# Time Series (MATH5845)

Dr. Atefeh Zamani
Based on the notes by Prof. William T.M. Dunsmuir

T2 2025

# Chapter 6

# Maximum Likelihood Estimation for ARMA models.

# Contents

As mentioned before, Yule-Walker estimator are efficient estimator for Autoregressive processes but not for general ARMA processes. For more on this topic, you can refer to Shumway et al. [2000], P. 113-116. In this chapter, we are going to consider maximum likelihood estimators for ARMA models in time series.

## 6.1 Innovations form of the Gaussian Likelihood.

- $\{X_t\}$ : A Gaussian time series with zero mean function and covariance function $\gamma(i,j) = E(X_i X_j)$.
- $\mathbf{X}_n = (X_1, \cdots, X_n)'$
- $\hat{\mathbf{X}}_n = (\hat{X}_1, \cdots, \hat{X}_n)'$.
- $\Gamma_n = E(\mathbf{X}_n \mathbf{X}_n^T) = [\gamma(i,j)]_{i,j=1}^n$ is the covariance matrix of $\mathbf{X}_n$ and we assume $\Gamma_n^{-1}$ exists.
- The likelihood of $\mathbf{X}_n$ is

multivariate normal dist

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp(-\mathbf{x}_n^T \underbrace{\Gamma_n^{-1}}_{n \times n} \mathbf{x}_n / 2).$$

In general this is a difficult expression to work with.

lower triangular

Remember from the Innovation algorithm we know that:

- innovation $\mathbf{U}_n = A_n \mathbf{X}_n$ $\textcircled{1}$
- $C_n = A_n^{-1}$ $\textcircled{2}$
- $\mathbf{U}_n = \mathbf{X}_n - \hat{\mathbf{X}}_n$ $\textcircled{3}$

$$X_n \overset{\textcircled{1}}{=} A_n^{-1} U_n \overset{\textcircled{2}}{=} C_n U_n \overset{\textcircled{3}}{=} C_n(X_n - \hat{X}_n) *$$

Using these points, we have

$$\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n) *$$

$$\underbrace{E(X_n - \hat{X}_n)(X_n - \hat{X}_n)^T}_{\text{covariance matrix of innovation}}$$

so that

$$\Gamma_n = E(\mathbf{X}_n \mathbf{X}_n^T) = E[C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n)(\mathbf{X}_n - \hat{\mathbf{X}}_n)^T C_n^T] = C_n D_n C_n^T$$

where $D_n$ $=$diag$(v_0, \ldots, v_{n-1})$. The quadratic form in the likelihood exponent is therefore

$$\mathbf{x}_n^T \Gamma_n^{-1} \mathbf{x}_n = (\mathbf{x}_n - \hat{\mathbf{x}}_n)^T D_n^{-1} (\mathbf{x}_n - \hat{\mathbf{x}}_n) = \sum_{j=1}^{n} (x_j - \hat{x}_j)^2 / v_{j-1}$$

and

$\Gamma_n = C_n D_n C_n^T$

$$\underline{\det\Gamma_n} = \overbrace{(\det C_n)^2}^{=1 \text{ because all elements on the main diagonal are } 1} \underbrace{\det(D_n)}_{\longrightarrow \prod \text{ elements on the main diagonal}} = v_0v_1\dots v_{n-1}.$$

$= \prod_{j=0}^{n-1} v_j$

$$x_n^T \Gamma_n^{-1} x_n = x_n^T \left(C_n D_n C_n^T\right)^{-1} x_n$$

$$= x_n^T \left(C_n^T\right)^{-1} D_n^{-1} C_n^{-1} x_n$$

$$= \left(C_n^{-1} x_n\right)^T D_n^{-1} C_n^{-1} x_n$$

$$= \left(x_n - \hat{x}_n\right) \underset{\sim}{D_n^{-1}} \left(x_n - \hat{x}_n\right)$$

since $D_n = \mathrm{Diag}\left(v_0, \dots, v_{n-1}\right)$

$\Rightarrow D_n^{-1} = \mathrm{Diag}\left(\frac{1}{v_0}, \dots, \frac{1}{v_{n-1}}\right)$

$$= \sum \frac{\left(x_j - \hat{x}_j\right)^2}{v_{j-1}} \quad \checkmark$$

Using these expressions the likelihood becomes

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n v_0 v_1 \dots v_{n-1}}} \exp\{-\frac{1}{2}\sum_{j=1}^{n} \frac{(x_j - \hat{x}_j)^2}{v_{j-1}}\}.$$

- This is the **innovations form of the likelihood**.
- It is a product of the distributions of $X_j - \hat{X}_j$ for $j = 1, \dots, n$ and can be re-written as

$$L(\Gamma_n) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi v_{j-1}}} \exp\{-\frac{1}{2}\frac{(x_j - \hat{x}_j)^2}{v_{j-1}}\}. \tag{6.1}$$

- Up to this point we have not expressed the $n(n+1)/2$ covariances $\gamma(i,j)$ as functions of a smaller set of parameters.
- Without this there are too many "unknown parameters" to estimate using the $n$ successive values $x_1, \dots, x_n$ observed on the time series.

- Let $\psi = (\psi_1, \ldots, \psi_r)^T$ where $r$ is fixed and finite

- Assume that all covariances, $\gamma(i,j)$, are defined in terms of $\psi$.

- Then the value $\hat{\psi}$ that maximises $L(\psi) = L(\Gamma_n(\psi))$ is called the maximum likelihood estimate of $\psi$.

- Even if $\{X_t\}$ is not Gaussian, you can use (6.1) as a measure of the goodness of fit of the covariance matrix $\Gamma_n(\psi)$ to the data, and still to choose the parameters $(\psi_1, \ldots, \psi_r)^T$ in such a way as to maximize (6.1). Refer to Brockwell and Davis [1991], Page 254.

We now apply these general results to estimation for ARMA$(p, q)$ time series. Let $X_t$ be a causal invertible ARMA$(p, q)$ process:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad Z_t \sim WN(0, \sigma^2)$$

Our first problem is to find maximum likelihood estimates of the following parameters:

- $\phi = (\phi_1, \cdots, \phi_p)'$,
- $\theta = (\theta_1, \cdots, \theta_q)'$,
- the white noise variance $\sigma^2$.

In Chapter 4, we showed that the one-step predictors $\hat{X}_{n+1}$ and their mean squared errors are given by,

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \overset{\text{innovations}}{\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j})}, & 1 \le n < m = \max(p, q) \\ \underset{\text{observation}}{\phi_1 X_n + \cdots \phi_p X_{n+1-p}} + \sum_{j=1}^{q} \underset{\text{innovation}}{\theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j})}, & n \ge m \end{cases} \tag{6.2}$$

Besides, we can show that $\nu_n = E(X_{n+1} - \hat{X}_{n+1})^2 = \sigma^2 r_n$, where $\theta_{ij}$ and $r_i$ are obtained using the innovation algorithm and $\theta_{ij}$ and $r_i$ are independent of $\sigma^2$, (Brockwell and Davis [1991], Page 175-176).

rewrite the Likelihood for ARMA

$v_j = \sigma^2 r_j$

Using this we can re-write the log-likelihood function (6.1) as follows.

maximize

$$
\begin{aligned}
l(\phi,\theta,\sigma^2) &= \log(L(\Gamma_n)) \\
&= \log\left((2\pi)^{-n/2}(\nu_0\cdots\nu_{n-1})^{-1/2}\right) - \frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-\hat{x}_j)^2}{v_{j-1}} \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{n}\log(r_{j-1}) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\underbrace{\sum_{j=1}^{n}\frac{(x_j-\hat{x}_j)^2}{r_{j-1}}}_{S(\phi,\theta)},
\end{aligned}
$$

and, consequently, $-2/n$ times the log-likelihood is

minimize

$$
-\frac{2}{n}\times l(\phi,\theta,\sigma^2) = \log(2\pi) + \frac{1}{n}\sum_{j=1}^{n}\log(r_{j-1}) + \log(\overset{b}{\underset{1}{\sigma^2}}) + \frac{1}{n\underset{b}{\sigma^2}}S(\phi,\theta), = A
$$

where

$$
S(\phi,\theta) = \sum_{j=1}^{n}\frac{(x_j-\hat{x}_j)^2}{r_{j-1}}.
$$

$$
A = \log(2\pi) + \frac{1}{n}\sum\log(r_{j-1}) + \log b + \frac{1}{nb}S(\phi,\theta)
$$

$$
\frac{\partial A}{\partial b} = \frac{1}{b} - \frac{1}{nb^2}S(\phi,\theta) = 0
$$

$$
\Rightarrow b = \frac{1}{n}S(\phi,\theta)
$$

- **Step 1**: For fixed $(\phi, \theta)$, $l(\phi, \theta, \sigma^2)$ is minimised by

$$\underline{\underline{\hat{\sigma}^2(\phi, \theta)}} = \frac{1}{n} S(\phi, \theta).$$

- **Step 2**: Substituting this back, we get the reduced expression

$$A = -\frac{2}{n} \times l(\phi, \theta, \sigma^2) = l(\phi, \theta, \hat{\sigma}^2(\phi, \theta))$$

⇒ take derivative to find $\phi, \theta$

$$= \log(2\pi) + \frac{1}{n} \sum_{j=1}^{n} \log r_{j-1} + \log(\frac{1}{n} S(\phi, \theta)) + 1.$$ ✓

- **Step 3**: The values of $(\phi, \theta)$ that minimise this are the maximum likelihood estimates $(\hat{\phi}, \hat{\theta})$. ✓

- **Step 4**: The maximum likelihood estimate of $\sigma^2$ is $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\phi}, \hat{\theta})$. ✓

- For many applications the “determinant” term $\frac{1}{n}\sum_{j=1}^{n}\log r_{j-1}$ is negligible for large $n$ and $l(\phi,\theta)$ can be minimised by minimising the sum of squares of scaled innovations $S(\phi,\theta)$.
  - For example, for the AR(1) time series, $r_0 = 1/(1-\phi^2)$ and $r_n = 1$ for $n \geq 1$ giving

$$\frac{1}{n}\sum_{j=1}^{n}\log r_{j-1} = -\frac{1}{n}\log(1-\phi^2) \to 0$$

see the next page for calculations

provided $|\phi| < 1$ and uniformly on any closed subset of this interval.

- Least squares estimates are those obtained by choosing $(\tilde{\phi},\tilde{\theta})$ to minimise $S(\phi,\theta)$ and forming

$$\tilde{\sigma}^2 = \frac{1}{n-p-q}S(\tilde{\phi},\tilde{\theta}).$$

AR(1) $X_t = \phi X_{t-1} + Z_t$ $\quad Z_t \sim WN(0,\sigma^2)$

$\hat{X}_1 = E(X_1) = 0$

$\hat{X}_t = \phi X_{t-1} \quad t \geqslant 2$

$$\Rightarrow X_t - \hat{X}_t = \begin{cases} X_1 & \text{if } t=1 \\ X_t - \phi X_{t-1} = Z_t & \text{if } t \geqslant 2 \end{cases}$$

$$\Rightarrow v_{t-1} = E(X_t - \hat{X}_t)^2 = \begin{cases} E(X_1^2) = \gamma_X(0) = \dfrac{\sigma^2}{1-\phi^2} & t=1 \\ E(Z_t^2) = Var(Z_t) = \sigma^2 & t \geqslant 2 \end{cases}$$

$\Rightarrow v_0 = \dfrac{\sigma^2}{1-\phi^2} = \sigma^2 r_0$ where $r_0 = \dfrac{1}{1-\phi^2}$ $\Rightarrow \log r_0 = -\log 1-\phi^2$, $t=1$

for $t \geqslant 2$, $v_{t-1} = \sigma^2 = \sigma^2 r_{t-1} \Rightarrow r_{t-1} = 1 \quad t \geqslant 2 \Rightarrow \log r_{t-1} = 0$, $t \geqslant 2$

## 6.2 Optimizing the likelihood

- It can be shown that (Exercise 5.2), even for the simplest ARMA models, the likelihood is not a quadratic function of the parameters and equating first derivatives with respect to parameters to zero will not result in easily solved linear equations.

- There are many alternative ways to optimize non-linear functions and a proper discussion of this is beyond the scope of this course.

- However, we will review some key underlying concepts because these provide some insights also about how to obtain standard errors for the maximum likelihood estimates.

For this section, we refer to Shumway et al. [2000] (Pages 119-122) for more details and, as they do, let $\beta = (\phi, \theta, \sigma^2)$ denote all the parameters of the ARMA equation. Let $l(\beta)$ denote $-1/n$ times the log-likelihood which is to be *minimized* over $\beta$.

Assume that there is unique global extremum of $l(\beta)$ in the interior of the parameter space allowed for $\beta$ to range over – this is typically a subset of $\mathbb{R}^{r=p+q+1}$ in the ARMA$(p,q)$ case.

- The negative of the score function is the vector of first derivatives

$$l^{(1)}(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \left(\frac{\partial l(\beta)}{\partial \beta_1}, \dots, \frac{\partial l(\beta)}{\partial \beta_r}\right)'.$$

- To find the extremum, we can solve the first order condition that $l^{(1)}(\beta) = 0$ to obtain the maximum likelihood estimator $\hat{\beta}$.

- In a neighbourhood of $\hat{\beta}$ the matrix of second derivatives (called the Hessian) given by

$$l^{(2)}(\beta) = \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} = \left\{\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k}\right\}_{j,k=1}^{r}.$$

  is strictly negative/positive definite and therefore invertible.

- Since the score equation is non-linear, methods such as Newton-Raphson and Gauss-Newton can be used.

Consider Newton-Raphson.

- Let $\hat{\beta}^{(k)}$ be the current attempt to solve the score equation and $\hat{\beta}^{(k+1)}$ be an improvement on that attempt.

- If this improved solution was perfect we would have

  Taylor expansion

  if we assume that $\hat{\beta}^{k+1}$ is the final solution

  $$0 = l^{(1)}(\hat{\beta}^{(k+1)}) \underset{=}{\approx} l^{(1)}(\hat{\beta}^{(k)}) + l^{(2)}(\hat{\beta}^{(k)}) \left[\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\right]$$

  where the approximation is based on Taylor expansion.

- Assuming equality throughout and solving for $\hat{\beta}^{(k+1)}$ gives

$$\underline{\hat{\beta}^{(k+1)}} = \underline{\hat{\beta}^{(k)}} + \left[-l^{(2)}(\underline{\hat{\beta}^{(k)}})\right]^{-1} l^{(1)}(\hat{\beta}^{(k)}) \qquad (6.3)$$

Of course, for non-linear score functions, we will not get perfection in a single step and (6.3) needs to be iterated by replacing the old guess by the new one until convergence to the MLE $\hat{\beta}$ to some required accuracy is achieved.

**Note 6.1** *At convergence, we must have* $l^{(1)}(\hat{\beta}) = 0$ *as required.*

- For implementation of the Newtown-Raphson method first and second derivatives of the likelihood with respect to parameters $\beta$ need to be derived.
  - For ARMA models these can be readily computed using recursions based on the ARMA model.

- Variations in methods typically involve how starting values for the recursions are calculated and whether or not the log determinant term is included.

- All these variations provide the same estimates in the limit as $n \to \infty$ and have the same asymptotic properties.

## 6.3 Large Sample Distribution of MLE's

Under specific regularity conditions it can be shown that, as $n \to \infty$,

$$\hat{\beta} \to \beta_0$$

asymptotic unbaisedness

$$\sqrt{n}(\hat{\beta} - \beta_0) \to^{d} \mathrm{MVN}(0, \Omega(\beta_0))$$

where $\beta_0$ is the true parameter vector.
Also, it can be shown that

$$\hat{\Omega}(\beta_0) = \left[-l^{(2)}(\hat{\beta})\right]^{-1} \to \Omega(\beta_0)$$

so that standard errors of the estimates can be calculated as

$$\text{s.e.}(\hat{\beta}_j) = \left[\frac{1}{n}\hat{\Omega}(\beta_0)_{j,j}\right]^{1/2}$$

The large sample multivariate normal distribution applies to the maximum likelihood estimates $(\hat{\phi}, \hat{\theta})$ as well as to the least squares estimates $(\tilde{\phi}, \tilde{\theta})$ and other variations of approximation to the likelihood.

**Example 6.1** *AR(p).*

$$\sqrt{n}(\hat{\phi} - \phi_0) \rightarrow^{d} MVN(0, \Omega(\phi_0))$$

*where*

$$\Omega(\phi_0) = \sigma_0^2 \Gamma_p^{-1}(\phi_0)$$

*and in particular when* $p = 1$

$$\Omega(\phi_0) = (1 - \phi_{0,1}^2) \qquad \checkmark \qquad \textcircled{1}$$

*and when* $p = 2$

$$\Omega(\phi_0) = \begin{bmatrix} (1 - \phi_{0,2}^2) & -\phi_{0,1}(1 + \phi_{0,2}) \\ -\phi_{0,1}(1 + \phi_{0,2}) & (1 - \phi_{0,2}^2) \end{bmatrix} \qquad \checkmark \qquad \textcircled{2}$$

$$\Gamma = \left[E\left(X_i X_j\right)\right]_{i,j=1}^{p}$$

$$p=1 \Rightarrow \Gamma = E(X_1 X_1) = Var(X_1) = \gamma_X(0) = \frac{\sigma^2}{1-\phi^2}$$

$$\Omega(\phi_0) = \sigma^2 \Gamma_p^{-1} = \sigma^2 \times \frac{1-\phi^2}{\sigma^2} = 1-\phi^2 \qquad ①$$

---

AR(2)

$$\gamma(0) = \frac{1-\phi_2}{1+\phi_2} \, \frac{\sigma^2}{(1-\phi_2)^2-\phi_1^2}$$

$$\gamma(1) = \frac{\phi_1}{1-\phi_2} \, \gamma(0)$$

$$\Gamma_2 = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix} \Rightarrow \Gamma_2^{-1} = \frac{1}{(\gamma(0))^2-(\gamma(1))^2} \begin{pmatrix} \gamma(0) & -\gamma(1) \\ -\gamma(1) & \gamma(0) \end{pmatrix}$$