# 25 Spring 439/639 TSA: Lecture 16

Dr Sergey Kushnarev

## Table of contents



## MLE for AR(1)

Last time, for the model

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

we derived that

$$\mathcal{L}(\mu, \phi, \sigma_e^2 \mid Y_1, Y_2, \dots, Y_n) = \left(2\pi\sigma_e^2\right)^{-\frac{n}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_e^2} S(\mu, \phi)\right],$$

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

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

We have noticed that $S_c$ is just the objective function in the conditional LS for AR(1). We call $S_c(\mu, \phi)$ the conditional sum of squares, and $S(\mu, \phi)$ the unconditional sum of squares, i.e.

$$\underbrace{S(\mu, \phi)}_{\text{unconditional sum of squares}} = \underbrace{S_c(\mu, \phi)}_{\text{conditional sum of squares}} + (1-\phi^2)(Y_1 - \mu)^2.$$

In the MLE, maximizing the likelihood function

$$\mathcal{L}(\mu, \phi, \sigma_e^2) = \left(2\pi\sigma_e^2\right)^{-\frac{n}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_e^2} S(\mu, \phi)\right]$$

is equivalent to maximizing the log likelihood function

$$l(\mu, \phi, \sigma_e^2) = \log \mathcal{L}(\mu, \phi, \sigma_e^2) = -\frac{n}{2} \log\left(2\pi\sigma_e^2\right) + \frac{1}{2} \log\left(1-\phi^2\right) - \frac{1}{2\sigma_e^2} S(\mu, \phi).$$

Take the partial derivative of $\sigma_e^2$:

$$\frac{\partial l}{\partial \sigma_e^2} = -\frac{n}{2} \frac{1}{\sigma_e^2} + 0 + \frac{1}{2(\sigma_e^2)^2} S(\mu, \phi).$$

By setting $\frac{\partial l}{\partial \sigma_e^2} = 0$, we know that the MLE estimator must satisfy

$$\hat{\sigma}_e^2 = \frac{1}{n} S(\hat{\mu}_{\text{MLE}}, \hat{\phi}_{\text{MLE}})$$

**Remark:** the MLE estimator for $\sigma_e^2$ is biased. To make it unbiased, we can use

$$\hat{\sigma}_e^2 = \frac{1}{n-2} S(\hat{\mu}_{\text{MLE}}, \hat{\phi}_{\text{MLE}})$$

since we lost 2 degree of freedoms when estimating the 2 parameters $\mu$ and $\phi$. In general, we replace $n$ by $n-p$ to get an unbiased estimator for variance, where $p$ is the number of other parameters to be estimated.

The partial derivative of $\mu$ has a very simple form

$$\frac{\partial l}{\partial \mu} = 0 + 0 - \frac{1}{2\sigma_e^2} \frac{\partial S(\mu, \phi)}{\partial \mu}.$$

**Exercise:** find $\hat{\mu}_{\text{MLE}}$ from $\frac{\partial l}{\partial \mu} = 0$

The partial derivative of $\phi$ is

$$\frac{\partial l}{\partial \phi} = \frac{\partial}{\partial \phi} \left( \frac{1}{2} \log(1-\phi^2) - \frac{1}{2\sigma_e^2} S(\mu, \phi) \right).$$

In general, $\frac{\partial l}{\partial \phi} = 0$ needs to be solved numerically.

## Unconditional sum of squares

In general, MLE is hard to compute. Recall that MLE, i.e. maximizing $\mathcal{L}(\mu, \phi, \sigma_e^2)$, is equivalent to maximizing

$$l(\mu, \phi, \sigma_e^2) = -\frac{n}{2} \log(2\pi\sigma_e^2) + \frac{1}{2} \log(1-\phi^2) - \frac{1}{2\sigma_e^2} S(\mu, \phi).$$

If we only think about $(\mu, \phi)$, and (make a bold move to) ignore the first two terms above, then the optimal $(\mu, \phi)$ are just the solutions to $\min_{\mu,\phi} S(\mu, \phi)$. We call this alternive method **unconditional sum of squares (UCSS)**:

$$\min S(\mu, \phi).$$

Of course, the solution of UCSS is different from the solutions of MLE in general, i.e.

$$\arg\max l(\mu, \phi) \neq \arg\min S(\mu, \phi).$$

But for large sample sizes, they are close

$$\arg\max l(\mu, \phi) \approx \arg\min S(\mu, \phi).$$

And the optimization problem of UCSS $\min S(\mu, \phi)$ is easier to solve (numerically) than the MLE $\max l(\mu, \phi, \sigma_e^2)$.

Also recall that

$$\mathcal{L}(\mu, \phi, \sigma_e^2) = (2\pi\sigma_e^2)^{-\frac{n}{2}} (1-\phi^2)^{\frac{1}{2}} \exp\left[ -\frac{1}{2\sigma_e^2} S(\mu, \phi) \right],$$
$$S(\mu, \phi) = S_c(\mu, \phi) + (1-\phi^2)(Y_1 - \mu)^2,$$

the UCSS method can be seen as a compromise between MLE and the CSS(conditional sum of squares, i.e., conditional LS) method.

## Comparison and comments of the methods

Let's compare the three methods

- MLE, maximizing the likelihood function $\mathcal{L}$, or equivalently the log likelihood function $l$.
- UCSS, minimizing the unconditional sum of squares $S$.
- CSS, minimizing the conditional sum of squares $S_c$ (i.e. the conditional LS method in earlier lectures).

(*Note:* MLE contains $\sigma_e^2$ in the objective function, while the other two do not. We only compare the behavior of the estimated parameters other than $\sigma_e^2$, like $\mu, \phi_i, \theta_i$.)

For large samples, the three methods MLE, UCSS, CSS have very similar results.

In general, UCSS is different from MLE,

$$\text{argmin}\, S(\mu, \phi_i, \theta_j) \neq \text{argmax}\ \ l(\mu, \phi_i, \theta_j)$$

especially if ARMA models are close to non-stationary.

If sample size is small or medium, MLE is preferred. (See the pros and cons of MLE in last lecture.)

But MLE is hard to compute, only viable through numerical methods (no closed form solutions in general). *Note:* As part of this, MoM estimates (which is much easier to get) are often used as the initial guesses in the numerical computing of MLE.

- MLE is *conceptually* better than the other since it uses "all information" and do not assume something is zero (like $e_i = 0$)
  - MOM only uses the first $k$ moments. (And it may not exist as we have seen.)
  - CSS assumes something, like $e_1 = \cdots = e_m = 0$ (so it throws away some information from the likelihood function.)
  - UCSS is a compromise between MLE and the CSS.
- MLE is hard to compute even numerically.

## Asymptotic theory

In this part, we think about the asymptotic, i.e., large sample properties for MLE/UCSS/CSS (since they are very similar when sample size is large). In parameter estimation, the estimated $\hat{\phi}_i, \hat{\theta}_j$ are all random variables that depend on the random sample.

Via asymptotic MLE theory, estimators are unbiased, asymptotically normal, and have some certain variance (see the examples below).

For AR(1),

$$\text{Var}(\hat{\phi}) \approx \frac{1-\phi^2}{n},$$

where the $\phi$ is the true parameter of the AR(1).

For AR(2),

$$\text{Var}(\hat{\phi}_1) \approx \text{Var}(\hat{\phi}_2) \approx \frac{1-\phi_2^2}{n},$$
$$\text{and} \quad \text{corr}(\hat{\phi}_1, \hat{\phi}_2) \approx -\frac{\phi_1}{1-\phi_2}.$$

For MA(1) and MA(2), the results have the same form, just replace the $\phi_1$ to $\theta_i$. (See the textbook.) For example, for MA(1), we have

$$\hat{\theta} \sim \mathcal{N}\left(\theta,\ \frac{1-\theta^2}{n}\right).$$

For ARMA$(1,1)$,

$$\mathrm{Var}(\hat{\phi}) \approx \frac{1-\phi^2}{n}\left(\frac{1-\phi\theta}{\phi-\theta}\right)^2, \quad \mathrm{Var}(\hat{\theta}) \approx \frac{1-\theta^2}{n}\left(\frac{1-\phi\theta}{\phi-\theta}\right)^2,$$
$$\mathrm{corr}(\hat{\phi}, \hat{\theta}) \approx \frac{\sqrt{(1-\phi^2)(1-\theta^2)}}{1-\phi\theta}.$$

*Note:* if $\phi \approx \theta$ in the ARMA$(1,1)$, then both the variance $\mathrm{Var}(\hat{\phi})$ and $\mathrm{Var}(\hat{\theta})$ are increasing to infinity. This is because the parameters in ARMA$(1,1)$ are redundant if $\phi = \theta$ (so the model itself is not correctly specified):

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

# Overfitting as a tool

We can use overfitting to check correctness of the fit:

For example, suppose the time series is actually an AR(1) $Y_t - \phi Y_{t-1} = e_t$. We fit two models, AR(1) and AR(2).

For the AR(1) fitting, we have an estimated parameter $\hat{\phi}$, and

$$\mathrm{Var}(\hat{\phi}) \approx \frac{1-\phi^2}{n}.$$

For the AR(2) fitting, we get two estimated parameters $\hat{\phi}_1, \hat{\phi}_2$. Note that the true model can be seen as an AR(2), $Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} = e_t$ where $\phi_2 = 0$. By the asymptotic variance property,

$$\mathrm{Var}(\hat{\phi}_1) \approx \frac{1-\phi_2^2}{n} = \frac{1}{n} > \frac{1-\phi^2}{n}.$$

So the variance of $\hat{\phi}_1$ in the second fitting is larger than the variance of $\hat{\phi}$ first fitting. We should go back to the AR(1) fitting.

In general, if corresponding parameter(s) have larger variance/standard error after fitting a larger model than a smaller model, then this suggests overfitting and we should go back to the smaller model.

For a concrete example, suppose we have $n = 100$ samples from an AR(1) model with true parameter $\phi = 0.7$. If we fit an AR(1), we know $\hat{\phi} \sim \mathcal{N}\left(\phi, \frac{1-\phi^2}{n}\right)$. Then the 95% CI for $\hat{\phi}$ is

$$\left[\hat{\phi} \pm 2\sqrt{\frac{1-\phi^2}{n}}\right] = \left[\hat{\phi} \pm 2\sqrt{\frac{1-0.49}{100}}\right] \approx \left[\hat{\phi} \pm 0.14\right].$$

If we fit an AR(2), then $\hat{\phi}_1 \sim \mathcal{N}\left(\phi_1, \frac{1-0^2}{n}\right)$. tTe 95% CI for $\hat{\phi}_1$ becomes

$$\left[\hat{\phi}_1 \pm 2 \times \sqrt{\frac{1}{n}}\right] = \left[\hat{\phi}_1 \pm 0.2\right].$$

So the standard error increased.