Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 03

Lecture 16: Model Identification

Hello all, and welcome to this course on time series modeling and forecasting using R. So now we will start with a new week and a few fresh topics to begin with. So, again, just to give you a brief overview of what we'll see this week, we'll focus more on something called model identification, right? So how do you identify the correct orders of a model, or even before that, how do you identify the correct model that could be applied to some real data, and so on and so forth, right? And then we'll talk about a few other aspects of, let's say, model identification using some information criteria, right? Okay?

So let's start. So just to give you a short overview of where we are at this stage is that so far we've looked at several different time series models, right? So if you remember, we covered something called a random walk to begin with, or let's say a very simple moving average, right? Or we talked about white noise. Or slightly more advanced processes such as autoregressive, moving average, then the combination of AR and MA, which is ARMA.

So, autoregressive moving average model. And then if the model has a trend, or if the real data has a trend, then one actually needs to model that using ARIMA. So, ARIMA means autoregressive integrated moving average, and then of course, the last extension that we studied last week was seasonal ARIMA. So, alongside the trend, if you also have seasonality, or let's say some repetitive nature in the time series, then probably one cannot fit a proper ARIMA model for that. So, for that, you require a slightly more advanced technique, which is called SARIMA.

Now the whole question, and again the important question is, how do you find the best suitable model given any practical scenario or given any practical data? So essentially, what this question is asking is, how do you find the optimal orders, let us say P and Q, of a particular model? So, again, let us say, why are we saying P and Q? Because again, let us say if you have an ARMA process. So, all of you know that the ARMA process kind of depends on two orders, right? So, P and Q, and then of course, ARIMA depends on three orders: P, D, Q, but then again, that middle order D is kind of coming from the number of times you difference the series, right?

So, essentially speaking, these two are the important ones, right? So, P and Q. So, how do you identify the optimal number of P and Q or the optimal lags up to which one should actually extend the AR part and the MA part, OK? So, we will talk about a few techniques in this regard and so on and so forth, OK? So, now before we start with anything new, I will again go back to one of my earlier lectures and then I will specify. And then I will elaborate more on this particular technique, which is using ACF and PACF plots.

So, I will say this is more of a visual kind of check that, based on both these plots, let us say ACF and PACF, how do you come up with a rough idea about what orders are true for the underlying AR structure or the MA structure. Now again, what exactly are the full forms of these two? So ACF stands for autocorrelation function, and then PACF stands for partial autocorrelation function. So again, just to quickly summarize, PACF is nothing but partial, or this is kind of similar to a conditional kind of correlation. So, we talked about the expression.

In one of the earlier lectures, but how exactly do you condition that? So, let us say this is something like the correlation between, for example, let us say y t and y t plus 2, and then let us say given the intermediate term which is y t plus 1. So, something like that, right? So, such correlations are called partial autocorrelations or conditional correlations, right? And if you vaguely remember, I gave you a description of ACF plots and PACF plots in one of my earlier lectures also.

So, as seen earlier, one can actually find the optimal orders, which are nothing but P and Q of a particular ARMA structure, using both these plots. So, specifically using the ACF plot and the PACF plot. So again, probably in the next slide, we'll talk more about or we'll kind of elaborate more using the plots themselves. But then here you see a kind of a revision table here. And the first row tells you which plot you are having.

So, either ACF or PACF. And then the first column gives you the appropriate model. So, the underlying models are or underlying choices are it could be either ARP or MAQ. And then what you see inside this table are the tendencies. So, let me explain a bit here.

So let us say for a particular ACF plot for an AR model. Right, which is nothing but this top left comment here. So again, just to repeat, for an ACF plot for a particular AR model, you actually see a tailing kind of tendency. So, the plot has to tail off after the lag peak. And then, at the same time, if you observe the ACF plot of an MA process, it shows a cutting-off tendency.

So, I would say that cutting off is much more predominant than tailing off. So, cutting off means that the correlations abruptly cut off after some lag or after some point. And tailing off means they kind of gradually decrease. And then one can actually observe an exact opposite kind of tendency when it comes to PACF. So, if you focus on a PACF plot for a particular AR model, it should show a cutting-off tendency after lag P, while if you see a PACF plot for a particular MA model, it should show a tailing-off tendency after lag Q.

So, in a way, if you observe just for a few more minutes on this entire table here. So, whatever happens for an underlying AR model is exactly opposite to what you observe in an underlying MA model. Alright. Now, just to explain this a step further. So, we will take up the actual plots, and then I think I showed this slide in my earlier lecture, as mentioned a short while back.

But then again, probably just to elaborate on how you can actually invoke the ACF plots and the PACF plots for a particular AR model and the MA model. Now again, this is the difference between ACF and PACF plots. So, you can actually mention PACF here. So, ACF and PACF. Because we are analyzing both of them, right, and not just one.

So, the difference between ACF and PACF plots for a particular AR model and a particular MA model. So, in front of you again, we have four plots, all right. Now, the first column gives you the ACF plot and the PACF plot for a particular AR(2) model. So, we are fixing the order to be 2, and similarly, on the right-hand side or the second column, it gives you the ACF plot and the PACF plot for a particular MA model of order 2, right. So, in general, we are fixing the order to be 2 in both cases, for the AR structure as well as the MA structure.

And then, of course, the top row gives you the ACF plots for both models. And then the bottom row gives you the PACF plots for both models. Now again, one can actually observe the tendencies we listed in the prior table that we just saw. So again, just to repeat, now if you focus on the ACF plot for an MA(2) model for a second. So, the ACF plot for an MA(2) model, which is this one right here on the top right.

Then again, as discussed in the earlier table, you can actually see a cutting of tendency after lag 2. So, by the way, you should remember one thing: the x-axis on each of these four plots shows you the lag, right? And then again, this is true for any of these four plots. So, on the x-axis, you are actually plotting the lag, and on the y-axis, what do you have? So, on the y-axis, you have either the autocorrelation or the partial autocorrelation being plotted, right?

All right. Now, let us say that you start at lag 1. So, this is lag 1, this is lag 2, this is lag 3, 4, 5, etc. Okay. So, for each lag, there will be some correlation because you are finding the correlation between, let us say, yt and yt plus 1.

So, the lag is 1. Right. Or, let us say, if you are finding the correlation between yt and yt plus 2, so the lag is 2, right? And then, hence, we can actually get more than 2 lags also. So, let us say, the correlation between yt and some yt plus 8, right?

So, lag is 8, okay? So, one can actually get this plot, I mean, just to explain what is going on in this plot, right? So, for each lag, I can actually plot the corresponding autocorrelation or the partial autocorrelation, okay? Alright, so now again coming back to the tendencies, can you see that your ACF plot is kind of showing a cutting off tendency after what lags? After lag 2, alright.

And then, since it is showing a cutting off tendency after lag 2, the model could be MA2 in that case, alright. Now, one more thing is, how do you sense that something is cutting off? So, in each of these plots, again, you can actually see these hypothetical confidence bands, which are given using horizontal blue lines in each of the plots. So, for example, these lines, right? So, what exactly are they?

So, these lines are nothing but 95% confidence bands for those particular correlations, right? And if any of the correlations are outside those bands, then those particular lags would be significant. Otherwise, if the correlations are inside the bands entirely, then those lags are not significant, right? And hence, if you again observe this plot for a second, can you now see that after lag 2 here, which is shown by the second spike, all the further correlations are really small, which are not significant, and hence we sense that such a plot is showing a cutting off tendency, alright? And then, similarly, on the other hand, if you observe this bottom left plot, which is nothing but PACF for an AR2 model, you can actually see the exact same thing, okay?

Now, how do you sense a tailing-off tendency? So, a tailing-off tendency could be seen using these two plots here. So, the top right one and then the bottom right one. So, the top left one and the bottom right one. So, in both these plots, you can actually sense the tailing-off tendency.

Now, again for a second, if you observe this plot which is on the top left. Now, again, as the x-axis gives you all the lags. So, let us say 1, 2, 3, 4, etc. So, I can actually conclude that such a plot is showing a tailing-off tendency after lag 2. And then the exact same thing could be seen in this plot here, which is on the bottom right.

So, again, just to summarize the entire idea about how you fix particular orders for the ARMA model using ACF and PACF plots is to Draw both the plots and conclude using both the plots. So, one should not jump to a conclusion only based on a single plot, either ACF or PACF. So, always, if you are analyzing any practical data, you should make this habit of drawing both the plots simultaneously, ACF and PACF, and then observe the patterns and behaviors of the correlation which are happening in both the plots. So, if in one of the plots you are showing a cutting-off tendency and in the other plot you are seeing a tailing-off tendency, then accordingly you can choose the model to be either AR or MA, and then after what lag you are seeing the cutting-off tendency, that sort of fixes the order.

Make sense so far? All right. Okay. So, this idea is entirely based on a collection of, let us say, AR plots and MA plots, right? And the corresponding ACF and PACF plots for these two models.

So, particularly the AR model and then the MA model. Okay. So, apart from ACF and PACF plots, you have several other criteria or several other techniques to identify the models. Now, I will tell you one disadvantage of ACF and PACF plots. So, for example, besides ACF and PACF plots, we actually have multiple other tools for model identification.

Now, again, just to summarize, this entire idea about model identification is nothing but identifying the orders. So, identify the orders of the model. Alright. So, if you keep this one statement in mind, what are we discussing, basically, right? I mean, the role of the ACF plot and PACF plot really is to identify the appropriate orders of the underlying model.

So, let us say ARMA or even ARIMA, etc. Alright. But what happens is if you have really large and messy data, right? So, let's say if you bring in a very, very heavy kind of practical data containing a lot of features and a lot of repetitions, right? Then what happens is ACF and PACF plots become complicated and harder to interpret.

So, in such a case, ACF and PACF plots may not show the exact tendency that we saw earlier. So, let us say cutting off or tailing off, etc. And this could be a disadvantage. So, if you bring in really heavy kind of structured data, which would be kind of messy and really large, then probably basing or identifying the orders from ACF and PACF plots becomes difficult. Okay, and here the last point is since many different models can be fit to the same data, right? So, let us say if you bring in some practical data, you plot the ACF and PACF. Now, based on after what lags are you seeing the cutting off and the tailing of tendencies, I can actually have slightly different models also in mind, right?

For example, let us say you are seeing cutting off and tailing of tendencies after lag 2 in both the ACF plots and the PACF plots. So, one possible model in our case could be, let us say, ARMA 2.2, right? But here, people or practitioners can play around slightly. So, what do you mean by play around? So, let us say ARMA 2.2. If they think that this might not be the model, they can actually play around slightly with the order.

So, let us say 1, 2, or they can actually look for fitting something like ARMA 2, 1, right? Or even they can actually think of fitting something like ARMA 1, 1. So, again, just to understand that even though you are basing conclusions using the ACF and PACF plots and then after what lag are they cutting off or tailing off, but then this may not be the 100 percent correct model. So, for that, you have to play around with the order slightly. So, here, playing around only means that you kind of take orders which are in the neighborhood.

of let us say 2, 2, right. So, here I am not saying that one should actually look for something like ARMA 6, 6, right, because 6, 6 is way away from 2, 2, which was seen from the plots, right. So, hopefully, this idea is clear. So, since many different models can be fit to the same data, precisely all these, right? I mean, you have a collection of models which you can actually try fitting on the data. But we should choose the most appropriate one having fewer parameters, firstly, and the information criteria will actually help us to decide this.

So, once you kind of fixate a bunch of models or a handful of models, then you kind of go ahead and then find out the information criteria, which we will discuss soon. And then, using the information criteria, you can actually fix on one particular model. Now, firstly, what do you mean by a model having fewer parameters? So, fewer parameters obviously means that, for example, ARMA 1, 1 would have fewer parameters than ARMA 2, 2, right? Because if you have more orders, then eventually, if you bring the ARMA equation in your mind, right?

So, if you keep on increasing the orders on any side, be it AR or MA, the number of parameters also increases in that case, right? The number of phi coefficients and the number of theta coefficients, they go on increasing. So, in this case, ARMA 1, 1 would have a lesser number of parameters as compared to all the above models, basically, okay? So, one should not jump to conclusions based on, let us say, if you are seeing something like ARMA 6, 6, which is really good or the fit is really good, but you may be overfitting the data, right?

Because ARMA 6, 6 contains a lot of coefficients here. So, my point is if the same job could be done with approximately the same accuracy, then you should always go with the model having a lesser number of parameters. So, now we will talk briefly about what we mean by information criteria and then how you can choose a particular model based on a set of information criteria. Okay, so these are predominantly some of the famous or some of the important information criteria which people use. Now, the first one is called Akaike's information criteria or, in short, we call it AIC.

So, by the way, AIC, the first one, and the second one are kind of more applied or more used as compared to the third one. So, by the way, this was introduced by a person called Akaike back in 1974. Then, similarly, the second one is called Schwartz-Bayesian criteria or, in short, SBC. And again, some people or some textbooks write it down as Bayesian information criteria or BIC. And then this particular criterion was introduced by Schwartz, a person called Schwartz, back in 1978.

Okay. So, Akaike's information criteria came slightly before Schwartz. And then very soon, we will see why. Because Schwartz is a slight extension of AIC. And then the last one was brought by these two people, Hannan and Quinn.

So, Hannan and Quinn criteria, or in short, HQIC. And then these were brought by a couple of people, Hannan and Quinn, in 1979. So, if you see the progression, right, I mean, the first kind of predominant information criteria was brought in 1974 by Akaike. Then Schwartz kind of brought an extension to that in 1978. And then Hannan and Quinn kind of again extended Schwartz's criteria and then brought other criteria in 1979.

Now, essentially, what do you actually mean by information criteria? So, all three are information criteria, but how can you apply them or how can you use them? Okay, so we will focus on the first one, which is Akaike's information criteria, or in short, AIC. So, assume that you have a statistical model with M parameters, and you are trying to fit some practical data using that model. So, again, how many parameters do you have?

$$AIC = -2\log\,[\text{maximum likelihood}] + 2M$$

You have capital M parameters in that model. So, I can actually write down a short formula for finding the Akaike's information criteria or AIC. So, in this case, AIC is nothing but minus 2 into the log of the maximum likelihood plus 2 into capital M. So, in short, this is nothing but minus 2 log of my maximum likelihood function plus 2 into the number of parameters which are underlying in the model. Now, again, here I am assuming that you should have some knowledge about maximum likelihood or MLE technique, right?

The MLE technique is nothing but an estimation technique, right? So, again, if you are not very confident about what exactly you mean by the MLE technique, you might want to revise slightly, right? So, here essentially what we are doing is likelihood is kind of similar to a joint density. So, let us see if you have a lot of different random variables. So, if you are working with practical data, then how do you put forward a joint density combining all the observations?

So, that joint density is nothing but the likelihood, and again the idea is always to maximize the likelihood as much as possible, and then the corresponding estimator that we get of the parameter is called as MA or the maximum likelihood estimate, right? OK. So, now, we will see what goes on for a particular ARMA model. So, let us say if you have an ARMA model with n observations, right? Then we can actually write down the log likelihood using a pen and paper. So, the structure of the log likelihood function actually becomes something like that. Now again, I will not go into too much detail. I will not show you or probably I will not prove as to why this is the log likelihood, but you require some assumptions.

$$\ln L = -\frac{n}{2}\ln 2\pi\sigma_e^2 - \frac{1}{2\sigma_e^2} S(\phi_p, \theta_q, \mu),$$

Where $S(\phi_p, \theta_q, \mu)$ is the residual sum of squares, assuming e_t follows a Normal distribution with mean 0 and variance σ_e^2

So, what exactly are the assumptions? So, we assume that this S function, which you see in the log likelihood right here, is nothing but the residual sum of squares.

$$AIC = nlog\, \hat{\sigma}_e^2 + 2M$$

Now, again, all these terms are coming from a regression context as well. So, if you have taken a regression course, maybe sometime in your college days, university days, or even in your current degree that you are pursuing, and if you have taken a regression course, then all these terminologies are kind of borrowed from regression. So, what do you mean by residual sum of squares, right?

So, residual sum of squares is nothing but, once you fit any kind of model, be it regression or time series, whatever, then you look at the residuals, right? So, residuals are nothing but, what exactly? y minus y hat, right? So, the actual value minus the fitted value, so this is nothing but the residual, and essentially what you are doing is you are taking the sum of the residual squares, right? Something like this, so, this capital S represents nothing but that. So, residual sum of squares assumes all my residuals or all my errors follow a particular normal distribution with mean 0 and variance sigma square E. So, these are a few assumptions, again, all these assumptions are kind of consistent if you are applying a regression kind of technique or time series modeling kind of technique, right? And then, generally speaking, all these assumptions are put on the residual or the random error that you have in the model, which, according to our notation, is it. And since we are assuming a normal distribution with mean 0 and variance sigma square E, the first part that you see here in the log likelihood actually comes from the constant that you have in the normal density or the normal PDF.

So, again, like I said, we will not prove this entirely, but then just to give you an idea as to what individual ingredients you have. So, the first part in the log likelihood comes from the constant that you have in the normal PDF, assuming that the mean is 0 and the variance is sigma square E, and the second part, which is capital S of, let us say, phi p theta q and mu, is nothing but the residual sum of squares, okay? And here, we are assuming an ARMA model, right? So, the ARMA model contains p number of phi coefficients and q number of theta coefficients. And if you have an intercept, there will be an intercept, which is mu, okay?

So, these are nothing but the set of parameters, right? So, essentially, what is happening here is that you are applying a maximum likelihood technique to estimate all these parameters, the underlying parameters. And then, once you estimate the parameters, we get hold of the residuals. So, y minus y hat. So, the actual parameter values minus the fitted values give you the residuals, and then we take the summation of the residual squares, which you have here in terms of capital S. So, again, just to reiterate, if you have an ARMA model with n observations, the log-likelihood function can be actually written down in this format.

Then what? So, let us say once you write down the log-likelihood, then the next step is to maximize that because you want an MLE or you want a maximum likelihood estimator. So, after maximizing the log-likelihood function, we actually get this. So, AIC is nothing but N log of sigma square E hat plus 2 into M. Now, again, why sigma square e hat?

So, why do you see hat here? Because this is nothing but the estimator of the variance. So, since you are using an estimation technique which is MLE, you can actually replace that estimator now in terms of variance. But essentially, this is the AIC formula. So, n into log of sigma square e hat plus 2 into the number of parameters you have.

And here, regardless of whichever information criteria you pick, be it AIC, SBC, or Hannan-Quinn, the idea is always to minimize the particular information criteria. So, similarly, pick the model or the particular value of M which minimizes the AIC. So, the goal should always be to minimize any of the information criteria that you have. Now, here, one point I would like to make is: can you actually see that if you bring in a different number of parameters? So, bringing in different parameters means what?

So, you are kind of changing the order of the ARMA model, isn't it? So, let us say if you have ARMA (2,2), as discussed earlier, as compared to something like ARMA (2,1), then you will obviously have a different number of parameters, right? And when you have a different number of parameters, the value of AIC would also change because this depends on this M here. Alright. Then again, the idea should be to pick a model out of these two, or let us say out of the four that we had earlier, which gives you the minimum AIC or the least AIC.

Alright. Okay. So, this is the first one. Now, we will talk briefly about the second one, which is the Schwartz-Bayesian criteria, and then we will see how this is an extension of AIC. Okay.

So, firstly, SBC is a criterion for selecting models with different numbers of parameters again. So, similar to AIC, the idea is the same kind of that if you want to fit or if you want to try out different models with different numbers of parameters, right? Then we can actually use some information criteria, let us say AIC or SBC, and so on. But how exactly is this different? So, when estimating model parameters using maximum likelihood estimation, It is possible to increase the likelihood by adding additional parameters.

So, this is one problem that one can actually face. So, let us see if you keep on increasing the number of parameters, it is possible to increase the likelihood. And if you keep on increasing the likelihood, we will get higher and higher MLE. So, we are kind of reaching the maximum there. So, the estimators might be slightly better.

But one disadvantage here is that it may result in overfitting also. So, let us say if you keep on increasing the number of parameters blindly, that would, of course, result in overfitting. Hence, the BIC kind of solves this problem. So, BIC or SBC resolves this problem by introducing a penalty term for the number of parameters that you bring into the model. So, the whole game here, when you talk about either SBC or BIC, is that it should contain an extra penalty term which kind of focuses on the number of parameters one can bring into the underlying model.

So, now, we will see the formula for SBC. So, this is the formula for SBC. Now, again, the idea is the same: you get hold of the likelihood, you take the log likelihood, you try to maximize that, right? And then this is the formula for SBC. So, n log sigma square e hat plus m log n. Now, again, if you focus on this formula and the earlier one, which is AIC, the only difference is this M log N is mu here, and then in AIC, this was nothing but 2, right?

$$SBC = nlog\, \hat{\sigma}_e^2 + Mlog\, n$$

But here, can you now see that you are kind of giving a penalty term on how many parameters, which is nothing but given by capital N or capital M rather, one can bring in, right? So, if you keep on increasing this capital M, which means that you are bringing additional parameters, right? Then, obviously, this term would go on increasing because now you do not have a constant, unlike we had in AIC, which is 2, but now you have log n, right? And then, essentially speaking, log n is generally bigger than 2, of course, for some n, right? But then, since log n is generally bigger than 2, it is kind of giving a penalty for any extra parameters that you bring.

And a couple of other things: SBC has superior large sample properties, and then SBC is consistent, unbiased, and a sufficient kind of estimator. So, people actually use either

AIC or SBC when it comes to choosing a model based on some information criteria. And again, like we discussed, we will discuss very quickly the third one, which is HQIC or Hannan and Quinn. So, again, as you see the formula. So, Hannan and Quinn gives an extra additional term here.

$$HQIC = nlog\ \hat{\sigma}_e^2 + 2Mlog(log\ n)$$

So, n log sigma square e hat is the same for AIC or SBC. But now, instead of 2m or let us say m into log n, you have 2m log of log n. So, again, Hannah and Quinn have their own advantages or probably disadvantages when it comes to AIC and SBC. But as discussed, people kind of prefer either AIC or SBC because of simplicity. All right.

Now, so, we actually have other techniques for model identification. So, one can actually look at something called a sample inverse autocorrelation function. Now, again, we will not go into details, but the short form is SIACF, right. So, rather than ACF, one can actually have an inverse autocorrelation. So, if you have a model, you write down that model in terms of an inverse kind of structure and then have its correlation, okay.

And a couple of advantages that such a correlation, which is SIACF, generally captures orders better in seasonal models as compared to let us say PACF or even ACF. And one very important idea here is that it is also useful for detecting over-differencing. So, over-differencing means what? In one of the earlier lectures, we discussed this nabla operator, right. So, if you operate this differencing operator on y t, it actually gives you y t minus y t minus 1, and there we saw that one can actually difference in model or difference in time series multiple times. So, let us say nabla to the power d, something like that also, right.

So, but then again, if you keep on differencing unnecessarily or the amount of differencing that you pursue is much more than the expected number of differencing, then using this SIACF kind of function, it detects that over-differencing which is happening. Okay, so now probably in the next session, we will extend this idea, we will see some examples, and then possibly we will move ahead. Thank you.