). If $d > 0$, apply the filter $\Delta^d = (1 - \mathsf{L})^d$ to $Y_t$ and model $X_t = \Delta^d Y_t$.
  2. Identify the model for $X_t$, that is, determine $p$ and $q$
- **Estimation**
  - Estimate $\phi(\mathsf{L})$, $\theta(\mathsf{L})$ and $\sigma^2$
- **Check the model**
  - Apply misspecification tests to the estimated model

# Identification

- Step 2 of the Identification stage relies on sample autocorrelations and partial autocorrelations.
- Time series $\{y_1, ..., y_T\}$, after potential differencing $\{x_{1+d}, ..., x_T\}$.
- If the partial autocorrelation function has a cut-off point at lag $p$ choose an AR(p) model
- If the model is an MA or an ARMA, the partial autocorrelation function does not have a cut-off point.

# Identification

Box and Jenkins specification rules:

- If the autocorrelation function has a cut-off point, choose an MA model and select $q$ to equal the cut-off point.
- If the partial autocorrelation function has a cut-off point, choose an AR model and select $p$ to equal the cut-off point.
- If the neither function has a cut-off point, choose an ARMA model. Caution recommended in determining $p$ and $q$ because of the potential identification problem discussed in Section 3.5.
- How to choose $d$? Consider the autocorrelation function. If the first autocorrelation is close to unity and the autocorrelations decay slowly, choose $d = 1$. (Repeat the same considerations for the differenced series.)
- More modern techniques (unit root testing) exist.

SYDDANSK UNIVERSITET

# Identification

- Plot the series
  - Does it have a trend? No
  - Is it weakly stationary? Yes
- Do we model it with an AR, MA or ARMA process?
  - Plot the ACF and PACF
  - Ljung-Box test of independence

| | AR(p) | MA(q) | ARMA(p,q) |
|---|---|---|---|
| ACF | geometric decay | cut-off at $q$ | geometric after $q$ |
| PACF | cut-off at $p$ | geometric decay | geometric after $p$ |

- We will see later on what to do in case there is a trend in the mean or heterokedasticity

Syddansk Universitet

# Model checking

- Test the hypothesis that the errors are normal and independent (in the Box and Jenkins strategy it is assumed that $\{\varepsilon_t\} \sim \text{IIDN}(0, \sigma^2)$).
- Consider any sequence $\{X_t\}_{t=1}^{T}$. Want to test the hypothesis that $\{X_t\} \sim$ iid. Have to assume that $\mathrm{E}X_t^4 < \infty$. Let $\rho_j = \text{corr}(X_t, X_{t-j})$. It can be shown

$$\sqrt{T}\widehat{\rho}_j \xrightarrow{d} Z \quad \text{where} \quad Z \sim \text{N}(0,1) \tag{2}$$

  when $\{X_t\} \sim$ iid, as $T \to \infty$.
- Note: Notation '$\xrightarrow{d}$' means 'converges in distribution'.

# Model checking

- The result in (2) can be used to derive the test statistic for $H_0$: $\{X_t\} \sim iid$ or $\rho_1 = 0 = \ldots = \rho_k$.
- It is well known that if $Z_i \sim \text{IIDN}(0,1)$, $i = 1, ..., k$,then $Z_i^2 \sim \chi_1^2$ and independent, and $\sum_{i=1}^k Z_i^2 \sim \chi_k^2$.
- Thus, Box-Pierce statistics

$$S_{\text{BP}}(k) = T \sum_{i=1}^{k} \widehat{\rho}_i^2 \xrightarrow{d} \chi_k^2$$

  when $H_0$ is valid. $S_{\text{BP}}(k)$ is the Box-Pierce statistic.
- Ljung-Box statistic (with better small-sample properties than BP):

$$S_{\text{LB}}(k) = T(T-2) \sum_{i=1}^{k} \frac{\widehat{\rho}_i^2}{T-i} \xrightarrow{d} \chi_k^2.$$

# Model checking

- Problem: Box-Pierce and Ljung-Box statistics are only valid when the null hypothesis is that the original observations are iid.
- The asymptotic theory does not hold when the test is applied to the residuals of an estimated ARMA or ARIMA model.
- Remedy: A degrees of freedom correction. Let $\widehat{\rho}_i = \text{corr}(\widehat{\varepsilon}_t, \widehat{\varepsilon}_{t-i})$ where $\{\widehat{\varepsilon}_t\}$ are the residuals from an estimated ARMA$(p, q)$ model.
- The null hypothesis is that the errors $\{\varepsilon_t\} \sim$ iid.
- Under this null hypothesis,

$$S_{\mathsf{LB}}(k) = T(T-2) \sum_{i=1}^{k} \frac{\widehat{\rho}_i^2}{T-i} \xrightarrow{d} \chi^2(k-p-q)$$

so carrying out the test requires $k > p + q$.

# Identifying AR(p) in practice

The autocorrelation:

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \ldots + \phi_p \rho_{k-p}$$

The partial autocorrelation:

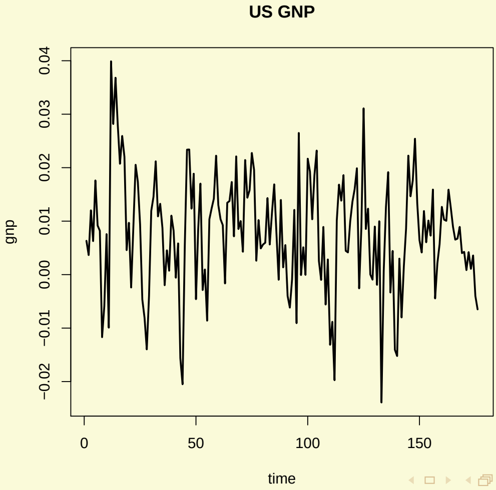
$$\alpha_k = \begin{cases} \neq 0 & k \leq p \\ = 0 & k > p \end{cases}$$

The partial autocorrelation cuts–off at lag $p$.

SYDDANSK UNIVERSITET

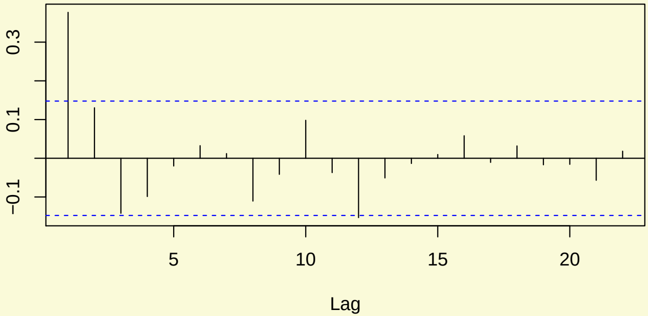
# Identifying AR(p) in practice

Figure: US. quarterly real GNP growth rate from 1947.II ti 1991.I



# Identifying AR(p) in practice

Figure: PACF of US. quarterly real GNP growth rate from 1947.II ti 1991.I



# Identifying AR(p) in practice

AIC criteria:

$$\text{AIC} = \frac{-2}{T}\ln(\text{ likelihood}) + \frac{2}{T} \times \text{ number of parameters}$$

For the Gaussian AR(p), the formula reduces to:

$$\text{AIC}(p) = \ln(\tilde{\sigma}^2) + \frac{2p}{T}$$

# Identifying AR(p) in practice

AIC criteria in R:

```
> options(width = 60)
> gnp.ar = ar(gnp, method = "mle")
> gnp.ar$aic
```

```
        0          1          2          3          4
27.8466897  2.7416324  1.6032416  0.0000000  0.3027852
        5          6          7          8          9
 2.2426608  4.0520840  6.0254750  5.9046676  7.5718635
       10         11         12
 7.8953337  9.6788727  7.1975452
```

```
> gnp.ar$order
```

```
[1] 3
```

# Model checking

- Once we have estimated the model, we must check its adequacy.
- The residuals $\hat{\epsilon}_t$ should behave like a white noise.
    - mean zero
    - no autocorrelation
    - homogeneity
- The ACF and the Ljung-Box statistics of the residuals can be used to check independence.

,

Syddansk Universitet

# The Ljung-Box statistics

$H_0 : \rho_0^\epsilon = \rho_1^\epsilon = \ldots = \rho_k^\epsilon$ vs. $H_1 : \rho_i^\epsilon \neq 0$ for some $i \leq k$

**Box-Pierce** or **Portmanteau test**

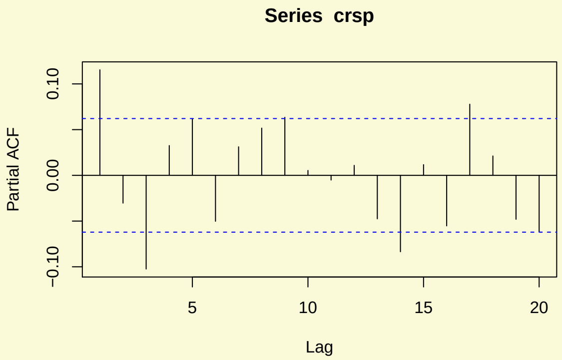
$$S_{\mathsf{BP}}(k) = T \sum_{i=1}^{k} \tilde{\rho}_i^2 \sim \chi_{k-p}^2$$

**Ljung-Box test** (higher power)

$$S_{\mathsf{LB}}(k) = T(T+2) \sum_{i=1}^{k} \frac{\tilde{\rho}_i^2}{T-i} \sim \chi_{k-p}^2$$

# Returns of CRSP value-waited index

Example: Monthly simple resturns of CRSP value-waited index.




AR(3) and AR(9)? Let us choose AR(3).

# Returns of CRSP value-waited index

```
> options(width = 60)
> crsp.2 <- arima(crsp, order = c(3, 0, 0))
> crsp.2

Call:
arima(x = crsp, order = c(3, 0, 0))

Coefficients:
         ar1      ar2      ar3  intercept
      0.1158  -0.0187  -0.1042     0.0089
s.e.  0.0315   0.0317   0.0317     0.0017

sigma^2 estimated as 0.002875:  log likelihood = 1500.86,
```

# Returns of CRSP value-waited index

The intercept in the AR model referst to $\mu$, so we have to find $c = (1 - \phi_1 - \phi_2 - \phi_3)\mu$.

```
> options(width = 60)
> phi <- round(crsp.2$coef[1:3], 3)
> c <- round((1 - sum(phi)) * crsp.2$coef[4], 3)
> c

intercept
    0.009

> sigma.2 <- round(crsp.2$sigma2, 3)
> sigma.2

[1] 0.003
```

# Returns of CRSP value-waited index

The model is

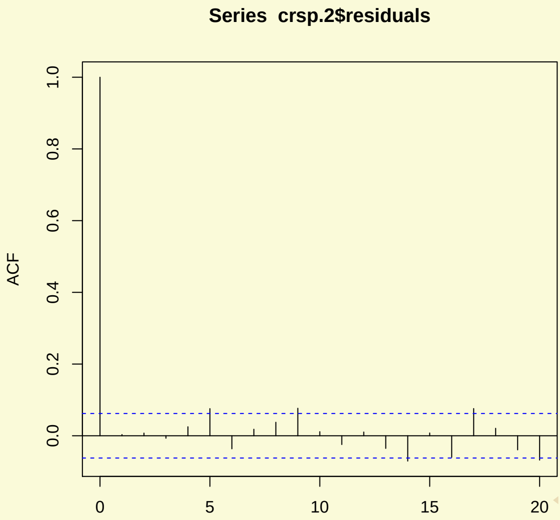
$$Y_t = 0.009 + 0.116\, Y_{t-1} - 0.019\, Y_{t-2} - 0.104\, Y_{t-3} + \epsilon_t \quad \epsilon_t \sim IID(0, 0.003)$$

The estimator of the variance of $\epsilon_t$:

$$\tilde{\sigma}^2 = \frac{\sum_{t=p+1}^{T} \hat{\epsilon}_T^2}{T - 2p - 1}$$

# Returns of CRSP value-waited index

Is there autocorrelation in the errors? let us plot their ACF:



# Returns of CRSP value-waited index

Let us test for autocorrelation for lags under 12.

```
> options(width = 60)
> test <- Box.test(crsp.2$residuals, lag = 12, type = "Ljung")
> pchisq(test$statistic, 9, lower.tail = F)

 X-squared
0.05987558
```

The null hypothesis of no residual serial correlation is not rejected at 5% level (p-value=0.06), but barely. Can we do better?
We noticed in the PACF that the AR(2) is not significant, let us remove it from the model.

# Returns of CRSP value-waited index

Try to remove second coefficient:

```
> options(width = 60)
> crsp.3 <- arima(crsp, order = c(3, 0, 0), fixed = c(NA,
+     0, NA, NA))
> test <- Box.test(crsp.3$residuals, lag = 12, type = "Ljung")
> pchisq(test$statistic, 10, lower.tail = F)

X-squared
0.0782661
```

The $H_0$ is not rejected and the model is better because there is a clear uncorrelated errors. We keep the model without the second coefficient.
Note: We use 10 degrees of freedom because now we are estimating only 2 coefficients instead of 3.

# AR(p) forecasting

Now that we found the proper AR model, let us forecast it.

- Suppose that we are at time $T$ (forecast origin)
- We want to forecast the value $Y_{T+h}$ for $h \geq 1$ (forecast horizon)
- We call $Y^*_{T+h}$ the forecast of $Y_{T+h}$.
- We define a loss function that measures the distance between the real value and the forecasted value.
- The MSE is a popular loss function, so the forecasted value $Y^*_{T+h}$ is chosen as the value that minimises

$$E[(Y_{T+h} - Y^*_{T+h})^2 | F_T) \leq \min_g E[(Y_{T+h} - g)^2 | F_T]$$

  where $F_T$ contains information until time $T$.
- Proof3.pdf

# Example: AR(p)

- We have neglected deterministic trends and have decided that our time series follows an AR(p) process.

$$y_T = \phi_1 y_{T-1} + \phi_2 y_{T-2} + \ldots + \phi_p y_{T-p} + \epsilon_T \quad \epsilon_T \sim IID(0, \sigma)$$

- 1-step-ahead forecast

$$y^*_{T+1} = E(y_{T+1} | y_T, y_{T-1}, \ldots) = \phi_1 y_T + \phi_2 y_{T-1} + \ldots + \phi_p y_{T+1-p}$$

- 2-step-ahead forecast, can be obtained recursively:

$$y^*_{T+2} = E(y_{T+2} | y_T, y_{T-1}, \ldots) = \phi_1 y^*_{T+1} + \phi_2 y_T + \ldots + \phi_p y_{T+2-p}$$

# Example: AR(p)

Forecast error:

- 1-step-ahead forecast

$$e_1 = \epsilon_{T+1}$$

  with variance $\sigma^2$

- 2-step-ahead forecast

$$e_2 = \phi_1(Y_{T+1} - Y^*_{T+1}) + \epsilon_{T+2} = \phi_1 e_1 + \epsilon_{T+2}$$

  with variance $(1 + \phi_1^2)\sigma^2$

Note that the variance of $e_2$ is greater than the variance of $e_1$. So the uncertainty in forecast increases as the horizon increases

# Example: AR(p)

Assuming that the shocks ($\epsilon$) are normally distributed, the forecast interval:

- 1-step-ahead

$$Y^*_{T+1} \pm 1.96\sigma$$

- 2-step-ahead

$$Y^*_{T+2} \pm 1.96\sqrt{(1+\phi_1^2)}\sigma$$

SYDDANSK UNIVERSITET

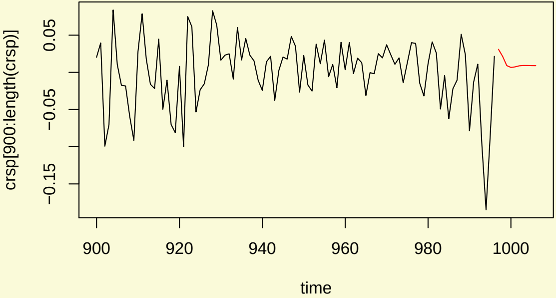
# Predict AR(p) in R

```
> crsp.star.1 = predict(crsp.3, 10)
> crsp.star.1$pred

Time Series:
Start = 997
End = 1006
Frequency = 1
 [1] 0.030956276 0.021454042 0.009034718 0.006617194
 [5] 0.007352650 0.008756259 0.009172633 0.009141748
 [9] 0.008989047 0.008927447
```

# Predict AR(p) in R

Figure: Prediction 10 lags of CRSP



# Identifying MA(q) in practice

The autocorrelation:

$$\rho_k = \begin{cases} \neq 0 & k \leq q \\ = 0 & k > q \end{cases}$$

The autocorrelation cuts–off at lag $q$.

# Identifying MA(q) in practice

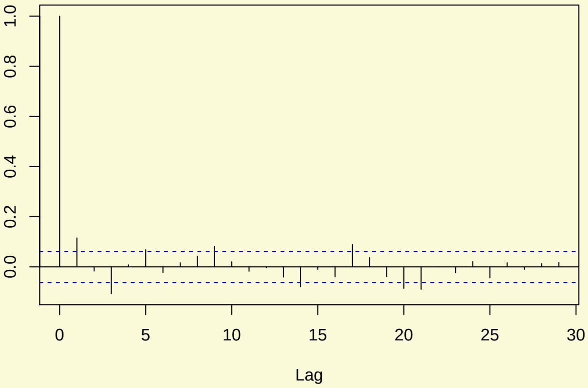



Figure: ACF of monthly simple returnts of CRSP equal-weighted index Jan. 1926 - Dec. 2008

# Identifying MA(q) in practice

In the example, lags 1, 3 and 9 seem significant. So the suitable model could be:

$$y_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_3 \epsilon_{t-3} - \theta_9 \epsilon_{t-9}$$

```
> crsp.4 <- arima(crsp, order = c(0, 0, 9), fixed = c(NA,
+     0, NA, 0, 0, 0, 0, 0, NA, NA))
> crsp.4$coef
```

```
        ma1          ma2          ma3          ma4
 0.110762094  0.000000000 -0.117357685  0.000000000
        ma5          ma6          ma7          ma8
 0.000000000  0.000000000  0.000000000  0.000000000
        ma9    intercept
 0.074343367  0.008910143
```

# MA(q) forecasting

$$Y_T = \mu + \epsilon_T + \theta_1 \epsilon_{T-1} + \ldots + \theta_q \epsilon_{T-q} \quad \epsilon_T \sim IID(0, \sigma^2)$$

- 1-step-ahead forecast

$$Y^*_{T+1} = E(Y_{T+1} | F_T) = \mu + \theta_1 \epsilon_T + \ldots + \theta_q \epsilon_{T+1-q}$$

- 2-step-ahead forecast

$$Y^*_{T+2} = E(Y_{T+2} | F_T) = \mu + \theta_2 \epsilon_T + \ldots + \theta_q \epsilon_{T+2-q}$$

- T+1-step-ahead forecast

$$Y^*_{2T+1} = E(Y_{2T+1} | F_T) = \mu$$

Syddansk Universitet

# MA(q) forecasting

Forecast error:

- 1-step-ahead forecast

$$e_1 = E(Y_{T+1} - Y^*_{T+1}) = \epsilon_{T+1}$$

  with variance $\sigma^2$

- 2-step-ahead forecast

$$e_2 = E(Y_{T+2}|F_T) = \epsilon_{T+2} + \theta_1 \epsilon_{T+1}$$

  with variance $\sigma^2(1+\theta_1^2)$

The forecast confident intervals are obtained as for the AR(p)

# Identifying ARMA models

- If the ACF and PACF both decrease then we have an ARMA model
- The ACF and PACF are not informative in determining the order
- This has to be done ad hoc.

# Forecasting ARMA(p,q)

$$Y_T = c + \sum_{i=1}^{p} \phi_i Y_{T-i} + \sum_{i=0}^{q} \theta_i \epsilon_{T-i}$$

with $\theta_0 = 1$.

- 1-step-ahead forecast

$$Y^*_{T+1} = E(Y_{T+1}|F_T) = c + \sum_{i=1}^{p} \phi_i Y_{T+1-i} + \sum_{i=0}^{q} \theta_i \epsilon_{T+1-i}$$

- 2-step-ahead forecast

$$Y^*_{T+2} = E(Y_{T+2}|F_T) = \mu + \phi_1 Y^*_{T+1} + \sum_{i=2}^{p} \phi_i Y_{T+2-i} + \sum_{i=0}^{q} \theta_i \epsilon_{T+2-i}$$

SYDDANSK UNIVERSITET

# Forecasting ARMA(p,q)

Forecast error:

- 1-step-ahead forecast

$$e_1 = E(Y_{T+1} - Y^*_{T+1}) = \epsilon_{T+1}$$

  with variance $\sigma^2$

- 2-step-ahead forecast

$$e_2 = E(Y_{T+2}|F_T) = \phi_1 e_1 + \epsilon_{T+2} + \theta_1 \epsilon_{T+1}$$

  with variance $(1 + \phi_1^2 + \theta_1^2)\sigma^2$

The forecast confident intervals are obtained as for the AR(p)

SYDDANSK UNIVERSITET