Linear Time Series Analysis
Lecture 3: Estimation of ARMA Models

Daniel Buncic

Institute of Mathematics & Statistics
University of St. Gallen
Switzerland

February 5, 2016
Version: [ltsa3]

Homepage
www.danielbuncic.com
University of St. Gallen

Outline

Fitting ARMA models

There are various statistical methods of fitting ARMA models

Most common ones are:

1) Method of Moments Estimation (MoM)
2) OLS Estimation
3) Maximum Likelihood Estimation (ML)

What method to use depends not only on what model is being fitted, but also on what the efficiency gain is that one can expect

- it is known that MoM estimation can be very inefficient for MA models

When estimating ARMA models (models that included ARMA terms) estimation problem becomes **non-linear**

- need to use numerical optimisation routines (see page 127 in Shumway and Stoffer (2011) for example)
- can be very sensitive to starting values

Finding the appropriate ARMA order

The determination of an appropriate ARMA(p,q) model to represent an observed stationary time series involves a number of interrelated problems:

- estimation of the mean μ, of the coefficients of interest $\{\phi_i\}_{i=1}^p$ and $\{\theta_i\}_{i=1}^q$, and of the white noise variance σ^2:
  - *preliminary estimation* (Yule-Walker estimation, Burg's algorithm, innovations algorithm, Hannan-Rissanen algorithm);
  - *maximum likelihood estimation*;
- choice of p and q (order selection);
- final selection of the model based on a series of goodness of fit tests (diagnostic checking).

MoM Estimation of AR(p) Models - Yule-Walker Estimation

Recall that principle of MoM estimation is to equate the population moments to the sample moments.

For general AR(p) model that means that we can use sample analogues of the population relations in the Yule Walker equation to back out the AR(p) coefficients.

For the population YW-relation

$$\begin{bmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ \vdots \\ \rho(j-1) \end{bmatrix} = \begin{bmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(j-1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(j-2) \\ \rho(2) & \rho(1) & 1 & \cdots & \rho(j-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(j-1) & \rho(j-2) & \rho(j-3) & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{j1} \\ \phi_{j2} \\ \phi_{j3} \\ \vdots \\ \phi_{jj} \end{bmatrix}$$

$$\boldsymbol{\rho}_j = \mathbf{R}_j \boldsymbol{\phi}_j \tag{1}$$

that means to replace population moments in (1) with sample moments, so that the MoM estimator is

$$\hat{\mathbf{R}}_j \hat{\boldsymbol{\rho}}_j = \hat{\boldsymbol{\phi}}_j$$

where $\gamma(j)$ in $\rho(j) = \gamma(j)/\gamma(0)$ is replaced by $\hat{\gamma}(j) = n^{-1} \sum_{t=1}^{n} (\tilde{x}_t \tilde{x}_{t-j})$ with $\tilde{x}_t = x_t - \bar{x}$ and $\bar{x} = n^{-1} \sum_{t=1}^{n} x_t$.

An estimate of the constant from the model once $\phi(L)$ has been estimated can be obtained from

$$\hat{c} = \hat{\phi}(1)\bar{x}.$$

An estimate of the variance of the white noise disturbance term σ^2 can be obtained from

$$\sigma^2 = \hat{\gamma}(0)[1 - \hat{\boldsymbol{\phi}}_j' \hat{\boldsymbol{\rho}}_j]$$

where

$$\hat{\gamma}(0)\hat{\boldsymbol{\rho}}_j = \begin{bmatrix}\hat{\gamma}(1) & \hat{\gamma}(2) & \hat{\gamma}(3) & \cdots & \hat{\gamma}(p)\end{bmatrix}',$$

Large-sample distribution of Yule-Walker estimators.

It is useful to remark that for a large sample from an AR(p) process

$$\hat{\phi} \approx \mathcal{N}\left(\phi, n^{-1}\sigma^2\Gamma_p^{-1}\right).$$

Under the assumption the the order p of the fitted model is the correct value, we can derive approximate large-sample confidence regions for the true coefficient vector ϕ_p:

$$\left\{\phi \in \mathbb{R}^p : (\hat{\phi}_p - \phi)'\hat{\Gamma}_p(\hat{\phi}_p - \phi) \leq n^{-1}\hat{v}_p\chi^2_{1-\alpha}(p)\right\} \tag{2}$$

contains ϕ_p with probability close to $(1-\alpha)$, where $\hat{v}_p = \hat{\gamma}(0)[1 - \hat{\boldsymbol{\phi}}_j'\hat{\boldsymbol{\rho}}_j]$.

Similarly, for large n the interval bounded by

$$\hat{\phi}_{pj} \pm \Phi_{1-\alpha/2} n^{-1/2} \hat{v}_{jj}^{1/2} \tag{3}$$

contains ϕ_{pj} with probability close to $(1-\alpha)$, where Φ_α denotes the α-quantile of the standard normal distribution and $\hat{v}_{jj}$ is the jth diagonal element of $\hat{v}_p \hat{\Gamma}_p^{-1}$.

Example: MoM estimation of AR(2) model

Suppose we have an AR(2) model that we would like to estimate and we get the following sample moment values of $\hat{\boldsymbol{\rho}}$ and $\hat{\mathbf{R}}$ from the data:

$$\hat{\boldsymbol{\rho}} = \begin{bmatrix} 0.85 \\ 0.52 \end{bmatrix} \text{ and } \hat{\mathbf{R}} = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}.$$

The MoM estimate of $\hat{\phi}$ is then given by

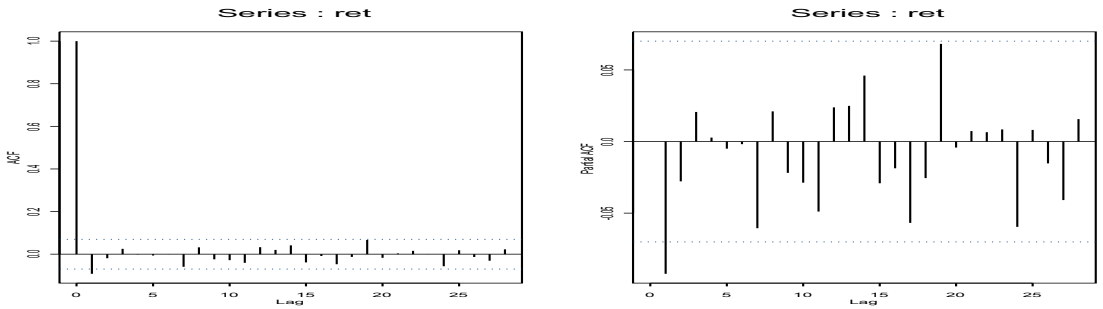
$$\begin{aligned}
\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{bmatrix} &= \begin{bmatrix} 1 & \hat{\rho}(1) \\ \hat{\rho}(1) & 1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.85 \\ 0.52 \end{bmatrix} \\
&= \begin{bmatrix} 1.4703 \\ -0.7297 \end{bmatrix}.
\end{aligned}$$

Given $\hat{\gamma}(0) = \mathrm{Var}(\tilde{x}_t)$, an estimate of the variance of Z_t would then be found from as

$$\begin{aligned}
\hat{\sigma}^2 &= \hat{\gamma}(0)[1 - \hat{\boldsymbol{\phi}}'\hat{\boldsymbol{\rho}}] \\
&= 8.900\left(1 - \begin{bmatrix} 1.4703 \\ -0.7297 \end{bmatrix}' \begin{bmatrix} 0.85 \\ 0.52 \end{bmatrix}\right) \\
&= 8.900 - 7.7454 \\
&= 1.1546.
\end{aligned}$$

The S&P500 Index, January 2003 to December 2005.
The very slowly decaying time series of S&P500 prices suggests differencing at lag 1. We apply first the operator $(1-B)$ to the time series of prices obtaining the new return series $Y_t = P_t - P_{t-1},\ t = 1, \ldots, 782$. The sample autocovariances of the series $y_1, \ldots, y_{782}$ are

$$\widehat{\gamma}(0) = 70.806, \quad \widehat{\gamma}(1) = -6.5396, \quad \widehat{\gamma}(2) = -1.3462.$$



The sample ACF and PACF of the return series support to fit an AR(1) model to the data $\{Y_t - 0.471\}$. The Yule-Walker estimator is

$$\widehat{\phi}_1 = \widehat{\gamma}(1)/\widehat{\gamma}(0) = -0.0924, \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0)[1 - \widehat{\phi}_1\widehat{\rho}(1)] = 70.202.$$

Therefore the fitted model equals

$$Y_t - 0.471 - 0.0924(Y_{t-1} - 0.471) = Z_t, \quad \{Z_t\} \sim WN(0, 70.202).$$

Assuming that the data are really generated by an AR(1) model, the approximate 95% bounds for ϕ_1 are $-0.0924 \pm \frac{1.96\sqrt{70.202}}{\sqrt{70.806}\sqrt{782}} = [-0.1622, -0.0226]$.

Problems with YW for AR models

One well known issue with Yule Walker estimation of the parameters of an AR model is that there can be:

1) a severe loss of efficiency and
2) heavy bias in the estimates in finite samples

for processes that are close to the non-stationary region of the parameter space (see the study by de Hoon *et al.* (2006)).

- we will see an example simulation of this in the Lecture

MoM estimation of MA Models

Estimation of MA models follows same principle of equating sample moments to population ones.

Here we use the relation between the ACF and the $\theta(L)$ parameters.

For an MA(1) process, we know that we have the population moment condition:

$$\begin{aligned} \rho(1) &= \frac{\gamma(1)}{\gamma(0)} \\ &= \frac{\theta_1 \sigma^2}{(1+\theta_1^2)\sigma^2} \\ \rho(1) &= \frac{\theta_1}{(1+\theta_1^2)}. \end{aligned}$$

Computing the sample analogue of $\rho(1), \hat{\rho}(1)$, we can then find $\hat{\theta}_1$ as the value that yields an invertible solution of the two possible MA(1) solutions

$$\hat{\theta}_1^{(1,2)} = \frac{1 \pm \sqrt{1 - 4\hat{\rho}\,(1)^2}}{2\hat{\rho}\,(1)}. \tag{4}$$

It should also be clear from (4) that, if $|\hat{\rho}(1)| > 1/2$, then $\hat{\theta}_1$ will be a complex number.

- have to impose the additional restriction that $|\hat{\rho}(1)| \leq 1/2$ which shows that the MA(1) process cannot generate a highly persistent series

- highest possible value is $|\hat{\rho}(1)| = 1/2$ (we can also see this from the plots of the ACF as a function of θ).
- for the relation in (4), the restriction $|\hat{\rho}(1)| \leq 1/2$ effectively means that the invertible root is (if $\hat{\rho}(1) < 0$)

$$\hat{\theta}_1^{(1)} = \frac{1 - \sqrt{1 - 4\hat{\rho}\,(1)^2}}{2\hat{\rho}\,(1)}. \tag{5}$$

The above outlined issues generalise to higher order MA(q) models and also to ARMA models in general due to the presence of the MA parameters.

Example: MoM estimation of MA(1) model

Suppose we get $\hat{\rho} = 0.32$ from the sample data. From the relation in (4) we get the two solutions

$$\hat{\theta}_1^{(1)} = \frac{1 - \sqrt{1 - 4(0.32)^2}}{2\,(0.32)} = 0.3619$$

$$\hat{\theta}_1^{(2)} = \frac{1 + \sqrt{1 - 4(0.32)^2}}{2\,(0.32)} = 2.7630$$

so we would keep the solution $\hat{\theta}_1^{(1)} = 0.3619$ as the invertible one. Note here that this is the characteristic root, ie., $\lambda = \left|\hat{\theta}_1\right| < 1$ and not the root of the Lag polynomial, which would be $1/\hat{\theta}_1$.

It should be pointed out here that MoM estimation of MA models is highly inefficient.

For the simple MA(1) example given above, it can be shown (see Shumway and Stoffer (2011), page 124) that the asymptotic distribution is

$$\sqrt{n}(\hat{\theta}_1 - \theta_1) \xrightarrow{d} \mathsf{N}(0, V_{\theta_1})$$

where

$$V_{\theta_1} = \frac{(1 + \theta_1^2 + 4\theta_1^4 + \theta_1^6 + \theta_1^8)}{(1 - \theta_1^2)^2}. \tag{6}$$

The ML estimator has $V_{\theta_1} = \left(1 - \theta_1^2\right)$. So, depending on the size of θ_1, the difference in the variance of the MoM estimator can be 3 to 4 times larger than that of the ML estimator.

In summary, we have the following weak points of MoM based estimation:

1) Highly non-linear estimation
2) Possibly, much larger variance than ML estimation
3) Multiple solutions, need to choose invertible one
4) Can get complex solutions for $\hat{\beta}$

The last point can be a real problem even when the true DGP has a population $\rho(1)$ of, say, 0.47 or so.

Preliminary Estimation: the Hannan-Rissanen Algorithm

(for general ARMA(p,q) models)

Problem with ARMA models is that Z_t of MA representation are not observed.

Hannan-Rissanen Algorithm proceeds by making the Z_t observable, by using the OLS residuals from a high order AR(m) regression.

Step 1. A higher-order AR(m) model (with $m > \max(p, q)$) is fitted to the data using the Yule-Walker technique $\rightarrow \hat{\phi}_m = (\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm})$.

Then, the estimated residuals are computed from the equations

$$\hat{Z}_t = X_t - \hat{\phi}_{m1} X_{t-1} - \ldots - \hat{\phi}_{mm} X_{t-m}, \quad t = m+1, \ldots, n. \tag{7}$$

Step 2. Estimate the vector of parameters $\beta = (\phi', \theta')'$ by least squares linear regression of X_t onto $(X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q})$, i.e.

$$\hat{\beta} = \arg\min_{\beta} S(\beta) \text{ where} \tag{8}$$

$$S(\beta) = \sum_{m+1+q}^{n} \left(X_t - \phi_1 X_{t-1} - \ldots \phi_p X_{t-p} - \theta_1 \hat{Z}_{t-1} - \ldots - \theta_q \hat{Z}_{t-q} \right)^2. \tag{9}$$

This gives the Hannan-Rissanen estimators

$$\hat{\beta} = (\mathbf{W}'\mathbf{W})^{-1}(\mathbf{W}'\mathbf{X}_n) \quad \text{and} \quad \hat{\sigma}^2_{\text{HR}} = \frac{S(\hat{\beta})}{n-m-q}, \tag{10}$$

where

$$\mathbf{X}_n = (X_{m+1+q}, \ldots, X_n)' \tag{11}$$

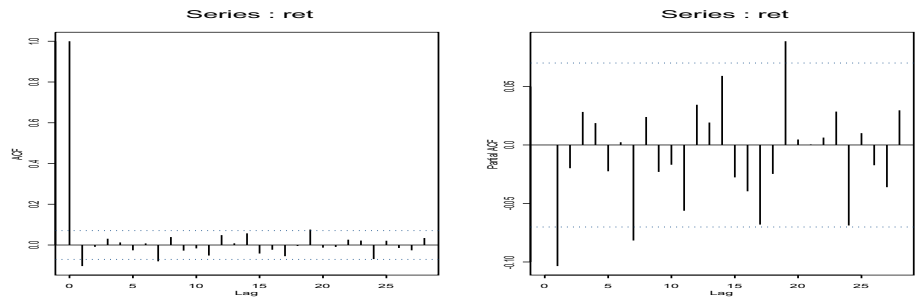
and $\mathbf{W}$ is the $(n-m-q)\times(p+q)$ matrix

$$\mathbf{W}=\begin{bmatrix} X_{m+q} & X_{m+q+1} & \cdots & X_{m+q+1-p} & \hat{Z}_{m+q} & \hat{Z}_{m+q-1} & \cdots & \hat{Z}_{m+1} \\ X_{m+q+1} & X_{m+q} & \cdots & X_{m+q+2-p} & \hat{Z}_{m+q+1} & \hat{Z}_{m+q} & \cdots & \hat{Z}_{m+2} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ X_{n-1} & X_{n-2} & \cdots & X_{n-p} & \hat{Z}_{n-1} & \hat{Z}_{n-2} & \cdots & \hat{Z}_{n-q} \end{bmatrix}.$$

Note here that we have assumed that constant c in ARMA model is zero, otherwise include vector of ones in $\mathbf{W}$.

The S&P500 Index, January 2003 to December 2005.
Let us consider once again the same S&P500 price series. Now, let we define the log-return series (in %) as

$$R_t = 100 \log\left(P_t/P_{t-1}\right), \quad t = 2, \ldots, 783.$$



We fit an ARMA(1,1) model to the mean-corrected series $\{R_t - 0.0447\}$. Let us choose $m = 10$. The Hannan-Rissanen estimators are given by

$$\widehat{\phi}_{\text{HR}} = -0.0403, \quad \widehat{\theta}_{\text{HR}} = -0.0613 \quad \text{and} \quad \widehat{\sigma}^2_{\text{HR}} = 0.6345.$$

Hence, the fitted model equals

$$R_t - 0.0447 - 0.0403(R_{t-1} - 0.0447) = Z_t - 0.0613 Z_{t-1}, \ \{Z_t\} \sim WN(0, 0.6345).$$

Maximum Likelihood Estimation

The most common estimation method of the parameters of ARMA models is Maximum Likelihood Estimation (MLE).

For MLE need to find the **join density function** of X_t, for all $t = 1, \ldots, n$.

Since the data are not independent, we will need to use conditioning rules to build up the joint density function for time series processes.

Note here also, that since we can always express the Z_t as a function of the observed X_t, this will also mean that if we condition on $\mathbf{X}_{1:n}$, that we also observe all Z_t for $t = 1, \ldots, n$.

Factorisation of joint density

The basic idea underlying the factorising of the joint density is the repeated use of conditional probability rules.

Recall that if we have two events A and B than we can form the joint probability $P(A, B) = P(B|A)P(A)$.

When working out the joint density to form the likelihood function, we fundamentally rely on this simple principle to write down the likelihood function.

AR models

To illustrate this, consider the standard simple AR(1) formulation:

$$X_t = c + \phi_1 X_{t-1} + Z_t \tag{12}$$

where $Z_t \sim \mathsf{N}\left(0, \sigma^2\right)$. We can recursively build up the joint density as follows.

Since X_t follows an AR(1) and $Z_t \sim \mathsf{N}\left(0, \sigma^2\right)$, the first observation is distributed as:

$$X_1 \sim \mathsf{N}\left(\frac{c}{(1-\phi_1)}, \frac{\sigma^2}{(1-\phi_1^2)}\right)$$

where $\frac{c}{(1-\phi_1)} = E(X_t)$ and $\frac{\sigma^2}{(1-\phi_1^2)} = \text{Var}(X_t), \forall t = 1, \ldots, n$ and hence holds also for X_1.

Let us denote the density of X_1 by $f(X_1)$. At time $t = 2$, the joint density $f(\mathbf{X}_{1:2})$, given that we have $f(X_1)$, is then simply computed from the conditional probability rules

$$f(\mathbf{X}_{1:2}) = f(X_2, X_1) \tag{13}$$

$$= f(X_2|X_1)f(X_1). \tag{14}$$

Now, since we know X_1, we can work out the conditional density to be:

$$f(X_2|X_1) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(X_2 - c - \phi_1 X_1)^2}{\sigma^2}\right\} \tag{15}$$

where we can think of (15) as being the **one step ahead forecast density** of X_t given information up to time $t-1$, where $t=2$.

The joint density for the two time periods is then simply put together from (14) as:

$$f(\mathbf{X}_{1:2}) = \underbrace{\left[(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\frac{(X_2-c-\phi_1 X_1)^2}{\sigma^2}\right\}\right]}_{f(X_2|X_1)} \tag{16}$$

$$\times \underbrace{\left[(2\pi\sigma^2/(1-\phi_1^2))^{-1/2}\exp\left\{-\frac{1}{2}\frac{(X_1-c/(1-\phi_1))^2}{\sigma^2/(1-\phi_1^2)}\right\}\right]}_{f(X_1)}. \tag{17}$$

With $f(X_2|X_1)$ as given in (15) we can see that this generalises to

$$\begin{aligned} f(X_t|\mathbf{X}_{1:t-1}) &= (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\frac{(X_t-c-\phi_1 X_{t-1})^2}{\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2}\frac{Z_t^2}{\sigma^2}\right\} \end{aligned}$$

where $Z_t = X_t - c - \phi_1 X_{t-1}$ from the AR(1) specification in (12).

That is, Z_t is the residual component and since it is constructed as $X_t - (c + \phi_1 X_{t-1})$ where $(c + \phi_1 X_{t-1}) = E_{t-1}(X_t)$ (conditional mean of X_t given time $t-1$), Z_t is also a **one step ahead prediction error** (or forecast error).

Due to this representation, forming the likelihood function in this way is frequently referred to as the **prediction error decomposition**.

Using the above principle, we can find

$$\begin{aligned} f(\mathbf{X}_{1:3}) &= f(X_3, X_2, X_1) && (18)\\ &= f(X_3|X_2, X_1)f(X_2, X_1) && (19)\\ &= f(X_3|X_2, X_1)f(X_2|X_1)f(X_1) && (20) \end{aligned}$$

and in general for the n observations that are available

$$f(\mathbf{X}_{1:n}) = f(X_1, X_2, \ldots, X_n) = \prod_{t=2}^{n} f(X_t|\mathbf{X}_{1:t-1})f(X_1) \quad (21)$$

where $f(\mathbf{X}_{1:n})$ in (21) represents the **exact likelihood function** of the AR(1) model above. Taking logs of (21) and explicitly writing down the dependence on a vector of

population parameters β, where $\beta = (c\ \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ for a general ARMA(p, q) model, yields the log-likelihood function $\mathscr{L}(\beta)$, which then has the following simple additive structure

$$\mathscr{L}(\beta) = \underbrace{\sum_{t=2}^{n} \log f(X_t | \mathbf{X}_{1:t-1}; \theta)}_{(1)\ [(n-1)\ \text{terms}]} + \underbrace{\log f(X_1; \theta)}_{(2)\ [1\ \text{term}]} \tag{22}$$

with the two components being:

1) the conditional density of $X_t | X_{t-1}$ and
2) the initial distribution of X_1.

We can see now how this generalises to an AR(p) process. There we need p initial values to condition upon, ie., we need to work out $f(\mathbf{X}_{1:p})$ to form the likelihood function as

$$f(\mathbf{X}_{1:n}) = \prod_{t=p+1}^{n} f(X_t|\mathbf{X}_{1:t-1}) f(\mathbf{X}_{1:p}). \tag{23}$$

Here, $f(X_t|\mathbf{X}_{1:t-1})$ with $t = p+1$ will have the from

$$f(X_t|\mathbf{X}_{1:t-1}) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(X_t - c - \sum_{i=1}^{p}\phi_i X_{t-i})^2}{\sigma^2}\right\} \tag{24}$$

and now $f(\mathbf{X}_{1:p})$ will be a multivariate distribution.

Given that the Z_t are N $\left(0, \sigma^2\right)$, the joint density of the first p initial values $\mathbf{X}_{1:p}$ is then distributed as a multivariate normal distribution, taking the form

$$f(\mathbf{X}_{1:p}) = (2\pi)^{-\frac{p}{2}} \left|\mathbf{\Gamma}^{-1}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})' \mathbf{\Gamma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\} \tag{25}$$

where $\mathbf{\Gamma} = E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right]$ is the covariance matrix as used before, that is,

$$\underset{(p\times p)}{\mathbf{\Gamma}} = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{bmatrix} \tag{26}$$

and $E(\mathbf{X}) = \boldsymbol{\mu}$ and $\underset{(p\times 1)}{\mathbf{X}} = (X_1, X_2, \ldots, X_p)'$ is a vector of initial (or conditioning) values.

What exactly the $\gamma(j), \forall j = 0, \ldots, p-1$ terms are in (26) needs to be worked out for the AR(p) process that is being considered.

Now it should be clear that the influence of the first say p observations or (initial values) that we have conditioned upon fades out for large samples.

Estimation of the model parameters often proceeds by only using the conditional density, where the conditioning is on the first p X_t values.

This is know as **conditional maximum likelihood estimation**.

Remark (OLS and Conditional MLE)

If $Z_t \sim \mathsf{N}\left(0, \sigma^2\right)$*, then maximising only the conditional density is the same as running an OLS regression of* X_t *on* $X_{t-1}, \ldots, X_{t-p}$ *and a constant. The corresponding log-likelihood function for an AR(*p*) process is then*

$$\mathscr{L}(\beta) = -\frac{1}{2}\left[(n-p)\log(2\pi) + (n-p)\log(\sigma^2) + \sigma^{-2}\sum_{t=p+1}^{n} Z_t^2\right] \tag{27}$$

where $Z_t = (X_t - c - \sum_{i=1}^{p}\phi_i X_{t-i})$*. Maximizing* $\mathscr{L}(\beta)$ *is thus the same as minimizing* $S(\beta)$ *from (9).*

MA models

Maximum likelihood estimation of MA(q) models follows the same principle as ML estimation of AR(p) models.

The issue with MA models in general is that the MA terms, ie., the Z_t are not observed and the likelihood function **must** be written down in terms of **observed quantities**.

When writing down the likelihood function, we thus need to express the Z_t terms as a function of all the observable variables.

This can be done through recursive substitution of the unobserved Z_t series with its observed counterpart, the X_t.

To illustrated this, suppose we have the following simple MA(1) model:

$$X_t = c + \theta_1 Z_{t-1} + Z_t \tag{28}$$

where $Z_t \sim \mathsf{N}(0, \sigma^2)$ as before. We then have the conditional distribution of X_t given all previous information (which is captured in $\mathbf{X}_{1:t-1}$)

$$X_t | \mathbf{X}_{1:t-1} = \mathsf{N}\left(c + \theta_1 Z_{t-1}, \sigma^2\right).$$

Thus, with a given general non-zero initial condition Z_0, we can build up Z_t recursively as:

$$\begin{aligned}
Z_1 &= X_1 - c - \theta_1 Z_0 \\
Z_2 &= X_2 - c - \theta_1 (X_1 - c) - \theta_1^2 Z_0 \\
&\vdots \\
Z_t &= \sum_{i=0}^{t-1} (-\theta_1)^i (X_{t-i} - c) + \underbrace{(-\theta_1)^t Z_0}_{\text{if } Z_0 \neq 0}.
\end{aligned} \tag{29}$$

The Z_t in (29) can then be used to form the likelihood function that will need to be maximized.

The representation in (29) is convenient here as it illustrates two important points that need to be kept in mind when estimating MA models in general.

1) we need $|\theta_1| < 1$ (or invertibility in general to hold) for ML estimation to be feasible. If this is not the case, then the sum involving θ_1^i will blow up.
2) since Z_t is polynomial in θ_1, estimation problem is highly non-linear and there will not be a closed form solution as there is with OLS for the AR(p) model.

Using the representation shown in (23), we can again split the exact likelihood function into two parts.

The conditional likelihood function $f(X_t|\mathbf{X}_{1:t-1})$, which is

$$f(X_t|\mathbf{X}_{1:t-1}) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(X_t - c - \sum_{i=1}^{q}\theta_i Z_{t-i})^2}{\sigma^2}\right\} \tag{30}$$

and the initial distribution of the q pre-sample values, denoted by $f(\mathbf{X}_{1:q})$. '

This a q dimensional Multivariate normal density, with mean c and autocovariance function $\mathbf{\Gamma}$ as for an MA(q) process.

ARMA models

Estimation of ARMA models follows same strategies as outlined above.

The conditional likelihood function for ARMA(p,q) models takes the from

$$f(X_t|\mathbf{X}_{1:t-1}) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(X_t - c - \sum_{i=1}^{p}\phi_i X_{t-i} - \sum_{i=1}^{q}\theta_i Z_{t-i})^2}{\sigma^2}\right\}. \tag{31}$$

So here we condition upon $X_1,\ldots,X_p$ and $Z_1,\ldots,Z_q$ so we need the conditional distribution $f(\mathbf{X}_{1:k})$, where $k = \max\{p,q\}$:

$$f(\mathbf{X}_{1:k}) = (2\pi)^{-\frac{k}{2}} \left|\mathbf{\Gamma}^{-1}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})'\mathbf{\Gamma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\} \tag{32}$$

and $\mathbf{\Gamma}$ is now a $(k \times k)$ covariance matrix which is a function of the ARMA(p,q) parameters that is being fitted.

Computation of the full or exact likelihood function for large order ARMA models can be tricky numerically.

It is thus common to disregard the influence of the term in (32) and focus on the conditional likelihood function in (31), as the influence of the k fixed terms in the full or exact likelihood function vanishes as the sample size increases.

Specification of initial values

To proceed with the estimation of the conditional likelihood function Box *et al.* (1994) recommend to set all $Z_1, \ldots, Z_q$ components equal to 0 and use actual observed values for the first p X_t values (ie., for $X_1, \ldots, X_p$) to form the conditional likelihood function.

Large sample distribution of maximum likelihood estimators

For a large sample from an ARMA(p,q) process,

$$\hat{\beta} = (\hat{\phi}, \hat{\theta})' \approx N\left(\beta, n^{-1}V(\beta)\right). \tag{33}$$

In some special cases, the general matrix $V(\beta)$ takes a simple form.

- AR(p) model: $V(\phi) = \sigma^2\Gamma_p^{-1}$ (like the one of Yule-Walker estimator).
  Special cases:

$$\text{AR(1): } V(\phi) = 1 - \phi_1^2; \tag{34}$$

$$\text{AR(2): } V(\phi) = \begin{bmatrix} 1-\phi_1^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix}. \tag{35}$$

- MA(q) model: Let Γ_q^* be the covariance matrix of $Y_1, \ldots, Y_q$, where $\{Y_t\}$ is the autoregressive process with autoregressive polynomial $\theta(B)$, i.e.

$$Y_t + \theta_1 Y_{t-1} + \ldots + \theta_q Y_{t-q} = Z_t, \quad \{Z_t\} \sim \text{WN}(0,1). \tag{36}$$

Then it can be shown that

$$V(\theta) = \Gamma_q^{*-1}. \tag{37}$$

Special cases (substitute $\phi_i = -\theta_i$ in the AR formulas):

$$\text{MA(1): } V(\phi) = 1 - \theta_1^2; \tag{38}$$

$$\text{MA(2): } V(\phi) = \begin{bmatrix} 1-\theta_1^2 & \theta_1(1-\theta_2) \\ \theta_1(1-\theta_2) & 1-\theta_2^2 \end{bmatrix}. \tag{39}$$

- stationary and invertible ARMA(1,1):

$$V(\phi,\theta) = \frac{1+\phi\theta}{(\theta+\phi)^2} \begin{bmatrix} (1-\phi^2)(1+\phi\theta) & -(1-\theta^2)(1-\phi^2) \\ -(1-\theta^2)(1-\phi^2) & (1-\theta^2)(1+\phi\theta) \end{bmatrix}. \tag{40}$$

Order Selection

How can we select the appropriate values for the orders p and q?

Direct answer: choose the orders as large as possible.

$\longrightarrow$ result in a small estimated white noise variance, but when the fitted model is used for forecasting this leads to poor forecasts due to overfitting.

Solution: introduce a penalty factor to penalise models with too many parameters.

Many criteria based on such penalty factors have been proposed in the literature, we will restrict our attention on the following three:

- the AIC (Akaike, 1973) and AICC criteria (Hurvich and Tsai, 1989);
- the BIC criterion (Akaike, 1978).

The AIC and AICC criteria

The AIC statistic is defined as

$$\text{AIC}(\beta) = -2\mathscr{L}(\beta) + 2(p+q+1),$$

where $\mathscr{L}(\beta)$ is the likelihood function.

Similarly, the AICC statistic is defined as

$$\text{AICC}(\beta) = -2\mathscr{L}(\beta) + 2(p+q+1)n/(n-p-q-2).$$

We select the values of p and q for our fitted model to be those that *minimize* $\text{AIC}(\hat{\beta})$ or $\text{AICC}(\hat{\beta})$, respectively.

The BIC criterion

The BIC is another criterion that attempts to correct for overfitting. The BIC is generally written as

$$\text{BIC} = -2\mathscr{L}(\beta) + (p+q+1)\ln(n).$$

We again select p and q for our fitted model to be those that *minimize* the BIC.

Example

We generate 200 observations from an ARMA(2,3) process

$$\begin{aligned} X_t - X_{t-1} + 0.24X_{t-2} &= Z_t + 0.4Z_{t-1} + 0.2Z_{t-2} + 0.1Z_{t-3}, \\ \text{with } \{Z_t\} &\sim N(0,1). \end{aligned}$$

We investigate whether using the different criteria for selecting the best order we get the true orders $(2, 3)$.

p	q	$-2 \ln L_X$	AIC	AICC	BIC
0	0	2185.68	2187.68	2187.70	2190.98
1	0	675.578	679.578	679.639	681.585
2	0	609.492	615.492	615.615	622.036
0	1	800.242	804.242	804.303	805.708
1	1	619.306	625.306	625.428	631.815
2	1	608.761	616.761	616.967	627.548
1	2	613.259	621.259	621.464	632.033
2	2	608.760	618.760	619.069	633.448
3	2	607.577	619.577	620.012	637.790
2	3	605.965	617.965	618.401	636.304
3	3	599.586	613.586*	614.170*	630.485*
4	3	597.657	613.657	614.411	634.537
3	4	605.902	621.902	622.656	646.975
4	4	599.132	617.132	618.079	640.132

The fitted model is therefore an ARMA(3,3) with parameters

$$\begin{aligned} X_t - 1.74X_{t-1} + 1.67X_{t-2} - 0.76X_{t-3} &= Z_t - 0.37Z_{t-1} + 0.6Z_{t-2} + 0.47Z_{t-3}, \\ \text{with } \{Z_t\} &\sim N(0, 1.13). \end{aligned}$$

Common Roots Problem

There exists a redundancy in ARMA(p, q) models when in fact the AR and MA lag polynomials have the same **common factors** or **common roots**.

- one will not be able to estimate the parameters of the model as the generated data will be observationally equivalent to data that was generated without the common roots.
  - model is then often referred to as **overfitting the data** in the sense that we are trying to include more *'right hand side variables'* in the model than we really need.
- estimation problems that this can create are quite different to the ones that one would encounter in a standard regression (or estimation) problem
  - extra variables would just get an insignificant and frequently close to zero coefficient attached to them.

In ARMA models, this is different and creates severe estimation problems.

To clarify this, suppose we have the following general ARMA(p, q) set up

$$\phi(L) X_t = \theta(L) Z_t \tag{41}$$

and we multiply both sides by $(1 - \zeta L)$ which then yields

$$(1 - \zeta L)\phi(L) X_t = (1 - \zeta L)\theta(L) Z_t$$
$$\phi^*(L) X_t = \theta^*(L) Z_t \tag{42}$$

and we proceed to estimate the model in (42) unknowing of this problem.

- the likelihood functions in (41) and (42) are the same,
  - we do not have any extra information in the data to estimate the ζ parameter and the likelihood function will be flat in the direction of ζ.
  - when estimating the models on the computer, numerical problems arise, with search algorithms simply stopping without having reached a maximum of the likelihood function.

Numerical Example

Suppose we have the following population ARMA$(2,1)$ model:

$$\begin{aligned} X_t &= 1.3X_{t-1} - 0.4X_{t-2} + Z_t - 0.5Z_{t-1} \\ \left(1 - 1.3L + 0.4L^2\right) X_t &= \left(1 - 0.5L\right) Z_t \\ \left(1 - 0.5L\right)\left(1 - 0.8L\right) X_t &= \left(1 - 0.5L\right) Z_t \\ \left(1 - 0.8L\right) X_t &= Z_t. \end{aligned}$$

- the common factor is $(1 - 0.5L)$ and the model is over-parameterised, ie. it is really only an ARMA(1,0) = AR(1) model
- there will be no information in the likelihood regarding the parameters of the ARMA(2,1) model.

Small Sample Bias in AR parameter estimates

All dynamic models (ie., ones with lags in them) have a small (or finite) sample bias when $E(Z_t|X_t) = 0$ is used as the FOC to estimate that parameters of the model.

- recall that the finite sample distribution of the OLS (or MLE or MoM) estimator requires that $E(Z_t|X_t) = 0$ holds for all leads and lags of Z_t and X_t, that is, for all $t = 1, \ldots, n$.
  - known the well known *strict exogeneity* assumption of OLS.
- in the AR(1), the moment condition $E(X_{t-1}Z_t) = 0$ is satisfied due to X_{t-1} being predetermined.
- But the strict exogeneity assumption requires that $E(X_{t-j}Z_t) = 0$, for all $j = 1, \ldots, n$ and for all $\{Z_t\}_{t=1}^n$.
  - so can always choose a combination of the $\{Z_t\}_{t=1}^n$ sequence and X_t to show that this does not hold.
  - bias is a small sample bias, so it disappears as the sample size n goes to infinity.

It should be pointed out here that the small sample bias is **not only related to the size of the sample** but also to the **persistence of the series**.

A more persistent series can, for example, have double the sample size, but still show the same amount of small sample bias.

Shaman and Stine (1988) provide analytic formulas for the small sample bias for some pure AR models.

Some rules of thumb are as follows that were computed from simulations: for an AR(1) parameter of $\phi = 0.9$ and samples of size $50, 100, 150$ and 200 the bias is $-0.08, -0.04, -0.025$ and -0.02, respectively.

The constants are upward biased and the ϕ_i coefficients will in general be downward biased.

Approximate analytic formulas of the bias in AR(1) models have been derived by Kendall (1954).

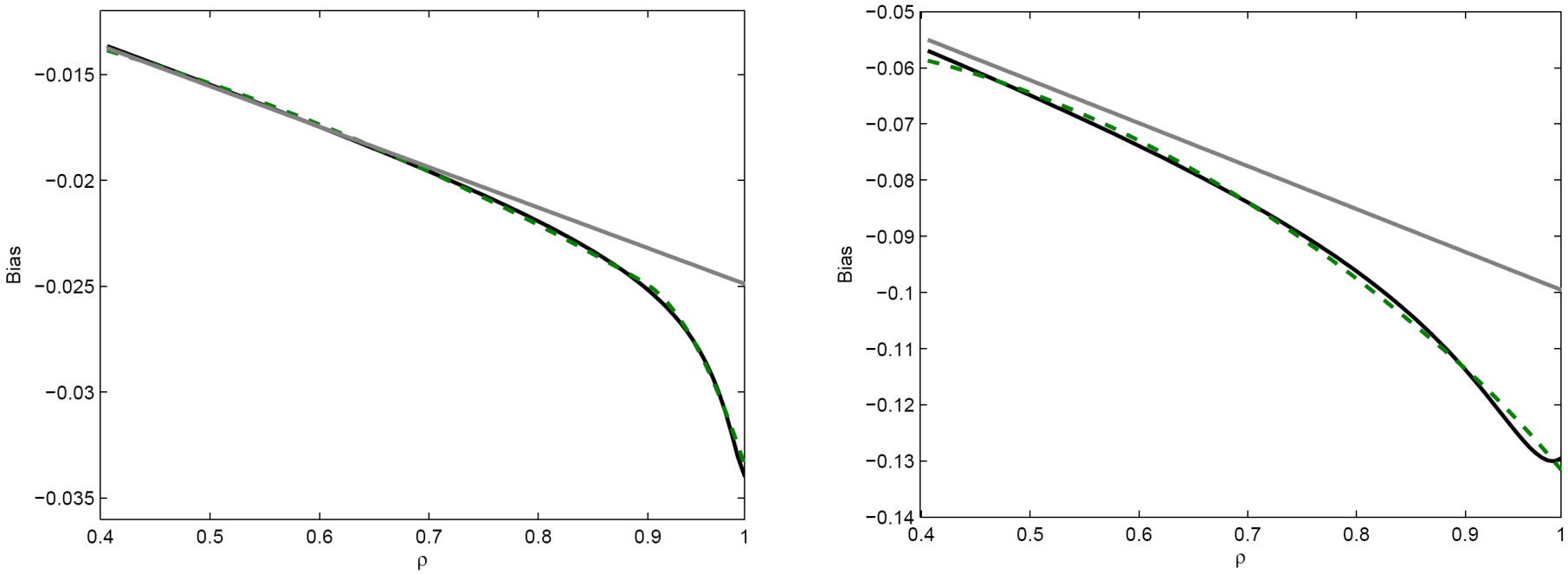
Kendall showed that the bias is (approx. linear in ϕ and of order $1/n$):

$$\frac{-(1+3\phi)}{n}, \tag{43}$$

bias [bias$(\hat{\phi}) = E(\hat{\phi}) - \phi$] around -0.025 and -0.1 when ϕ close to 1 ($n = 160,\ 40$)

- formula in (43) only good approx. of non-linear relation between estimator bias and ϕ over $[0, 0.7]$ interval.
- for values closer to 1, it breaks down pretty badly. simulation in Jardet *et al.* (2011) show the bias of the OLS estimator of ϕ, with true $\phi = [0.4,\ 0.99]$ [using $50\,000$ AR(1)].

Plots in Figure 1 show, depending on the sample size, non-linearity in the bias function can be substantial the closer one gets to the non-stationary region of unity.



(a) $n = 160$ (b) $n = 40$

Figure 1: Bias of the OLS estimator of ϕ. Kendall's linear approximation (gray line), true bias (black line).

References

Box, George E.P., Gwilym M. Jenkins and Gregory C. Reinsel (1994): *Time Series Analysis: Forecasting and Control*, *3rd Edition*, Prentice Hall.

de Hoon, M. J. L., T. H. J. J. van der Hagen, H. Schoonewelle and H. van Dam (2006): *"Why Yule-Walker should Not be used for Autoregressive Modelling,"* Manuscript, Delft University of Technology. Available from: http://www-stat.wharton.upenn.edu/~steele/Courses/956/ResourceDetails/YWSourceFiles/WhyNotToUseYW.pdf.

Jardet, Caroline, Alain Monfort and Fulvio Pegoraro (2011): "Persistence, Bias, Prediction and Averaging Estimators," *Banque de France Working Paper*.

Kendall, Maurice G. (1954): "A note on bias in the estimation of autocorrelation," *Biometrika*, **41**(3-4), 403–404.

Shaman, Paul and Robert A. Stine (1988): "The Bias of Autoregressive Coefficient Estimators," *Journal of the American Statistical Association*, **83**(403), 842–848.

Shumway, Robert H. and David S. Stoffer (2011): *Time Series Analysis and Its Applications: With R Examples*, *3rd Edition*, Springer.