25 Spring 439/639 TSA: Lecture 15

Dr Sergey Kushnarev

Table of contents

Method of Moments (MoM) for other models and other parameters

Recall that in MoM (or GMoM), we solve

$$\begin{aligned} &\text{theoretical moment} = \text{sample moment} \\ \text{or} \quad &\text{theoretical ACF} = \text{sample ACF} \end{aligned}$$

where the theoretical moment μ_k, theoretical ACF ρ_k are functions of the parameters to estimate (like the ϕ_i, θ_i), and the sample moment m_k, sample ACF r_k are functions of the observed samples $(Y_1, ..., Y_n)$.

MoM for MA(1)

Last time we used MoM for AR(p). Let's look at another example.

Suppose $Y_1, \dots, Y_n$ are from an MA(1) model

$$Y_t = e_t - \theta\, e_{t-1}.$$

Consider the GMoM method, we need to solve

$$\rho_1 = r_1 \implies -\frac{\theta}{1+\theta^2} = r_1 \implies r_1\theta^2 + \theta + r_1 = 0.$$

If $r_1 = 0$, we get $\hat{\theta}_{\text{MOM}} = 0$. If $r_1 \neq 0$, we get

$$\theta_{1,2} = \frac{-1 \pm \sqrt{1-4r_1^2}}{2r_1}$$

MoM estimator does not always exist. The theoretical ACF ρ_1 for MA(1) always satisfy $|\rho_1| \leq \frac{1}{2}$, but the sample ACF may have $|r_1| > \frac{1}{2}$. If $|r_1| > \frac{1}{2}$ happens (which depends on the randomness in observations), then we cannot find the MoM estimator in this scenario (since the solutions above are not real).

Suppose $0 < |r_1| < \frac{1}{2}$, then we have two solutions for θ. In this case, we always have

$$\left|\frac{-1+\sqrt{1-4r_1^2}}{2r_1}\right| < 1 < \left|\frac{-1-\sqrt{1-4r_1^2}}{2r_1}\right|.$$

We choose the $|\theta| < 1$ one to make the estimated MA(1) invertible. So for $0 < |r_1| < \frac{1}{2}$, we get

$$\hat{\theta}_{\text{MOM}} = \frac{-1+\sqrt{1-4r_1^2}}{2r_1}.$$

In general, for MA(q), MoM method results in highly nonlinear equations with possibly many solutions, but only one $(\hat{\theta}_1, \ldots, \hat{\theta}_q)$ of them corresponds to an invertible model.

MoM estimate for μ

Suppose we want to estimate the mean of the time series, μ.

$$\hat{\mu}_{\text{MOM}} = \frac{1}{n}\sum_{t=1}^{n} Y_t = \overline{Y}.$$

MoM estimate for σ_e^2

Suppose we want to estimate the variance of the noise, σ_e^2.

The basic idea is:

1. Express γ_0 in terms of ϕ_i, θ_i, ρ_i and σ_e^2
2. Then we can solve σ_e^2 from step 1, i.e., express σ_e^2 in terms of γ_0, ϕ_i, θ_i, ρ_i.
3. Obtain the MoM estimates $\hat{\phi}_i$, $\hat{\theta}_i$ for the parameters ϕ_i's and θ_i's.
4. From step 2, replace the theoretical parameters/ACF/ACVFs by the corresponding estimated/sample version to get the MoM estimate $\hat{\sigma}_e^2$. To be specific, we do the following plug-in:

$$\begin{aligned}&\phi_i \to \hat{\phi}_i, \quad \theta_i \to \hat{\theta}_i, \quad \rho_i \to r_i,\\ &\text{and} \quad \gamma_0 \to \hat{\gamma}_0 = s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \overline{Y})^2.\end{aligned}$$

Example. Consider AR(p):

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p} = e_t.$$

We first express γ_0 in terms of other parameters (including σ_e^2) and ACFs. The 0-th YW equation is

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma_e^2 = \gamma_0(\phi_1\rho_1 + \phi_2\rho_2 + \cdots + \phi_p\rho_p) + \sigma_e^2.$$

So we can express σ_e^2 in terms of γ_0 and other parameters/ACFs

$$\sigma_e^2 = \gamma_0\left(1 - \phi_1\rho_1 - \phi_2\rho_2 - \cdots - \phi_p\rho_p\right).$$

By the plug-in rule we stated above, the MoM estimate for σ_e^2 is

$$\hat{\sigma}_e^{2\,\text{MOM}} = s^2\left(1 - \hat{\phi}_1^{\text{MOM}} r_1 - \hat{\phi}_2^{\text{MOM}} r_2 - \cdots - \hat{\phi}_p^{\text{MOM}} r_p\right).$$

Least-Squares Estimation (conditional LS)

The idea is to construct an objective function/loss function that is a sum of squares, then obtain the estimated parameters by minimizing this sum of squares.

Example: AR(1)

Consider an AR(1) with mean mu:

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + e_t, \quad e_t \sim \text{iid}(0, \sigma_e^2).$$

The goal is to estimate the parameters ϕ and μ. Given observations $(Y_1, ..., Y_n)$, we can define the objective function

$$S_c(\phi, \mu) = \sum_{t=2}^{n} e_t^2 = \sum_{t=2}^{n} \left[(Y_t - \mu) - \phi(Y_{t-1} - \mu)\right]^2$$

Remark: the subscript "c" stands for conditional. In this example, the summation starts from $t = 2$, so it can be thought as we are assuming/conditioning on $e_1 = 0$. This explanation is not very satisfying here, since we can only compute e_2 to e_t given $(Y_1, ..., Y_n)$ and the AR(1) model. But the sense of "conditional" will be more apparent in the MA example later.

The conditional LS estimator is the minimizer of the objective function S_c, i.e.,

$$\left(\hat{\phi}_{\text{LS}}, \hat{\mu}_{\text{LS}}\right) = \underset{\phi,\, \mu}{\arg\min}\, S_c(\phi, \mu).$$

To minimize it, we take the partial derivatives, set them to zero, and solve the equations.

$$\frac{\partial S_c}{\partial \mu} = 2\sum_{t=2}^{n} \left(Y_t - \mu - \phi(Y_{t-1} - \mu)\right)(-1 + \phi),$$
$$\frac{\partial S_c}{\partial \phi} = 2\sum_{t=2}^{n} \left((Y_t - \mu) - \phi(Y_{t-1} - \mu)\right)(-Y_{t-1} + \mu).$$

We want to get a stationary AR(1), so $\phi \neq 1$. Then setting $\frac{\partial S_c}{\partial \mu} = 0$ gives

$$\sum_{t=2}^{n} (Y_t - \mu - \phi(Y_{t-1} - \mu)) = 0$$
$$\implies \left(\sum_{t=2}^{n} Y_t\right) - (n-1)\mu = \phi\left(\sum_{t=1}^{n-1} Y_t\right) - \phi(n-1)\mu,$$

so the conditional LS estimator $\left(\hat{\phi}_{\text{LS}}, \hat{\mu}_{\text{LS}}\right)$ satisfies

$$\hat{\mu}_{\text{LS}} = \frac{\sum_{t=2}^{n} Y_t - \hat{\phi}_{\text{LS}}\left(\sum_{t=1}^{n-1} Y_t\right)}{(n-1)(1 - \hat{\phi}_{\text{LS}})} = \frac{\frac{1}{n-1}\left(\sum_{t=2}^{n} Y_t\right) - \hat{\phi}_{\text{LS}}\frac{1}{n-1}\left(\sum_{t=1}^{n-1} Y_t\right)}{1 - \hat{\phi}_{\text{LS}}} \approx \overline{Y},$$

where the last step is because $\frac{1}{n-1}\left(\sum_{t=2}^{n} Y_t\right) \approx \frac{1}{n-1}\left(\sum_{t=1}^{n-1} Y_t\right) \approx \frac{1}{n}\left(\sum_{t=1}^{n} Y_t\right) = \overline{Y}$ when n is large. For this phenomenon, we also say "$\hat{\mu}_{\text{LS}} \approx \overline{Y}$ **except for end effects**".

Setting $\frac{\partial S_c}{\partial \mu} = 0$ gives

$$\sum_{t=2}^{n} \left((Y_t - \mu) - \phi(Y_{t-1} - \mu)\right)(-Y_{t-1} + \mu) = 0.$$

Consider the large sample setting, where we just got $\hat{\mu}_{\text{LS}} \approx \overline{Y}$. Then $\hat{\phi}_{\text{LS}}$ (approximately) satisfies

$$\sum_{t=2}^{n} \left((Y_t - \overline{Y}) - \phi(Y_{t-1} - \overline{Y})\right)\left(-Y_{t-1} + \overline{Y}\right) = 0$$
$$\implies \sum_{t=2}^{n} (Y_t - \overline{Y})(Y_{t-1} - \overline{Y}) = \phi\sum_{t=2}^{n} (Y_{t-1} - \overline{Y})^2.$$

So $\hat{\phi}_{\text{LS}}$ (under large sample setting) is approximately

$$\hat{\phi}_{\text{LS}} = \frac{\sum_{t=2}^{n}(Y_t - \overline{Y})(Y_{t-1} - \overline{Y})}{\sum_{t=2}^{n}(Y_{t-1} - \overline{Y})^2} \approx r_1.$$

So when sample size n is large (except for end effects), the conditional LS estimator for AR(1) is approximately

$$\hat{\mu}_{\text{LS}} \approx \overline{Y}, \quad \hat{\phi}_{\text{LS}} \approx r_1 = \hat{\phi}_{\text{MOM}}.$$

Remark: For comparison, last lecture we showed the MoM for AR(1) is $\hat{\phi}_{\text{MOM}} = r_1$. So $\hat{\phi}_{\text{LS}} \approx \hat{\phi}_{\text{MOM}}$ when sample size n is large.

Example: AR(2)

Consider an AR(2) with mean mu:

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t.$$

Given observations $(Y_1, ..., Y_n)$, we define the objective function as

$$S_c(\phi_1, \phi_2, \mu) = \sum_{t=3}^{n} e_t^2 = \sum_{t=2}^{n} \left[(Y_t - \mu) - \phi_1(Y_{t-1} - \mu) - \phi_2(Y_{t-2} - \mu)\right]^2.$$

It can be shown that, (similar to the previous AR(1) example,) when n is large,

$$\frac{\partial S_c}{\partial \mu} = 0 \overset{n \text{ large}}{\Longrightarrow} \hat{\mu}_{\text{LS}} \approx \overline{Y}$$

$$\begin{cases} \frac{\partial S_c}{\partial \phi_1} = 0 \\ \frac{\partial S_c}{\partial \phi_2} = 0 \end{cases} \overset{n \text{ large}}{\Longrightarrow} \begin{cases} r_1 \approx \hat{\phi}_1^{\text{LS}} + \hat{\phi}_2^{\text{LS}} r_1 \\ r_2 \approx \hat{\phi}_1^{\text{LS}} r_1 + \hat{\phi}_2^{\text{LS}} \end{cases}$$

Note that the latter system (as equations for $(\hat{\phi}_1^{\text{LS}}, \hat{\phi}_2^{\text{LS}})$) is exactly same as the "sample YW equations" we saw in last lecture (see the AR(2) or AR(p) example of MoM). So the solution to the system above is same as the MoM estimator of AR(2).

So when sample size n is large (except for end effects), the conditional LS estimator for AR(2) is approximately

$$\hat{\mu}_{\text{LS}} \approx \overline{Y}, \quad \hat{\phi}_1^{\text{LS}} \approx \hat{\phi}_1^{\text{MOM}}, \quad \hat{\phi}_2^{\text{LS}} \approx \hat{\phi}_2^{\text{MOM}}.$$

Example: MA(1)

Consider an MA(1):

$$Y_t = e_t - \theta e_{t-1}.$$

From the previous examples, we know that we hope to construc some objective functions in the form $S_c = \sum e_t^2$. The question is how to express these e_t using observed data $(Y_1, ..., Y_n)$.

One idea: using the invertible (assuming $|\theta| < 1$) representation of MA(1), i.e., $e_t = Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \cdots$. Given $(Y_1, ..., Y_n)$, we only look at e_1 through e_n and truncate these infinite sums. So we get

$$S_c(\theta) = \sum_{t=1}^{n} e_t^2 = \left(Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \cdots + \theta^{t-1} Y_1\right)^2.$$

Another way to think about it: Assume $e_0 = 0$, then using the MA(1) equation, we have

$$\begin{aligned} e_1 &= \theta e_0 + Y_1 = Y_1, \\ e_2 &= \theta e_1 + Y_2 = Y_2 + \theta Y_1, \\ e_3 &= \theta e_2 + Y_3 = Y_3 + \theta Y_2 + \theta^2 Y_1, \\ &... \\ e_n &= Y_n + \theta Y_{n-1} + \theta^2 Y_{n-2} + \cdots + \theta^{n-1} Y_1. \end{aligned}$$

In this way, we can also construct the objective function

$$S_c(\theta) = \sum_{t=1}^{n} e_t^2 = \left(Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \cdots + \theta^{t-1} Y_1\right)^2.$$

And from the second way, we can see that the S_c is a square of sum **conditional on** $e_0 = 0$. As we promised earlier, this example explains the "conditional" in the name of the method.

After the construction of S_c, the conditional LS method computes $\hat{\theta}_{\text{LS}} = \arg\min_\theta S_c(\theta)$. This is a highly nonlinear function (in fact it's a high degree polynomial) of θ, so we can only solve it numerically (by software).

Example: ARMA(p, q)

Consider an ARMA$(1, 1)$:

$$Y_t - \phi Y_{t-1} = e_t - \theta e_{t-1}.$$

Given observed data $(Y_1, ..., Y_n)$, the idea to construct the objective function is similar to the MA(1) example.

Assume $e_1 = 0$, then

$$\begin{aligned} e_2 &= \theta e_1 + Y_2 - \phi Y_1 = Y_2 - \phi Y_1 \\ e_3 &= \theta e_2 + Y_3 - \phi Y_2 = \cdots \\ &\cdots \\ e_n &= \theta e_{n-1} + Y_n - \phi Y_{n-1} = \cdots \end{aligned}$$

So the e_2 through e_n can be written in terms of $(Y_1, ..., Y_n)$. The objective function is set to be

$$S_c(\phi, \theta) = \sum_{t=2}^{n} e_t^2$$

where e_2 through e_n are defined recursively above. The conditional LS method then minimizes this function $S_c(\phi, \theta)$ (numerically).

Remark: The reason we assume $e_1 = 0$ is that we can make use of all the observed data. If we assume $e_0 = 0$ like the previous example, then we cannot compute $e_1 = \theta e_0 + Y_1 - \phi Y_0$ since Y_0 is not observed.

In general, for ARMA(p, q), we condition on $e_p = e_{p-1} = \cdots = e_{p-q+1} = 0$, then define

$$S_c(\phi_1, ..., \phi_p, \theta_1, ..., \theta_q) = \sum_{t=p+1}^{n} e_t^2.$$

Remark: this covers the previous two examples MA(1) (which can be seen as ARMA$(0, 1)$) and ARMA$(1, 1)$.

Maximum Likelihood Estimation (MLE)

- Pros: Use all the "data/information" (don't assume something is zero like $e_i = 0$). Relevant for small datasets. We have distributional results on estimates
- Cons: No closed form solution. Numerical optimization is hard.

The idea of MLE: define the likelihood function of the parameters as the joint pdf of the observed data,

$$\underbrace{L\left(\text{parameters} \mid Y_1, Y_2, \ldots, Y_n\right)}_{\text{function of parameters}} \overset{\text{def}}{=} \underbrace{f\left(Y_1, Y_2, \ldots, Y_n \mid \text{parameters}\right)}_{\text{joint pdf of } (Y_1, \ldots, Y_n)}.$$

Then maximize the likelihood function L over the paramters.

Main assumption: When we use MLE in time series, we assume

$$e_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2).$$

Remark: This enables us to write out the pdf of $(Y_1, ..., Y_n)$. Since the general white noise setting does not require a specific distribution.

Under this normality assumption, the pdf for a single e_t is

$$f(e_t) = (2\pi\sigma_e^2)^{-\frac{1}{2}} \exp\left(-\frac{e_t^2}{2\sigma_e^2}\right),$$

and the joint pdf for $(e_1, ..., e_n)$ is

$$\prod_{t=1}^{n}\left[(2\pi\sigma_e^2)^{-\frac{1}{2}} \exp\left(-\frac{e_t^2}{2\sigma_e^2}\right)\right] = (2\pi\sigma_e^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_e^2}\sum_{t=1}^{n} e_t^2\right).$$

Example. Consider an AR(1) with mean μ:

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + e_t, \quad e_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2).$$

Suppose we observe $(Y_1, ..., Y_n)$. The likelihood function of the parameters is defined as the joint pdf of $(Y_1, ..., Y_n)$. For simplicity, we omit the dependence on parameters in the joint pdf of $(Y_1, ..., Y_n)$:

$$\mathcal{L}(\phi, \mu, \sigma_e^2 \mid Y_1, Y_2, \dots, Y_n) = f(Y_1, Y_2, \dots, Y_n) = \underbrace{f(Y_2, Y_3, \dots, Y_n \mid Y_1)}_{\text{(ii)}} \underbrace{f(Y_1)}_{\text{(i)}}.$$

For part (i), we need to find the pdf of Y_1. Recall the GLP repesentation for AR(1), we have

$$\begin{aligned} Y_1 - \mu &= e_1 + \psi_1 e_0 + \psi_2 e_{-1} + \psi_3 e_{-2} + \cdots \\ &= e_1 + \phi e_0 + \phi^2 e_{-1} + \phi^3 e_{-2} + \cdots. \end{aligned}$$

Since $e_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$, we have

$$Y_1 \sim \mathcal{N}\left(\mu, \sum_{k=0}^{\infty}\phi^{2k}\sigma_e^2\right) = \mathcal{N}\left(\mu, \frac{\sigma_e^2}{1-\phi^2}\right).$$

So the pdf of Y_1 is

$$f(Y_1) = \left(2\pi\frac{\sigma_e^2}{1-\phi^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{1-\phi^2}{2\sigma_e^2}(Y_1 - \mu)^2\right).$$

For part (ii), we need to find the joint pdf of $(Y_2, ..., Y_n)$ conditional on Y_1. From the AR(1) model, we know that $Y_t \sim \mathcal{N}\left(\mu + \phi(Y_{t-1} - \mu),\ \sigma_e^2\right)$, and Y_t depends on $(Y_{t-1}, Y_{t-2}, ...)$ only through Y_{t-1}. So we have

$$\begin{aligned} &f(Y_2, \cdots, Y_n | Y_1) = \prod_{t=2}^{n} f(Y_t | Y_{t-1}, \cdots, Y_1) = \prod_{t=2}^{n} f(Y_t | Y_{t-1}) = f(Y_2|Y_1)\ f(Y_3|Y_2) \cdots f(Y_n|Y_{n-1}) \\ &= \prod_{t=2}^{n}\left[(2\pi\sigma_e^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_e^2}\left(Y_t - \mu - \phi(Y_{t-1} - \mu)\right)^2\right)\right] \\ &= (2\pi\sigma_e^2)^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2\sigma_e^2}\underbrace{\sum_{t=2}^{n}\left(Y_t - \mu - \phi(Y_{t-1} - \mu)\right)^2}_{S_c(\phi,\mu)}\right). \end{aligned}$$

Note: the square of sum in the last line is same as the conditional LS objective function for AR(1).

Combining part (i) and (ii):

$$\begin{aligned}
\mathcal{L}(\phi, \mu, \sigma_e^2 \mid Y_1, Y_2, \ldots, Y_n) &= \underbrace{f(Y_2, Y_3, \ldots, Y_n \mid Y_1)}_{\text{(ii)}} \; \underbrace{f(Y_1)}_{\text{(i)}} \\
&= \left(2\pi\sigma_e^2\right)^{-\frac{n}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_e^2} S_c(\phi, \mu) - \frac{1}{2\sigma_e^2}(1-\phi^2)(Y_1-\mu)^2\right] \\
&= \left(2\pi\sigma_e^2\right)^{-\frac{n}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_e^2} S(\phi, \mu)\right].
\end{aligned}$$

where $S(\phi, \mu) = S_c(\phi, \mu) + (1-\phi^2)(Y_1-\mu)^2$.