Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 09

Lecture 45: Practical Session in R - 9

Hello all welcome to this course on time series modeling and forecasting using R. Now again we are almost towards the end of this week lectures and then the focus area towards the entire this week has been spectral analysis or Fourier transform inverse Fourier transform etcetera. So how can one transition from a time domain approach to a frequency domain approach? And again, over the last two or three sessions, we have sort of seen a lot of examples where one can actually transition from a time domain approach to a frequency domain approach. And then what are the examples that one can have where analyzing a practical time series using frequencies or sort of putting forward some sinusoidal functions on top of the practical time series makes much more sense.

Now, this would be the very last session this week and then you have a R session open in front of you and then we will tie all the theory you have seen this week with a small practical session in this session. So, by the way this is code 9 since we are sitting in week 9. And then here we will sort of break it down today's code into three parts. So the first part would be generating time series data which is synthetic data. So synthetic data means you are basically simulating some data.

And then here the heading as you see is spectral analysis in R or using R. And here, let me just scroll down before I run any of the lines just to give you a feel of what we will try to cover in the practical session today. So, we will try to first superimpose a particular time series model on different frequencies and then that collective frequency kind of a domain there. And then let us say here you have to compute the periodogram. So, all the ideas we have covered so far.

So, let us say estimating the spectral density, be it a parametric approach or a non-parametric approach. So, we will Try to tie all the theory we have seen this week with a practical kind of setting here. Now again, just to summarize very quickly, the periodogram is a very famous technique to estimate the spectral density using a non-parametric approach. And this we have seen in the last couple of lectures as well.

And again, down the line, let us say how do you plot the periodogram and then how do you highlight some of the dominant frequencies, right? And then the third aspect could be, let us say, how do you smooth the periodogram. So, again, if you remember from the last session, the periodogram has a difficulty or rather a disadvantage that it is not consistent, right? So, rather, we have to use an approach where consistency could be met, right? So, consistency means what? That variance, if it is not approaching zero or if the overall variance is not getting reduced, then we will say that it is not consistent.

On the other hand, if variance approaches zero as we take capital T to infinity or as we collect more and more observations, if we see that the variance approaches zero as well, then we will say that the estimator or the underlying estimator is consistent, okay? And again, since the periodogram is not consistent, we will opt for the smoothed periodogram. And then, how do you smooth it, and so on and so forth. And then, if you remember, we discussed Welch's method last time. So, we will implement very quickly Welch's method.

And then let us say how do you plot the Welch's method. And now we pay attention to the second example. So, second example is generating slightly different time series data with different frequencies. And down the line, the very last example we will deal with today would be an actual real data example or a practical data example. And then again, as probably all of you might know that we have been using this air passengers data.

So, I thought of why not we will use the same air passengers data and then try to implement from a frequency domain approach on that data set. So, again let me go back and then we will start with the first example. So, the heading again is spectral analysis in R. Now, the first thing is you have to generate some sort of a data first right. So, the very first step is to generate or rather simulate some synthetic data. So, again synthetic means is not a real data.

So, synthetic means something which is simulated ok. So, let us create a synthetic time series with two sinusoidal components and some noise. So, this is the first example. So, the length of the time series we are fixing at 20. So, let me run capital T is 20 and then we will keep the two frequencies at 0.1 hertz and then 0.3 hertz.

Now, this is the exact same example we took in the last session if you remember. So, if you remember the example from the last session where we manually found out the significant frequencies, then we ran a Fourier transform. From the Fourier transform, we found out the periodogram, and eventually, we tried to smooth the periodogram, right? So, all those things we will now try to do using R. So, in this case, my first frequency would be 0.1 hertz, which is the lower frequency. Let me run this, and then the second frequency would be slightly higher, which is 0.3 hertz.

So, 0.1 hertz is the low frequency, 0.3 hertz is the higher frequency, and then the length of the overall time series is 20. And now my small t would be running from 0 to t minus 1. So, let me run all the lines so far. Now, the time point. So, how do you fix the time point?

So, again, my small t would run from 0 to capital T minus 1. So, rather 0 to 19, and once you sort of fix all these things, then you can generate the signal. So, this is my signal, and again, as you see clearly, my signal contains two sinusoidal components. So, the first one is sine of 2 pi into F1 into t, and then the second is sine of 2 pi into F2 into t, of course, with some amplitude here, which is 0.5. By the way, the constant that you see in front of the sinusoidal components is called amplitude because it will decide the height of each of the frequencies.

Make sense? So, if you have if you change this 0.5 to some other number let us say 2 or 3. So, accordingly the height of each frequency or that repetition would sort of change. So, we will generate the signal now which contains two sinusoidal components and the last step is to add some random noise. So, since it is not we do not want an exact kind of a frequency based approach containing exactly two sinusoidal components there should be some random component as well.

So, we will again add some random noise and then again how are we doing it? So, we will generate some random noise like this that you see in line 17. So, noise equals 0.2 into R norm of the length which is capital T. So, here what we are doing is we are generating a random sample of length 20 from a standard normal distribution and we are simply multiplying that vector by 0.2. So, 0.2 into a set of standard normal distributed sample variables with length 20 would be nothing but the noise vector.

And once you generate the noise vector, then the last thing is simply add the signal and the noise. So, signal plus noise would be the actual time series in our case, ok. So, now the overall time series if you ask me would be nothing but the signal containing the two

sign components plus this random noise term, ok. So, the very initially I will quickly show you as to how the plot of the generated time series looks like, ok. So, before and again this is a very simple

strong suggestion that before you proceed with any of the formal analysis, let's say modeling, forecasting, whatever, you have to actually visually try to see exactly as to how the generated time series looks like, okay? And then, so sort of a visual kind of a check has to be made, which is also beneficial to sort of tell you that are there any problems in the generated time series or are there no problems, and all these things could be answered, okay? So, in the same spirit we will draw a plot of the generated series now. So, plot of T and then comma time series type is O. So, type is O means you will see some circles here as pointers color is blue and then we can give it some headings. So, let us say synthetic time series and then the x axis is time and then y axis is value.

Let me let me zoom in the plot for a better and clearer picture. So, this is exactly how the generated synthetic time series looks like ok. Now, again clearly one can see that you do not have any trend here there is no seasonality right it is completely random basically. So, essentially what you have you have two sine components with those frequencies 0.1 and 0.3 hertz and plus the random noise all right. So, let me close the plot and then go back to the code.

Now, once you generate the signal, add the noise and generate the synthetic time series, then the next thing is computing the periodogram. So, again just to summarize very quickly is that periodogram kind of tells you in which positions the significant frequencies lie or periodogram is one idea to sort of identify the significant frequencies. So, the second step is computing the periodogram and here in this period we will use the FFT. So, FFT stands for fast Fourier transform. So, use the fast Fourier transform to compute the periodogram here.

And again some of you may remember that we covered very briefly as to what do you mean by FFT in the last session. So, all these ideas right FFT are kind of extensions of the general non-parametric periodograms that we have seen. So, we will use we will apply the FFT to generate the periodogram and then again in R you have this very simple command called as FFT applied on the generated time series ok. So, fast Fourier transform applied on the generated time series. So, we will we will give it a name as FFT underscore values which is nothing, but FFT applied on the generated synthetic time series ok.

Now, now that now that that part is done. Now, the next step is to fix some frequencies. So, we have to fix some frequencies of interest, right. So, frequencies of interest, again a small disclaimer here that up to the Nyquist frequency. Now, again if you remember what was the idea behind Nyquist frequency is that one has to only focus up to half the frequency length.

So, if the overall time series length is capital T, one can only focus or one should rather only focus up to half that length. So, 0 to T by 2. And then overall, we are kind of standardizing by this t. So, my frequencies would be nothing but 0 to capital T by 2 and then divided by capital T. So, divided by capital T is simply we are sort of standardizing the frequency range or frequency values. So, I will run the frequencies now, frequencies command and the last thing which will generate is power. So, here, so what exactly do you mean by power?

So, power means the actual periodogram values. So, once you generate the frequencies, the x-axis in the periodogram plot contains the frequencies, and the y-axis contains the actual power. So, the actual periodogram values. And this is exactly the definition of power. Again, some of you may remember the formula.

So, this is nothing but the formula where, when you form the frequencies, you take the absolute value, square that, and then standardize by 1 by T. So, this is the exact same formula we are applying here, OK. So, we will generate the frequencies, we will generate the power, and now the last step is to plot the periodogram. So, this is exactly how the periodogram looks, OK. Again, let me zoom in. So again, this is a typical-looking periodogram.

So again, on the x-axis, you have the different frequencies. Let's say 0, then 0.1, 0.2, 0.3, etc. Again, all of these are in hertz. And then the y-axis contains the power. So, here clearly you can see that a couple of frequencies are significant, right?

So, one is at 0.1, the other one is at 0.3 and this is exactly according to our belief or according to what we fixed earlier right. So, if you again if you remember my F1 was 0.1, my F2 was 0.3 and this is exactly what is being reflected in the estimated spectral density through a periodogram. So, one should always do this elementary check that whatever the frequencies you are fixing initially are they being outputted in the periodogram or not make sense. So, here clearly we are seeing the same thing of course, you have some other spikes also which, but of course, all those spikes are not that significant right. So, the 0.1 and 0.3 are highly significant.

So, let me close this. So, one can actually highlight the dominant frequencies also. So, just for some clarity. So, one can actually highlight the peaks. So, let us say highlighting the dominant frequencies.

So, the first highlighting vertical line would be at F1. The second highlighting vertical line would be at F2. So, AB line V equals F1. Now, this V equals F1 command places a simple vertical line at the position F1. and color is blue and then the line style is number 2.

So, again, of course, I can change the line style to, let us say, 1, 2, 3, 4, whatever. So, you have different ways of plotting the line. Do you want a solid line? Do you want a dashed line? Do you want a dotted line?

So, all these things or all these beautifications can be done. So, now the first highlighted frequency is at f1, and the second highlighted frequency is at f2, basically. Now, again, remember here that we are basically highlighting the dominant frequencies, and since we have fixed these two frequencies to be dominant, F1 and F2, at 0.1 hertz and 0.3 hertz. Then again, let me zoom in the plot.

So, apart from the actual periodogram, you will now see the highlights also. So, the first highlighted vertical line is at F1. The second highlighted vertical line is at 0.3. And now, clearly, you can see that the actual true values of the frequency plot kind of superimpose on the generated estimated periodogram plot. Now, again, remember one important thing: the periodogram is nothing but the estimate of the actual spectral density.

So, whatever you think or whatever you fix initially has to be replicated by the periodogram. Make sense? So, again, let us close this plot and then move on. So now, the third point is: why not generate a smoother periodogram? Now again, just to reiterate the same thing that we studied or talked about at the beginning of today's lecture is that a general periodogram has a disadvantage: it is not consistent.

So, if you want to add the consistency property as well, then we have to smooth the periodogram. So, we will create a smooth periodogram, and this is essentially to reduce the noise in the periodogram; we have to apply a smoothing technique. And for this, you require a slightly different package called 'stats'. So, make sure that you install this first and then run it. So, library stats, and then we'll give it a name as smooth_underscore_power, which is nothing but filter.

So, filter is the inbuilt command in the stats package, and filter is applied on the power because power contains the periodogram values. Isn't it? And we have to replicate at these two points. So, one-third and three. So, we will see what you get from this.

So, you generated the smooth periodogram, and then the final step is to plot it. So, plot the frequencies, the smoothed power type is L. L means line color is purple, and then you can give it some title, label the axis, and so on and so forth. So, let me zoom in, let me zoom the plot. So, this is exactly what the smooth periodogram looks like. So, again, the smooth periodogram is nothing but an extension of the earlier periodogram you saw.

So, even here, if you observe, you have a tendency of slightly higher powers at, let us say, 0.1, and again, if you come across 0.3, the same thing could be seen. Okay, and now a slightly different form of a smooth periodogram is called Welch's method. So, Welch's method will again be implemented on the same generated synthetic series, and in short, this is called a segmented periodogram. So again, if you do not remember what you mean by Welch's method, you can always go back to the earlier lecture, try to see the video again, and figure out what Welch's method is and how it differs from any other smooth periodogram technique. So, again, just to reiterate very quickly, Welch's method contains some segmentation.

So, you have to divide the frequencies into baskets, okay. So, divide the time series into overlapping segments and average the periodograms, okay. So, for this, you require a slightly different package called bspec. So, make sure you install bspec, and since I have already installed it, I will not run this command again. I will simply run line 53, which is 'library bspec,' just to bring that package into the R environment. And once you install the BSPEC package and apply the library command on BSPEC, now I can simply apply Welch's PSD estimation.

Now, PSD is nothing but spectral density. So, PSD underscore Welch is the name we are giving it and then this is the inbuilt command in the BSPEC package. So, Welch PSD in capital letters. So, Welch PSD applied on Now, again you have to do a small change here or rather a small extension here that you have to ensure that this time series is a time series object.

right. So, as of now we are not treating this time series as a particular time series object right. So, what we can do is we can simply input this Ts before time series just to make sure that R understands that the generated time series is in fact a time series object. And of course, this kind of a command that adding Ts before any data set you must have seen

in almost all the earlier codes as well. So, whenever you simulate something or you are playing around with some real data, how do you ensure or how do you rather tell R that

A particular vector or particular series has to be considered to be a time series vector or a time series itself, right. So, we simply add this Ts behind any series like this. And here we can, one can actually specify the length of each of the segments. So, we will have the length of each of the segments to be 10. So, these are some technicalities when it comes to Welch PSD command.

So, let me run this, all right. And now the next step is to extract the frequencies and power from the Welch's method. So, the first thing is we will extract the frequencies. So, PSD underscore Welch dollar frequencies and the next thing is to extract the actual power. Now again remember power is nothing but the actual periodogram values.

So, Welch underscore power is nothing but PSD underscore Welch dollar power. So, these are the two things we want. So, first is frequency, the second is power. So, on the x axis again if you remember you plot the frequencies and corresponding to each and every frequency how much is the height of those frequencies is given by the power. So, now once you specify these two things now the next thing is plotting the Welch's periodogram.

So, you plot the Welch's periodogram like this and again let me zoom in the plot this is exactly how the Welch's periodogram looks like. And again clearly you can see that at point 1 and point 3 you see some peaks. This is exactly what we want. But rather you do not see some vertical lines which are broken because you are kind of smoothing the periodogram. So, the entire idea behind Welch's periodogram is you are trying to average out over those overlapping baskets and it should give you a smooth kind of a graph rather than some discrete graph.

But again, clearly the same tendency would be seen here, that at frequencies point 1 and point 3, you see dominant frequencies basically. So now, the next thing is we will pay attention to the second example. The second example's idea is a slightly different time series, that is all. So, the length of this time series is 100. Let me run this.

Before running any of the lines, let me just quickly give you a background of the second example. So, what we are doing here is we are creating a synthetic time series with exponential decay and sinusoidal components. So, in this generation, there will be one

sinusoidal component at 0.15 frequency, which is F1. And then there will be some decay rate. So, which is an exponential decay factor.

So, the decay rate is 0.05 for us. Again, the same thing. So, capital T is 100. So, my small t would run from 0 to t minus 1. So, rather 0 to 99.

And now I can move ahead and generate the signal. So, the signal is exp and then minus decay rate into t into sin of 2 pi f1 into t. And lastly, again as done before, we can add that random noise here as well. So, noise is again 0.3 into R norm of capital T, and then the final time series is nothing but signal plus noise there. All right, signal plus noise. And then now, eventually, we can plot the time series.

So, where you are plotting the time series, it is nothing but a plot between T and then the actual time series again, the type is O. Let me zoom in and then show you the actual plot. And this is exactly how the second generated time series looks like. So, again, if you observe, this time series has a tendency of, let us say, much more oscillations or much more repetitions than before. Make sense? A slightly different example just to differentiate.

Now, we will sort of repeat the exact same thing we did earlier. So, the second step is computing the periodogram again using the FFT technique or the fast Fourier transform technique. So, FFT values is nothing but the FFT command applied on the generated time series. Then we will specify the frequencies, and then again we will generate the power and then we will plot the periodogram. And in this case, this is exactly how the periodogram looks like.

Let me zoom in. And here as you see that you should see a peak at 0.15 roughly. Because again if you remember that the dominant frequency we fixed earlier was at 0.15. But again remember that along with that 0.15 combined inside the sinusoidal component you also had the decay rate. So decay rate might shift the dominant frequency here and there.

So here for example. But again, rather here you see that the significant peak is in fact at 0.15. This 0.15 is the midpoint of 0.1 and 0.2, right? And how does the decay rate come into play is that apart from this 0.15, you can also see some other slightly significant peaks here. So, such a structure is kind of different from the structure we had earlier.

So, earlier we only saw two significant peaks, but now you see multiple of course, one at 0.15 is highly significant, but apart from that you see some lower peaks also. So, again the same thing, if you want to highlight the dominant frequency, you can do that and then

add a legend, something like that. Let me again zoom in. So, this is the highlighted dominant frequency and then legend is F1 is 0.15 hertz. And now again the smooth program.

So, apply a smoothing technique again for the second example. Right. And, by the way, this is one more way of incorporating any function available in a particular package. So, the package name is stats. The function name is filter.

So, how do you incorporate both things in a single line of code? So, stats::filter. And then apply it on power, repeating at 1/5, comma, 5, size equals 2. So, all these are technical terms. So, smoothed power, and then plot the smoothed periodogram, and in this case, this is exactly how the smoothed periodogram looks.

Now, again, clearly you can see a significant peak at 0.15, which we essentially want. And now, the fourth thing is spectral density estimation using this command called spectrum. So, density estimation, and then we will apply the spectrum command to the generated time series, and we can give it some name, let us say, spectral density estimation, something like that. So, all these things could be done, and now we will pay attention to the last example, which contains the actual practical data. Or the real data.

So, again we will play around with the air passengers data slightly. So, we load the data set first. So, data air passengers and then time series and nothing, but the actual air passengers data. And now we will try to plot the actual time series. Again, if you might have forgot as to how the air passengers data set looked like, then again we can give a small plot of that.

So, again if you remember vaguely that this is exactly how the air passengers data looked like. So, it contains both a trend as well as seasonality. Now, before you try to transition into frequency domain and try to put forward any sinusoidal kind of components on this air passengers data, we have to ensure that we sort of remove the trend first ok. We try to make the time series to be roughly stationary. So, the second step is to detrend the time series removing the trend using differencing operator ok.

So, my detrended time series would be nothing but differencing applied on the actual time series. Now, let us plot the actual detrended time series. So, now you can clearly see that the trend component has gone, but you still have some seasonality, some changing variance all of that. Nevertheless, our task is to focus on let us say generating a

periodogram and then identifying the frequencies. So, in that spirit the next step is to compute the periodogram again using the FFT approach.

So, in this case, my capital T is nothing but the length of the actual air passengers time series, right? And FFT values are nothing but FFT applied on the detrended series. Now, we fix some frequencies, we fix some power, and we plot the actual periodogram. So, this is exactly how the periodogram looks like, and here, remember one thing: since we are playing with real data, we did not specify the actual frequency, right? But through this periodogram, you can clearly see that the significant frequencies or the peaks appear. The first one is here.

This could be, let us say, 0.8 roughly, and 1 appears here, which is, let us say, 0.8. This is 0.08, right? Because this is 0, and this is 0.1. So, the first significant frequency appears at 0.08 roughly. The second appears somewhere around 0.17, 0.18 in that range, okay?

Alright, and then again, as before, we can highlight the significant frequencies with the legend and so on and so forth. And then the last thing is spectral density estimation. So, we will estimate the density using spectral density estimation, and this is exactly how the plot looks like, basically. So, this is the spectral density applied on the spectrum generated by the air passengers data. Okay.

And now the very last thing is small interpretation. So, peaks around 1 by 12 might suggest annual seasonality. So, since we deep rendered, but did not ignore the seasonality. So, there is still seasonality in the air passengers data. So, if you see peaks around 1 by 12 that sort of suggest that there is some annual seasonality as expected in this data set.

And additional peaks may indicate other cyclic behaviors in the data. So, all these things would be judged when you try to find out the spectral density estimation of any real data set. So, I think this was a very short attempt to sort of tie all the theory down with the practical data where we generated synthetic data initially. So, first example was the one which contained two sinusoidal components. The second example was the one where there was one sinusoidal component and one decay rate.

And we also worked with a real data set containing the air passengers data.

Thank you.