PSTAT 174/274: Time Series

Prof. Dr. Gareth W. Peters

(CStat-RSS, CMath-IMA, FIOR, SIRM, FRSS, FIMA, YAS-RSE)

Professor of Statistics for Risk and Insurance &
Janet and Ian Duncan Endowed Chair in Actuarial Science
University of California Santa Barbara

February 23, 2023

PART VI: Parameter Estimation and Model Selection

Complete Framework: Box-Jenkins

In general in this course we will adopt the widely used approach to modelling and forecasting of Box-Jenkins.

The Box-Jenkins methodology for forecasting

1. **Model identification**
2. **Parameter estimation**
3. **Verification**

   Check model obtained from 1 & 2

   - Good? Goto 4
   - Bad? Goto 1 & decide on new model

4. **Forecasting**

We will now discuss specific details of sub-aspects of this framework and associated automated methods to assist when performing this process.

Modelling Guide

When fitting an ARIMA model to a set time series data, the following procedure provides a useful general approach.

- Plot the data and identify any unusual observations.
- If necessary, consider transformations pf the data (using a Box-Cox transformation) to stabilise the variance.
- If the data are non-stationary, take first differences of the data until the data are stationary.
- Examine the ACF/PACF: Is an ARIMA(p,d,0) or ARIMA(0,d,q) model appropriate or ARIMA(p,d,q)?
- Try your chosen model(s), and use the AICc to search for a better model.
- Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals. If they do not look like white noise, try a modified model.
- Once the residuals look like white noise, calculate forecasts.

The Hyndman-Khandakar algorithm only takes care of steps 3–5.

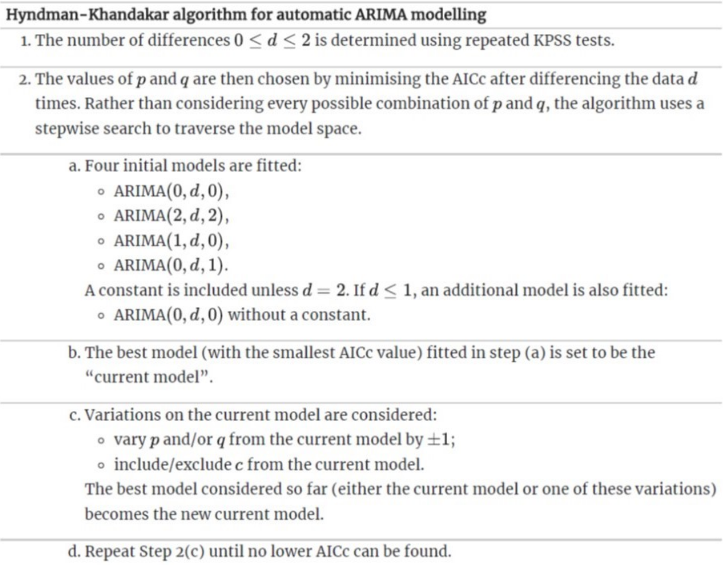
R package auto.arima()

The auto.arima() function in R uses a variation of the Hyndman-Khandakar algorithm which combines:

- unit root tests,
- minimisation of the AICc and
- MLE to obtain an ARIMA model.

The arguments to auto.arima() provide for many variations on the algorithm.

R package auto.arima()



Hyndman-Khandakar algorithm for automatic ARIMA modelling

1. The number of differences $0 \leq d \leq 2$ is determined using repeated KPSS tests.

2. The values of p and q are then chosen by minimising the AICc after differencing the data d times. Rather than considering every possible combination of p and q, the algorithm uses a stepwise search to traverse the model space.

   a. Four initial models are fitted:
      - ARIMA$(0, d, 0)$,
      - ARIMA$(2, d, 2)$,
      - ARIMA$(1, d, 0)$,
      - ARIMA$(0, d, 1)$.

      A constant is included unless $d = 2$. If $d \leq 1$, an additional model is also fitted:
      - ARIMA$(0, d, 0)$ without a constant.

   b. The best model (with the smallest AICc value) fitted in step (a) is set to be the “current model”.

   c. Variations on the current model are considered:
      - vary p and/or q from the current model by ± 1;
      - include/exclude c from the current model.

      The best model considered so far (either the current model or one of these variations) becomes the new current model.

   d. Repeat Step 2(c) until no lower AICc can be found.

Figure: Source: Rob J. Hyndman - Forecasting: Principles and Practice (2nd ed)

R package auto.arima()

Often you may want to also fit your own model to compare to the auto.arima optimal models.

NOTE: on fitting packages -

- If you want to choose the model yourself, use the Arima() function in R.
- There is another function arima() in R which also fits an ARIMA model. However, it does not allow for the constant level term so make sure you first subtract the mean if you use this package. Furthermore, it does not return everything required for other functions in the forecast package to work. Finally, it does not allow the estimated model to be applied to new data (which is useful for checking forecast accuracy).
- Consequently, it is recommended that Arima() be used instead

Parameter Estimation of Linear Time Series Models (Preliminary)

Summary of TS Modelling Process

Typical procedure for time series modelling which can be described as follows:

1. Plot the time series (look for trends, seasonal, components, step changes, outliers).
2. Transform data so that residuals are stationary
   - Estimate and subtract time trend m_t and seasonal trend s_t.
   - Difference
   - Nonlinear transformations $(\log, \sqrt{\cdot})$
3. Fit model to residuals.

Summary of TS Modelling Process

In this next section we will cover a class of estimation methods known as Moment Matching Estimators (MMEs)

In general the MMEs for MA and ARMA models are complicated.

- In general, regardless of AR, MA or ARMA models, the MMEs are sensitive to rounding errors. They are usually used to provide initial estimates needed for a more efficient nonlinear estimation method such as Maximum Likelihood Estimators (MLE)
- The moment estimators are not recommended for final estimation results and should not be used if the process is close to being nonstationary or noninvertible.

$ARMA(p, q)$ Model Estimation

Consider a generic ARMA(p,q) model for observed time series data $y_1, \ldots, y_T$ and the estimation task is in several parts

- Determine orders p,q,
- process mean μ_Y,
- Estimate the AR polynomial parameters $\phi_1, \ldots, \phi_p$,
- Estimate the MA polynomial parameters $\theta_1, \ldots, \theta_q$, and
- process white noise variance σ^2,

which will be combined once presented. We will begin with estimation of the AR coefficients via the Yule-Walker equations, Durbin-Levinson recursion and the MA coefficients via the Innovations algorithm, which we will then combine.

Method of Moments Estimation AR Models - Yule-Walker

Yule-Walker AR Parameter Estimation

Let $\{Y_t\}$ be the zero-mean causal autoregressive process

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \epsilon_t, \quad \{\epsilon_t\} \sim WN(0, \sigma^2). \qquad (0.1)$$

and we want to estimate the coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ and the white noise variance σ^2 based on observations $y_1, \ldots, y_T$ where we require $T > p$.

Via causality assumption, one has

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

where $\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = 1/\phi(z)$, $|z| \leq 1$. Now we multiply each side of Eqn 0.1 by Y_{t-j}, $j = 0, \ldots, p$ and then take expectation. This produces

$$\begin{aligned}
&\mathbf{E}\left[Y_t Y_{t-j} - \phi_1 Y_{t-1} Y_{t-j} - \cdots - \phi_p Y_{t-p} Y_{t-j}\right] \\
&= \mathbf{E}\left[\epsilon_t Y_{t-j}\right] \\
&= \mathbf{E}\left[\sum_{i=0}^{\infty} \psi_i \epsilon_{t-j-i} \epsilon_t\right] \\
&= \sum_{i=0}^{\infty} \psi_i \mathbf{E}\left[\epsilon_{t-j-i} \epsilon_t\right] \\
&= \sigma^2 \mathbb{I}[j = 0], \quad \text{since } \psi_0 = 1
\end{aligned}$$

Yule-Walker AR Parameter Estimation

Thus we have the system of equations for each $j = 0, 1, 2, \dots$

$$\begin{aligned} \gamma_Y(0) - \phi_1\gamma_Y(1) - \cdots - \phi_p\gamma_Y(p) &= \sigma^2, \\ \gamma_Y(1) - \phi_1\gamma_Y(0) - \cdots - \phi_p\gamma_Y(p-1) &= 0, \\ &\vdots \\ \gamma_Y(p) - \phi_1\gamma_Y(p-1) - \cdots - \phi_p\gamma_Y(0) &= 0. \end{aligned}$$

Where we can now obtain the Yule-Walker equations

$$\Gamma_p \boldsymbol{\phi} = \boldsymbol{\gamma_p}$$

with $\Gamma_p = [\gamma_Y(i-j)]_{i,j=1}^p$, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)^T$, $\boldsymbol{\gamma}_p = (\gamma_Y(1), \dots, \gamma_Y(p))^T$ and

$$\sigma^2 = \gamma_Y(0) - \boldsymbol{\phi}^T \boldsymbol{\gamma}_p.$$

This produces the Yule-Walker estimators

$$\begin{aligned} \widehat{\boldsymbol{\Gamma}}_p \widehat{\boldsymbol{\phi}} &= \widehat{\boldsymbol{\gamma}}_p, \\ \widehat{\sigma}^2 &= \widehat{\gamma}_Y(0) - \widehat{\boldsymbol{\phi}}^T \widehat{\boldsymbol{\gamma}}_p. \end{aligned}$$

Yule-Walker AR Parameter Estimation

Now observe that we have

$$\widehat{\gamma}_Y(k) = \begin{cases} T^{-1}\sum_{t=1}^{T-|k|}\left(Y_{t+|k|} - \overline{Y}_T\right)\left(Y_t - \overline{Y}_T\right), & \text{if } |k| < T, \\ 0, & \text{if } |k| \geq T, \end{cases}$$

furthermore, we know that $\widehat{\gamma}_Y$ is nonnegative definite. Thus $\gamma_Y(\cdot)$ is the autocovariance function of some stationary process i.e.

"*A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and non-negative definite.*"

Furthermore, we have learnt that

" *If $\{Y_t\}$ is a stationary q-correlated time series (i.e. $Cov(Y_s, Y_t) = 0$ whenever $|s-t| > q$ with mean 0, then it can be represented as an MA(q) process.* "

this stationary process must be an MA(T-1) process. Finally, based on the fact that we learnt

" *For a stationary process, if $\gamma_Y(0) > 0$ and $\gamma_Y(k) \to 0$ as $k \to \infty$, then the covariance matrix $\mathbf{\Gamma}_T = [\gamma_Y(i-j)]_{i,j=1}^T$ is positive definite for every T* "

Yule-Walker AR Parameter Estimation

Hence, if $\widehat{\gamma}_Y(0) > 0$, then it can be shown that $\widehat{\boldsymbol{\Gamma}}_p$ is non-singular.

Dividing by $\gamma_Y(0)$, we therefore obtain

$$\begin{aligned} \widehat{\boldsymbol{\phi}} &= \widehat{\boldsymbol{R}}_p^{-1} \widehat{\boldsymbol{\rho}}_p \\ \widehat{\sigma}^2 &= \widehat{\gamma}_Y(0) \left(1 - \widehat{\boldsymbol{\rho}}_p^T \widehat{\boldsymbol{R}}_p^{-1} \widehat{\boldsymbol{\rho}}_p \right) \end{aligned}$$

where $\widehat{\boldsymbol{\phi}} = (\widehat{\rho}(1), \ldots, \widehat{\rho}(p)) = \frac{\widehat{\gamma}_p}{\widehat{\gamma}_Y(0)}$

Remark

A feature of the Yule-Walker estimator is that the fitted model

$$Y_t - \widehat{\phi}_1 Y_{t-1} - \cdots - \widehat{\phi}_p Y_{t-p} = \epsilon_t, \;\; \{\epsilon_t\} \sim WN\left(0, \widehat{\sigma}^2\right)$$

is also causal. And the fitted models ACVF is $\widehat{\gamma}_Y(k)$ *for* $k = 0, 1, \ldots, p$ *(but in general different for higher lags).*

Yule-Walker AR Parameter Estimation

Proposition (Reminder on Performing Parameter Estimation)

From Yule-Walker:

$$\boldsymbol{\rho} = \boldsymbol{R}\boldsymbol{\phi}\,.$$

In practice, we can use sample ACF $\hat{\boldsymbol{\rho}}$ *and solve for* $\hat{\boldsymbol{\phi}}$ *to estimate* $AR(p)$ *parameters:*

$$\hat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{\rho}}.$$

Example

Consider $AR(1)$*:*

$$\hat{\phi}_1 = \hat{\rho}(1)\,.$$

Example

Consider $AR(2)$*:*

$$\begin{aligned}\hat{\phi}_1 &= \frac{\hat{\rho}(1) - \hat{\rho}(1)\hat{\rho}(2)}{1 - \hat{\rho}(1)^2}\\ \hat{\phi}_2 &= \frac{\hat{\rho}(2) - \hat{\rho}(1)^2}{1 - \hat{\rho}(1)^2}\,.\end{aligned}$$

Yule-Walker AR Parameter Estimation

Theorem

If $\{Y_t\}$ *is the causal AR(p) process with* $\{\epsilon_t\} \sim IID\left(0, \sigma^2\right)$, *then the Yule-Walker estimator* $\widehat{\phi}$ *enjoys that*

$$\sqrt{T}\left(\widehat{\phi} - \phi\right) \xrightarrow{d} N\left(0, \sigma^2 \mathbf{\Gamma}_p^{-1}\right)$$

and

$$\widehat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

Remark

The Yule-Walker estimator is based on a moment matching method and generally, moment matching can be far less efficient (higher variance estimator) than the MLE. However, one important feature of Yule-Walker estimator is that it is asymptotically ($T \to \infty$*) the same as the MLE estimator.*

Yule-Walker AR Parameter Estimation

One also has the following approximated 95% confidence interval for ϕ_j given by

$$\left[\phi_j - 1.96\sqrt{\widehat{\nu}_{jj}}/\sqrt{T}, \phi_j + 1.96\sqrt{\widehat{\nu}_{jj}}/\sqrt{T}\right]$$

where $\widehat{\nu}_{jj}$ is the j-th diagonal element of $\widehat{\sigma}^2\widehat{\boldsymbol{\Gamma}}_p^{-1}$ and a 95% confidence ellipsoid region for $\boldsymbol{\phi}$ given by

$$\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right)^T \widehat{\boldsymbol{\Gamma}} \left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \leq \chi^2_{0.95}(p)\frac{\sigma^2}{T}.$$

Yule-Walker AR Parameter Estimation

What if we don't know the correct order of the AR i.e. p and we select $m > p$?

Theorem

If $\{Y_t\}$ is the causal AR(p) process with $\{\epsilon_t\} \sim IID(0, \sigma^2)$ and if

$$\widehat{\boldsymbol{\phi}}_m = (\phi_{m1}, \ldots, \phi_{mm})^T = \widehat{\boldsymbol{R}}_m^{-1} \widehat{\boldsymbol{\rho}}_m, \quad m > p,$$

then

$$\sqrt{T}\left(\widehat{\boldsymbol{\phi}}_m - \boldsymbol{\phi}\right) \xrightarrow{d} N\left(0, \sigma^2 \boldsymbol{\Gamma}_m^{-1}\right)$$

where $\boldsymbol{\phi}_m$ is the coefficient vector of the best linear predictor $\boldsymbol{\phi}_m^T \boldsymbol{Y}_m$ of Y_{m+1} based on $\boldsymbol{Y}_m = (Y_m, \ldots, Y_1)^T$ i.e. $\boldsymbol{\phi}_m = \boldsymbol{R}_m^{-1} \boldsymbol{\rho}_m$. In particular for $m > p$

$$\sqrt{T}\widehat{\phi}_{mm} \xrightarrow{d} N(0, 1).$$

Remark

Hence, when fitting an AR model, the order p will be unknown. However, if the true order is p and we want to fit by order m, $m > p$, then we should expect the estimated coefficient vector $\widehat{\boldsymbol{\phi}} = (\phi_{m1}, \ldots, \phi_{mm})^T$ to have small values of $\widehat{\phi}_{mm}$ for each $m > p$.

Yule-Walker AR Parameter Estimation

We can also note that ϕ_{mm} is the PACF of $\{Y_t\}$ at lag m, recall, PACF is a good tool to identify AR series (while ACF is for MA processes). If $m > p$, we have from PACF that $\phi_{mm} = 0$.

As a general approach, one can select the p to be the smallest positive integer (say p_0) such that

$$|\widehat{\phi}_{mm}| < 1.96/\sqrt{T}, \quad \text{for } p_0 < m \leq K$$

where K is the maximum lag for a reasonable estimator of γ_Y i.e. $K \leq T/4$ and $T \geq 50$.

Durbin-Levinson Recursive Estimation

Just like in least squares regression estimation there is a recursive least squares method, there is also a recursive AR(p) parameter estimation method known as the Durbin-Levinson recursion.

Proposition

If $\widehat{\gamma}_Y(0) > 0$ then the fitted autoregressive model for $m = 1, 2, \ldots, T-1$ can be determined from the relations,

$$\begin{aligned}
\widehat{\phi}_{mm} &= \left\{ \widehat{\gamma}_Y(m) - \sum_{j=1}^{m-1} \widehat{\phi}_{m-1,j} \widehat{\gamma}_Y(m-j) \right\} \widehat{\nu}_{m-1}^{-1}, \\
\begin{pmatrix} \widehat{\phi}_{m1} \\ \widehat{\phi}_{m2} \\ \vdots \\ \widehat{\phi}_{m,m-1} \end{pmatrix} &= \begin{pmatrix} \widehat{\phi}_{m-1,1} \\ \widehat{\phi}_{m-1,2} \\ \vdots \\ \widehat{\phi}_{m-1,m-1} \end{pmatrix} - \widehat{\phi}_{mm} \begin{pmatrix} \widehat{\phi}_{m-1,m-1} \\ \widehat{\phi}_{m-1,m-2} \\ \vdots \\ \widehat{\phi}_{m-1,1} \end{pmatrix} \\
\widehat{\nu}_m &= \widehat{\nu}_{m-1} \left(1 - \widehat{\phi}_{mm}^2 \right).
\end{aligned}$$

with initialisation $\widehat{\phi}_{11} = \widehat{\gamma}_Y(1)/\widehat{\gamma}_Y(0)$ and $\widehat{\nu}_0 = \widehat{\gamma}_Y(0)$, where $\widehat{\phi}_{11}, \widehat{\phi}_{22}, \ldots, \widehat{\phi}_{mm}$ is the sample partial autocorrelation function at lags $1, 2, \ldots, m$.

Least Squares Estimation of AR Models

Least Squares Estimation AR Models

Example

Assume identification process implies data is $AR(1)$ (with possibly non-zero mean):

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t$$

In the estimation step of Box-Jenkins we try to estimate the parameters μ and ϕ_1 that fit the model, given some observations.

We have seen that a common approach is that of least squares

Least squares

Given observations, find 'best' values of parameters such that sum of squared differences between the expected ('predicted' c.f. regression) values of the model and the actual observations are minimised.

Least Squares Estimation AR Models

Example

Consider $AR(1)$ model $Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t$, and realisation (observed) $y_t = \mu + \phi_1(y_{t-1} - \mu) + \varepsilon_t$.

The value at time t expected by the model, given observations $\{y_1, y_2, \dots y_{t-1}\}$ is:

$$\mathbb{E}\left(Y_t \middle| Y_1, \dots, Y_{t-1}\right) = \mathbb{E}\left(\mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t \middle| Y_1, \dots, Y_{t-1}\right)$$

$$\begin{aligned}
&= \mu + \phi_1 \mathbb{E}\left(Y_{t-1} - \mu \middle| Y_1, \dots, Y_{t-1}\right) + \cancelto{0}{\mathbb{E}\left(\epsilon_t \middle| Y_1, \dots, Y_{t-1}\right)} \\
&= \mu + \phi_1 \\
&= \mu + \phi_1 \mathbb{E}(y_{t-1} - \mu) \\
&= \mu + \phi_1 (y_{t-1} - \mu)
\end{aligned}$$

The actual observed value at time t is, of course, y_t. The difference between observed and expected is:

$$y_t - (\mu + \phi_1(y_{t-1} - \mu)).$$

But this is equal to ε_t (see $AR(1)$ model). I.e.

$$\varepsilon_t = y_t - (\mu + \phi_1(y_{t-1} - \mu)).$$

Least Squares Estimation AR Models

$$\varepsilon_t = y_t - (\mu + \phi_1(y_{t-1} - \mu)) .$$

Remark

As we have seen previously: estimating the AR(1) parameters μ and ϕ_1 via least squares is equivalent to minimising the sum of square errors:

$$S(\mu, \phi_1) := \sum_{t=2}^{T} \left| y_t - (\mu + \phi_1(y_{t-1} - \mu)) \right|^2 = \sum_{t=2}^{T} \varepsilon_t^2$$

Note errors $\{\varepsilon_t\}$ can be written as:

$$\begin{bmatrix} \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_T \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_T \end{bmatrix} - \begin{bmatrix} 1 & y_1 - \mu \\ 1 & y_2 - \mu \\ \vdots & \vdots \\ 1 & y_{T-1} - \mu \end{bmatrix} \begin{bmatrix} \mu \\ \phi_1 \end{bmatrix}$$

I.e.

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} ,$$

where...

Least Squares Estimation AR Models

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}\,,$$

where $\boldsymbol{\varepsilon} := [\varepsilon_2, \ldots, \varepsilon_T]^T$, $\mathbf{y} := [y_2, \ldots, y_T]^T \in \mathbb{R}^{T-1}$, $\boldsymbol{\beta} := [\mu, \phi_1]^T \in \mathbb{R}^2$,

$$\text{and} \qquad \mathbf{X} := \begin{bmatrix} 1 & y_1 - \mu \\ 1 & y_2 - \mu \\ \vdots & \vdots \\ 1 & y_{T-1} - \mu \end{bmatrix}$$

Hence, the least squares approach finds the parameters $\boldsymbol{\beta}$ that minimise the sum of squares of

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}\,,$$

i.e. that minimise

$$S(\boldsymbol{\beta}) := \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{t=2}^{T} \left| y_t - \sum_{j=1}^{2} X_{t,j}\beta_j \right|^2,$$

with respect to $\boldsymbol{\beta}$.

Least Squares Estimation AR Models

Theorem (Time Series Least Squares)

Let $S(\hat{\boldsymbol{\beta}}) := \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$*, i.e. let* $\hat{\boldsymbol{\beta}}$ *be the parameter values that minimise the sum of squares, of* S*. Then,*

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\,.$$

Reminder of details (lecture 4-5) “Proof” Plan: find $\boldsymbol{\beta}$ s.t. $\dfrac{\partial S}{\partial \beta_j} = 0$.

$$\begin{aligned}
\frac{\partial S}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\sum_{t=2}^{T}\varepsilon_t^2 \\
&= \sum_{t=2}^{T}\frac{\partial \varepsilon_t^2}{\partial \beta_j} \\
&= \sum_{t=2}^{T}\frac{\partial \varepsilon_t^2}{\partial \varepsilon_t}\frac{\partial \varepsilon_t}{\partial \beta_j} \qquad \text{(chain rule)} \\
&= 2\sum_{t=2}^{T}\varepsilon_t\frac{\partial \varepsilon_t}{\partial \beta_j}
\end{aligned}$$

Least Squares Estimation AR Models

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_{t=2}^{T} \varepsilon_t \frac{\partial \varepsilon_t}{\partial \beta_j}$$

Now

$$\begin{aligned}
\frac{\partial \varepsilon_t}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j} \left(y_t - \sum_{k=1}^{2} X_{t,k} \beta_k \right) \\
&= -\sum_{k=1}^{2} X_{t,k} \frac{\partial \beta_k}{\partial \beta_j} \\
&= -\sum_{k=1}^{2} X_{t,k} \delta_{k,j} \\
&= -X_{t,j} \, ,
\end{aligned}$$

I.e.

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{t=2}^{T} \varepsilon_t X_{t,j} \, .$$

Least Squares Estimation AR Models

$$
\begin{aligned}
\frac{\partial S}{\partial \beta_j} &= -2 \sum_{t=2}^{T} \varepsilon_t X_{t,j} \\
&= -2 \mathbf{X}^T \boldsymbol{\varepsilon}\,, \qquad \text{[LHS = jth row of RHS]} \\
&= -2 \mathbf{X}^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)
\end{aligned}
$$

Let $\hat{\boldsymbol{\beta}}$ satisfy $\frac{\partial S}{\partial \beta_j} = 0$. Then

$$
\cancel{-2}\mathbf{X}^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = 0\,,
$$

i.e.

$$
\begin{aligned}
\mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}\,, \qquad \text{(a.k.a. ‘normal equations’)} \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}\,. \qquad \blacksquare
\end{aligned}
$$

Least Squares Estimation AR Models

Note, $\mathbf{X}^T\mathbf{X}\hat{\beta}$ is given by

$$\begin{aligned}\mathbf{X}^T\mathbf{X}\hat{\beta} &= \begin{bmatrix} 1 & \dots & 1 \\ y_1-\hat{\mu} & \dots & y_{T-1}-\hat{\mu} \end{bmatrix} \begin{bmatrix} 1 & y_1-\hat{\mu} \\ \vdots & \vdots \\ 1 & y_{T-1}-\hat{\mu} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\phi}_1 \end{bmatrix} \\ &= \begin{bmatrix} T-1 & \sum_{t=1}^{T-1} y_t-\hat{\mu} \\ \sum_{t=1}^{T-1} y_t-\hat{\mu} & \sum_{t=1}^{T-1}(y_t-\hat{\mu})^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\phi}_1 \end{bmatrix}.\end{aligned}$$

Also

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & \dots & 1 \\ y_1-\hat{\mu} & \dots & y_{T-1} \end{bmatrix} \begin{bmatrix} y_2 \\ \vdots \\ y_T \end{bmatrix}.$$

Then, the first row of the normal equations $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$, gives

$$\begin{aligned}(T-1)\hat{\mu} + \hat{\phi}_1 \sum_{t=1}^{T-1}(y_t-\hat{\mu}) &= \sum_{t=2}^{T} y_t \\ (T-1)\hat{\mu} - \hat{\phi}_1(T-1)\hat{\mu} + \hat{\phi}_1 \sum_{t=1}^{T-1} y_t &= \sum_{t=2}^{T} y_t\end{aligned}$$

Least Squares Estimation AR Models

$$
\begin{aligned}
(T-1)\hat{\mu} - \hat{\phi}_1 (T-1)\hat{\mu} + \hat{\phi}_1 \sum_{t=1}^{T-1} y_t &= \sum_{t=2}^{T} y_t \\
(T-1)(1-\hat{\phi}_1)\hat{\mu} + \hat{\phi}_1 \sum_{t=1}^{T-1} y_t &= \sum_{t=2}^{T} y_t \,.
\end{aligned}
$$

I.e.

$$\hat{\mu} = \frac{1}{(1-\hat{\phi}_1)(T-1)} \left(\sum_{t=2}^{T} y_t - \hat{\phi}_1 \sum_{t=1}^{T-1} y_t \right)$$

But, for large T,

$$\sum_{t=2}^{T} y_t \approx \sum_{t=1}^{T-1} y_t \,.$$

i.e. for large T,

$$\hat{\mu} \approx \frac{\cancel{(1-\hat{\phi}_1)}}{\cancel{(1-\hat{\phi}_1)}(T-1)} \sum_{t=2}^{T} y_t = \overline{Y}$$

Least Squares Estimation AR Models

Recall

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} T-1 & \sum_{t=1}^{T-1} y_t - \hat{\mu} \\ \sum_{t=1}^{T-1} y_t - \hat{\mu} & \sum_{t=1}^{T-1} (y_t - \hat{\mu})^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\phi}_1 \end{bmatrix},$$

and

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & \dots & 1 \\ y_1 - \hat{\mu} & \dots & y_{T-1} - \hat{\mu} \end{bmatrix} \begin{bmatrix} y_2 \\ \vdots \\ y_T \end{bmatrix}.$$

Then, second row gives:

$$\sum_{t=2}^{T} y_t(y_{t-1} - \hat{\mu}) = \hat{\mu} \sum_{t=1}^{T-1} (y_t - \hat{\mu}) + \hat{\phi}_1 \sum_{t=1}^{T-1} (y_t - \hat{\mu})^2$$

$$\hat{\phi}_1 = \frac{\sum_{t=2}^{T} y_t(y_{t-1} - \hat{\mu}) - \hat{\mu} \sum_{t=1}^{T-1} (y_t - \hat{\mu})}{\sum_{t=1}^{T-1} (y_t - \hat{\mu})^2}$$

$$\hat{\phi}_1 = \frac{\sum_{t=2}^{T} (y_t - \hat{\mu})(y_{t-1} - \hat{\mu})}{\sum_{t=1}^{T-1} (y_t - \hat{\mu})^2} \approx \hat{\rho}(1)$$

Not too surprising; recall $\rho(k) = \phi_1^{|k|}$ for (1).

Least Squares Estimation AR Models

Corollary

Let $\{Y_t\}$ *be* $AR(p)$. *Define*

$$\begin{aligned}
\boldsymbol{\varepsilon} &:= [\varepsilon_{p+1}, \varepsilon_{p+2}, \dots, \varepsilon_T]^T \in \mathbb{R}^{T-p} \\
\mathbf{y} &:= [y_{p+1}, y_{p+2}, \dots, y_T]^T \in \mathbb{R}^{T-p} \\
\boldsymbol{\beta} &:= [\mu, \phi_1, \dots, \phi_p]^T \in \mathbb{R}^{p+1},
\end{aligned}$$

and

$$\mathbf{X} := \begin{bmatrix} 1 & y_p - \mu & y_{p-1} - \mu & \dots & y_1 - \mu \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_{p+T-1} - \mu & y_{p+T-2} - \mu & \dots & y_T - \mu \end{bmatrix} \in \mathbb{R}^{(T-p)\times(T-p)},$$

with $T \gg 2p$. *Then,*

$$S(\hat{\beta}) = \min_{\beta} S(\beta) \Leftrightarrow \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,.$$

Least Squares Estimation AR Models

Alternative way to handle non-zero mean

Consider $AR(p)$: $Y_t - \mu = \epsilon_t + \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu)$. Then

$$\begin{aligned} Y_t &= \epsilon_t + \mu(1 - \phi_1 - \ldots - \phi_p) + \sum_{j=1}^{p} \phi_j Y_{t-j} \\ &=: \epsilon_t + \phi_0 + \sum_{j=1}^{p} \phi_j Y_{t-j} \,. \end{aligned}$$

Now, minimise

$$\begin{bmatrix} y_{p+1} \\ \vdots \\ y_{p+T} \end{bmatrix} - \begin{bmatrix} 1 & y_p & y_{p-1} & \ldots & y_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_{p+T-1} & y_{p+T-2} & \ldots & y_T \end{bmatrix} \begin{bmatrix} \phi_0 \\ \vdots \\ \phi_p \end{bmatrix},$$

in the least squares sense with respect to $\boldsymbol{\beta} := [\phi_0, \ldots, \phi_p]^T$. Then,

$$\hat{\mu} = \frac{\hat{\phi}_0}{1 - \sum_{j=1}^{p} \hat{\phi}_j} \,.$$

Least Squares Estimation AR Models

Remark

Recall, for $AR(1)$, $\hat{\mu} \approx \overline{Y}$, and $\hat{\phi}_1 \approx \hat{\rho}(1)$. I.e. least squares and moments methods give approx. same result. This also also holds for the general $AR(p)$ case.

Remark

It can be shown, asymptotically , that t-statistics can be used to test whether or not the 'true' value of a parameter is zero (c.f. regression).

Example ($AR(1)$ process with zero-mean)

This will compute summary statistics, which include:

N	*Mean*	*SEMean*	*StDev*	*Minimum*	*Q1*	*Median*	*Q3*
512	0.2179	0.0616	1.3946	−3.9501	−0.6794	0.2181	1.1252

Note, plot of series (and ACF and PACF) $\Rightarrow$ stationary. Note also mean $\approx$ median and quartiles roughly equidistant from median $\Rightarrow$ symmetric (Gaussian assumption might be appropriate).

Unit Root Tests for Stationarity

Now that we have seen the least squares estimator for AR(1) models, we can see how this can be applied to also derive a method to test for stationarity and therefore whether any differencing is required on a process.

Unit Root tests help determine if differencing is required, for example consider Y_t follows an ARIMA(0,1,0) model:

$$(1-B)Y_t = \epsilon_t$$

This can be views as a special case of $AR(1)$ process as follows:

$$(1-\phi B)Y_t = \epsilon_t, \quad \text{where } \phi = 1.$$

Hence, one can design a test for null hypothesis $\phi = 1$. Dickey and Fuller in 1979 designed a test statistic for this based on least squares which they used to estimate not ϕ but a reparameterisation $\phi^* = \phi - 1$ and they tested the null hypothesis $\phi^* = 0$. The method is based on equation

$$\begin{aligned} \nabla Y_t &= Y_t - Y_{t-1} \\ &= (\phi Y_{t-1} + \epsilon_t) - Y_{t-1} \\ &= (\phi - 1) Y_{t-1} + \epsilon_t \\ &= \phi^* Y_{t-1} + \epsilon_t \end{aligned}$$

Unit Root Tests for Stationarity

The Dickey-Fuller unit root test involves using ordinary least squares (OLS) to regress Y_t on Y_{t-1}. Note that if the mean of Y_t is μ but not 0, then one has slight extension

$$\nabla Y_t = \phi_0^* + \phi_1^* Y_{t-1} + \epsilon_t$$

where $\phi_0^* = \mu(1-\phi)$ and $\phi^* = \phi - 1$.

The goal is then to test

$$H_0 : \phi_1^* = 0 \quad \text{vs } H_a : \phi_1^* < 0.$$

Note: we don't need to consider $\phi_1^* > 0$ as this corresponds to the non-causal AR(1) model.

A small change of variable: define $\widetilde{Y}_t = Y_t - Y_{t-1}$ and $Z_t = Y_{t-1}$ to rewrite the regression model as

$$\widetilde{Y}_t = \phi_0^* + \phi_1^* Z_t + \epsilon_t, \quad t = 1, 2, \ldots, T$$

Unit Root Tests for Stationarity

We can the perform least squares estimation to obtain estimator for parameters of interest

$$\widehat{\phi}_1^* = \frac{\sum_{t=2}^T \left(\widetilde{Y}_t - \overline{Z}\right)\left(\widetilde{Y}_{t-1} - \overline{Z}\right)}{\sum_{t=2}^T \left(\widetilde{Y}_{t-1} - \overline{Z}\right)^2} - 1$$

and we can derive the standard error of this estimator to obtain

$$\widehat{SE}\left(\widehat{\phi}_1^*\right) = \left\{\frac{\sum_{t=2}^T \left(\nabla Y_t - \widehat{\phi}_0^* - \widehat{\phi}_1^* Y_{t-1}\right)^2}{(n-3)\sum_{t=2}^T \left(\widetilde{Y}_{t-1} - \overline{Z}\right)^2}\right\}^{1/2}$$

In addition, we can obtain estimator for the intercept given by

$$\widehat{\phi}_0^* = \frac{1}{T-1}\left(\widetilde{Y}_T - \widetilde{Y}_1 - \widehat{\phi}_1^* \sum_{t=1}^{T-1} \widetilde{Y}_t\right).$$

Unit Root Tests for Stationarity

Note the test statistic for

$$H_0 : \phi_1^* = 0 \quad \text{vs} \quad H_a : \phi_1^* < 0$$

is given by a Wold type statistic (analogous to familiar t-test statistic but not t-distributed in this case)

$$t = \frac{\widehat{\phi_1^*} - \phi_1^*(H_0)}{\widehat{SE}\left(\widehat{\phi_1^*}\right)} = \frac{\widehat{\phi_1^*}}{\widehat{SE}\left(\widehat{\phi_1^*}\right)}$$

We reject null (have a unit root) in favor of alternative (AR(1) is appropriate) at level α is t falls below $100(1-\alpha)$ percentage point established for Dickey-Fuller test statistic under assumption that T is large.

The 1%, 5% and 10% critical points are -3.43,-2.86, and -2.57, respectively.

Method of Moments Estimation MA Models - Innovations Algorithm

Estimation of MA via Innovations

Consider estimation of θ_1, given the $MA(1)$ model

$$Y_t = \epsilon_t - \theta_1 \epsilon_{t-1} \, .$$

Could try to represent this as AR and us same least squares approach we used to estimate the AR parameters. $MA(1)$ as an AR is:

$$\epsilon_t = Y_t + \sum_{j=1}^{\infty} \theta_j Y_{t-j} \, .$$

Unfortunately, this is not practicable! Instead, from model, we have:

$$\epsilon_t = Y_t + \theta_1 \epsilon_{t-1} \, .$$

Hence, we try to mininimise the sum of squares:

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t + \theta_1 \varepsilon_{t-1})^2 \, .$$

Unfortunately, we cannot observe ε_t.

Estimation of MA via Innovations

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t + \theta_1 \varepsilon_{t-1})^2 \,.$$

However, we can substitute in:

$$\begin{aligned} \varepsilon_1 &= y_1 + \theta_1 \varepsilon_0 \\ \varepsilon_2 &= y_2 + \theta_1 \varepsilon_1 \\ &\vdots \\ \varepsilon_T &= y_T + \theta_1 \varepsilon_{T-1} \,, \end{aligned}$$

and then minimise sum of squares over a range of values for ε_0 (or simply assume $\varepsilon_0 = 0$.)

This can be extended to the general $MA(q)$ case but then either assume $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$ (or otherwise).

Estimation of MA via Innovations

Summary: We know that we can fit autoregressive models of orders $1, 2, \ldots$ to the data $y_1, \ldots, y_T$ by applying the Durbin-Levinson algorithm based on the sample autocovariances. Just like this, we can also fit moving average models,

$$Y_t = \epsilon_t + \widehat{\theta}_{m1}\epsilon_{t-1} + \cdots + \widehat{\theta}_{mm}\epsilon_{t-m}, \quad \epsilon_t \sim WN\left(0, \widehat{\nu}_m\right)$$

of orders $m = 1, 2, \ldots$ by using what is know as the innovations algorithm, where $\widehat{\boldsymbol{\theta}}_m = \left(\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mm}\right)^T$ and white noise variance $\widehat{\nu}_m$, $m = 1, 2, \ldots$ are given as follows

Proposition (Innovation Estimate of MA Parameters)

If $\widehat{\gamma}_Y(0) > 0$, we define the innovation estimates $\boldsymbol{\theta}_m$, ν_m for $m = 1, \ldots, T-1$ by the recursive relationships

$$\begin{aligned} \widehat{\theta}_{m,m-k} &= \widehat{\nu}_k^{-1}\left[\widehat{\gamma}_Y(m-k) - \sum_{j=0}^{k-1}\widehat{\theta}_{m,m-j}\widehat{\theta}_{k,k-j}\widehat{\nu}_j\right], \quad k = 0, \ldots, m-1 \\ \widehat{\nu}_m &= \widehat{\gamma}_Y(0) - \sum_{j=0}^{m-1}\widehat{\theta}^2_{m,m-j}\widehat{\nu}_j, \end{aligned}$$

with initialisation $\widehat{\nu}_0 = \widehat{\gamma}_Y(0)$.

Estimation of MA via Innovations

The asymptotic behaviour of the MA model parameters $\widehat{\boldsymbol{\theta}}_m$ is given as follows.

Theorem (Consistency Conditions)

Let $\{Y_t\}$ *be the causal invertible ARMA process* $\phi(B)Y_t = \theta(B)\epsilon_t$ *with* $\{\epsilon_t\} \sim IID(0, \sigma^2)$, *with* $\mathbb{E}\left[\epsilon_t^4\right] < \infty$ *and let* $\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, $|z| \leq 1$, *(with* $\psi_0 = 1$ *and* $\psi_j = 0$ *for* $j < 0$*). Then for any sequence of positive integers* $\{m(T) : T = 1, 2, \ldots\}$ *such that* $m < T$, $T \to \infty$ *and* $m = o\left(T^{1/3}\right)$ *as* $n \to \infty$, *we have for each* k

$$\sqrt{T}\left(\widehat{\theta}_{m1} - \psi_1, \widehat{\theta}_{m2} - \psi_2, \ldots, \widehat{\theta}_{mk} - \psi_k\right)^T \xrightarrow{d} N(0, A)$$

$$\widehat{\nu}_m \xrightarrow{p} \sigma^2$$

where $A = [a_{ij}]_{i,j=1}^{k}$ *given by*

$$a_{ij} = \sum_{r=1}^{\min(i,j)} \psi_{i-r}\psi_{j-r}.$$

Estimation of MA via Innovations

Remark

The difference between this innovations approach and the one based on the Durbin-Levinson algorithm for AR model.

- *For an AR(p) process, the Yule-Walker estimator* $\widehat{\boldsymbol{\phi}}_p = \left(\widehat{\phi}_{p1}, \ldots, \widehat{\phi}_{pp}\right)^T$ *is consistent for* $\boldsymbol{\phi}_p$ *as the sample size* $T \to \infty$.
- *However, for an MA(q) process, the estimator* $\widehat{\boldsymbol{\theta}}_q = \left(\widehat{\theta}_{q1}, \ldots, \widehat{\theta}_{qq}\right)$ *is not consistent for the true parameter vector* $\boldsymbol{\theta}_q$ *as* $T \to \infty$.

For consistency it is necessary to use the estimator $\left(\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mq}\right)$ *with* $\{m(T)\}$ *satisfying conditions in the previous Theorem on Consistency Conditions.*

Method of Moments Estimation of ARMA Models

Estimation of ARMA(p,q) Process

Now we are ready to estimate the ARMA(p,q) model combining these ideas. We have seen that for a causal invertible ARMA process $\phi(B)Y_t = \theta(B)\epsilon_t$, we can use innovations algorithm to obtain consistent estimators $\left(\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mk}\right)$ for each k of $\psi_1, \ldots, \psi_k$, where $\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$.

So given the estimate of $\Psi(z)$ coefficients, can we co back to $\theta(z)$ and $\phi(z)$?

Let $\{Y_t\}$ be the zero-mean causal ARMA(p,q) process,

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \quad \{\epsilon_t\} \sim WN(0, \sigma^2),$$

where the causality assumption gave us

$$Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}.$$

Estimation of ARMA(p,q) Process

Hence, based on

$$\sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)},$$

we can match the coefficients as $\phi_0 = 1$ and

$$\psi_j = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \psi_{j-i}, \quad j = 1, 2, \ldots$$

and by convention $\theta_j = 0$ for $j > q$ and $\phi_j = 0$ for $j > p$. So when $j > p$ we actually have

$$\psi_j = \sum_{i=1}^{p} \phi_i \psi_{j-i}, \quad j = q+1, q+2, \ldots, q+p$$

So we know how to estimate ψ_j's. Replacing them with their estimates we have

$$\widehat{\theta}_{mj} = \sum_{i=1}^{p} \phi_i \widehat{\theta}_{m,j-i}, \quad j = q+1, q+2, \ldots, q+p.$$

Then solving for ϕ_i, we have the estimator $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$

Estimation of ARMA(p,q) Process

The estimates of $\theta_1, \dots, \theta_q$ are then found easily from

$$\widehat{\theta}_j = \widehat{\theta}_{mj} - \sum_{i=1}^{\min(j,p)} \widehat{\phi}_i \widehat{\theta}_{m,j-i}, \;\; j = 1, 2, \dots, q.$$

Finally, the white noise variance σ^2 is estimated by

$$\widehat{\sigma}^2 = \widehat{\nu}_m.$$

Now by the consistency of $\widehat{\theta}_{mj} \xrightarrow{p} \psi_j$, where $m = m(T)$ satisfying the previously stated conditions in Theorem on Consistency Conditions to get

$$\widehat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}, \;\; \widehat{\theta} \xrightarrow{p} \boldsymbol{\theta}, \quad \text{and } \widehat{\sigma}^2 \xrightarrow{p} \sigma^2, \quad \text{as } T \to \infty.$$

Remark

However, the efficiency (asymptotic variance) of this moment-matching type estimator is somewhat poor. A more efficient estimation procedure (strictly more efficient if q $\geq 1)of\,(\phi;\boldsymbol{\theta})$ *is based on maximization of the Gaussian likelihood. Noting that, when* $q = 0$*, we have AR process in which we know that the Yule-Walker (based on moment-matching) estimator is the same efficiency as the MLE.*

Least Squares Estimation of ARMA Models

Estimation of ARMA(p,q) Process

Least squares estimation of *ARMA* models works in a similar way to the *MA* case.

Example

Consider zero mean $ARMA(1,1)$: $Y_t = \phi_1 Y_t + \epsilon_t - \theta_1 \epsilon_{t-1}$. *Then, we minimise*

$$S(\phi_1, \theta_1) = \sum_{t=1} (y_t - \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1})^2 .$$

Now $\varepsilon_t = y_t - \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1}$. *Hence, we use*

$$\begin{aligned} \varepsilon_1 &= y_1 - \phi_1 y_0 + \theta_1 \varepsilon_0 \\ &\vdots \\ \varepsilon_T &= y_T - \phi_1 y_{T-1} + \theta_1 \varepsilon_{T-1} \end{aligned}$$

and assume $y_0 = \varepsilon_0 = 0$.

Estimation of ARMA(p,q) Process

Remark

For $ARMA(p,q)$, like $MA(q)$, this approach requires us to assume $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{-q+1} = 0$.

Remark

This approach can be refined by fitting an $ARMA$ model to the time reversed version of $\{Y_t\}$ to predict past observations (a.k.a 'backcasting'), then refit model with these updated values.

Remark

Note, that many software packages estimate parameters using maximum likelihood-based methods instead of least squares. CAVEAT!: Some software

functions will assume data is normal and apply maximum likelihood methods.

Computationally Efficient Representation of a Gaussian Process Likelihood

Gaussian Process Likelihood

We will start with the case in which we assume that $\{Y_t\}$ is an arbitrary zero-mean Gaussian Process with covariance function $k(i,j) = \mathbb{E}[Y_i Y_j]$. Denote by $\boldsymbol{Y}_T = (Y_1, \ldots, Y_T)^T$ and Γ_T as the covariance of $\boldsymbol{Y}$ given by $\Gamma_T = \mathbb{E}\left[\boldsymbol{Y}_T, \boldsymbol{Y}_T^T\right]$ which is assumed to be non-singular. Then one has

$$\boldsymbol{Y}_T \sim N(\mathbf{0}, \Gamma_T)$$

and the likelihood of $\boldsymbol{Y}_T$ is given by

$$L(\Gamma_T) = (2\pi)^{-T/2} (\det \Gamma_T)^{-1/2} \exp\left(-\boldsymbol{Y}_T^T \Gamma_T^{-1} \boldsymbol{Y}_T / 2\right)$$

Gaussian Process Likelihood

Efficient Estimation of detΓ_T and Γ_T^{-1} can be achieved using the one-step predictors and their mean squared errors as follows.

Denote the one-step predictors of $\boldsymbol{Y}$ as $\widehat{\boldsymbol{Y}} = \left(\widehat{Y}_1, \dots, \widehat{Y}_T\right)$ where $\widehat{Y}_1 = 0$ and $\widehat{Y}_j = \mathbb{E}\left[Y_j | Y_1, \dots, Y_{j-1}\right]$, $j \geq 2$ and $\nu_{j-1} = \mathbb{E}\left[\left(Y_j - \widehat{Y}_j\right)^2\right]$, for $j = 1, 2, \dots, T$.

We know from innovations algorithm in previous section that

$$\widehat{Y}_{T+1} = \begin{cases} 0, & T = 0 \\ \sum_{j=1}^{T} \theta_{nj} \left(Y_{T+1-j} - \widehat{Y}_{T+1-j}\right), & T \geq 1, \end{cases}$$

now define the matrix

$$C = \begin{pmatrix} 1 & & & & \\ \theta_{11} & 1 & & & \\ \theta_{22} & \theta_{21} & 1 & & \\ \vdots & \vdots & \dots & \vdots & \\ \theta_{T-1,T-1} & \theta_{T-1,T-2} & \dots & \theta_{T-1,1} & 1 \end{pmatrix}$$

Gaussian Process Likelihood

We can now express

$$\widehat{\boldsymbol{Y}}_T = (C - I_T)\left(\boldsymbol{Y}_T - \widehat{\boldsymbol{Y}}_T\right),$$

hence

$$\boldsymbol{Y}_T = \boldsymbol{Y}_T - \widehat{\boldsymbol{Y}}_T + \widehat{\boldsymbol{Y}}_T = C\left(\boldsymbol{Y}_T - \widehat{\boldsymbol{Y}}_T\right).$$

Note that $\boldsymbol{Y}_T - \widehat{\boldsymbol{Y}}_T \sim N(0, D)$ where $D = \mathrm{diag}\left(\nu_0, \ldots, \nu_{T-1}\right)$, then we can express $\Gamma_T = CDC^T$ which gives

$$\boldsymbol{Y}_T^T \Gamma_T^{-1} \boldsymbol{Y}_T = \left(\boldsymbol{Y}_T - \widehat{\boldsymbol{Y}}_T\right)^T D^{-1}\left(\boldsymbol{Y}_T - \widehat{\boldsymbol{Y}}_T\right) = \sum_{j=1}^{T}\left(Y_j - \widehat{Y}_j\right)/\nu_j$$

and

$$\det \Gamma_T = \det C \times \det D \times \det C = \nu_0 \nu_1 \ldots \nu_{T-1}.$$

Hence, we can re-express the likelihood as follows

$$L\left(\Gamma_T\right) = (2\pi)^{-T/2}\left(\prod_{j=0}^{T-1}\nu_j\right)^{-1/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(Y_j - \widehat{Y}_j\right)^2}{\nu_{j-1}}\right)$$

Parameter Estimation of Linear Time Series Models via Maximum Likelihood Estimation

ARMA Process Likelihood

We will begin with the general process for the likelihood speicifcation for an ARMA process and the MLE under 6 basic steps

- Step 1: assume a distribution for the WN errors. Typically i.i.d. Normal (or sometimes student-t), lets assume i.i.d. Normal as default $\epsilon_t \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$.
- Step 2: write down the joint pdf for ϵ_t:

$$f\left(\epsilon_1, \ldots, \epsilon_T\right) = \prod_{t=1}^{T} f\left(\epsilon_t\right)$$

  Note: we are not writing the joint pdf in terms of the y_t's as a multiplication of the marginal pdfs because of the dependency in y_t.

- Step 3: Get ϵ_t for teh general stationary ARMA(p,q) model

$$\epsilon_t = y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

  Note: if $\mu_Y \neq 0$ demean Y_t.

ARMA Process Likelihood

- Step 4: the joint pdf for $\epsilon_1, \ldots, \epsilon_T$ is under Normal assumption

$$f\left(\epsilon_1, \ldots, \epsilon_T | \mu, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\epsilon_t^2\right\}$$

- Step 5: let $\boldsymbol{Y} = (y_1, \ldots, y_T)$ and assume that initial conditions $\boldsymbol{Y}_* = (y_{1-p}, \ldots, y_0)$ and $\epsilon_* = (\epsilon_{1-q}, \ldots, \epsilon_0)$ are known
- Step 6: the conditional log-likelihood function is given by

$$l\left(\mu, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2\right) = -\frac{T}{2}\ln\left(2\pi\sigma^2\right) - \frac{S_*\left(\mu, \boldsymbol{\phi}, \boldsymbol{\theta}\right)}{2\sigma^2}$$

  where $S_*\left(\mu, \boldsymbol{\phi}, \boldsymbol{\theta}\right) = \sum_{t=1}^{T}\epsilon_t^2\left(\mu, \boldsymbol{\phi}, \boldsymbol{\theta} | \boldsymbol{Y}, \boldsymbol{Y}_*, \boldsymbol{\epsilon}_*\right)$ is the conditional sum-of-squares (SS).

Note: usual initial conditions $\boldsymbol{y}_* = \overline{y}$ and $\epsilon_* = \mathbb{E}\left[\epsilon_t\right] = 0$.

For the numerical optimization problem, the initial values $\boldsymbol{y}_*$ matter.

AR(1) Likelihood Example

Example: AR(1) model

$$Y_t = \phi Y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$$

We can write down the joint density for ϵ_T

$$f\left(\epsilon_1, \ldots, \epsilon_T\right) = \left(2\pi\sigma^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\epsilon_t^2\right\}$$

Solve for ϵ_t:

$$\begin{aligned}
y_1 &= \phi y_0 + \epsilon_1, \quad \text{lets take } y_0 = 0 \text{ so } \epsilon_1 = y_1 \\
y_2 &= \phi y_1 + \epsilon_2 \Rightarrow \epsilon_2 = y_2 - \phi y_1 \\
y_3 &= \phi y_2 + \epsilon_3 \Rightarrow \epsilon_3 = y_3 - \phi y_2 \\
&\vdots \\
y_T &= \phi y_{T-1} + \epsilon_T \Rightarrow \epsilon_T = y_T - \phi y_{T-1}
\end{aligned}$$

AR(1) Likelihood Example

Now to change the joint density from $\boldsymbol{wn}_T$ to $\boldsymbol{y}_T$, we need the Jacobian

$$|J| = \begin{vmatrix} \frac{\partial \epsilon_2}{Y_2} & \frac{\partial \epsilon_2}{Y_3} & \cdots & \frac{\partial \epsilon_2}{Y_T} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \epsilon_T}{Y_2} & \frac{\partial \epsilon_T}{Y_3} & \cdots & \frac{\partial \epsilon_T}{Y_T} \end{vmatrix} = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ -\phi & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{vmatrix} = 1$$

which gives

$$f\left(y_2, \ldots, y_T | y_1\right) = f\left(\epsilon_2, \ldots, \epsilon_T\right) |J| = f\left(\epsilon_2, \ldots, \epsilon_T\right)$$

Then the likelihood function can be written as

$$\begin{aligned} L\left(\phi, \sigma^2\right) &= f\left(y_1, \ldots, y_T\right) = f\left(y_1\right) f\left(y_2, \ldots, y_T | y_1\right) = f\left(y_1\right) f\left(\epsilon_2, \ldots, \epsilon_T\right) \\ &= \left(\frac{1}{2\pi\gamma_0}\right)^{1/2} \exp\left\{-\frac{\left(y_1 - 0\right)^2}{2\gamma_0}\right\} \left(\frac{1}{2\pi\sigma^2}\right)^{(T-1)/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(y_t - \phi y_{t-1}\right)^2\right\} \end{aligned}$$

where $Y_1 \sim N\left(0, \gamma_0 = \frac{\sigma^2}{1-\phi^2}\right)$

AR(1) Likelihood Example

One can then group these terms as follows to get the log-likelihood

$$
\begin{aligned}
l\left(\phi,\sigma^2\right)=\ln L\left(\phi,\sigma^2\right) &= -\frac{T}{2}\ln(2\pi)-\frac{T}{2}\ln\sigma^2-\frac{1}{2}\ln\left(1-\phi^2\right) \quad (0.2)\\
&- \frac{1}{2\sigma^2}\underbrace{\left[\underbrace{\sum_{t=2}^{T}(y_t-\phi y_{t-1})^2}_{S_*(\phi)}+(1-\phi^2)y_1^2\right]}_{S(\phi)} \quad (0.3)
\end{aligned}
$$

where $S_*(\phi)$ is the unconditional sum-or-square errors and $S(\phi)$ is the conditional sum-of-square errors.
We now evaluation the Fisrt Order Conditions (F.O.C.'s) as follows

$$
\begin{aligned}
\frac{\partial \ln L\left(\phi,\sigma^2\right)}{\partial\phi} &= 0\\
\frac{\partial \ln L\left(\phi,\sigma^2\right)}{\partial\sigma} &= 0
\end{aligned}
$$

AR(1) Likelihood Example

Note: if we neglect $\ln(1-\phi^2)$ term, then the MLE estimator will instead become the unconditional Least-squares-estimator (LSE).

$$\max_{\phi} L\left(\phi,\sigma^2\right) = \min S(\phi)$$

If we neglect both $\ln(1-\phi^2)$ and $(1-\phi^2)y_1^2$, then the estimator becomes the conditional LSE (CLSE)

$$\max_{\phi} L\left(\phi,\sigma^2\right) = \min S_*(\phi_*)$$

ARMA Process Likelihood

We will now build upon the likelihood specification in the previous section to extend it from the Guassian process setting to the ARMA process context where we have the process given by

$$\phi_p(B)Y_t = \theta_q(B)\epsilon_t.$$

The one step predictor in this case is given again by the innovations algorithm and takes the form in the ARMA setting of

$$\widehat{Y}_{i+1} = \sum_{j=1}^{i} \theta_{ij}\left(Y_{i+1-j} - \widehat{Y}_{i+1-j}\right), \quad 1 \leq i < m = \max(p,q).$$

and

$$\widehat{Y}_{i+1} = \phi_1 Y_1 + \cdots + \phi_p Y_{i+1-p} + \sum_{j=1}^{q} \theta_{ij}\left(Y_{i+1-j} - \widehat{Y}_{i+1-j}\right), \quad i \geq m,$$

with

$$\mathbb{E}\left[\left(Y_{i+1} - \widehat{Y}_{i+1}\right)^2\right] = \sigma^2 r_i.$$

ARMA Process Likelihood

The likelihood for the ARMA(p,q) case can then be derived as follows

$$L\left(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2\right)=(2\pi)^{-T/2}\left(\prod_{j=0}^{T-1} r_j\right)^{-1/2}\exp\left(-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(Y_j-\widehat{Y}_j\right)^2}{\sigma^2 r_{j-1}}\right)$$

which can be differentiated to find a system of equations to solve for the MLE estimators of the ARMA(p,q) model parameters as follows, differentiate w.r.t σ^2 and set equal zero, to get

$$\widehat{\sigma}^2=T^{-1}S\left(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\theta}}\right)$$

where

$$S\left(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\theta}}\right)=\sum_{j=1}^{T}\frac{\left(Y_j-\widehat{Y}_j\right)^2}{r_{j-1}}$$

and $\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\theta}}$ are the values of $\boldsymbol{\phi},\boldsymbol{\theta}$ which minimize what is known as the reduced likelihood (function of $\boldsymbol{\phi},\boldsymbol{\theta}$ only), given by

$$l(\boldsymbol{\phi},\boldsymbol{\theta})=\ln\left\{T^{-1}S\left(\boldsymbol{\phi},\boldsymbol{\theta}\right)\right\}+T^{-1}\sum_{j=1}^{T}\ln r_{j-1}.$$

ARMA Process Least Squares

We start with the causal condition, so it is better to search for ϕ that makes the sequence casual. However, invertible is not required, but you can also look for parameters that make the process invertible also.

Often people will minimise an alternative to the reduced likelihood given simply by the simple weighted sum of squares

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{j=1}^{T} \frac{\left(Y_j - \widehat{Y}_j\right)^2}{r_{j-1}}$$

to obtain estimators, denoted $\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}$ which are known as the **least squares estimators** and which then produce WN variance estimator

$$\widetilde{\sigma}^2 = \frac{S\left(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}\right)}{T - p - q}$$

ARMA MLE Estimator Properties

We assume we have a causal invertible process $\{Y_t\}$ and that $\{\epsilon_t\} \overset{i.i.d.}{\sim} N(0,\sigma^2)$. Then if we denote the MLE estimators for the AR and MA parameters by $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\phi}}^T, \widehat{\boldsymbol{\theta}}^T\right)^T$, we can state the following CLT as sample size $T \to \infty$

$$\sqrt{T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta})\right)$$

where for $p \geq 1$ and $q \geq 1$ one has asymptotic covariance given by

$$\boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \begin{pmatrix} \mathbb{E}\left[\boldsymbol{U}_t \boldsymbol{U}_t^T\right] & \mathbb{E}\left[\boldsymbol{U}_t \boldsymbol{V}_t^T\right] \\ \mathbb{E}\left[\boldsymbol{V}_t \boldsymbol{U}_t^T\right] & \mathbb{E}\left[\boldsymbol{V}_t \boldsymbol{V}_t^T\right] \end{pmatrix}^{-1}$$

where $\boldsymbol{U}_t = (U_t, \ldots, U_{t+1-p})$ and $\boldsymbol{V}_t = (V_t, \ldots, V_{t+1-q})$ and

$$\begin{aligned} \phi(B)\boldsymbol{U}_t &= \epsilon_t \\ \theta(B)\boldsymbol{V}_t &= \epsilon_t \end{aligned}$$

Model Assessment and Residual Analysis for Fitted ARIMA(p,d,q) Model

ARIMA Residual Analysis

Suppose a candidate model has been chosen and that the unknown parameters have been estimated. The residuals:

$$(\text{residuals} = \text{actual observation} - \text{fitted value}) ,$$

can be used to help verify whether the fitted model is appropriate.

Example

Consider zero-mean $AR(1)$ model $Y_t = \phi_1 Y_{t-1} + \epsilon_t$. The parameter ϕ_1 is estimated, by least squares, to be $\hat{\phi}_1$. Then, the residual is

$$\hat{\varepsilon}_t = y_t - \hat{\phi}_1 y_{t-1} ,$$

i.e. an estimate of the white noise sequence $\{\varepsilon_t\}$. If the model is good, then $\hat{\varepsilon}$ will

1. *have constant zero mean*
2. *have constant variance*
3. *be uncorrelated*

1 and 2 can be checked visually (plot $\hat{\varepsilon}_t$). 3 can be checked in various ways...

Box-Pierce statistics

Recall, that the sample ACF of white noise $\hat{\rho}_\epsilon \overset{approx}{\sim} \mathcal{N}(0, 1/T)$. A similar result holds for the sample PACF.

Hence, plot the ACF and PACF of the residuals $\{\hat{\varepsilon}_t\}$, together with the approximate 95% confidence intervals at $\pm 2/\sqrt{T}$, where T is the length of the sequence after any differencing has been applied (e.g., note performing the difference operator ∇ reduces the number of data points by one).

Assuming $\hat{\rho}_\epsilon \sim \mathcal{N}(0, 1/T)$, then

$$Q_K := T \sum_{k=1}^{K} \hat{\rho}_\epsilon^2(k) \sim \chi_K^2 \,,$$

i.e. Q has a χ^2 distribution with K many degrees of freedom and is used to test

$$H_0 \colon \; \rho_\epsilon(1) = \rho_\epsilon(2) = \cdots = \rho_\epsilon(K) = 0 \,.$$

Box-Pierce statistics

Assume $\hat{\rho}_\epsilon \sim \mathcal{N}(0, 1/T)$:

Box-Pierce statistic

Box & Pierce showed that the statistic

$$Q_K := T \sum_{k=1}^{K} \hat{\rho}_\epsilon^2(k) \sim \chi^2_{K-p-q} \, ,$$

is a better approximation. I.e. Q has a χ^2 distribution with $K - p - q$ many degrees of freedom, where p and q are the number of and terms in the model being tested.

Ljung-Box-Pierce (modified Box-Pierce) statistic

Ljung & Box then showed that a better approximation is:

$$Q^*_K := T(T+2) \sum_{k=1}^{K} \frac{\hat{\rho}_\epsilon^2(k)}{T-k} \sim \chi^2_{K-p-q} \, ,$$

Box-Pierce statistics

The Ljung-Box-Pierce statistic Q_K^* (and others) are used to test the null hypothesis

$$H_0: \; \rho_\epsilon(1) = \rho_\epsilon(2) = \cdots = \rho_\epsilon(K) = 0 \, .$$

A value of Q_K^* greater than, say, the 95 percentile of χ^2_{K-p-q} would cast doubt (at the 5% level) on the null hypothesis.

Example

Consider data captured monthly. Then compute Q_K^*, for $K = 12, 24, 36, 48$. If H_0 is accepted for $K = 12$ and 24 but rejected for $K = 36$ we might suspect some seasonal autocorrelations between the months of year 2 and 3. The next step would be to add these autocorrelations into an improved model and try again.

Box-Pierce statistics

Example of interpretation for an MA(1) model:

Type	*Coef*	*SECoef*	*T*	*P*
MA 1	0.7510	0.0293	25.65	0.000
Constant	0.99083	0.01292	76.70	0.000

p-values of parameters look good (can reject $\theta_1 = 0$).

Modified Box-Pierce (Ljung-Box) Chi-Square statistic:

Lag	12	24	36	48
Chi-Square	57.6	72.5	85.7	101.4
P-Value	0.000	0.000	0.000	0.000

But Ljung-Box looks bad! We must reject, e.g., $\rho(1) = \ldots \rho(12) = 0$. Hence, autocorrelations still exist.

After looking at the ACF of **residuals**: you see a spike at lag 2

Conclusion: $\Rightarrow$ might want to try an $MA(2)$ model...

Box-Pierce statistics

For the same data you now fit an MA(2) model with a non-zero constant mean. It gives outcomes:

Type	*Coef*	*SECoef*	*T*	*P*
MA 1	0.4922	0.0415	11.87	0.000
MA 2	0.3489	0.0417	8.38	0.000
Constant	0.990203	0.007823	126.57	0.000

Modified Box-Pierce (Ljung-Box) Chi-Square statistic:

Lag	12	24	36	48
Chi-Square	4.5	15.1	31.9	44.0
P-Value	0.877	0.820	0.520	0.516

p-values of parameters look good (can reject $\theta_1, \theta_2 = 0$). Now Ljung-Box looks much better! We can accept null hypothesis that no significant autocorrelations exist in residuals. You should also check the ACF and PACF!

Overfitting Models

If $MA(2)$ seems appropriate, it might be a good idea to try to 'overfit' the data with an $MA(3)$ model or an $ARMA(1,2)$ model (depending on how ambiguous the ACF and PACF of y_t or the residuals are).

The parameter t- statistics can be used as a guide as to when to stop overfitting, c.f. Example **??**.

However, t-tests are sometimes ambiguous...

Example

A plausible practical example

- *You fit an $AR(2)$ model and find $\hat{\phi}_1$, $\hat{\phi}_2$ are significantly different from zero.*
- *You overfit with an $ARMA(2,1)$. Now $\hat{\phi}_2$ and $\hat{\theta}_1$ are* **not** *significantly different from zero.*
- *You overfit again with $ARMA(2,2)$. Now, all parameters are significantly different from zero except $\hat{\theta}_2$.*

Model Selection Criterion

To help deal with such ambiguity, we need a way to compare models quantitatively. A naive way would be to compute...

Definition (R^2-statistic)

Let s_y^2 be the sample variance of the data and s_ε^2 be the sample variance of the residuals, after some model is fitted. The R^2 statistic is defined as

$$R^2 := 1 - \frac{s_\varepsilon^2}{s_y^2} .$$

Remark

- *But: more model parameters ⇒ better R^2. Hence, basing model choice on R^2 will tend to lead to overfitting.*
- *We need to penalise number of parameters.*
- *C.f. principle of parsimony: if two models fit the data with (approx.) the same error, choose the most simple one (fewest parameters, in our case).*

Model Selection in the ARIMA(p,d,q) Family

ARMA(p,q) process ACF Example

SUMMARY: model identification with the ACF

Model	ACF
$AR(1)$	$\rho(k) = \phi_1^k$: exponential decay for $0 < \phi_1 < 1$ (alternating exponential decay if $-1 < \phi_1 < 0$)
$AR(p)$	exponential decay or damped sine wave
$MA(1)$	$\rho(1) = \frac{-\theta_1}{1+\theta_1^2}$: 'spike' at lag 1, then 0 for lags ≥ 2 (spike is positive if $\theta_1 < 0$ and negative if $\theta_1 > 0$)
$MA(q)$	spikes at lags 1 to q and 0 for lags $\geq q+1$
$ARMA(p,q)$	exponential decay or damped sine wave

ARMA(p,q) process IACF Example

The Inverse Autocorrelation Function (IACF) and plays much the same role in ARIMA modelling as the PACF. The IACF of the ARMA(p,q) model

$$\phi(B)Y_t = \theta(B)\epsilon_t$$

is defined to be (assuming invertibility) the ACF of the inverse (or dual) process

$$\Theta(B)Y_t^{-1} = \phi(B)\epsilon_t$$

- The IACF has the same property as the PACF: AR(p) is characterised by an IACF that is non-zero at lag p but zero for higher lags.
- The IACF can also be used to detect over-differencing. If the data come from a nonstationary or nearly nonstationary model, the IACF has the characteristics of a noninvertible moving-average.

ARMA Model Selection Criterion

One needs to select the model orders p, q and the parameters $\boldsymbol{\phi}_p, \boldsymbol{\theta}_q$ which is often done by minimising a penalised information criterion such as the Akaike Information Criterion (AIC) or small sample corrected version AICC.

Definition (Akaike Information Criterion (& small sample corrected AICC))

One should select p, q and the parameters $\boldsymbol{\phi}_p, \boldsymbol{\theta}_q$ to minimise the penalised information criterion given by the log-likelihood and penalty term as follows ($M = (p + q + 1)$ number of fitted model parameters):

$$AICC = -2 \ln L \left\{ \boldsymbol{\phi}_p, \boldsymbol{\theta}_q, \frac{S\left(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q\right)}{T} \right\} + 2M.$$

or small sample corrected version (AICC)

$$AICC = -2 \ln L \left\{ \boldsymbol{\phi}_p, \boldsymbol{\theta}_q, \frac{S\left(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q\right)}{T} \right\} + 2\frac{M(M+1)}{T - M - 1}.$$

ARMA Model Selection Criterion

Another popular information criterion often used to select the model orders p, q and the parameters $\phi_p, \boldsymbol{\theta}_q$ is the Bayesian Information Criterion (BIC)

Definition (Bayesian Information Criterion)

One should select p, q and the parameters $\phi_p, \boldsymbol{\theta}_q$ to minimise the penalised information criterion given as follows ($M = (p + q + 1)$ number of fitted model parameters):

$$BIC = -2 \ln L \left\{ \phi_p, \boldsymbol{\theta}_q, \frac{S\left(\phi_p, \boldsymbol{\theta}_q\right)}{T} \right\} + M \ln(T)$$

ARMA Model Selection Criterion

A third popular information criterion often used to select the model orders p, q and the parameters $\phi_p, \boldsymbol{\theta}_q$ is the Hannan-Quinn (HQIC)

Definition (Hannan-Quinn Criterion)

One should select p, q *and the parameters* $\phi_p, \boldsymbol{\theta}_q$ *to minimise the penalised information criterion given as follows (*$M = (p + q + 1)$ *number of fitted model parameters):*

$$HQIC = -2 \ln L \left\{ \phi_p, \boldsymbol{\theta}_q, \frac{S\left(\phi_p, \boldsymbol{\theta}_q\right)}{T} \right\} + 2M \ln\left[\ln(T)\right]$$

ARMA Model Selection Criterion

There is no agreement on which criteria is best. The AIC is the most popular, but others are also used regularly such as the BIC.

- The **AIC is not consistent**, generally producing too large a model, but is more efficient –i.e., when the true model is not in the candidate model set the AIC asymptotically chooses whichever model minimizes the MSE/MSPE.
- Asymptotically, the **BIC is consistent** –i.e., it selects the true model if, among other assumptions, the true model is among the candidate models considered. For instance in the case of common roots in the AR and MA polynomials, the BIC still select the correct orders p and q consistently. But, **BIC it is not efficient**.
- It can be shown that in the case of common roots in the AR and MA polynomials, the Hannan-Quinn and Schwarz criteria (BIC) still select the correct orders p and q consistently.

Model Criticism and Selection Framework

1. plot series
2. plot ACF, PACF
3. Check mean (constant terms? — if this is 0 after differencing, you might have overdifferenced) & variance (if, e.g. $\text{var}(\nabla Y) >> \text{var}(Y)$ you might have overdifferenced), symmetry of distribution (e.g. t-tests and especially AIC work 'better' on normal distribution)
4. Check non-stationarity. If non-stat., take $\nabla_{?}^{??}$ and goto 1
5. Make a shortlist of a few candidate models. Fit models and output residuals for assessment.
6. Check p-values of estimated parameters to check whether you have overfitted (i.e. used a model with too many parameters)
7. check residuals (plot time series, ACF, PACF), R^2, Ljung-Box-Pierce. (ACF or PACF might suggest why your model is wrong.)
8. If all looks good, try overfitting (and goto 6)
9. Pick a few of the better models and compute AIC, AIC_C, and/or BIC.
10. Give one, and only one, answer! Don't just follow this checklist! Use (a little bit of) your own initiative!!!