Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 04

Lecture 17: Model Estimation

Thank you. Hello all, welcome to this new session in this course on time series modeling and forecasting using R. Now, again, just to quickly revise or tell you where we stopped in the last lecture. So, again, the focus of the last lecture was revolving around the idea of model identification. We will extend the same idea in a few more slides, and then we will talk about a slightly newer idea, okay? Now, again, just to give you a brief idea about what exactly you mean by model identification is that, let's say if you're working with practical data,

and if you want to fit some time series model on that practical data, then how do you identify the particular model that could be fitted on the data? And once you identify the model, how do you specify the underlying orders of the model? So, let us say if you are fitting an ARMA model, ARIMA model, or SARIMA model, right? And so on and so forth. So, how do you fix the underlying orders of these models?

Precisely, how do you fix P and Q? In the last lecture, we saw multiple ideas. The first idea was based on some visual plots. So, ACF plots, PACF plots, and then the underlying tendencies in both plots. And then, later on, in the latter half of the lecture, we talked about information criteria, if you remember.

So let us say Akaike's information criteria, which is AIC, or Schwartz-Bayesian information criteria, or SBC, or Hannon and Quinn information criteria, or HQIC. So, let us say if you have a pool of models that could be fitted on the data based on, let us say, some visual checks or some other technique, then how do you proceed with identifying the best model? Then this is exactly where the information criteria come into the picture. And again, just to summarize the idea of all the information criteria, so be it AIC, SBC, or HQIC, the idea is that one should always tend to minimize the information criteria as much as possible. So, let us say if you have ARMA 2.2 and then ARMA 2.1, and then you want to compare which model could be slightly better, then you find out the AIC for both the fits, and whichever AIC is the least or whichever AIC is minimum, you go with that model.

And I think in the last lecture, we talked about one idea which was based on the sample inverse autocorrelation function, or SIACF. And then we talked briefly about a couple of advantages of such a function, which is SIACF. And one of the important advantages was that it kind of prevents or it kind of detects over-differencing. Okay. So, again, in today's lecture, we will kind of extend that, and then we will propose one more idea which is based on this extended sample autocorrelation function, or ESACF, right.

So, last class we talked about SIACF, and then this one is ESACF, okay. So, the advantage of applying such a function is that the extended sample autocorrelation function, or ESACF, can tentatively identify the orders of a stationary or non-stationary ARMA process based on a slightly different idea, which is iterated least squares estimates of the autoregressive parameters. Now, probably I will pause here just for a second, and then again, if you go back to the last lecture, the main idea of how we estimated the parameters was based on MAE entirely, right, or maximum likelihood. So, be it AIC, SBC, or HQIC. So, in all these ideas, the estimation technique was kind of similar, which was maximum likelihood estimation, right, but can you somehow find the estimators by changing the technique?

This is exactly what ESACF considers: rather than estimating using maximum likelihood, can we estimate the underlying parameters based on iterated least squares estimate? So, I think this is the only difference between either AIC or SIACF or SBC and all those techniques, and whether you should go with something like ESACF. Now, probably what we will do is I will show you a table, which is a practical example, and then we will identify which model is the best one. So, here we have something called the minimum information criteria or MINIC. So, M I N I C, right?

So, minimum information criteria. So, this minimum information criteria uses a combination of AIC and SBC, right? So, if you want to evaluate which model is the best one, you use a combination of AIC and SBC in some sense to minimize the overall information in the underlying model structure. Now, what do you mean by that? So, let us say you have an AR structure.

So, all these are the orders. So, if you see the first row in this table. So, the model orders for an AR model are given. So, let us say 0, 1, 2, 3, and so on. And if you look at the first row, this is MA, and just below that, you can see the MA orders, right.

So, the first column gives you AR orders, and the first row gives you the MA orders. So, 0, 1, 2, 3, and it can kind of extend, right? Now, what is the idea about this table here? So, for each pair of orders, let us say AR0MA0, AR1MA1, AR2MA2, and so on. So, for each pair, I am actually finding the SBC value for that.

So, SBC is the Schwartz-Bayesian criteria. So, I am finding an information criterion. Alright. So, whatever value I have based on the practical data, I will input that in this cell. So, SBC 0 0 because the order is 0 0, then 0 1, 0 2, 0 3, etc.

So, this table will kind of give you a better idea really quickly as to which cell gives you the minimum kind of value. And then you kind of fix those orders accordingly. So, we'll see an example here. So, let's say this table is produced using software, of course. So, let's say if you're analyzing a practical dataset and you get this MINIC table, right?

So, again, like I said, the first row gives you all the orders for possible MA structures, and then the first column gives you all the orders for possible AR structures, right. So, 0, 1, 2, 3, 4, 5, and then similarly 0, 1, 2, 3, 4, 5 for MA as well, okay. And the values inside this table kind of give you that MINIC value, okay. And here, the idea is that you want to identify the minimum possible value, alright. So, you want to identify the minimum possible value.

And then here, you do it using a slightly different kind of technique. So, if you observe all these values, right. So, this software kind of identified this value that you see in red to be the minimum in terms of both AIC and SBC collectively, all right. So, this value is the minimum based on both AIC and SBC collectively, okay. And again, you can kind of see that this value is, in fact, the minimum, right, because this is minus 0.03571, right.

So, all the other values are higher than that. Obviously, many of the values are positive, and then you have a handful of values which are negative. For example, this one, that one, and then this one. But then, even out of those negative values, this is the most negative. But then, you should remember one thing: these values are not just based on SBC alone, but they use a combination of AIC and SBC.

So, here you can see the comment in red that the minimum table value based on BIC 1 0 is nothing but minus 0.03571, which corresponds to nothing but AR 1 and then MA 0 as the orders, right. So, essentially, the model using this MINIC criteria comes out to be ARMA 1, 0, right. Now, again, ARMA 1, 0 is nothing but an AR model of order 1 because here you do not have any MA component. Because the order is 0. Make sense?

So, one can actually implement this MINIC criteria in any software. Let us say R or any other software. So, let us say SAS, MINITAB, or Python. Right? To find out which value gives you the least kind of MINIC value.

And then, the MINIC value is nothing but a combination of AIC and SBC together. Okay? So, this is just an idea about how you identify models based on some information criteria or, let's say, some visual plots, etc. Alright, so I think now we move a step ahead and then we enter this plethora of different estimation techniques. So, and then what we call model estimation criteria or model estimation techniques.

So, how do we estimate particular models? Okay. So, again, after specifying the model orders. Right. So, how do you do that?

Using some information criteria and so on. Right. So, once you specify the model orders, we need to estimate the underlying parameters. Isn't it? So, let us say, by using some information criteria, you obtain the best orders to be, let us say, ARMA 1, 2.

So, once you know that the ARMA 1, 2 model is the best in that scenario, then, if you remember that equation of ARMA 1, 2, you have certain parameters in that underlying model. Now, the idea is, how do you estimate the underlying parameters? So, here we can actually assume that the order is known, of course, and then the model has a 0 mean. So, we kind of assume that there is no intercept. So, the overall model has a 0 mean.

And this idea could be easily tweaked, by the way. So, even if the mean is not 0, we can do a very simple manipulation or a simple transformation to ensure that the transformed mean is 0. So, what do you mean by that? So, let us say if the mean is not 0. We can always subtract the sample mean, so y bar is the sample mean, and then fit a zero-mean ARMA process, so let us say xt, and then somehow get back to yt by applying this transformation.

So, again, just to tell you, y bar is what? So, y bar is the sample mean. xt is what? So, xt is a zero-mean ARMA process. So, xt follows an ARMA process with zero mean, so there is no mean. Or there is no intercept there. So, using this combination, you can actually get back to the original time series. So, even if the mean is not 0, you can actually replace that by something like xt plus y bar to get hold of the actual mean.

All right. Now, how about some estimation techniques? So, or in other words, what are all the ways of estimating the underlying parameters in a model? So, the first one is the method of moments. So, in short, MOM or MOME.

Now, the second one we discussed in the last lecture, MAE. So, maximum likelihood estimation or one can actually go with something called OLS or ordinary least squares or let us say least squares estimation, either conditional or unconditional. So, in short, you have several different techniques of estimating the parameters. So, method of moments, maximum likelihood, OLS, least squares, etc. By the way, OLS is exactly similar to what we have in regression, isn't it?

So, even in regression, we have the OLS technique. So, we estimate the underlying parameters in, let us say, simple regression or multiple regression using an OLS technique, kind of okay. Alright, so now we will probably elaborate a bit more on each of the estimation techniques and then probably see how exactly you can deploy those, okay. So, now the first one is called the method of moments technique. And by the way, this method of moments technique has a slightly different name also.

So, it is called Yule-Walker estimation. So, it is called Yule-Walker estimation also. But then you have a certain drawback here. So, such a kind of technique, which is the method of moments or Yule-Walker, works best for only AR models. And then not, let us say, ARMA models or ARIMA models and so on.

And then for large N. So, in a way, this is not an efficient method, right? Because you have a lot of restrictions here. So, it works only for AR models, and then n should be really large, right? So, in a way, or in short, this is not a very efficient kind of method, okay? But what exactly is the basic idea underlying this method of moments technique?

So, the basic idea is that you simply create some equations, right? And what exactly are the equations? So, you basically equate the population moments to the corresponding sample moments and then keep on solving for the parameters. Now, before we go into any of the details, I will give you one very easy example. So, let us say you have a random variable x. So, I will not give you an example from time series because we have one example in the next slide. But let us say if x is a random variable which follows, let us say, some distribution normal with mean mu and variance sigma square.

Okay. And then let us say the idea is you want to estimate both these parameters because both these parameters are unknown. Right. So, the mean is unknown. The variance is unknown.

So, how do you estimate that using the method of moments technique? So, for that, what we will do is, or rather what I will assume is that you have a collection of random variables. So, let us say x1, x2 up to, let us say, xn. So, you have a sample basically, and then not just one random variable. So, you have a sample x1, x2 up to xn, and then each of these random variables are iid.

So, independent and identical, and then the distribution is normal. So, this is the problem. Now, the question is, how do you estimate the underlying parameters which are mu and sigma square? So, here, what we will do is, again, if you go back to this basic idea that we discussed a short while back. So, you have to create those equations now.

And what exactly are the equations? So, the first equation would be you look at the first parameter which is mu. This is the population version of the mean. So, you equate this to the sample version. Now, the sample version is what?

Sample version is nothing but the sample mean, which is x bar, clearly. So, this would be my first equation. And then the second equation would be to equate this sigma square to the sample variance. Sample variance is written as 1 by n minus 1 summation xi minus x bar whole square. Isn't it?

So, the moment you create these two equations, the idea is that you should solve for mu and sigma square. Based on, of course, x. Because x is the data that you have. So, the idea is clear, hopefully, that depending on the number of parameters you would want to estimate. So, here in this case, you are estimating 2. So, mu and sigma square.

So, the idea is that you create those many equations. So, the first equation is mu equals x bar, a very simple equation, and then the second equation is sigma square equals the sample variance, and then you keep on equating and solving for the parameters. So, this is broadly the idea underlying the method of moments technique. So, we will take up an example. So, let us say you have a time series yt.

So, you have a time series yt. So, the expected value of yt would be nothing but 1 by n summation of yt. Now, the expected value of yt is a population version. So, this is nothing but the population mean. And what you have here is nothing but the sample.

So, 1 by n summation yt, t goes from 1 to n, is the sample. So essentially, what you are doing is you are again doing the same thing that we discussed in the last slide. So, you are equating the population mean and the sample mean, which is y bar. So, this would be my first equation. And of course, if you have more parameters, then you have to create those many equations.

$$E(Y_t) = \frac{1}{n}\sum_{t=1}^{n} Y_t \Rightarrow \mu = \bar{Y}$$

$$E(Y_t Y_{t+k}) = \frac{1}{n}\sum_{t=1}^{n} Y_t Y_{t+k} \Rightarrow \gamma_k = \hat{\gamma}_k$$

Similarly, $\rho_k = \hat{\rho}_k$

So, this could be a generalized kind of equation. So, let us say the expectation of yt into something like yt plus k. Again, this is the population version. And what does this give you? By the way, this gives you the gamma k or the covariance, is it not? Because the expectation of yt into yt plus k is nothing but the covariance.

And again, assuming that the process has a zero mean, by the way. So, if the process has a zero mean, then you can actually create this equation. And then the sample version of the covariance is nothing but this, right? So, 1 by n summation y t into y t plus 3. So, essentially you are equating the population auto-covariance and the sample auto-covariance.

And similarly, one can actually create equations based on correlations also, right? So, something like rho k equals rho k hat. So, here wherever you see hats. So, hats represent estimators, right? Every time. So, gamma k hat or rho k hat, so both these are the estimated quantities or these are the estimators of the underlying parameters, basically.

So, in this sense, we can create those many equations, keep on solving them, and then solve for the underlying parameters. So, we will take one particular example of an AR model of order 1. So, this is my AR 1 model structure, right? So, this is AR 1. So, y t equals phi y t minus 1 plus e t. Now, again, if you go long back to one of our earlier lectures where we kind of elaborated more about AR 1.

So, from that lecture, you can actually borrow a couple of facts. So, if you talk about correlation. So, rho 1. So, correlation is nothing but phi, which is this coefficient here. So, correlation happens to be exactly equal to phi.

$$Y_t = \phi Y_{t-1} + e_t$$

It is known that $\rho_1 = \phi$

Thus, $\hat{\phi} = \hat{\rho}_1 = \frac{\sum_{t=1}^{n}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$

So, using this idea, we can quickly create an equation. So, phi hat would be nothing but rho 1 hat because phi 1 equals, or rather rho 1 equals phi. So, the estimated versions can be replaced by the actual versions here. So, phi hat would be nothing but equal to rho 1 hat, and the formula that you see here is nothing but a sample version of the correlation. Because this is summation y t minus y bar into y t minus 1 minus y bar divided by the variance, the underlying variance.

So, this is my sample correlation, and then you basically equate the sample correlation to something like phi hat in this case and then solve. So, in such a technique, you can actually equate a set of equations and then keep on solving for the underlying parameters. Now, on the other hand, we actually know a few other things also, right? So, let us say the error variance. So, by the way, this is one more parameter in any model.

So be it, let us say AR1 or AR2 or any other model; you also have this variance of the errors. So, this is another parameter, right? So, again, if you come back to the AR1 structure. So, if you see this AR1 model structure, you have this error term here, is it not? So, ET. So, there must be some variance of that ET, which is given by sigma square E, and sigma square E is another parameter for us.

So, in fact, one needs to estimate both parameters. So, phi as well as sigma square E. So, if you want to estimate two parameters, how many equations would you require? You would require two equations. So, this is one of them, and then the other equation comes from this fact, by the way. So, but we know for a particular AR1 model, gamma 0.

$$\gamma_0 = \frac{\sigma_e^2}{(1 - \phi^2)}$$

Thus, $\hat{\sigma}_e^2 = (1 - \hat{\phi}^2)\hat{\gamma}_0 = (1 - \hat{\phi}^2)\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2$

$$= (1 - \hat{\rho}_1{}^2)\frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2$$

So, gamma 0 is a variance. So, gamma 0 is nothing but the covariance of yt and yt, which is nothing but the variance of yt. So, from earlier lectures, we know that the variance of an AR1 model is nothing but sigma square E divided by 1 minus phi square. So, here we can replace the estimated version. So, sigma square E hat would be nothing but 1 minus phi square hat into gamma 0 hat.

So, if you cross multiply, if you bring this to the right-hand side, then this is the equation that you have. And now I can essentially replace my gamma 0 hat because gamma 0 hat is the estimated variance of the AR1 model. So, I can replace that by the sample variance, which is 1 by n summation y t minus y bar square. And then, basically, now I can replace my phi hat by rho 1 hat, which we saw from the earlier equation. Because phi is nothing but rho 1, so I can actually replace phi hat square by rho 1 hat square.

So, the value for this estimator here, in terms of variance or the error variance, is nothing but this 1 minus rho 1 hat square into the sample variance, basically, all right. So, here we talked about the method of moments or, in particular, Yule-Walker, right. Now, the next idea is MLE, or the maximum likelihood methodology, right. So, here we require some assumptions that, let us say, the error vector or all the random errors follow some normal distribution again. So, let us say the mean is 0 and the variance is sigma square E, and on top of that, all these errors are also independent and identical, right.

So, these are some assumptions that we require. Now, the idea is that, again, like we discussed in the last lecture, what exactly do you mean by likelihood. So, likelihood is nothing but some sort of a joint density or joint distribution, okay. Now, the whole question is, can we actually use the joint distribution applied on the actual time series values. So, something like f of y 1, y 2, up to y n.

So, by the way, this joint density cannot be broken down into the product of marginals, and why? So, why do you think? So, again, the idea is that if you use the joint density of the actual time series values, right, or the observations. So, y1, y2, y3 up to yn, would this be equal to, let us say, the product of the marginals? So, f of y1 into f of y2 into dot dot dot f of yn, would this be true?

The answer is no, because these are not independent. So, since yi's are not independent, this equation does not hold true. So, this is not equal. So, then, can we somehow simplify the problem? The answer is yes. So, instead of looking at the actual time series values, we can look at creating a likelihood function or a joint density based on the errors, because errors are independent.

So, rather than using the actual y values, we can actually use the e values. So, e1, e2, en, and here, clearly, we can write down this equation because all the errors are independent, as per the assumption. So, this joint density, which is nothing but the likelihood, can now be broken down into the product of the marginals. Now, again, you may ask, why do we require this? I mean, why do we require any joint density to be broken down into the product of marginals? Just for simplicity.

Just for simplicity. So, we need not make the problem too complicated, right? So, if you are able to kind of break down a joint density into the product of the marginal density, then why not, right? I mean, we have an advantage here, okay? So, this is the idea.

And now, so, we write down the conditional log likelihood. So, why conditional? We will talk about that, but then, as discussed even in the last lecture, we somehow get hold of a log likelihood function. And, by the way, all these are the underlying estimators or the parameters that need to be estimated, right. So, mu, phi, theta, and then sigma square E, okay. And then, probably, as seen from the last lecture, can you recollect this formula here?

$$\ln L(\mu, \phi, \theta, \sigma_e^2) = -\frac{n}{2} \ln\left(2\pi\, \sigma_e^2\right) - \frac{S_*(\mu,\phi,\theta)}{2\sigma_e^2},$$

Where $S_*(\mu, \phi, \theta)$ is the conditional sum of squares, based on some starting values

So, the log likelihood takes this form. So, minus n by 2 ln of 2 pi sigma square E minus a certain sum of squared residuals divided by some constant. But here, I am putting a star just to denote that this S star is some sort of a conditional sum of squares and not exactly the sum of squares. And why conditional? Because when you are finding MLEs using some numerical approximation kind of technique, you have to kind of fix some starting values of the parameters.

So, I am not sure how many of you have taken a numerical analysis course where, let us say, you have all these Newton-Raphson techniques or all these convergence techniques, right. And if you are applying any convergence techniques or numerical approximation techniques, you have to consider some starting values of the parameters, right. So, once you feed in those starting values to the algorithm, then the algorithm runs and it gives you all the estimated values once the algorithm kind of converges, right. So, this is the idea of Newton-Raphson or, for that matter, any other convergence algorithm. So, a similar kind of technique comes handy here.

So, rather than taking the actual sum of squares, we are taking a conditional sum of squares. So, conditional why? Because these sums of squares are based on some starting values. Alright, so once you feed in that sum of squares, the conditional sum of squares, in this equation, now I can actually get hold of a complete log likelihood function. So, once you have a complete log likelihood function, then the next step is to maximize it, as always, right.

So, now at the end of the day, we can find out the estimators of all the parameters upon maximizing my conditional log likelihood function. And generally speaking, like I said, one cannot solve this using pen and paper because the structure of the underlying MLE or the likelihood function is too complicated. So, the structure is too complicated in a way. So, usually, some numerical approximation techniques or numerical maximization techniques are needed to solve this. And here I was referring back to, let us say, Newton-Raphson or probably some other convergence techniques or some other numerical maximization techniques.

And so on and so forth, right. So, rather than solving or rather than maximizing all these conditional log-likelihood functions or likelihood functions in general using pen and paper or even using some basic codes in a software, one has to apply some numerical maximization techniques such as Newton-Raphson or something similar, right. And the moment you apply some numerical maximization technique; you require to feed the algorithm with some starting values. Okay. And hence, if you go back a slide, then we have this conditional sum of squares because these are conditioned upon what starting values you feed the algorithm with.

Hopefully, this idea is clear. So, how do you write down the log-likelihood? Then how do you maximize that? And then a few advantages of MLE. So, once you find out the MLE, this estimator is asymptotically unbiased, efficient, sufficient, and consistent for large sample sizes.

Right. But then again, you have a problem also that it slightly becomes difficult to deal with joint PDF because you have to deal with joint PDF here. Because likelihood is nothing but a joint PDF. So, if you have a very, very complicated kind of a model structure, then how would you deal with a joint PDF combining all the observations and so on and so forth. So, this could be a challenge here.

But nevertheless, the MLE does have some exciting aspects to it, as it is basically an asymptotically unbiased estimator, efficient, sufficient, and consistent. Alright, so now, on the other hand, we have one more idea, which is called the conditional least squares technique. So, what do you mean by that? So, the conditional least squares technique. So, here we will take up an example using an AR1 kind of model structure.

$$\text{AR(1): } Y_t = \phi_1 Y_{t-1} + e_t. \quad \text{Where } e_t \sim N(0, \sigma_e^2)$$

$$\Rightarrow e_t = Y_t - \phi_1 Y_{t-1}$$

$$\text{SSE} = \sum_{t=1}^{n} e_t^2 = \sum (Y_t - \phi_1 Y_{t-1})^2 = S_*(\phi) \text{ for observed } Y_1, \dots, Y_n$$

$$\frac{dS_*(\phi)}{d\phi} = -2 \sum_{i=1}^{n} Y_{t-1}(Y_t - \phi Y_{t-1}) = 0$$

$$\Rightarrow \hat{\phi} = \frac{\sum Y_t Y_{t-1}}{\sum Y_{t-1}^2}$$

So, let us say this is the AR1 equation. So, y t equals phi 1 y t minus 1 plus e t, and again, the famous assumption is still there. So, the errors have a normal distribution with mean 0 and variance sigma square e. So, from here, I can get hold of E t, right. So, can you solve for E t from this equation?

Yes, we can. So, E t is nothing but y t minus phi 1 y t minus 1, right. Now, you get hold of the sum of squares of the errors. So, summation E t square. So, summation E t square would be nothing but summation of this structure, square of that.

So, y t minus phi 1 y t minus 1 whole square, and then this is nothing but my s star, right? This is nothing but my s star. Now, what do you do? So, you kind of differentiate this, right?

So, you kind of differentiate this, right? So, the derivative of this s star with respect to phi would be nothing but this, right? And then you equate it to 0, basically. And then you solve for phi. So, eventually the MLE or the maximum likelihood estimator in this case for an AR1 model would be nothing but phi hat which is summation y t y t minus 1 divided by summation y t minus 1 squared.

So, essentially this slide contains an example of how do you apply MLE for a basic AR1 model. And then this is based on a conditional least square, right? Because you are finding out the sum of squared errors and then you are kind of taking the sum of squared errors and then kind of using that to find out the maximum likelihood by taking the derivative and then equating to 0 and eventually finding out the estimator of the underlying parameter. So, by the way, here, why are we taking a derivative? Because you want to maximize the likelihood, right?

So, if you want to maximize any function, you would want to first take the derivative of that and then equate it to 0 and then solve for the parameter. Isn't it? So, if you want to maximize or minimize any function, you differentiate it, set it equal to 0, and then solve, right? So, this is exactly what is going on here. So, you find out SSE, right?

And then you differentiate that, set it equal to 0, and then solve for the parameters. So, hopefully, the idea about the method of moments and MLE is clear, right? So, down the line, we will talk more about, and try to extend on, let us say, a few more estimation techniques when it comes to the practical session we will have towards the end of this week. And then there will be one more idea which we will cover probably in the next lecture, and the lecture after that is all about diagnostic checking. So, once you identify the orders, once you perform the estimation, then what?

So, how would you go ahead and then do some diagnostic checking is what we will see next. Thank you.