Applied Time-Series Analysis

Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 106
Lecture 46A - MLE and Bayesian Estimation -1

Good evening in the last week of classes, today we will close down on the estimation in the sense of a discussion on estimation methods and then very quickly, we will discuss how to apply these methods to estimating signal properties and time series models. Now before I begin, I know we kind of rush through the non-linear least squares, but as I said there is not much to analyze there so easily so it was more of learning, what is the difference between linear least squares and non-linear least squares and they said the difference is that you move from the notion of regressors to gradients of the predictors, otherwise the remaining results more or less follow having said that I have really condensed a lot of extensive work on the non-linear least squares literature and normally you will see in some preliminary text or preliminary material on non-linear regression on this idea of transforming variables explanatory variables.

If you know the type of non-linearity up front so going back to the ideal gas law suppose I want to predict the temperature using pressure and volume, if I know through the physics or the chemistry of the process that at least for an ideal gas temperatures, temperature varies proportional to the product of pressure and volume then I can construct a regressor up front or sometimes if I am estimating let us say a some Arrhenius constant from the rate of reaction then I, if in other words if there is an exponential relation between the predictor and the explanatory variable I take the logarithm and so on.

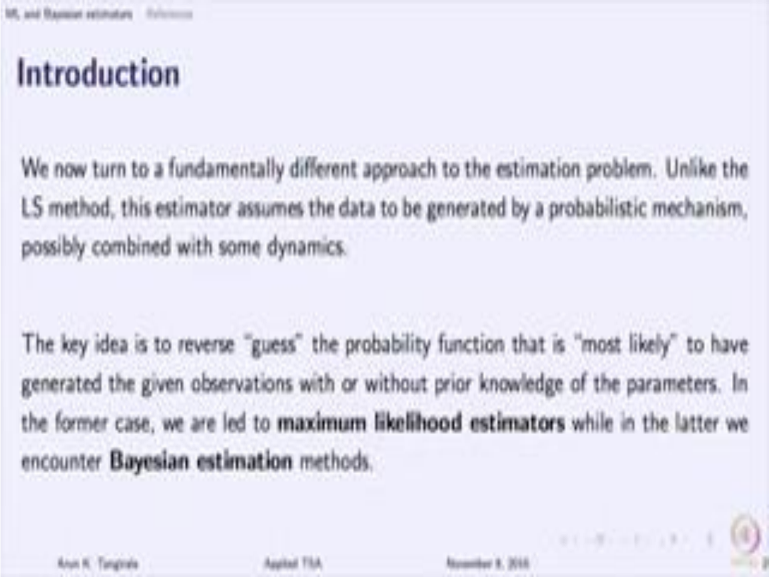
These are the class of transformation methods where you transform a non-linear regression problem to a linear regression problem, in doing so, it is alright when you are doing it at high school and so on because basically the idea had to be given to you that you could transform a non-linear regression problem to a linear regression problem, but it is now time to know that in doing. So, you could be transforming the errors in the data as well suppose y is the measured variable and x or u is the in independent variable and y and u share an exponential relation then I could take the logarithm of y and then rewrite a everything in terms as a linear relation between the transformed y and the explanative variable.

But remember since we said y is a measured variable the errors in y also a transformed. So, question is now whether you assume Guassianity of the errors in a measurement or Guassianity of the errors in transform measurement and that can make a big difference to the quality of the estimates. So, whether they are efficient and so on. So, those are simple approaches normally you do not apply such methods in a setting like this, where you are learning some advance stuff as far as non-linear least square system is concerned as far as time series analysis is concerned you are learning the very basic stuff fine.

Let us now go ahead with the maximum likelihood and Bayesian estimation; we will go through a quick round up I have already explained to you. So, in the last few classes we have learnt extensively at least reasonable extensively sensibly on the principle of least squares method when they give you efficient consistent estimates and so on, they have gone through the idea. So, weighted least squares and non-linear least squares we have already looked at maximum likelihood estimation principle when we were learning Fisher's information.

So, the idea is not new as far as Fisher's MLE is concerned.

(Refer Slide Time: 04:21)



ML and Bayesian estimators References

Introduction

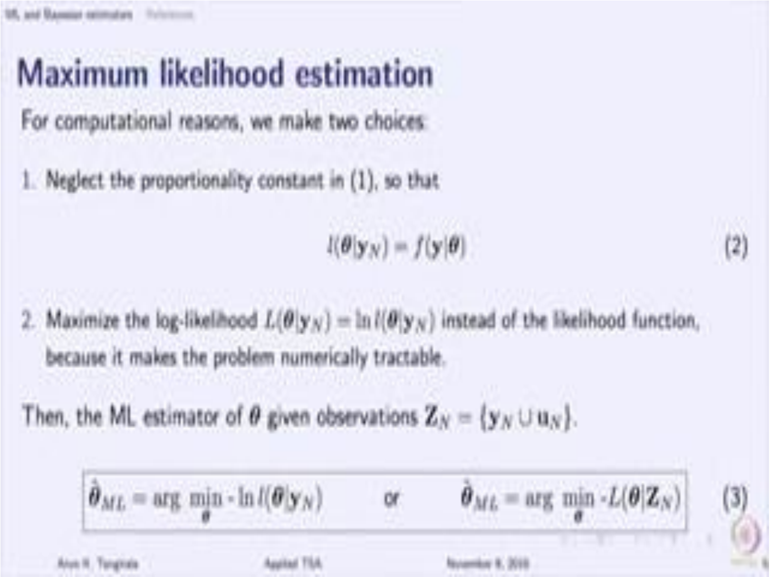
We now turn to a fundamentally different approach to the estimation problem. Unlike the LS method, this estimator assumes the data to be generated by a probabilistic mechanism, possibly combined with some dynamics.

The key idea is to reverse "guess" the probability function that is "most likely" to have generated the given observations with or without prior knowledge of the parameters. In the former case, we are led to **maximum likelihood estimators** while in the latter we encounter **Bayesian estimation** methods.

Arun K. Tangirala Applied TSA November 8, 2016

But let us take a closer look at MLE today again using our classic example of estimating the mean and then quickly move on to Bayesian estimation. Again we will study Bayesian estimation in the context of estimating mean so that estimation of mean is the beautiful example that helps us to illustrate several methods. So, we know we I am going to skip the concept of likelihood and just to refresh likelihood function is nothing, but the pdf, but fundamentally or philosophically they are two different things.

(Refer Slide Time: 04:58)



ML and Bayesian estimators Preliminaries

Maximum likelihood estimation

For computational reasons, we make two choices

1. Neglect the proportionality constant in (1), so that

$$l(\boldsymbol{\theta}|\mathbf{y}_N) = f(\mathbf{y}|\boldsymbol{\theta}) \qquad (2)$$

2. Maximize the log-likelihood $L(\boldsymbol{\theta}|\mathbf{y}_N) = \ln l(\boldsymbol{\theta}|\mathbf{y}_N)$ instead of the likelihood function, because it makes the problem numerically tractable.

Then, the ML estimator of $\boldsymbol{\theta}$ given observations $\mathbf{Z}_N = \{\mathbf{y}_N \cup \mathbf{u}_N\}$.

$$\boxed{\hat{\boldsymbol{\theta}}_{ML} = \arg\min_{\boldsymbol{\theta}} -\ln l(\boldsymbol{\theta}|\mathbf{y}_N) \qquad \text{or} \qquad \hat{\boldsymbol{\theta}}_{ML} = \arg\min_{\boldsymbol{\theta}} -L(\boldsymbol{\theta}|\mathbf{Z}_N)} \qquad (3)$$

Arun K. Tangirala Applied TSA November 8, 2016

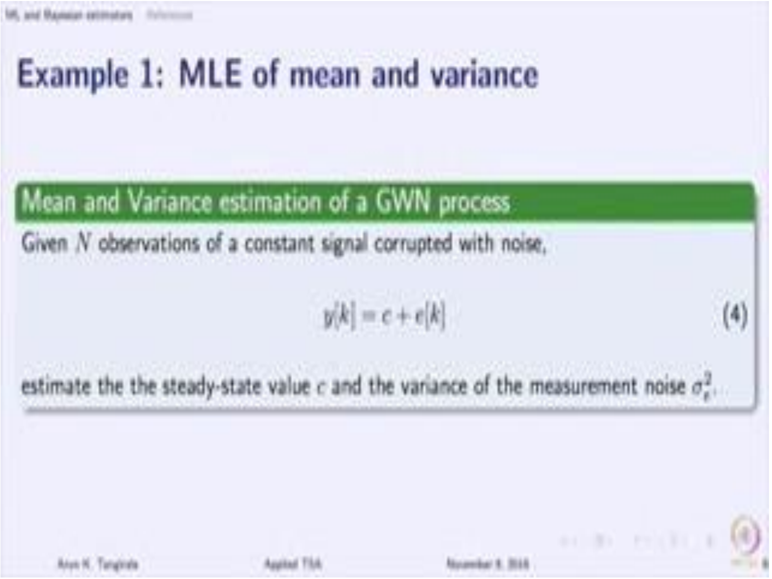
The pdf is a function of y and the parameters are fixed. Where as a likelihood function is conditioned on a given data set it is fixed and the parameters are free to vary and maximum likelihood principle believes that the winner among the unknown p d fs, candidate p d fs is the 1 that produces the observations with the maximum probability that happens to be the pdf proportion to the pdf and that is why you have the likelihood function. So, instead of maximizing the likelihood we maximize log likelihood for mathematical tractability or rather numerical tractability and therefore, we come up with this kind of an objective function although we say maximize log likelihood typically all optimization problems are presented as minimizing something.

We say minimize negative log likelihood now there are some packages in r which will do the MLE for you, but I do not know see why you need a package for MLE unless you say well here is the data and I mention the type of pdf for example, there is a routine call MLE I think it is in stats four package if I am right all it says is give me the log likelihood function if I am going to pass on the log likelihood function why do I need a MLE well there is a reason, but still after all once I construct the log likelihood function it is only an optimization problem there is a reason why there must be there is an exclusive routine for MLE even though you have to manually construct the log likelihood and I will tell you a bit later, but by enlarge you do not need an exclusive function as long as you have access to an optimizer a non-linear optimizer.

If this package or routine, sorry somehow helps you overcome this step of construct in the p d f all you have to say is I have Gaussian I assume that the data as a Gaussian distribution and so on.

Then it is, but you may not find such specialized packages, at least to the best of my knowledge, it is better for you to manually construct what I am going to show you, we will skip the procedure, we know that I am going to take.

(Refer Slide Time: 07:13)



ML and Bayesian estimators References

Example 1: MLE of mean and variance

Mean and Variance estimation of a GWN process

Given N observations of a constant signal corrupted with noise,

$$y[k] = c + e[k] \tag{4}$$

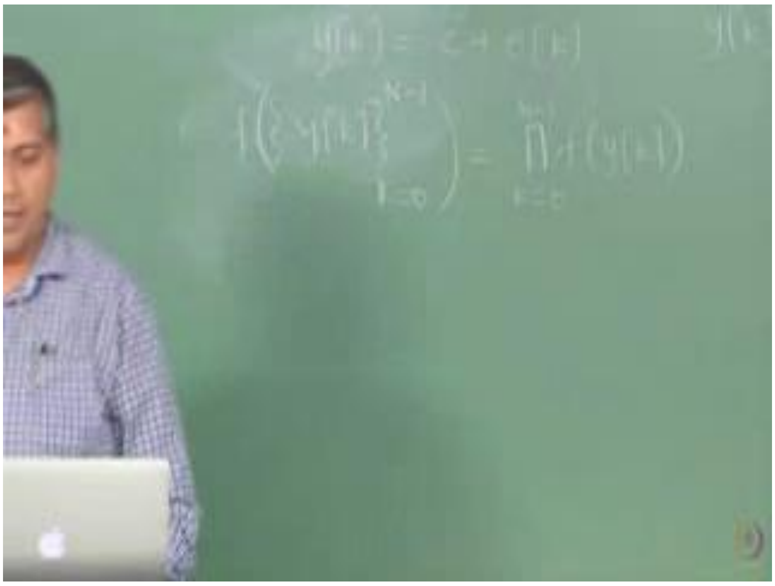
estimate the the steady-state value c and the variance of the measurement noise σ_e^2.

Arun K. Tangirala Applied TSA November 8, 2016

You straight to the example, what I am going to show you now is how to set up the MLE for the estimation of mean, we have seen that already almost half 75 percent of it, we have seen when we were computing Fisher's information, but now we will look at it from an estimation view point rather than the Fisher's information view point and observe close similarities with least squares later on that is tomorrow or bit day after when I show you how these methods are apply to estimating time series models then you realize that a significant effort from the users side in MLE is in setting up the likelihood function.

Once you have set up the likelihood function then it is the computer baba's work, not our work, it has to find the optimum our role kind of ends there. So, the procedure for any MLE at least there are different ways, but this is the very unique standard procedure that you can adopted for solving any MLE problem or setting up likelihood function first given the observations you have to write a model that way which we have already said the model maps the parameter space to the known space. So, here I am given observations of a Gaussian white noise process and I am interested in estimating now the mean and variance.

(Refer Slide Time: 08:42)



The model that we write as I have written on the screen is c plus e k or you can say m u plus e k does not matter you can think of this as mu y does not matter, now the first step in likelihood method is in setting up the likelihood of this n observations and remember since we are going to work with n observations to collectively we have to set up a joint p d f that is where the story begins, here it is a lot easy, in this example when I show you how to set up MLE for AR, one you will feel the flame turned on a bit more right, now there is no flame per say it is it is fairly easy. So, the joint p d f of these n observations has to be written first by identifying the source of randomness in y. So, source of randomness in y is this e k one begins from that point you say this what you make a assumptions on the source of randomness in first you identify the source which is e k here and then you have to make assumptions on the pdf that is generating the source in this already given to us it is a Gaussian white. So, e k is a Gaussian white noise process.

Therefore, I know. So, now, then the next step is asking how does the source uncertainty propagate to the measurements right the step one is to identify the nature of uncertainty in source that is generating the measurements and then determining how the source uncertainty propagates to the measurements. So, here the source uncertainty is Gaussian white and if you ask how does it propagate to white, it is fairly straight forward it just propagates and results in a Gaussian white right if e k Gaussian y k is also Gaussian only mean shifted. So, that is that makes it easy, but that is a single as far as the single observation is concerned.

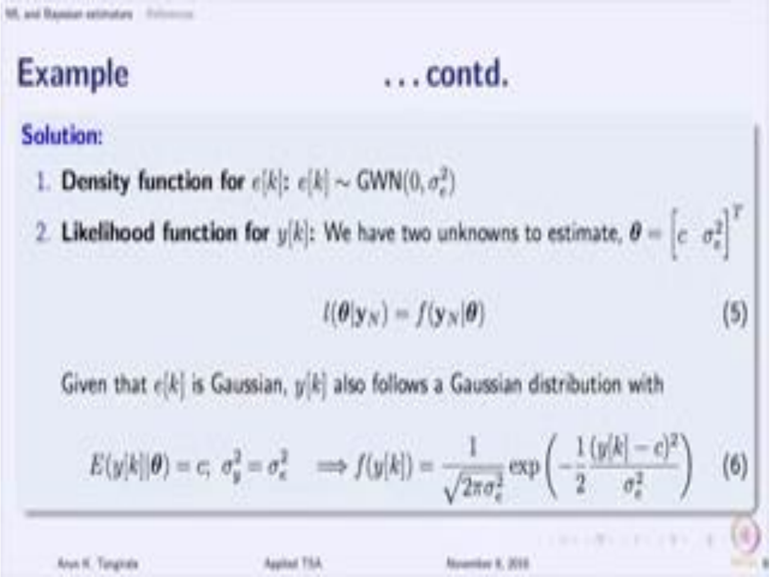
Now, I have to considered n observations jointly when I am looking at n observations jointly I have to worry about the correlation or the dependence between the observing. In fact, specifically the dependence because that is what determines the joint pdf this case we are lucky because it is Gaussian white noise we know that uncorrelated since e k is uncorrelated y k is also going to be uncorrelated on the other hand if you had an AR1 just to give you glance, suppose y was being generated by an AR1 process against step one. So, the goal here would be estimate the model parameter d one if necessary the mean of y otherwise you say well y 0 mean, but definitely d one and sigma square e.

In this case again you start with the same question, what is the source of randomness in white e k? What is the nature of such randomness Gaussian white? Let us see you have given that now you have to ask if e k is a Gaussian white, what is going to be the nature of y k in terms of distribution that we know because it is a linear process we know that a Gaussian nature is going to be transmitted to y k as well, but the difference between y and e is that e k's are uncorrelated where as y k's are not. So, that brings a dependence and writing the join p d fs is not going to be straight forward it is going to be not difficult, but it definitely it is not going to be straight forward.

We will learn a bit later as to how to handle correlated series here there is no correlation among y. So, it is fairly straight forward we have done this before we say that this is a product since you have to be careful here since you are looking at Gaussian uncorrelated series you might as well call the series as independent if it was some other distribution and still being uncorrelated you cannot claim independence remember only for Gaussian processes uncorrelatedness means independence.

You have to keep that in mind that is it. So, you say this a product that is exactly what it is.

(Refer Slide Time: 13:10)



ML and Bayesian estimators Reference

Example ...contd.

Solution:

1. **Density function for** $e[k]$: $e[k] \sim \text{GWN}(0, \sigma_e^2)$
2. **Likelihood function for** $y[k]$: We have two unknowns to estimate, $\boldsymbol{\theta} = \begin{bmatrix} c & \sigma_e^2 \end{bmatrix}^T$

$$l(\boldsymbol{\theta}|\mathbf{y}_N) = f(\mathbf{y}_N|\boldsymbol{\theta}) \qquad (5)$$

Given that $e[k]$ is Gaussian, $y[k]$ also follows a Gaussian distribution with

$$E(y[k]|\boldsymbol{\theta}) = c; \ \sigma_y^2 = \sigma_e^2 \quad \Longrightarrow f(y[k]) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{1}{2}\frac{(y[k]-c)^2}{\sigma_e^2}\right) \qquad (6)$$

Arun K. Tangirala Applied TSA November 8, 2016

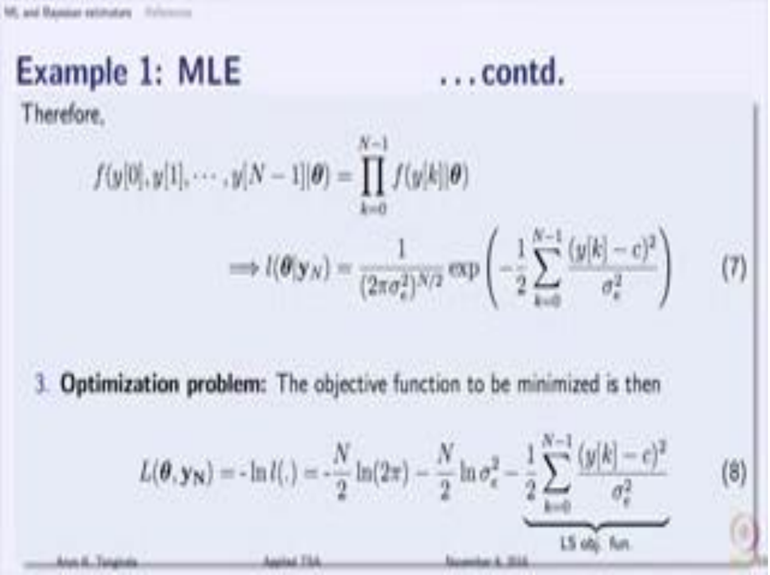
And the p d f of a single observation is written on the screen for it is a standard general generate Gaussian distribution with the mean c and variance sigma square, how did I write this? In doing this we had to do first determined the p d f of y since i given e k is Gaussian y k is Gaussian step one then remember that i in order to write the pdf of single observation I need to know the mean of y and variance of y. So, mean of y is cleared it is see I have to be now alert when it comes to calculation of variance of y whether it is a same as e or something else what do you think the variance of y what is it this is same as sigma square e again it is a simple one here.

So, no brainer, but here one has to be careful, assume that this is the stationary process, assume a e k is 0 mean, we have shown that y k is 0 mean, but the variance of y is not the same as variance of e, what is the variance of y in the AR1 case?

Student: (Refer Time: 14:27)

By 1 minus d 1 square, so, you have to be careful when you write the pdf of a single observation here it straight forward. So, here we know that mu y is c and sigma square y is sigma square e. So, using the fact that y k is Gaussian it as a mean c and variance sigma square e we write the Gaussian pdf here that is it and then we put together the p d fs as a product.

(Refer Slide Time: 14:58)



Example 1: MLE ...contd.

Therefore,

$$f(y[0], y[1], \cdots, y[N-1]|\boldsymbol{\theta}) = \prod_{k=0}^{N-1} f(y[k]|\boldsymbol{\theta})$$

$$\Longrightarrow l(\boldsymbol{\theta}|\mathbf{y}_N) = \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left(-\frac{1}{2}\sum_{k=0}^{N-1} \frac{(y[k]-c)^2}{\sigma_e^2}\right) \qquad (7)$$

3. **Optimization problem:** The objective function to be minimized is then

$$L(\boldsymbol{\theta}, \mathbf{y_N}) = -\ln l(.) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma_e^2 - \underbrace{\frac{1}{2}\sum_{k=0}^{N-1} \frac{(y[k]-c)^2}{\sigma_e^2}}_{\text{LS obj. fun}} \qquad (8)$$

Arun K. Tangirala Applied TSA November 4, 2016

.

And that is what that is expression that you get in equation 7 and the goal and now this is my likelihood function where I am given y k’s and I do not know c and sigma square. The optimization problem is that of maximizing the log likelihood or minimizing. In fact, there should be a minus l here in equation it I will correct that at the front minus big l.

The optimization problem is that of minimizing the negative log likelihood. So, here you see the negative log likelihood now if you look at this expression eight carefully particularly pay attention to the last term that is something that we have seen before right we have seen this before this expression that last term which as a summation we have seen that before in the context of least squares estimation if you recall when we setup the when we introduce estimation with this example and we said we would like to be minimize the sum square error which is nothing, but the least squares idea you encounter this term.

Naturally therefore, least squares is embedded in MLE we did not start with a least squares idea at all look at this gauss preceded fisher at least by hundred plus years it order of least squares, but when it comes to Fisher term by when gauss at actually already atomized you see MLE which embeds least squares on the other hand if you look at Bayesian, Bayesian precedes MLE at least the idea of Bayesian Bayes rules and so on.

But we will see shortly that Bayesian ideas include MLE as a natural thing. So, in the order of hierarchy you see least squares and then MLE as a super set and then Bayesian as even a super super set. So, always remember this that you will naturally the working out on least squares methods the moment you work with MLE, but with the Gaussian distributed errors you wouldn't run into this expression if I were to tell you that e k is i i d with the poisson pass on the distribution or with the uniform distribution or some other distribution I tell you that e k's are independent, but I it is a non Gaussian distribution let us say.

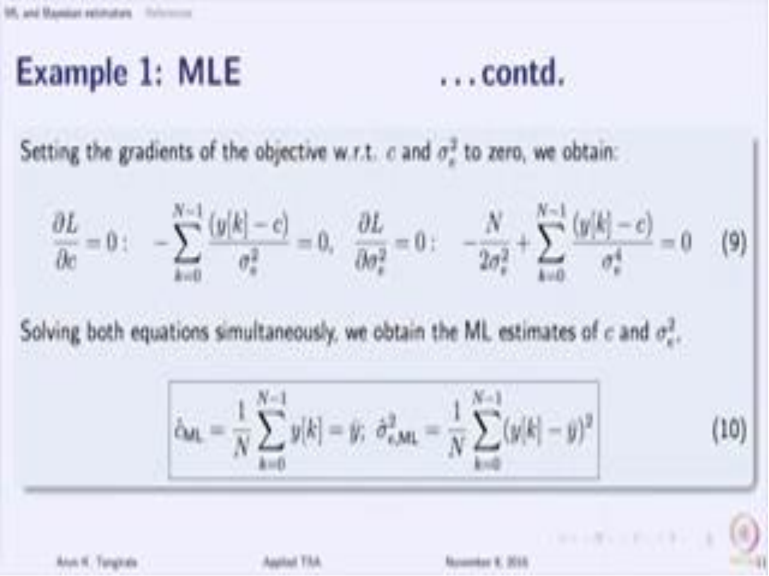
Then you would not run into this kind of expression which means that least squares is a special case of a MLE with Gaussian distribution that is very well known and just telling you giving you this information and probably that is coming to you as a first time in your life now apart from this summation here there is another term that as to be optimized in total there are three terms, but the first term is a constant. So, we can ignore that it is independent of theta.

The second term is only dependent on sigma square e and the third term is dependent on both c and sigma square e; however, you can pull out you may say it is not exactly identical to least squares because in least squares I had you I had only sigma y k minus c square, but you can pull out the sigma square e out of the summation and basically see that you are giving or you can include that and you can say that you are including a waiting of sigma square e which is the same for all observations.

Gradually you should sense you should now smell that even weighted least squares is perhaps a part of MLE, it is not just on ordinary least square even weighted least squares can be shown to be a part of MLE we will come to that very quickly. So, the point here is least squares is a special case of a MLE with Gaussian distributed errors and that MLE looks at solving a different kind of optimization problem with least squares embedded into it now; obviously, in this case this is so special that you can come up with an analytical solution straight away.

What is the estimate of sigma square e in this case see we have already solved it is a sample mean you can see straight away minimization of this with respect to c we will get to the sample mean.

(Refer Slide Time: 19:38)



ML and Bayesian estimators References

Example 1: MLE ...contd.

Setting the gradients of the objective w.r.t. c and σ_e^2 to zero, we obtain:

$$\frac{\partial L}{\partial c} = 0: \quad -\sum_{k=0}^{N-1} \frac{(y[k]-c)}{\sigma_e^2} = 0, \quad \frac{\partial L}{\partial \sigma_e^2} = 0: \quad -\frac{N}{2\sigma_e^2} + \sum_{k=0}^{N-1} \frac{(y[k]-c)}{\sigma_e^4} = 0 \qquad (9)$$

Solving both equations simultaneously, we obtain the ML estimates of c and σ_e^2.

$$\boxed{\hat{c}_{ML} = \frac{1}{N}\sum_{k=0}^{N-1} y[k] = \bar{y}; \; \hat{\sigma}_{e,ML}^2 = \frac{1}{N}\sum_{k=0}^{N-1}(y[k]-\bar{y})^2} \qquad (10)$$

Arun K. Tangirala Applied TSA November 8, 2016

A sample mean is also the maximum likelihood estimator sigma square e estimate optimal estimate turns out to be what I show you here in equation ten it is a it is our un it is of biased estimator of variance it is not your least squares if it were to be a least squares estimator of variance you to you would see a one over n minus 1.

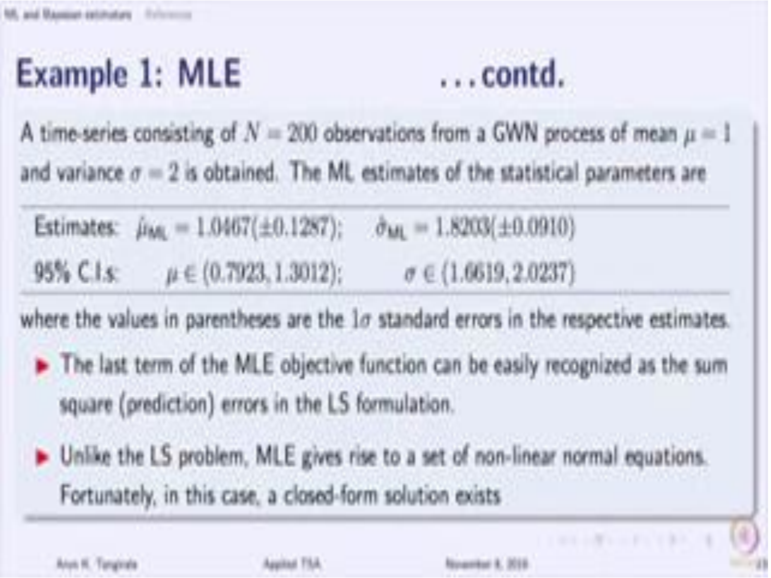
What do we do in least square method? We first estimate the unknowns parameters, unknown here is c and that terms out to be sample mean and then you estimate sigma square e as sum square error by n minus p the sum square error is sigma y k minus y bar square divided by n minus p in least squares where p is the number of parameters you would have estimated which is one that is why the we have a one over n minus one in MLE you have one over n right. So, in MLE; in general MLE will lead to bias estimates sample mean is a un biased, but that is just a coincidence in general you can expect maximum likelihood to give to you bias estimates, but we know that asymptotically un biased. So, as n grows large the bias vanishes.

So, which is very good that is why maximum likelihood estimators are good in the sense and for as n becomes large it will give you asymptotically un bias estimators, but there is a more important property why maximum likelihood estimators are preferred because it gives you the most efficient estimators that is point number one point number two is it naturally accommodates heteroskedastic errors as well what I mean by that is if I were to tell you go back to this problem and tell you that e k is white in the sense it is uncorrelated, but it is heteroskedastic; that means, it is variance changes with time where do you see the change coming in equation six you would have sigma square k right the variance of the kth observation and all this equations would be affected instead of the some weightage to all observation you would have observation dependent weighting that is nothing your optimal weighted least squares method.

Weighted least squares with the optimal waiting already incorporated when we formulated weighted least squares we had to take two steps first we had to modify the ols by introducing the weighting matrix and then ask the question what kind of weighting matrices would give me efficient estimates in MLE you do not have to ask such questions at all straight away it is taken care of the weighting is taken care of the optimal weighting is also incorporated as a result you should expect efficient estimates of course, you can solve this assuming sigma square e k is known if it is not known then you have to estimate them iteratively that is another issue, but what I am trying to tell you here is weighted least squares is a special case of MLE with heteroskedastic Gaussian errors uncorrelated Gaussian errors that is it.

MLE incorporates least squares weighted least squares and if necessary even non-linear least squares depending on the predictor that you have naturally.

(Refer Slide Time: 23:04).



Example 1: MLE ...contd.

A time-series consisting of $N = 200$ observations from a GWN process of mean $\mu = 1$ and variance $\sigma = 2$ is obtained. The ML estimates of the statistical parameters are

Estimates: $\hat{\mu}_{ML} = 1.0467(\pm 0.1287)$; $\hat{\sigma}_{ML} = 1.8203(\pm 0.0910)$

95% C.I.s: $\mu \in (0.7923, 1.3012)$; $\sigma \in (1.6619, 2.0237)$

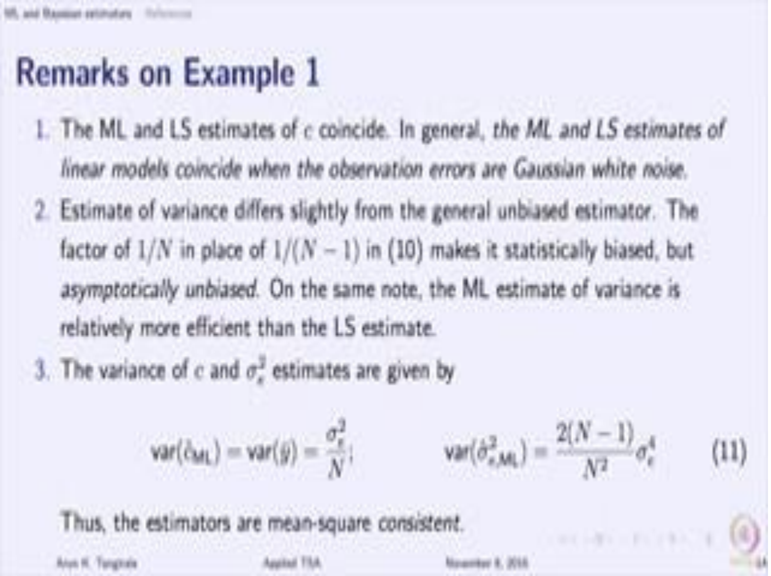
where the values in parentheses are the 1σ standard errors in the respective estimates.

- The last term of the MLE objective function can be easily recognized as the sum square (prediction) errors in the LS formulation.
- Unlike the LS problem, MLE gives rise to a set of non-linear normal equations. Fortunately, in this case, a closed-form solution exists

Some of the just as a quick numerical example I am showing you just have randomly simulated about two hundred observations of a Gaussian white noise process with mean one and variance two and I am reporting to you the maximum likelihood estimates is nothing, but your sample mean with this standard error and this is the estimate of the standard deviation standard deviation is sorry it should be standard deviation not variance here.

The standard deviation turns out to be 1.82 pretty close to the true value with this error and the 95 percent confidence intervals are included here once you derive the estimator then you go through the standard questions whether it is consistent efficient what is the distribution and all of that we have already discussed what sample mean is bias un biased estimator and so on, we have also discussed about the sigma MLE as well now of course, in this case you could use least squares as well.

(Refer Slide Time: 24:11)



Remarks on Example 1

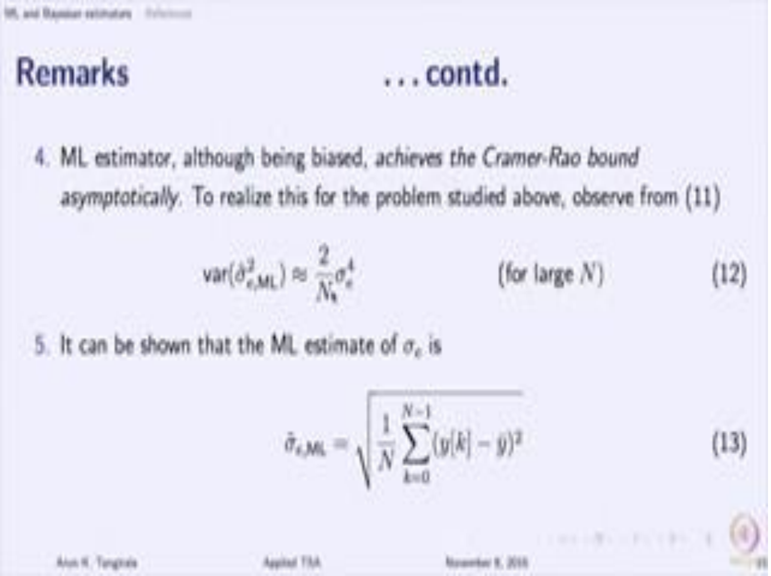
1. The ML and LS estimates of c coincide. In general, *the ML and LS estimates of linear models coincide when the observation errors are Gaussian white noise.*
2. Estimate of variance differs slightly from the general unbiased estimator. The factor of $1/N$ in place of $1/(N-1)$ in (10) makes it statistically biased, but *asymptotically unbiased.* On the same note, the ML estimate of variance is relatively more efficient than the LS estimate.
3. The variance of c and σ_e^2 estimates are given by

$$\text{var}(\hat{c}_{\text{ML}}) = \text{var}(\bar{y}) = \frac{\sigma_e^2}{N}; \qquad \text{var}(\hat{\sigma}^2_{e,\text{ML}}) = \frac{2(N-1)}{N^2}\sigma_e^4 \tag{11}$$

Thus, the estimators are mean-square *consistent.*

But in general MLE is preferred. So, some remarks on this example we have talked about a few things already I am going to actually go pass this.

(Refer Slide Time: 24:21)



Remarks ...contd.

4. ML estimator, although being biased, *achieves the Cramer-Rao bound asymptotically.* To realize this for the problem studied above, observe from (11)

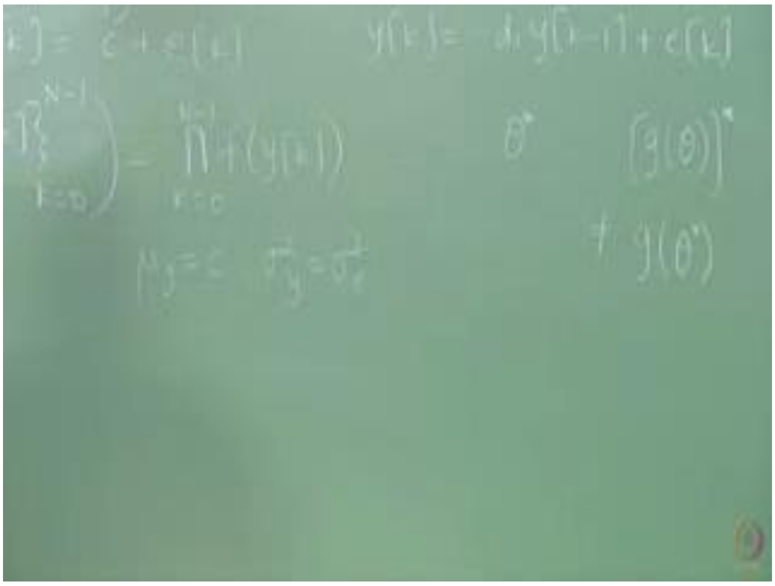
$$\text{var}(\hat{\sigma}^2_{e,\text{ML}}) \approx \frac{2}{N}\sigma_e^4 \qquad \text{(for large } N\text{)} \tag{12}$$

5. It can be shown that the ML estimate of σ_e is

$$\hat{\sigma}_{e,\text{ML}} = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1}(y[k]-\bar{y})^2} \tag{13}$$

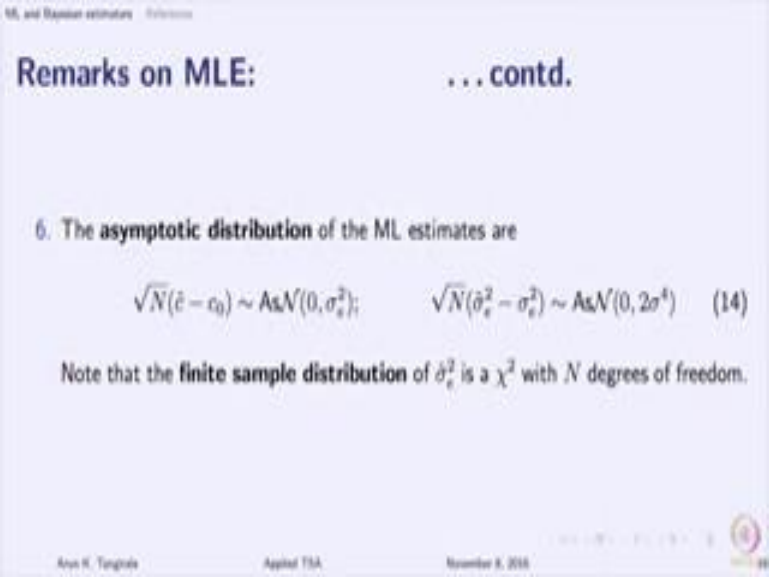
There is 1 point remark that I want to draw your attention to which is that the maximum likelihood estimate of sigma e is the same as first estimating sigma square e and then taking a sigma square root remember when we talked of Fisher's information we said the Fisher's information of theta is different from Fisher's information of g of that is if if there is a parameter theta and then you are looking at estimating g of theta then in general the estimation of g of theta is not necessarily optimal estimate of g of theta is not necessarily the same as first optimal estimating theta and then taking g right if theta star is an estimate of theta and I am looking at estimating g of theta optimally then g star of theta.

(Refer Slide Time: 25:11)


$y[k] = -d_1 y[k-1] + e[k]$
$\prod_{k=0}^{N-1} f(y[k])$
$\mu_y = c \quad \sigma_y^2 = \sigma_e^2$
θ^*
$[g(\theta)]^*$
$\neq g(\theta^*)$

If you said this star is not necessarily equal to g of theta star in general, but with maximum likelihood it as a beautiful invariance property which says that you do not have to worry if you want to estimate g of theta simply obtain ml of theta if that simplifies life for you estimate theta first and then take the transformation that is what this expression is telling you, you do not have to re again solve the MLE with sigma e as the unknown you do not have to do that you can simply estimate sigma square e like we did and then take the square root this is called the invariance property of MLE I am giving you certain things without proofs if you are interested in proof certainly you should go and look up standard text book on MLE.

(Refer Slide Time: 26:16)



ML and Bayesian estimators

Remarks on MLE: ...contd.

6. The **asymptotic distribution** of the ML estimates are

$$\sqrt{N}(\hat{c} - c_0) \sim \text{As}\mathcal{N}(0, \sigma_e^2); \qquad \sqrt{N}(\hat{\sigma}_e^2 - \sigma_e^2) \sim \text{As}\mathcal{N}(0, 2\sigma^4) \qquad (14)$$

Note that the **finite sample distribution** of $\hat{\sigma}_e^2$ is a χ^2 with N degrees of freedom.

Arun K. Tangirala Applied TSA November 8, 2016

.

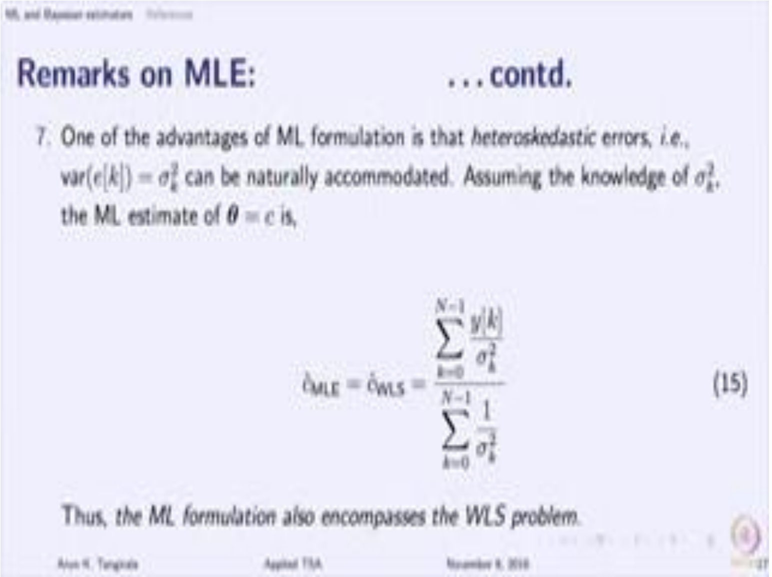
And finally, you can show that the asymptotic distributions of all maximum likelihood estimates are Gaussian distributions in this case the parameters of interest are c and sigma square e these are asymptotic distributions which means large sample cases only although I am giving you these distribution results for this example in general if you take any maximum likelihood estimate it is a asymptotic distribution is going to be Gaussian, let me ask you simple question, what is a finite sample distribution of sigma square hat?

Student: (Refer Time: 26:58)

What is it? If you have to if the observations fall from a Gaussian white noise process then you write sigma square hat sigma square e hat will follow chi square distribution, but that is finite sample distribution this is large sample distribution and we know that if we look at a chi square distribution. In fact, the answer is not complete if you would say sigma square hat is a chi square distribution with n minus one degree of freedom as n grows the chi square distribution tends to look like a Gaussian so that is one way at least verifying this statement that I have made in equation 14.

Because these are asymptotic distributions these are large sample distributions therefore, you should not be surprised that root ten sigma square e hat follows a Gaussian distribution where as we are bombarded all the time with the chi square distribution so much about MLE I have already talked about the weighted least square part very quickly.

(Refer Slide Time: 27:54)



ML and Bayesian estimators Reference

Remarks on MLE: ...contd.

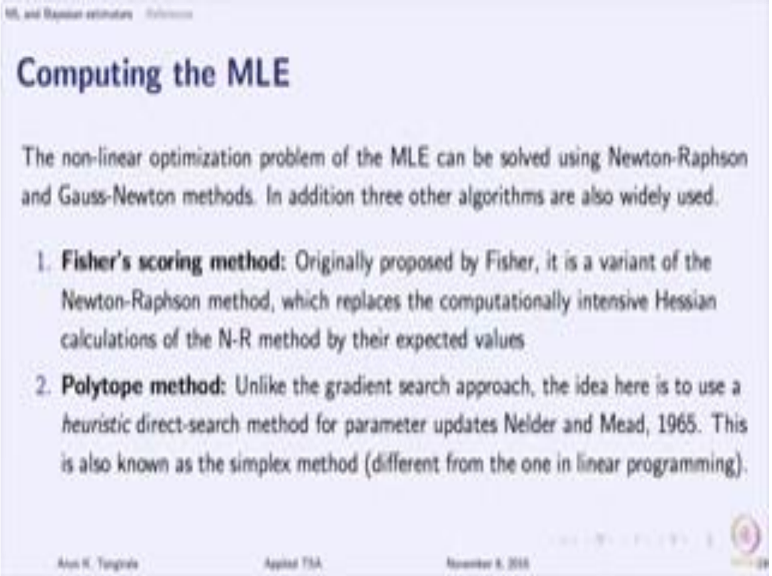
7. One of the advantages of ML formulation is that *heteroskedastic* errors, *i.e.*, $\text{var}(e[k]) = \sigma_k^2$ can be naturally accommodated. Assuming the knowledge of σ_k^2, the ML estimate of $\boldsymbol{\theta} = c$ is,

$$\hat{c}_{\text{MLE}} = \hat{c}_{\text{WLS}} = \frac{\displaystyle\sum_{k=0}^{N-1} \frac{y[k]}{\sigma_k^2}}{\displaystyle\sum_{k=0}^{N-1} \frac{1}{\sigma_k^2}} \qquad (15)$$

Thus, *the ML formulation also encompasses the WLS problem.*

Arun K. Tangirala Applied TSA November 8, 2016

(Refer Slide Time: 28:00)



ML and Bayesian estimators Reference

Computing the MLE

The non-linear optimization problem of the MLE can be solved using Newton-Raphson and Gauss-Newton methods. In addition three other algorithms are also widely used.

1. **Fisher's scoring method:** Originally proposed by Fisher, it is a variant of the Newton-Raphson method, which replaces the computationally intensive Hessian calculations of the N-R method by their expected values
2. **Polytope method:** Unlike the gradient search approach, the idea here is to use a *heuristic* direct-search method for parameter updates Nelder and Mead, 1965. This is also known as the simplex method (different from the one in linear programming).

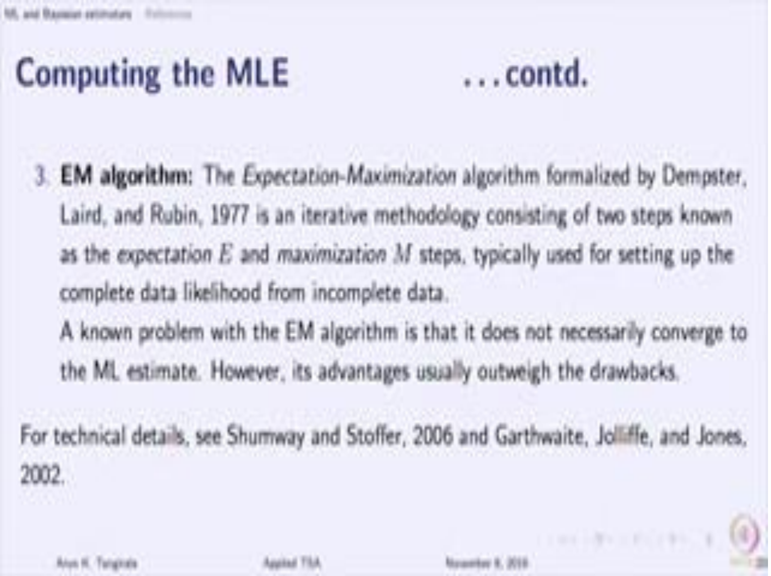
Arun K. Tangirala Applied TSA November 8, 2016

I just want to talk about computing MLE, I would not go in to any details, I just want to tell you that in general any MLE will lead to a non-linear optimization problem, there is no doubt about it.

However there are certain specialized algorithms. So, you can use your standard Newton Raphson method Gauss Newton method and so on to solve the non-linear optimization problem, but it turns out that there are some specialized algorithms that are written for MLE because of the nature of the likelihood function and the kind of problems that you expect to see some of these at least three of these are called the one is the Fisher's scoring method developed by fisher it is a variant of Newton Raphson method the other is called the polytope method which uses a gradient such kind of approach this is also incidentally called as simplex algorithm.

But this is not the same simplex that you encounter in l p in linear program.

(Refer Slide Time: 28:59).



ML and Bayesian estimators References

Computing the MLE ...contd.

3. **EM algorithm:** The *Expectation-Maximization* algorithm formalized by Dempster, Laird, and Rubin, 1977 is an iterative methodology consisting of two steps known as the *expectation* E and *maximization* M steps, typically used for setting up the complete data likelihood from incomplete data.
   A known problem with the EM algorithm is that it does not necessarily converge to the ML estimate. However, its advantages usually outweigh the drawbacks.

For technical details, see Shumway and Stoffer, 2006 and Garthwaite, Jolliffe, and Jones, 2002.

Arun K. Tangirala Applied TSA November 8, 2016

And the third is called e m or the expectation maximization algorithm which is very popular that is used in solving MLE again I would not go in to the details, but these are perhaps the reason reasons why there are routines specialized for MLE which are asking you to give the law to supply the likelihood function and they use one of these algorithms to get the optimum value for you.

I have given you the references you can look at the details if you are interest in this algorithms, but just to let you know and I am going to skip the asymptotic properties one point that I want to make in passing is a asymptotic efficiency we have talked about asymptotic normality asymptotic efficiency which is got to do with the variance of the estimator you can show that the maximum likelihood estimator results in the Cramer's laws lower bound as n goes to infinity and that is one of the reasons apart from consistency the asymptotic efficiency is the one that makes ml methods very attractive maximum likelihood methods very attractive for large samples.

For small samples it as poor performance and there is no assurance, but for large samples ml is good on and the one point that you want to keep in mind with maximum likelihood estimators is that it is applicable only if your pdf satisfies so called regularity conditions we have talked about that when we discussed Fisher's information one of the conditions that for the p d f to be called as regular is the range of the parameters value should not determined the range of possibility. So, range of outcomes and the classic example is the uniform distribution.

You should make sure that your p d f is regular by enlarge you will seen in the literature ml problems being set up for Gaussian or a few non Gaussian if nothing is known always maximum likelihood problems are solved for Gaussian because it easy to solve that is all it does not mean that it is a truth and we have talked about invariance and that kind of brings us close to the MLE. So, ML estimators large sample properties are very attractive it is consistent efficient it gives you Gaussian distribution it has invariants property very good, but remember that ml methods are not so great for finite samples, remember that.