

# Lecture 6: Estimation of ARMA Models

---

Raul Riva

FGV EPGE

October, 2025

# Intro

---

- So far, we took parameters as given when working with ARMA models;
- In practice, we need to *estimate* these parameters from data;
- There many ways to estimate ARMA models: maximum likelihood, method of moments, Kalman filter, etc;
- We will focus on MLE estimation;
- Usually, good software for ARMA estimation gives you several options;
- More than mastering math tricks and details, it is important to understand the *big picture*;

- It is always the MA part that will complicate things;
- A natural estimator for  $AR(p)$  models is just the OLS estimator: regress  $y_t$  on  $y_{t-1}, \dots, y_{t-p}$ ;
- Mild conditions will guarantee consistency, asymptotic normality, bla, blah, blah...
- But for  $ARMA(p, q)$  models, we cannot do that! We do not observe  $\varepsilon_t$ !!!

- It is always the MA part that will complicate things;
- A natural estimator for  $AR(p)$  models is just the OLS estimator: regress  $y_t$  on  $y_{t-1}, \dots, y_{t-p}$ ;
- Mild conditions will guarantee consistency, asymptotic normality, bla, blah, blah...
- But for  $ARMA(p, q)$  models, we cannot do that! We do not observe  $\varepsilon_t$ !!!
- MLE will require a *distributional assumption* for  $\varepsilon_t$ ;
- We will relax that later when we touch on “*quasi-MLE*”;
- We will start with *given* values of  $p$  and  $q$  and discuss model choice later;

## Preliminaries

- Assume that  $\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$ ;
- This will imply that  $y_t$  is a Gaussian random variable. Why?

# Preliminaries

- Assume that  $\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$ ;
- This will imply that  $y_t$  is a Gaussian random variable. Why?
- We will denote by  $\Theta$  the vector of parameters to be estimated;
- First step: characterize the joint distribution of the sample  $\mathbf{y} = (y_1, \dots, y_T)'$ ;
- Denote this distribution by  $f_{y_T, y_{T-1}, \dots, y_1}(\mathbf{y}; \Theta)$ ;

# Preliminaries

- Assume that  $\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$ ;
- This will imply that  $y_t$  is a Gaussian random variable. Why?
- We will denote by  $\Theta$  the vector of parameters to be estimated;
- First step: characterize the joint distribution of the sample  $\mathbf{y} = (y_1, \dots, y_T)'$ ;
- Denote this distribution by  $f_{y_T, y_{T-1}, \dots, y_1}(\mathbf{y}; \Theta)$ ;
- Recall:  $f_{Y|X}(y, x) = f_{Y,X}(y, x)/f_X(x) \implies f_{Y,X}(y, x) = f_{Y|X}(y, x)f_X(x)$



# Preliminaries

- Assume that  $\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$ ;
- This will imply that  $y_t$  is a Gaussian random variable. Why?
- We will denote by  $\Theta$  the vector of parameters to be estimated;
- First step: characterize the joint distribution of the sample  $\mathbf{y} = (y_1, \dots, y_T)'$ ;
- Denote this distribution by  $f_{y_T, y_{T-1}, \dots, y_1}(\mathbf{y}; \Theta)$ ;
- Recall:  $f_{Y|X}(y, x) = f_{Y, X}(y, x) / f_X(x) \implies f_{Y, X}(y, x) = f_{Y|X}(y, x) f_X(x)$
- For any integer  $k \geq 1$ :

$$f_{y_T, y_{T-1}, \dots, y_1}(\mathbf{y}; \Theta) = f_{y_k, \dots, y_1}(y_k, \dots, y_1; \Theta) \cdot \prod_{t=k+1}^T f_{y_t|y_{t-1}, \dots, y_1}(y_t | y_{t-1}, \dots, y_1; \Theta)$$

**Questions?**

## The AR( $p$ ) Case

---

## The AR( $p$ ) Case

- Consider the AR( $p$ ) model below and let  $\Theta = (c, \phi_1, \dots, \phi_p, \sigma^2)$ :

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$$

- Notice that  $y_t | y_{t-1}, \dots, y_{t-p} \sim N(c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p}, \sigma^2)$ . Therefore:

$$f_{y_t | y_{t-1}, \dots, y_1}(y_t | y_{t-1}, \dots, y_1; \Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - c - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p})^2}{2\sigma^2}}$$

## The AR( $p$ ) Case

- Consider the AR( $p$ ) model below and let  $\Theta = (c, \phi_1, \dots, \phi_p, \sigma^2)$ :

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$$

- Notice that  $y_t | y_{t-1}, \dots, y_{t-p} \sim N(c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p}, \sigma^2)$ . Therefore:

$$f_{y_t | y_{t-1}, \dots, y_1}(y_t | y_{t-1}, \dots, y_1; \Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - c - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p})^2}{2\sigma^2}}$$

- The likelihood of the first  $p$  observations,  $f_{y_p, \dots, y_1}(y_p, \dots, y_1; \Theta)$ , is more involved;
- Notice that the  $p \times 1$  vector  $\mathbf{y}_{1:p} = (y_1, \dots, y_p)'$  is multivariate normal;

$$\mathbf{y}_{1:p} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \quad \boldsymbol{\mu} = \frac{c}{1 - \sum_{i=1}^p \phi_i} \mathbf{1}, \quad \boldsymbol{\Omega}_{ij} = \gamma(|i - j|) \quad \forall i, j \in \{1, \dots, p\}$$

## The AR( $p$ ) Case

- The likelihood of the first  $p$  observations is given by:

$$f_{y_p, \dots, y_1}(y_p, \dots, y_1; \Theta) = (2\pi)^{-p/2} |\mathbf{\Omega}^{-1}|^{1/2} e^{-\frac{1}{2}(\mathbf{y}_{1:p} - \boldsymbol{\mu})' \mathbf{\Omega}^{-1} (\mathbf{y}_{1:p} - \boldsymbol{\mu})}$$

## The AR( $p$ ) Case

- The likelihood of the first  $p$  observations is given by:

$$f_{y_p, \dots, y_1}(y_p, \dots, y_1; \Theta) = (2\pi)^{-p/2} |\mathbf{\Omega}^{-1}|^{1/2} e^{-\frac{1}{2}(\mathbf{y}_{1:p} - \boldsymbol{\mu})' \mathbf{\Omega}^{-1}(\mathbf{y}_{1:p} - \boldsymbol{\mu})}$$

- From here, we can write the full likelihood function:

$$\begin{aligned} f_{y_T, \dots, y_1}(\mathbf{y}; \Theta) &= f_{y_k, \dots, y_1}(y_k, \dots, y_1; \Theta) \cdot \prod_{t=k+1}^T f_{y_t | y_{t-1}, \dots, y_1}(y_t | y_{t-1}, \dots, y_1; \Theta) \\ &= (2\pi)^{-p/2} |\mathbf{\Omega}^{-1}|^{1/2} e^{-\frac{1}{2}(\mathbf{y}_{1:p} - \boldsymbol{\mu})' \mathbf{\Omega}^{-1}(\mathbf{y}_{1:p} - \boldsymbol{\mu})} \cdot \prod_{t=p+1}^T \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - c - \sum_{i=1}^p \phi_i y_{t-i})^2}{2\sigma^2}} \\ &= (2\pi)^{-T/2} \sigma^{-(T-p)} |\mathbf{\Omega}^{-1}|^{1/2} e^{-\frac{1}{2}(\mathbf{y}_{1:p} - \boldsymbol{\mu})' \mathbf{\Omega}^{-1}(\mathbf{y}_{1:p} - \boldsymbol{\mu})} \cdot e^{-\frac{1}{2\sigma^2} \sum_{t=p+1}^T (y_t - c - \sum_{i=1}^p \phi_i y_{t-i})^2} \end{aligned}$$

# The Log-Likelihood Function

- We always optimize the log-likelihood function  $\mathcal{L}(\Theta|\mathbf{y}) = \log \left( f_{y_T, \dots, y_1}(\mathbf{y}; \Theta) \right)$

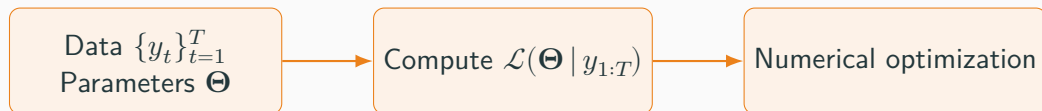
$$\begin{aligned}\mathcal{L}(\Theta|\mathbf{y}) &= \log \left( f_{y_T, \dots, y_1}(\mathbf{y}; \Theta) \right) \\ &= -\frac{T}{2} \log(2\pi) \\ &\quad - (T - p) \log(\sigma) + \frac{1}{2} \log(|\mathbf{\Omega}^{-1}|) \\ &\quad - \frac{1}{2} (\mathbf{y}_{1:p} - \boldsymbol{\mu})' \mathbf{\Omega}^{-1} (\mathbf{y}_{1:p} - \boldsymbol{\mu}) - \frac{1}{2\sigma^2} \sum_{t=p+1}^T (y_t - c - \sum_{i=1}^p \phi_i y_{t-i})^2\end{aligned}$$

- The blue part looks like the OLS objective function;
- The red part is “distorting” this objective function;



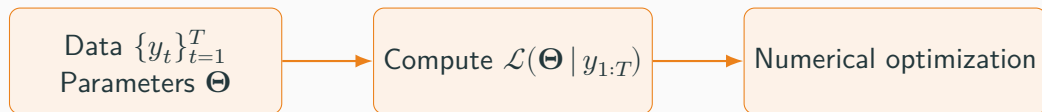
# The Log-Likelihood Function

- Full ML estimation requires optimizing this function w.r.t.  $\Theta$ ;
- Notice that this requires inverting a  $p \times p$  matrix any time we evaluate the function;



# The Log-Likelihood Function

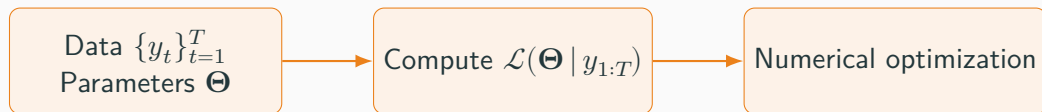
- Full ML estimation requires optimizing this function w.r.t.  $\Theta$ ;
- Notice that this requires inverting a  $p \times p$  matrix any time we evaluate the function;



- Wait a minute... what if  $T \gg p$ ?
- In that case the main contribution to the log-likelihood function comes from the blue part;

# The Log-Likelihood Function

- Full ML estimation requires optimizing this function w.r.t.  $\Theta$ ;
- Notice that this requires inverting a  $p \times p$  matrix any time we evaluate the function;



- Wait a minute... what if  $T \gg p$ ?
- In that case the main contribution to the log-likelihood function comes from the blue part;
- This suggests a simpler approach: *conditional* MLE;
- Assume that the first  $p$  observations are fixed (non-random);
- Approximate  $\mathcal{L}(\Theta | \mathbf{y}_{1:T})$  by  $\log \left( f_{y_{p+1}, \dots, y_T | y_{1:p}}(\mathbf{y}; \Theta) \right)$

## The Numerical Shortcut for the AR( $p$ ) Case

- Recall that, up to a constant, we have:

$$\log \left( f_{y_{p+1}, \dots, y_T | y_{1:p}}(\mathbf{y}; \boldsymbol{\Theta}) \right) = - \sum_{t=p+1}^T \frac{(y_t - c - \sum_{i=1}^p \phi_i y_{t-i})^2}{2\sigma^2} - (T-p) \log(\sigma)$$

# The Numerical Shortcut for the AR( $p$ ) Case

- Recall that, up to a constant, we have:

$$\log \left( f_{y_{p+1}, \dots, y_T | y_{1:p}}(\mathbf{y}; \boldsymbol{\Theta}) \right) = - \sum_{t=p+1}^T \frac{(y_t - c - \sum_{i=1}^p \phi_i y_{t-i})^2}{2\sigma^2} - (T-p) \log(\sigma)$$

- Estimators for  $c$  and  $\phi_i$ 's are the same as the OLS from regressing  $y_t$  on  $y_{t-1}, \dots, y_{t-p}$ ;
- Super simple closed-form solutions! 😎
- The estimator for  $\sigma^2$  is just the (biased) sample variance of the OLS residuals;

# The Numerical Shortcut for the AR( $p$ ) Case

- Recall that, up to a constant, we have:

$$\log \left( f_{y_{p+1}, \dots, y_T | y_{1:p}}(\mathbf{y}; \Theta) \right) = - \sum_{t=p+1}^T \frac{(y_t - c - \sum_{i=1}^p \phi_i y_{t-i})^2}{2\sigma^2} - (T-p) \log(\sigma)$$

- Estimators for  $c$  and  $\phi_i$ 's are the same as the OLS from regressing  $y_t$  on  $y_{t-1}, \dots, y_{t-p}$ ;
- Super simple closed-form solutions! 😎
- The estimator for  $\sigma^2$  is just the (biased) sample variance of the OLS residuals;
- If  $T$  is large, this is a very good approximation to the full MLE;
- $\mathcal{L}(\Theta | \mathbf{y})$  is efficiently computed using the Kalman filter – darker magic for the next year!

**Questions?**

## The MA( $q$ ) Case

- Consider the MA( $q$ ) model below and let  $\Theta = (\mu, \theta_1, \dots, \theta_q, \sigma^2)$ :

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$$

- There is no hope to get an “OLS”-type trick... we do not see the shocks...
- There are again two main approaches: full MLE and conditional MLE;
- We will focus on the conditional MLE approach;
- You can see the full MLE approach in Hamilton's book (Chapter 5);
- If  $T$  is large, the two approaches will give very similar results;
- Similar to the forecasting exercise in the last lecture!



## The MA( $q$ ) Case

- The key observation is that  $y_t | \varepsilon_{t-1}, \dots, \varepsilon_{t-q} \sim N(\mu + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \sigma^2)$ ;
- But how is that useful if we do not observe  $\varepsilon_t$ ?

## The MA( $q$ ) Case

- The key observation is that  $y_t | \varepsilon_{t-1}, \dots, \varepsilon_{t-q} \sim N(\mu + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \sigma^2)$ ;
- But how is that useful if we do not observe  $\varepsilon_t$ ?
- Let's assume that  $\varepsilon_{-q+1} = \varepsilon_{-q+2} = \dots = \varepsilon_0 = \mathbb{E}[\varepsilon_t] = 0$ ;
- We can start a recursion, like in the forecasting case:

$$\varepsilon_1 = y_1 - \mu$$

$$\varepsilon_2 = y_2 - \mu - \theta_1 \varepsilon_1$$

$$\varepsilon_3 = y_3 - \mu - \theta_1 \varepsilon_2 - \theta_2 \varepsilon_1$$

$$\vdots$$

$$\varepsilon_t = y_t - \mu - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

$$\vdots$$

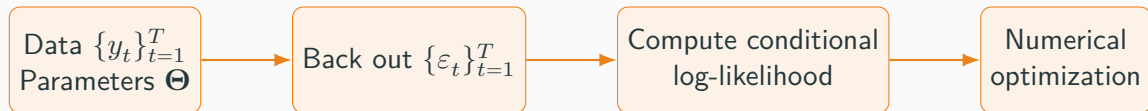
$$\varepsilon_T = y_T - \mu - \theta_1 \varepsilon_{T-1} - \dots - \theta_q \varepsilon_{T-q}$$

# The Conditional Log-Likelihood Function

- From here, we can write the conditional log-likelihood function:

$$\begin{aligned}\log \left( f_{y_t, \dots, y_1 | \varepsilon_{-q+1} = \varepsilon_{-q+2} = \dots = \varepsilon_0 = 0}(\mathbf{y}; \Theta) \right) &= \sum_{t=q+1}^T \log \left( f_{y_t | \varepsilon_{t-1}, \dots, \varepsilon_{t-q}}(y_t | \varepsilon_{t-1}, \dots, \varepsilon_{t-q}; \Theta) \right) \\ &= - \sum_{t=1}^T \frac{(\varepsilon_t)^2}{2\sigma^2} - (T - q) \log(\sigma)\end{aligned}$$

- When there is an MA component, the logical flow is:



Questions?

## The ARMA( $p, q$ ) Case

---

## The ARMA( $p, q$ ) Case

- Consider a Gaussian ARMA( $p, q$ ) model and let  $\Theta = (c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2)$ :

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2)$$

- We can combine the two previous approaches;
- Given  $\Theta$ , we will back out  $\varepsilon_t$  recursively;
- We also note that  $y_t | y_{t-1}, \dots, y_1, \varepsilon_{t-1}, \dots, \varepsilon_{t-q} \sim N(c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p}, \sigma^2)$
- Then we are ready to use the conditioning trick once again!

# The Recursion

- As we did with the  $AR(p)$ , assume  $y_1, \dots, y_p$  are fixed;
- Assume that  $\varepsilon_p = \varepsilon_{p-1} = \dots = \varepsilon_{p-q+1} = 0$
- The first shock to be backed out is  $\varepsilon_{p+1} = y_{p+1} - c - \sum_{i=1}^p \phi_i y_{p+1-i}$
- Then we get  $\varepsilon_{p+2} = y_{p+2} - c - \sum_{i=1}^p \phi_i y_{p+2-i} - \theta_1 \varepsilon_{p+1}$
- And so on...
- You might be skeptical of “assuming values” for the shock... but usually  $p$  and  $q$  are small compared to  $T$ !
- You will almost never see  $q > 10$  and  $p > 20$  in practice!

## The Conditional Log-Likelihood Function

- The conditional log-likelihood function, up to a constant, is given by:

$$\mathcal{L}(\Theta|\mathbf{y}) = \log \left( f_{y_t, \dots, y_1 | \varepsilon_{-q+1} = \varepsilon_{-q+2} = \dots = \varepsilon_0 = 0}(\mathbf{y}; \Theta) \right) = - \sum_{t=p+1}^T \frac{(\varepsilon_t)^2}{2\sigma^2} - (T-p) \log(\sigma)$$

- The logical flow is the same as in the MA( $q$ ) case;



# The Conditional Log-Likelihood Function

- The conditional log-likelihood function, up to a constant, is given by:

$$\mathcal{L}(\Theta|\mathbf{y}) = \log \left( f_{y_t, \dots, y_1 | \varepsilon_{-q+1} = \varepsilon_{-q+2} = \dots = \varepsilon_0 = 0}(\mathbf{y}; \Theta) \right) = - \sum_{t=p+1}^T \frac{(\varepsilon_t)^2}{2\sigma^2} - (T-p) \log(\sigma)$$

- The logical flow is the same as in the MA( $q$ ) case;

## Regarding numerical optimization:

- Do we have guarantees the numerical method will converge to the global maximum? No.
- Is it much harder as we increase  $p$  and  $q$ ? Yes and no: increasing  $p$  is fine, but  $q$  is hell;
- Where to start the optimization? OLS estimates for  $\phi$  are a good shot;
- What about  $\theta$ ? Start with zeros or small values;
- Try several different starting points and make sure you get similar answers;

# Inference

---

# Inference

- Ok great, we can estimate  $\text{ARMA}(p, q)$  models;
- How to do inference?

- Ok great, we can estimate  $\text{ARMA}(p, q)$  models;
- How to do inference?
- We will use standard MLE results;
- Important assumptions: a correctly specified model and  $\Theta_0$  must be an interior point;
- Recall that, if the model is correctly specified, then:

$$\sqrt{T}(\hat{\Theta} - \Theta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\Theta_0))$$

where  $\mathcal{I}(\Theta)$  is the Fisher information matrix;

- Recall that, in this case,  $\mathcal{I}(\Theta) = -\mathbb{E} \left[ \frac{\partial^2 l_t(\Theta)}{\partial \Theta \partial \Theta'} \right] = \mathbb{E} \left[ \frac{\partial l_t(\Theta|\mathbf{y})}{\partial \Theta} \frac{\partial l_t(\Theta|\mathbf{y})}{\partial \Theta'} \right]$ , where  $l_t(\Theta|\mathbf{y}) = \log \left( f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \Theta) \right)$ ;

## Feasible Inference

- Of course we do not know  $\mathcal{J}(\Theta_0) \implies$  it needs to be estimated!

- Of course we do not know  $\mathcal{J}(\Theta_0) \implies$  it needs to be estimated!
- Theory suggests two equally valid ways of estimating it. Let us define two objects:

1. The Hessian:

$$\mathcal{H}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \frac{\partial^2 \mathcal{L}(\hat{\Theta}|\mathbf{y})}{\partial \Theta \partial \Theta'} = \frac{1}{T-p} \cdot \sum_{t=p+1}^T \frac{\partial^2 \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta \partial \Theta'}$$

# Feasible Inference

- Of course we do not know  $\mathcal{J}(\Theta_0) \implies$  it needs to be estimated!
- Theory suggests two equally valid ways of estimating it. Let us define two objects:

1. The Hessian:

$$\mathcal{H}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \frac{\partial^2 \mathcal{L}(\hat{\Theta}|\mathbf{y})}{\partial \Theta \partial \Theta'} = \frac{1}{T-p} \cdot \sum_{t=p+1}^T \frac{\partial^2 \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta \partial \Theta'}$$

2. The score function and its associated *outer product*:

$$\mathcal{S}(\hat{\Theta})_t \equiv \frac{\partial \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta}; \quad \mathcal{O}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \sum_{t=p+1}^T \mathcal{S}(\hat{\Theta})_t \mathcal{S}(\hat{\Theta})_t'$$

## Feasible Inference

- Of course we do not know  $\mathcal{J}(\Theta_0) \implies$  it needs to be estimated!
- Theory suggests two equally valid ways of estimating it. Let us define two objects:

1. The Hessian:

$$\mathcal{H}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \frac{\partial^2 \mathcal{L}(\hat{\Theta}|\mathbf{y})}{\partial \Theta \partial \Theta'} = \frac{1}{T-p} \cdot \sum_{t=p+1}^T \frac{\partial^2 \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta \partial \Theta'}$$

2. The score function and its associated *outer product*:

$$\mathcal{S}(\hat{\Theta})_t \equiv \frac{\partial \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta}; \quad \mathcal{O}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \sum_{t=p+1}^T \mathcal{S}(\hat{\Theta})_t \mathcal{S}(\hat{\Theta})_t'$$

- Then, we can estimate  $\mathcal{J}(\Theta_0)$  by either  $[-\mathcal{H}(\hat{\Theta})]$  or  $[\mathcal{O}(\hat{\Theta})]$ ;



# Feasible Inference

- Of course we do not know  $\mathcal{J}(\Theta_0) \implies$  it needs to be estimated!
- Theory suggests two equally valid ways of estimating it. Let us define two objects:

1. The Hessian:

$$\mathcal{H}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \frac{\partial^2 \mathcal{L}(\hat{\Theta}|\mathbf{y})}{\partial \Theta \partial \Theta'} = \frac{1}{T-p} \cdot \sum_{t=p+1}^T \frac{\partial^2 \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta \partial \Theta'}$$

2. The score function and its associated *outer product*:

$$\mathcal{S}(\hat{\Theta})_t \equiv \frac{\partial \log(f_{y_t|\mathbf{y}_{t-1}}(y_t|\mathbf{y}_{t-1}; \hat{\Theta}))}{\partial \Theta}; \quad \mathcal{O}(\hat{\Theta}) \equiv \frac{1}{T-p} \cdot \sum_{t=p+1}^T \mathcal{S}(\hat{\Theta})_t \mathcal{S}(\hat{\Theta})_t'$$

- Then, we can estimate  $\mathcal{J}(\Theta_0)$  by either  $[-\mathcal{H}(\hat{\Theta})]$  or  $[\mathcal{O}(\hat{\Theta})]$ ;
- (Adjust the starting point of the sum as needed, it doesn't matter asymptotically);

- What if  $\varepsilon_t$  is not Gaussian?
- The MLE will converge to a different  $\Theta_0$ ;
- This parameter is the *pseudo-true* value that minimizes the Kullback-Leibler divergence between the true model and the assumed Gaussian model;
- The idea, and the term *Quasi-MLE*, is due to White (Econometrica, 1982);

- What if  $\varepsilon_t$  is not Gaussian?
- The MLE will converge to a different  $\Theta_0$ ;
- This parameter is the *pseudo-true* value that minimizes the Kullback-Leibler divergence between the true model and the assumed Gaussian model;
- The idea, and the term *Quasi-MLE*, is due to White (Econometrica, 1982);
- The asymptotic distribution is now:

$$\sqrt{T}(\hat{\Theta} - \Theta_0) \xrightarrow{d} N\left(0, \underbrace{\mathcal{H}^{-1}(\Theta_0) \mathcal{J}(\Theta_0) \mathcal{H}^{-1}(\Theta_0)'}_{\text{the "sandwich" variance}}\right)$$

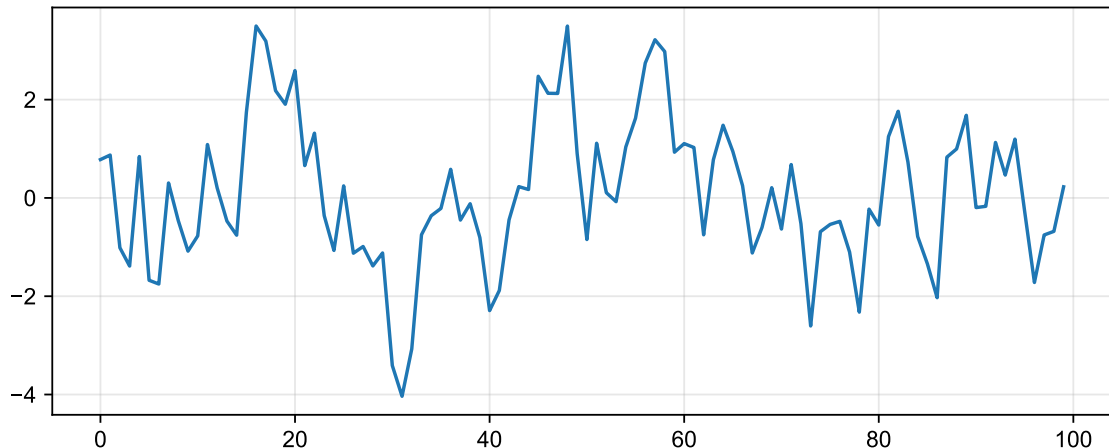
- The “bread” uses the Hessian and the “meat” uses the outer product of the score;
- The estimator for the sandwich is  $\left[-\mathcal{H}^{-1}(\hat{\Theta}) \mathcal{O}(\hat{\Theta}) \mathcal{H}^{-1}(\hat{\Theta})'\right]$

## Some Simulations

---

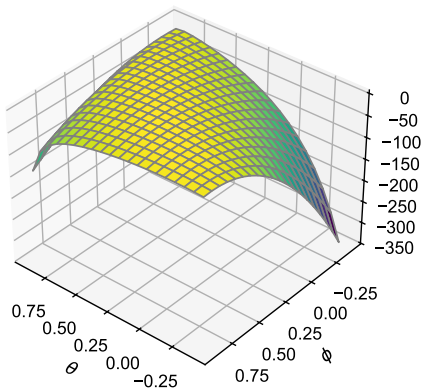
## The ARMA(1,1) Case - Simulated Path ( $\phi = 0.5, \theta = 0.2$ )

Let's say we have an ARMA(1,1), with  $\mu = 0$  and  $\sigma^2 = 1$ . Let's simulate one path:

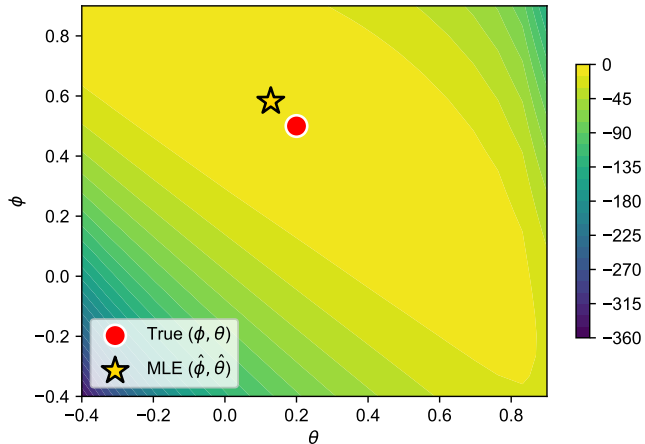


# The Likelihood Surface ( $T=100$ )

3D Likelihood

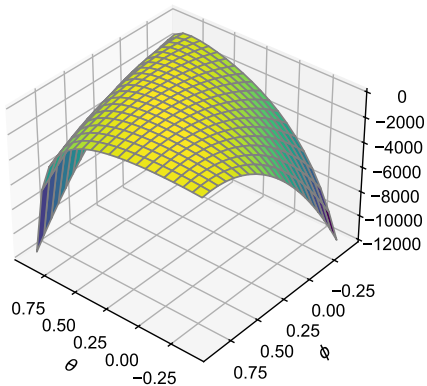


Contour: True vs MLE ( $T=100$ )

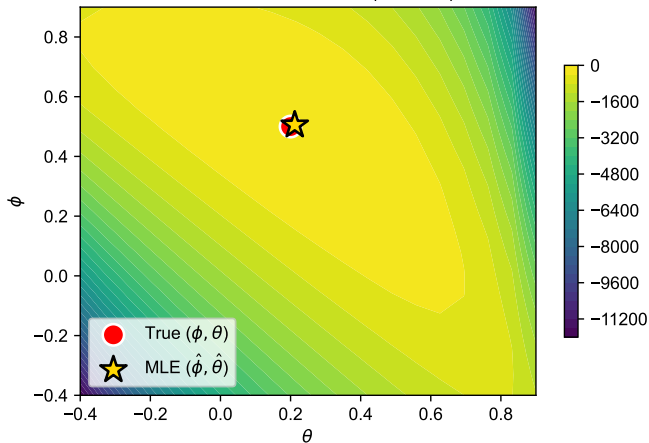


# The Likelihood Surface ( $T=5000$ )

3D Likelihood

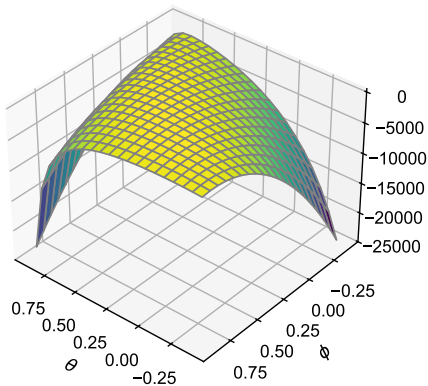


Contour: True vs MLE ( $T=5000$ )

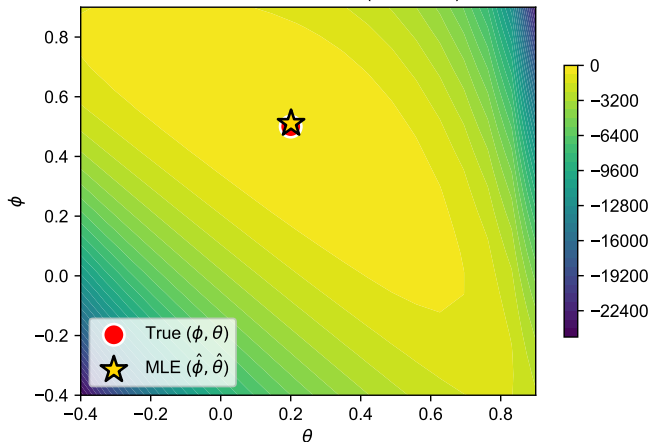


# The Likelihood Surface ( $T=10000$ )

3D Likelihood

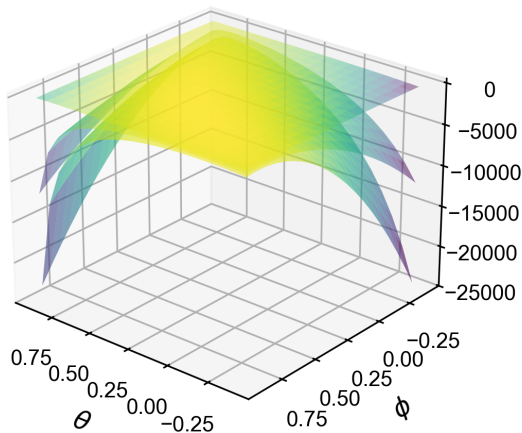


Contour: True vs MLE ( $T=10000$ )





# The Likelihood Gets More Concentrated!



- $\uparrow T \Rightarrow$  tighter likelihood. Why?
- How is the Hessian at the optimum related to this?
- What is the connection with the asymptotic distribution of the MLE?

**Questions?**

**How to choose  $p$  and  $q$ ?**

---

# The Model Selection Problem

- So far, we have assumed that  $p$  and  $q$  are known;
- In practice, we need to choose them from the data;
- Naive approach: maximize the log-likelihood over all possible  $(p, q)$  pairs;

# The Model Selection Problem

- So far, we have assumed that  $p$  and  $q$  are known;
- In practice, we need to choose them from the data;
- Naive approach: maximize the log-likelihood over all possible  $(p, q)$  pairs;
- But this will **always** favor larger models! Why?
  - Adding parameters can never decrease the maximized log-likelihood;
  - Converges to a perfect fit *in\_sample* as  $(p, q) \rightarrow \infty$  (overfitting);

# The Model Selection Problem

- So far, we have assumed that  $p$  and  $q$  are known;
- In practice, we need to choose them from the data;
- Naive approach: maximize the log-likelihood over all possible  $(p, q)$  pairs;
- But this will **always** favor larger models! Why?
  - Adding parameters can never decrease the maximized log-likelihood;
  - Converges to a perfect fit *in\_sample* as  $(p, q) \rightarrow \infty$  (overfitting);
- We need a formal criterion that **penalizes model complexity**;
- This leads to *information criteria*: balance fit vs. parsimony;

- The general form of information criteria is:

$$\text{IC} = -2 \cdot \mathcal{L}(\hat{\Theta}|\mathbf{y}) + \text{penalty}(k, T)$$

where  $k$  is the number of parameters and  $T$  is the sample size;

- The general form of information criteria is:

$$\text{IC} = -2 \cdot \mathcal{L}(\hat{\Theta}|\mathbf{y}) + \text{penalty}(k, T)$$

where  $k$  is the number of parameters and  $T$  is the sample size;

- The first term measures **goodness-of-fit** (we want it small);
- The second term **penalizes complexity** (increases with  $k$ );
- We choose the model that **minimizes** the IC;



## Information Criteria: General Framework

- The general form of information criteria is:

$$\text{IC} = -2 \cdot \mathcal{L}(\hat{\Theta}|\mathbf{y}) + \text{penalty}(k, T)$$

where  $k$  is the number of parameters and  $T$  is the sample size;

- The first term measures **goodness-of-fit** (we want it small);
- The second term **penalizes complexity** (increases with  $k$ );
- We choose the model that **minimizes** the IC;
- Different penalties lead to different criteria;
- The key trade-off: smaller penalty  $\implies$  more likely to select larger models;

## The Main Information Criteria

Let  $k = p + q + 2$  be the number of parameters in an  $\text{ARMA}(p, q)$  model (including  $c$  and  $\sigma^2$ ).

## The Main Information Criteria

Let  $k = p + q + 2$  be the number of parameters in an ARMA( $p, q$ ) model (including  $c$  and  $\sigma^2$ ).

**Akaike Information Criterion (AIC):**

$$\text{AIC} = -2 \cdot \mathcal{L}(\hat{\Theta}|\mathbf{y}) + 2k$$

- It will find that minimizes the expected KL divergence to the true model;
- It might deliver an over-parametrized model (remember the minimal representation?);

# The Main Information Criteria

Let  $k = p + q + 2$  be the number of parameters in an ARMA( $p, q$ ) model (including  $c$  and  $\sigma^2$ ).

**Akaike Information Criterion (AIC):**

$$\text{AIC} = -2 \cdot \mathcal{L}(\hat{\Theta}|\mathbf{y}) + 2k$$

- It will find that minimizes the expected KL divergence to the true model;
- It might deliver an over-parametrized model (remember the minimal representation?);

**Bayesian Information Criterion (BIC) or Schwarz Criterion (SIC):**

$$\text{BIC} = -2 \cdot \mathcal{L}(\hat{\Theta}|\mathbf{y}) + k \log(T)$$

- It approximates the model with the highest posterior probability (assuming equal priors);
- It is **consistent**: selects the true model (if it is in the candidate set) with probability  $\rightarrow 1$  as  $T \rightarrow \infty$ ;

## Comparing the Penalties

- Notice that for  $T > 8$ , we have  $\log(T) > 2$ , so BIC penalizes more heavily than AIC;

Sample Size	AIC penalty	BIC penalty
$T = 50$	$2k$	$3.91k$
$T = 100$	$2k$	$4.61k$
$T = 500$	$2k$	$6.21k$
$T = 1000$	$2k$	$6.91k$

## Comparing the Penalties

- Notice that for  $T > 8$ , we have  $\log(T) > 2$ , so BIC penalizes more heavily than AIC;

Sample Size	AIC penalty	BIC penalty
$T = 50$	$2k$	$3.91k$
$T = 100$	$2k$	$4.61k$
$T = 500$	$2k$	$6.21k$
$T = 1000$	$2k$	$6.91k$

- As  $T \rightarrow \infty$ : BIC penalty grows much faster than AIC;
- Implication: BIC tends to select **more parsimonious models** than AIC;

## Step-by-step procedure:

1. Choose a maximum order  $p_{\max}$  and  $q_{\max}$  (often based on theory or exploratory analysis);
2. Estimate all ARMA( $p, q$ ) models for  $p \in \{0, 1, \dots, p_{\max}\}$  and  $q \in \{0, 1, \dots, q_{\max}\}$ ;
3. Compute your chosen IC for each model;
4. Select the model with the **minimum** IC value;

# How to Use Information Criteria in Practice

## Step-by-step procedure:

1. Choose a maximum order  $p_{\max}$  and  $q_{\max}$  (often based on theory or exploratory analysis);
2. Estimate all ARMA( $p, q$ ) models for  $p \in \{0, 1, \dots, p_{\max}\}$  and  $q \in \{0, 1, \dots, q_{\max}\}$ ;
3. Compute your chosen IC for each model;
4. Select the model with the **minimum** IC value;

## Important notes:

- All models must be estimated on the **same sample** (same  $T$ );
- Start with reasonable  $p_{\max}$  and  $q_{\max}$  (e.g., 5-10 for quarterly data, 12-24 for monthly);
- If the selected model is at the boundary, consider increasing the maximum orders;



**The End**

- Chapter 5 from Hamilton's book for ARMA estimation;
- Chapter 28 from Hansen's book on model selection for MLE;