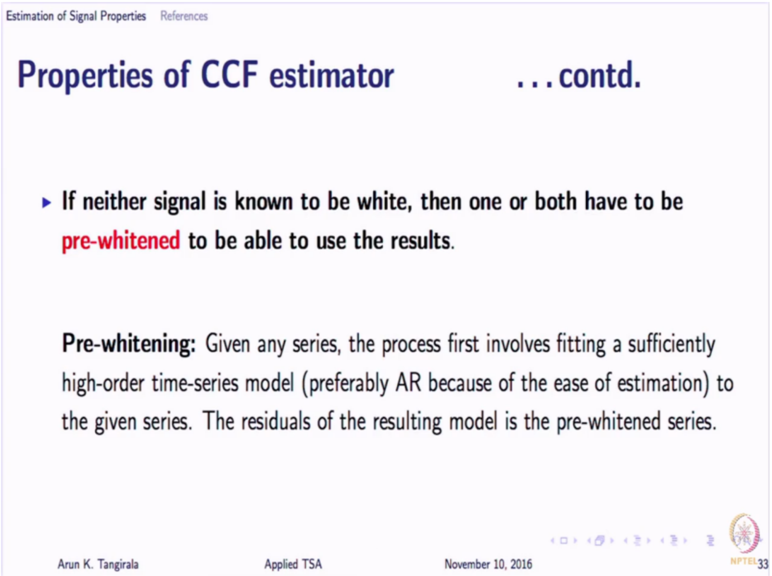
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

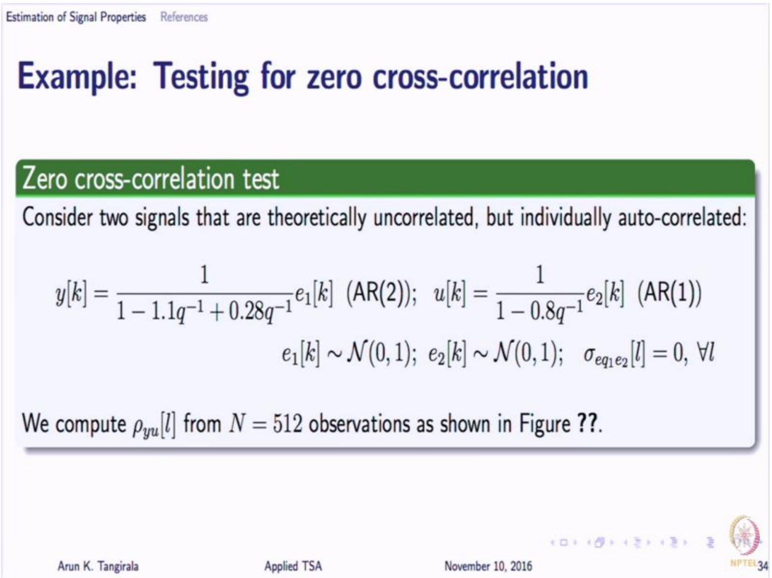
Lecture – 111
Lecture 48B - Estimation of Time Domain Statistics 2

(Refer Slide Time: 00:12)



So, let me show you an example and then I will come back to the statement. So, let us look at this example here where I am generating 2 signals both of them are uncorrelated.

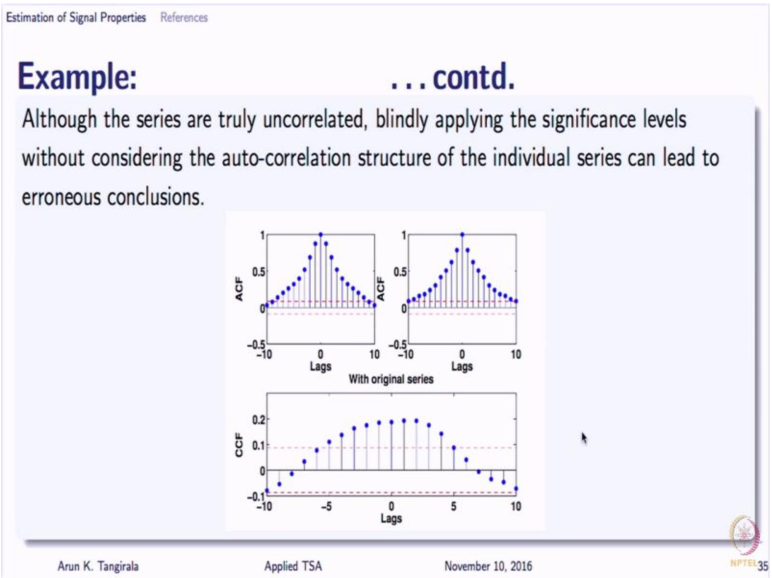
(Refer Slide Time: 00:17)



They are individually coming out of auto regressive processes, which means individually y and u are correlated internally there is a correlation between them this is called internal correlation, what we call as sorry auto correlation, it is also known as internal correlation. So, there is an internal correlation in y and u, but there is no cross correlation why? Because the generating white noises are uncorrelated e 1 and e 2 or e y and e u are uncorrelated.

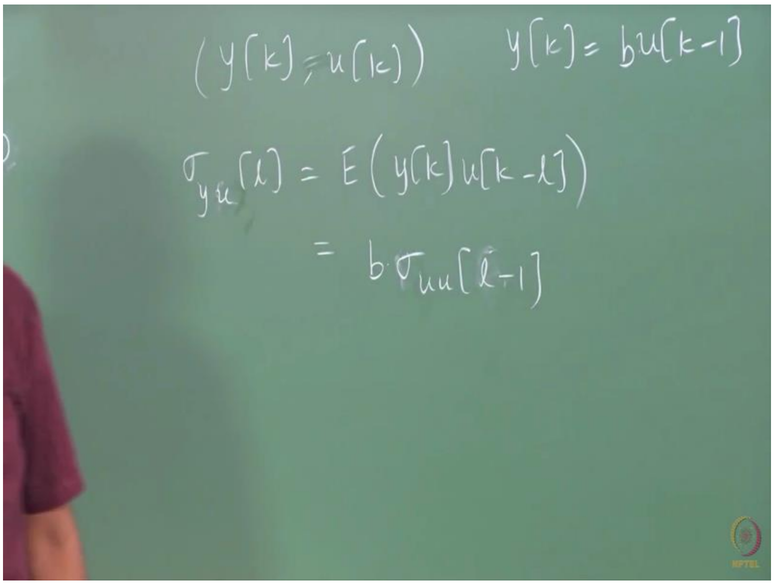
The sources themselves are uncorrelated or e 1 and e 2 whichever where you want to look at it. Now let us generate 512 observations just a number and this is a broken cross reference I am going to remove that.

(Refer Slide Time: 01:13)



So, here what you see on the bottom is the cross correlation, we know that the truth is that they are uncorrelated, then why do we have a cross correlation here, what do you think, what is a explanation for observing this kind of a correlation? We have generated uncorrelated series agree therefore, I should have expected to see a negligible correlation; cross correlation what you see on the bottom is a cross correlation function.

(Refer Slide Time: 01:56)



But we do see a significant correlation, why is that? That is because of the internal correlation you can see suppose I am just taking a different example, suppose that I am looking at now y k and u k right let us write a generic expression I am looking at these two series. So, I am looking at correlation between y and u at some lag, which is expectation of y k let us assume zero mean signals just for simplicity, this is the definition of cross covariance right we are looking at theoretical properties here.

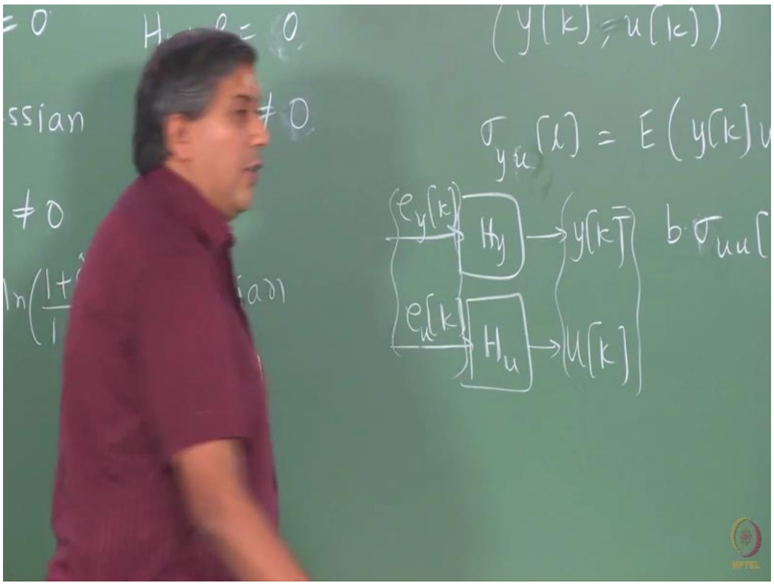
Then what you are essentially doing is, if you; when you are looking at the cross covariance, if internally y and u are correlated then that reflects in this definition, that is a primarily let us assume in the case of correlated, let us, it is easy to understand and so I am going to take a simple example assume that in this example y and u are uncorrelated, but the one that I am writing on the board y and u are correlated there is just a pure delay. So, u y is simple a shifted version of u, ideally based on this relation I should see the cross covariance or cross correlation peak at lag 1 right and b 0 at other lags, but that is not going to happen right. So, substitute this expression here.

Let us say that there is some b u k minus 1. So, I get here well I am going to get here l minus 1. So, what is it tell you? The cross covariance between y and u depends on the internal correlation and u, if the input is this u let us say if this u is white, then what we want to see we will see; what is it ideally we expect to see? The cross covariance peaking at lag 1 becoming being non zero at lag 1 and remaining 0 at all other lags, but that is not going to happen for a general u, for what kind of u will be have that cross covariance then u is white.

So, now you understand the restriction that one of the signals is being white, the actual restriction comes about in deriving the theoretical results that I have given, but I am just illustrating the need for one of the signals being white through a simple example. So, coming back to this example here, primarily you see this correlation significant correlation because of the internal correlation. Let me put it the other way round, what we call as significant is based on this band the significance band here, you can blame it on the significance band you say the significance band calculations are correct only if one of the signals is white, but none of the signals is white here therefore, the band that you see here the dash line that you see I do not know how will you can see. So, hope you are able to see the dash line now, the dash line is no longer valid.

It may be true that this correlation should be insignificant, it is only a numerical artifact, but this one here appears significant because I am using a wrong expression. The dash line is valid only when none of the signals is white. So, you can look at it either way you can say that correlation appears significant because one of the signals none of the signals is white or you can say the significance band itself which I am using to turn the correlation as significant is not valid any more. So, what should we do here? The remedy is to pre whiten the signal.

(Refer Slide Time: 05:58)



What we mean by pre whitening is, suppose I am given I want pre whiten u then I fit a time series model.

So, what I do is I fit some time series model, here there is no issue of ARMA, AR and so on you just fit some time series model and now what you do is, instead of working with u you work with e u. So, you assume imagine that u is being generated by e u and then compute the correlation between y and e u; what is a justification, how can I replace the cross correlation between y and u with y and e u? The justification is if y and u are linearly related then y and e u will also be linearly related, on the other hand if y and u are not linearly related then y and e u are also going to be not linearly related, but at least then I would have satisfied the conditions required for the significance band expressions.

So, when I do that, I am now fine. So, in this case I have pre whitened both y and e u, in other words I have done a similar thing for y as well and instead of computing cross correlation between y and u, now I compute cross correlation between the sources that are generating y and u, that is the point; essentially try to compute the correlation between the sources rather than the series themselves.

Why sources you mean white noise sources. So, this is a very important point to keep in mind, many miss out this point when examining cross correlation, you just simply come to the conclusion that yeah now these are the significance band the cross correlation is significant.

But one as to keep in mind this very important requirement that one of the signals is white, what you can do is you can you can rework this example in R right and see if you get a similar result even if you only pre whiten one of them. How do you pre whiten by the way? As I said you fit a time series model make sure that the residuals are white and the residuals happened to be e r your e u and estimate of e, how do you come to e u? We cannot recover the true e u that is not possible.

But whatever white residuals I achieve are suitable representatives of e u or e y as a case may be. So, this notion of pre whitening is quite prevalent in data analysis. So, therefore, it better to get use to this concept of pre whitening; any questions?

Student: Some of the coefficients (Refer Time: 08:43) error model somehow some of the abbreviations might be (Refer Time: 08:46).

Hm.

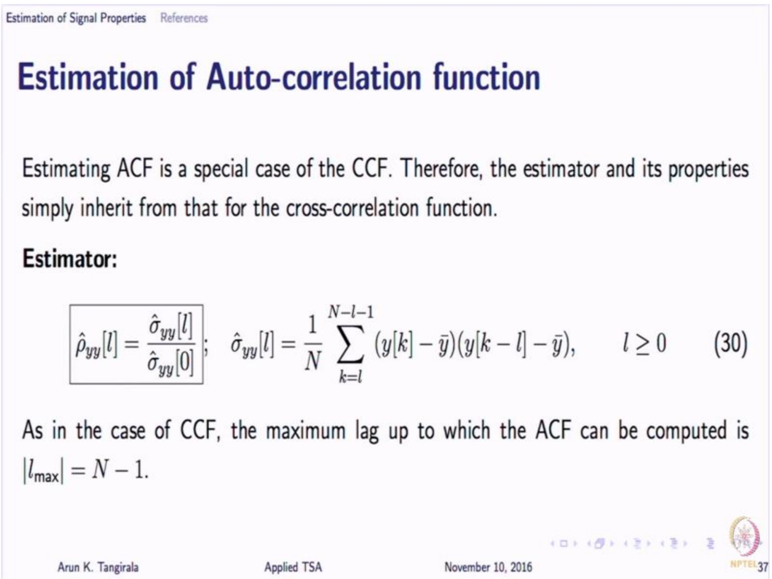
Student: So, the cross correlation (Refer Time: 08:48) what the same member of (Refer Time: 08:51).

Ah that is I mean you basically fit an equal high order model, may be tenth order AR model, just to make sure that both of them give white residuals and of the same length that is practical aspects that is a very minor issue all a.

Student: If you are writing only one?

Then you would to skip the then you yeah your that is very important point, you just throw away the initial few observations in y as well, if you have pre whitening u that yeah good point indeed; any other question? So, this is a story of CCF, just have to be careful always we are interested in the 0 the truth being 0 cases for whatever reasons. So, we will just close this discussion very quickly with estimation of ACF, where we will again come across the same whiteness test.

(Refer Slide Time: 09:44)



Now, this ACF of course, derives itself a estimator derives from the cross correlation function and this is once again a biased estimator yes by the way.

Student: (Refer Time: 09:52) correlation in y and e we have correlation in y and e (Refer Time: 09:59).

Yeah you would not; sorry?

Student: Of, we pre whitening y and u and some correlation.

Correct like the example there.

Student: It does not have a correlation it does not internal correlation

No here in this example y and u do have.

Student: So, how do we differentiate any when we use this CCF (Refer Time: 10:19).

So, when you pre whiten after pre whitening you still see the cross correlation being significant outside the band at any lag, then you conclude that y and u are correlated.

Student: If you distance at lag one in this?

Is there some correlation? So, I do not see anything.

Student: It should be correlation.

No no no, this is no no this is a example for this processing.

Student: This is (Refer Time: 10:42).

Correct.

Student: So, in this.

I did not simulate this case right

Student: In this case you (Refer Time: 10:46).

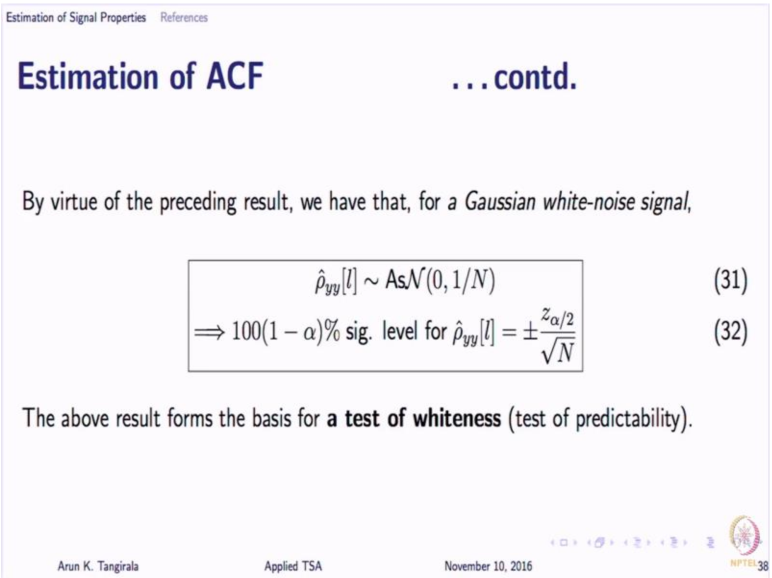
Yes

Student: p get 1.

p get 1 yeah. So, this is a biased the ACF ACVF estimator is a biased estimator and once again we know that it is consistent and so on, but why do we prefer this biased estimator? That is because you can show if you use the unbiased estimator, the unbiased estimator would have a 1 over N minus model or model minus 1 reach after 16 of my textbook carefully, it gives you all the details on this full details. When you work with the biased estimator of ACF, then you would actually end up with a not a non negative definite sequence, I am not proving that here, but proofs are available, but that is a main reason why we do not work with a unbiased estimator; because we want to guarantee non negative definiteness, where do we you know where do things get a beating, when I use the ACF to compute power spectral density?

At that point I have to guarantee that the power spectral density estimates turn out to be non negative that is what essentially non negative definiteness means, when you work with the unbiased estimator, you are not guaranteed that the power spectral density estimates are going to be non negative value where as this one ensures.

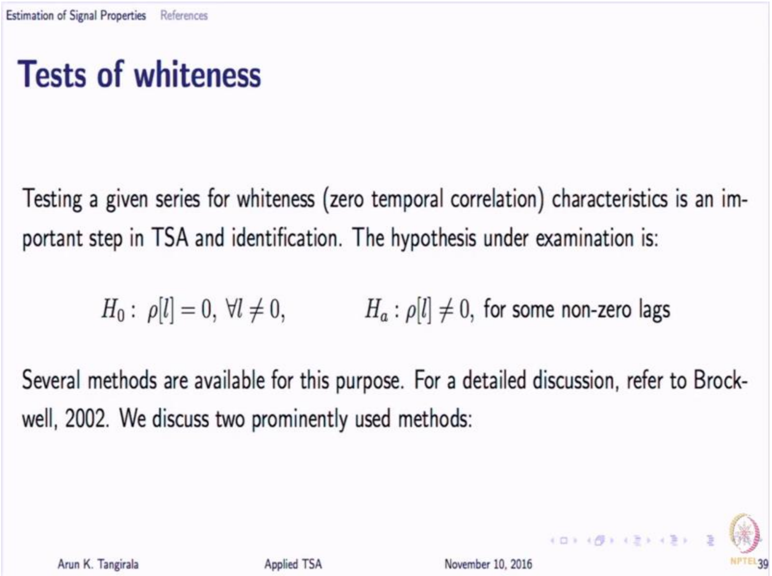
(Refer Slide Time: 12:12)



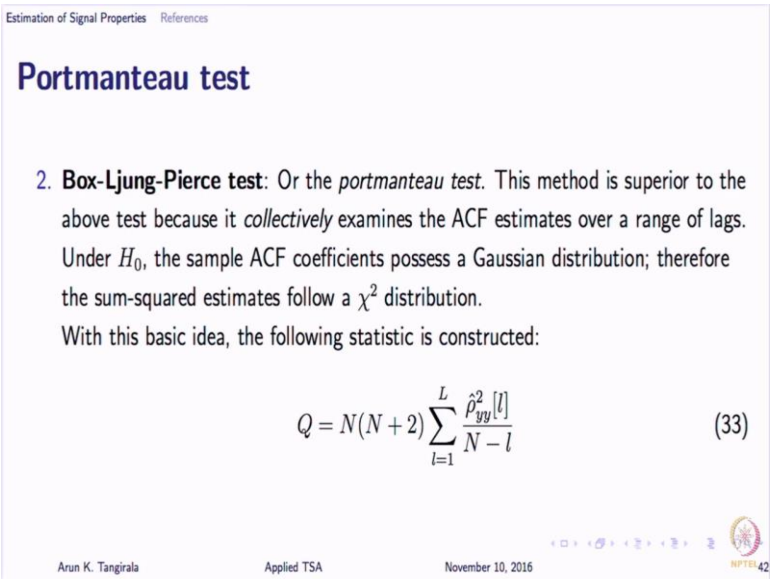
So, that is a reason why we preferred to work this biased estimator, as usual we have results and the result that are of interest to us are the of the case where the process is white, because that is the test that we repeatedly conduct in time series modeling.

Before modeling and during modeling; during modeling we are conducting the test on residuals, before modeling we are conducting the test on the series given series. So, in both cases I am testing for whiteness of the series, now this result says if the process with the data generating process is white, then the auto correlation function as an asymptotically Gaussian distribution with mean zero and variance 1 over N, this is a result that is responsible for the bands that you see in ACF. Long ago someone had asked how do you come up with the bands, significance bands for the ACF? This is a result that gives you the bands that you normally see in your ACF plots, which means that you can now use those bands to figure out if the series is white.

(Refer Slide Time: 13:21)



So, that is that constitutes one test of whiteness, there are apparently if you look at the literature there are several tests for whiteness, several I mean you can go and usually this results would have come out from econometrics of some you know statistic statistics. (Refer Slide Time: 13:46)



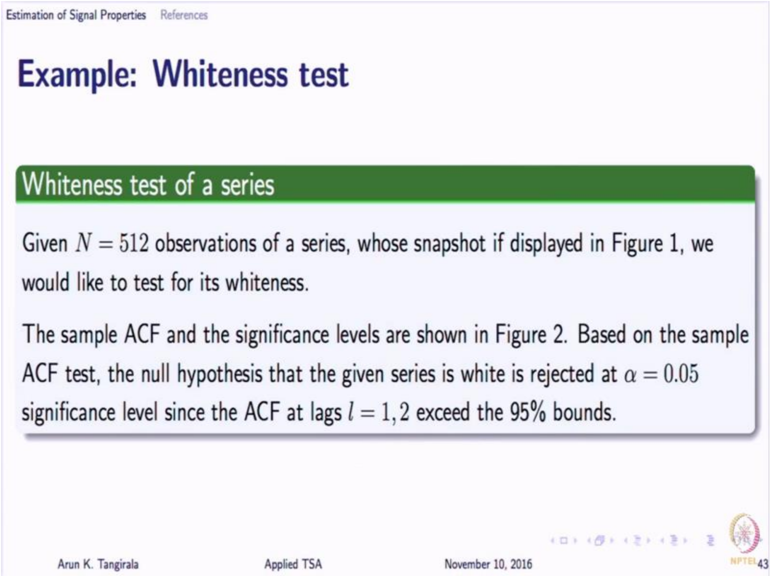
The one test for whiteness uses the significance band result, there is another test of whiteness which is called the portmanteau test or the Box-Ljung-Pierce test, which is based on the assumption that if the data generating process is Gaussian white or you know is uncorrelated it does not have to be Gaussian white, then you can show that the auto correlation estimates posses a Gaussian distribution that is fine.

But furthermore the auto correlation estimates a different lags are also uncorrelated remember we talked about a correlation between estimates at different lags, this result makes use of that fact where as the other one the other test of whiteness which is only based on the significance band does not look at the ACF estimates collectively. Where as this is a stronger version, it can be possible it may be it is it is possible that your auto correlation estimates are correlated at different lags, we are talking about correlation between rho hat of l not the correlation within the series.

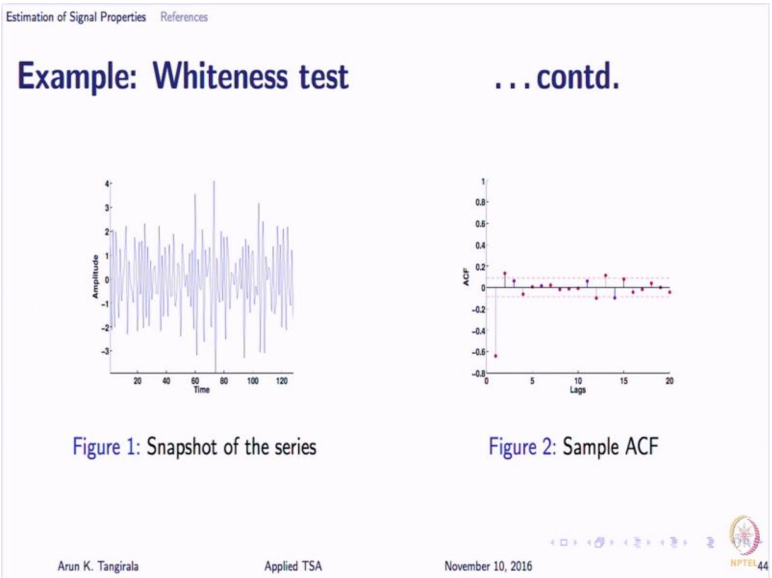
So, the Box-Ljung-Pierce test realize on this fact that when the generating process is white, then the sample ACF coefficients posses a Gaussian distribution and also are uncorrelated therefore, they construct a statistic based on squared auto correlation and we know when I take a bunch of squared uncorrelated Gaussian random variables and sum them up then the resulting one would have a chi square distribution. So, the Box-Ljung Pierce test essentially does not look at a single lag, it is looking at a collection of rows over different lags unlike the significance test which is giving you result at specific lags.

In that since the Box-Ljung-Pierce test is considered to be superior, but then there are results there are in the literature which point out to the short comings of this test as well, but then there is no end to it, if you are in econometric program or quantitative finance program then we would go further and discuss many more tests and so on. So, you can look up there are something there are many other tests for example, got free test and so on. So, I am not going to go over that I am just going to show you result here.

(Refer Slide Time: 16:13)



(Refer Slide Time: 16:21)

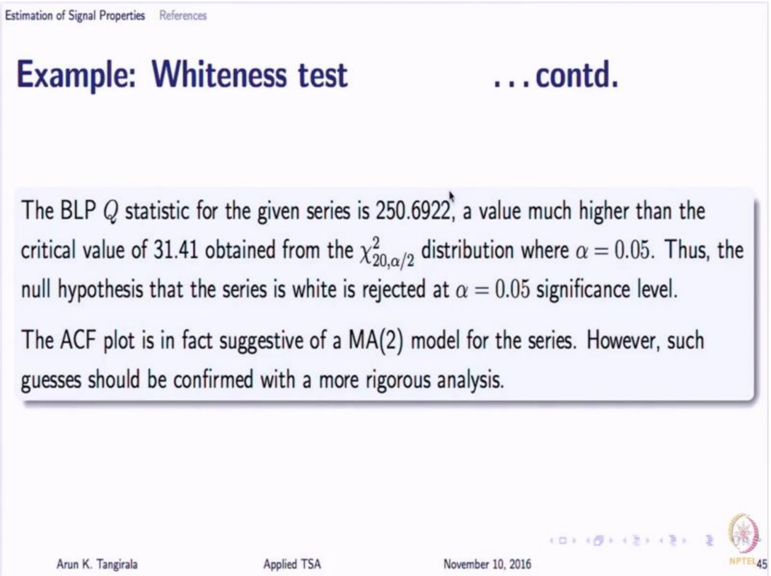


So, here you have 512 observations of a series white noise series, and the sample ACF is shown here. Suppose I am given the series I want to conduct a whiteness test, then I can use the significance band method to determine whether the series is white, as you can see here I do not know how well you can see I zoom in for you. So, this is, so what is happening here I am sorry I have not generated this out of a I mean I not basically telling you whether it is white.

Let us do the test significance band tells clearly that at lag 1 there is a significance correlation. Now at lag 2 also there is a significant correlation and then there are some that go outside and this is a one that bothers the one that look at this ACF very closely, since I saw one ACF estimate going out for a picnic small picnic outside the significance band at lag 20, what do you think about it? Sounds more like generalize (Refer Time: 17:17) and someone. So, I would say take these significance band sometimes with a bucket of salt, because they there are some assumptions involved particularly it assumes that the correlation in the estimates are not there and so on where as the Box-Ljung-Pierce test is much better because it looks at the estimates collectively.

So, what do the Box-Ljung-Pierce test what does a box BLP test say? By the way at the Box-Ljung-test in r does that for you.

(Refer Slide Time: 17:52)



It gives the Q statistic of this value which is much higher than the critical value of 31.41. So, the Box-Ljung pierce test clearly tells you that the statistics is chi square distributed and with how many degrees of freedom right and the critical the observed statistic is higher when the critical value therefore, you reject the null hypothesis that the series is white, the null hypothesis here is that the series is white.

Of course you can look at the ACF plot and may be come to the conclusion that this is MA 2, but you should not make any such conclusion that is definitely MA 2, it may give you a hint this significance band that you have drawn is for what as under what assumption?

Student: (Refer Time: 18:45).

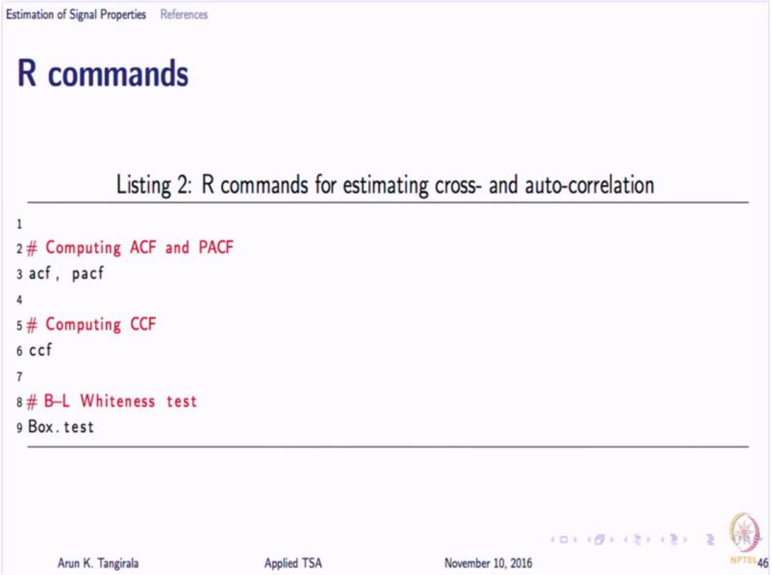
That the underlying process is.

Student: (Refer Time: 18:46).

White; it is only going to serve the purpose of testing for the whiteness, but not for MAQ 2, you have to re derive the significance bands at lags beyond 2, if you hypothesis that the data generating process is MA 2; what I mean is the significance bands are not valid for conducting the hypothesis test of the type whether if the series is MA 1, MA 2 and so on. If it is MA 1 theoretically what should be the case? The auto correlations at lags beyond 2 or what from 2 onwards should be 0. So, we should be drawing the significance band from 2 onwards, but you cannot use these expressions any more you have to re derive, those expressions are given you can refer box in to box and Jenkins. So, there is a worked out example in box and Jenkins, but I do not want to go into that because that is not the full proof way of determining whether the underlying process is MA 1 or MA 2 and so on.

We only want to know if the series is white or not. Once I determine the series is not white, I will go ahead and fit a model and I will fit the best model and I will look at the errors in the estimates of the coefficients look at the trade off and so on and then determine the order, this only gives me a starting guess. Therefore, that is unfortunate part with the working with estimates, theoretically we know what signature the ACF has, but we cannot fully use that theoretical result because this is you have to re derive this significance band, but that is a pain you do not want to do that. Rather use this significance band or the BLP test for determining whether the series is white and if the series fails the test of whiteness, then go head and fit a model of suitable order right.

(Refer Slide Time: 20:49)

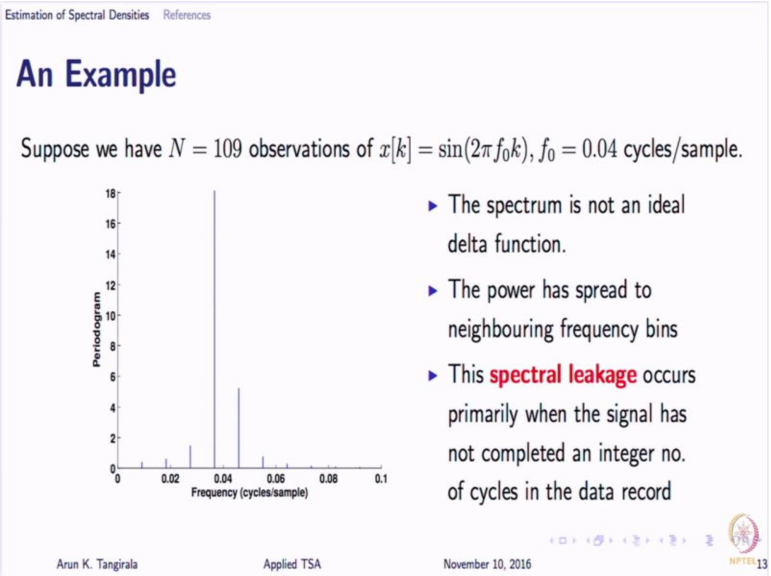


So, that is the sorry. So, that is a thing to remember these are the commands, I will talk about the PACF coefficients when I talk about the correlation test sorry when I talk about tomorrow what methods are available for fitting time series models, because the estimates of PACF are derived from estimates of AR models. So, we need to know the distribution of the AR estimate coefficients estimates, we PACF coefficients remember at any lag l is the last coefficient of an AR l model. So, the distribution of PACF estimate at any lag l will be the distribution of the AR coefficient at that is of that order therefore, we need to look at the time series modeling which I will definitely talk about tomorrow, there is I just spend the couple of minutes in talking about the spectral density estimation and then we will continue tomorrow.

So, now that we have looked at time domain properties, we need to look at the spectral spectrum estimation remember spectral density and spectrum are 2 different things, and one has to be really sure whether the underline signals is deterministic or stochastic and further whether it is a periodic signal or an aperiodic signal and so on. Now having said of all of that, in practice what are we going to use DFT; whether the underlying process is deterministic or stochastic in practice I am going to compute the DFT of the given data record. From the DFT of the data record what am I going to construct an empirical power spectral density, known as a periodogram this is fixed.

Now, whether this periodogram is giving me an estimate of the underlying spectral density or not depends on the generating process. If the generating process is deterministic and periodic and if the data has completed integer if your record as integer number of cycles then there is no issue; if the data as not completed integer number of cycles, I will I am just going to skip those recaps.

(Refer Slide Time: 23:07)

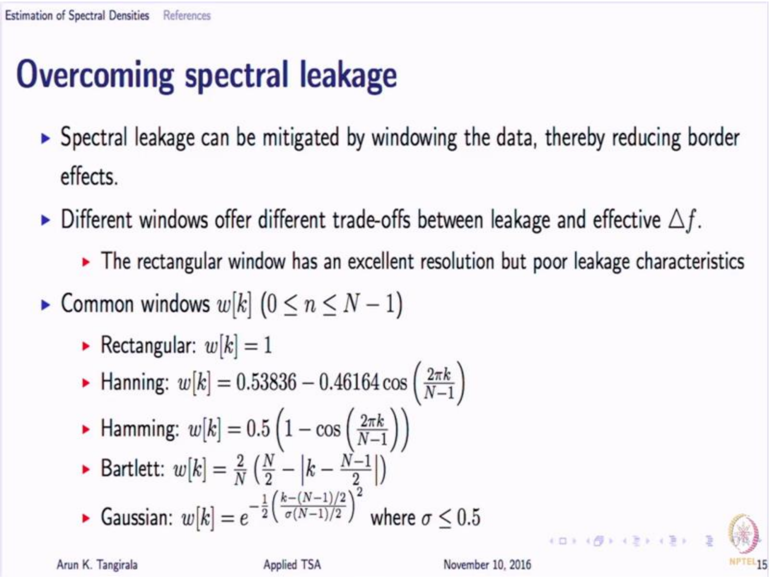


This is your periodogram, we know that finite length effects give raise to spectral equation I will come to the slides perhaps tomorrow. I just want to remind you of the spectral leakage, this is a sinusoidal signal of frequency 0.04 cycles per sample, but I have 109 observations.

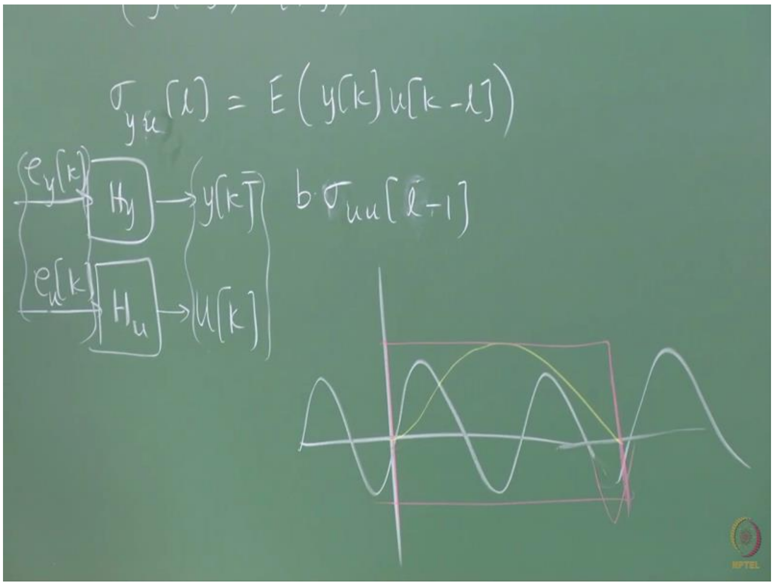
We have seen this example before; ideally I should be seeing a peak in the periodogram exactly at 0.04, but I do not see that because I have 109 observations right and in 109 observations this frequency component does not complete integer number of cycles, that fractional part is giving raise to this kind of a phenomenon known as spectral leakage. More over I do not even have a peak at 0.04 because the binning is different now, the binning is now 1 over 109 that the frequency spacing is 1 over 109 and we know bin with this frequency spacing will exactly hit 0.04.

So, what is a source of this spectral leakage? You can look at it in different ways; you can say that one the signal as not completed integer number of cycles therefore, the fractional thing is causing a problem or you can say that the binning is a problem or you can say that look the underlying signal is periodic. But I have just looked at the periodic signal through a window and that window is not so merciful, it can chop of the signal at any point in time. So, the solution to this spectral leakage is typically to re-window the signal; that means, you take the signal that you have the record that you have, apply a different window so that at the borders where the fractional cycles are being seen, that is being kind of mitigated and that is where the use of window functions are used.

(Refer Slide Time: 25:03)



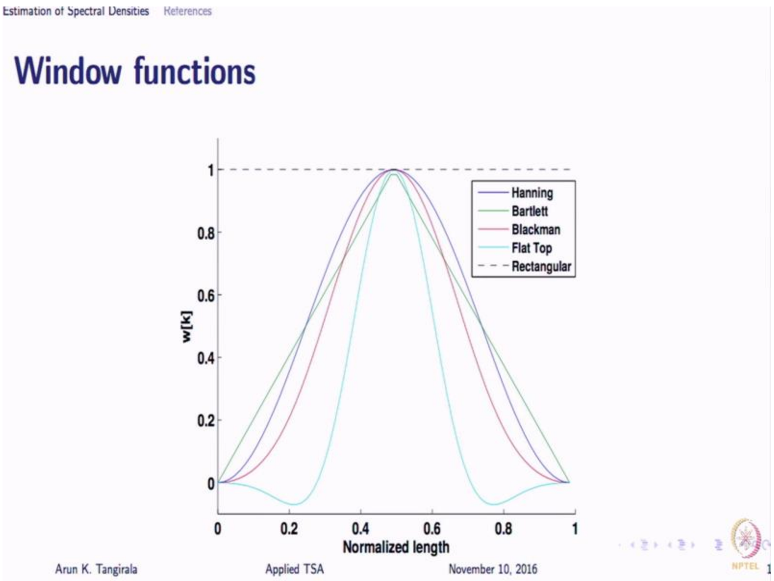
(Refer Slide Time: 25:14)



So, there are number of window functions that people use, rectangular is the default one that is you have this signal here, which you are observing this is an infinitely long signal, but perhaps you are only looking at this window. So, you have just abruptly cut off right of course, I have not drawn properly, but you have just abruptly cut off, this is essentially the window through which you are seeing this infinitely long procession of the periodic signal, the windowing methods essentially now apply a different window here to the same data record.

So, for example, I could now multiply this signal with the window so that you can see the window is gentle at the borders, so that the signal is not abruptly cut off at the borders. By default by observing the signal over a finite time, you are applying a rectangular window and this rectangular window is distorting the true spectrum.

(Refer Slide Time: 26:12)



The introduction of these windows so there are different windows you can see, there is a the dash line here is the corresponding to the rectangular window, all the other windows improve the spectral leakage behavior, but there is only one point that I want to make and we will continue tomorrow. As you use this window functions you may mitigate spectral leakage; that means, you may suppress the power in the neighboring frequency bins, but the effective frequency resolution, what is a frequency resolution here?

Student: 1 by (Refer Time: 26:42).

1 by N that is only when you use a rectangular window, when you start using these other windows which are like bell shaped functions there are number of windows Hamming, Hanning, Blackman, Kaiser and so on. So, all these windows are being gentle at the borders, but effectively if you look at it, it is reducing the number of observations you can look at it that way. You no longer have it is like a degrees of freedom are coming down, you no longer truly have N observations, but less than N; as a result the frequency resolution also drops, there is this is you can think of as a tradeoff between bias and variance. Although we do not talk of bias and variance in the deterministic world, but that is essentially yours; what you lose is your ability to resolve between 2 closely spaced frequencies.

The moment you start windowing, but you cannot help you want to reduce a spectral leakage you apply a window function, what you have to keep in mind is whenever you apply window function and construct the periodogram, by the way when you apply a window function and construct the periodogram it is called modified periodogram. So, when you work with this modified periodogram then and for example, you look at the spectrum command in r there is something called taper, how much fraction of the data you want to taper off it is asking? You can specify that, you can also allow you to specify the kind of window that you want to apply, there is a package called S A P A sapa for spectral analysis, but that is more meant stochastic signals which I will talk about tomorrow.

So, tomorrow for the first 10 minutes or 15 minutes we will talk of spectral density estimation or stochastic signals, this is for deterministic signals and then we will discuss finally, the topic of estimating time series model and a bit of forecasting.

Thank you.