Dynamic Data Assimilation
Prof. S Lakshmivaraham
School of Computer Science
Indian Institute of Technology, Madras

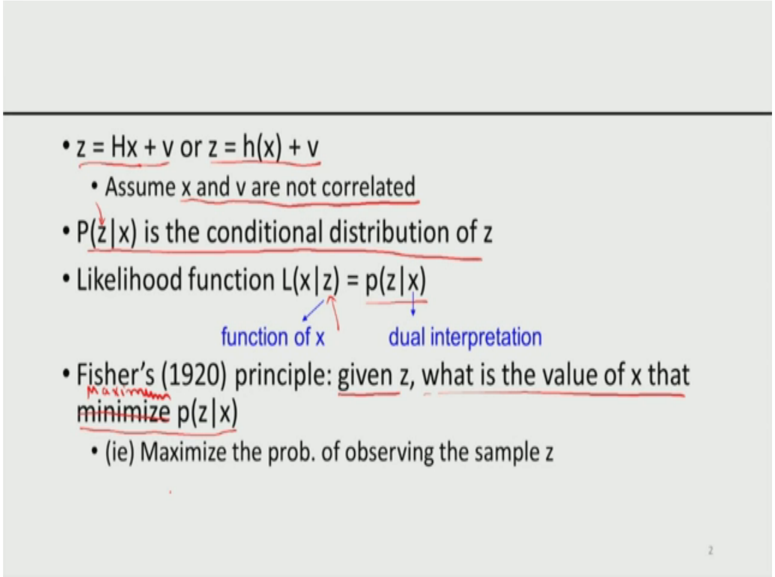
Lecture - 26
Maximum Likelihood Method

In module 6.1, we talked about two schools of estimation one is the Fisher school another is the Bayesian school. Fisher invented this notion of maximum likelihood estimation technique for point estimation of unknown constant vector or unknown scalar. We generally will not be using maximum likelihood estimation techniques in the parlance, in our discussion of data assimilation we have not used we largely depend on least squares, but I believe because of the underlying importance of this at least one should have a nodding understanding of what is maximum likelihood estimation technique. Once we talk about the some of the basic aspects of maximum likelihood techniques which belongs to the Fisher school then we will talk about the Bayesian estimation techniques in the next module 6.4.

So, 6.1, 2, 3 and 4 together contain an expose a of the basic idea of statistical least square principles illustrations some of the fundamental theorem, intrinsic properties of estimate, Gauss Markov theorem and there is also a couple of other fundamental theorem that comes out in the base in estimation. Once you understand these basic estimation you now know how to evaluate the goodness of the estimate once we do a data simulation procedure. That is the reason for us to be able to be, for us to include all these fundamental results from statistical estimation theory statistical estimation theory. These are no less important than tools from multivariate calculus tools from matrix theory tools from linear algebra and so on and so forth.

So, a quick expose of maximum likelihood method.

(Refer Slide Time: 02:23)



- $z = Hx + v$ or $z = h(x) + v$
  - Assume x and v are not correlated
- $P(z|x)$ is the conditional distribution of z
- Likelihood function $L(x|z) = p(z|x)$

function of x

dual interpretation

- Fisher's (1920) principle: given z, what is the value of x that maximum ~~minimize~~ $p(z|x)$
  - (ie) Maximize the prob. of observing the sample z

Let us z be equal to again Hx plus v or z is equal to h x plus v linear non-linear. Assume x could be a random v is always random u x is random, v and x are uncorrelated. In the Fishers case we assume x is an unknown constant. So, there is no prior distribution as in the case of Bayes x is an unknown constant given x z is random. So, given x z has a distribution that distribution is called the conditional distribution.

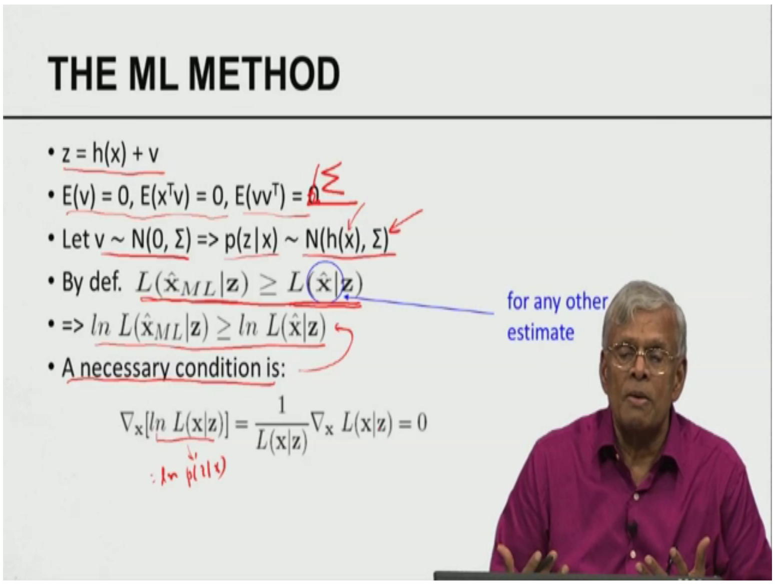
Conditional distribution essentially relates to the properties of the observation conditioned on the unknown nature plays a game with us. She picks a value of x and keeps it constant we do not know what x is, we are going to make our nature is teasing us we make measurements on the nature. So, the measurements are going to be providing information about x, but the measurements are random. So, z is a random vector it has an underlying distribution, but z the properties of z is conditioned on the value that mother nature has already chosen, but did not care to tell you our aim is to be able to uncover what mother nature had picked. So, the information of x is to be gleaned from z gained from the conditional distribution of z given x. So, that is what is the conditional distribution. So, conditional distribution is always the new information that arises because we are able to make observations about the system.

This p x of z has two ways of looking at it as a dual interpretation. If given x as a function of z is called a conditional distribution, but Fisher turned the table around for a given z. So, what is that he asked I am I have observed something observations given to you given z; what is the most likelihood value of x that mother nature must have picked that I observe z. So, let me talk about the differences here what is the conditional distribution the right hand side means given x mother nature has already picked, yes, but she did not tell you, but z exhibits randomness.

So, the randomness of z conditioned on the value of x is p z that is a function for a given x distribution over z that is the conditional distribution, but Fisher asked the turned the table around what is that he said yes I know mother nature has picked x, but she did not tell me I can, but I have the ability to make observations on mother nature which is going to give me a z I have got an z, z is ready to react. So, he asked the following question, what is the most probable value of the unknown x that the mother nature should have picked that will exhibit in my viewing the observation z. That is the difference the quantities are same, but if you turn the table around one is a function of x another function of z, one is called a likelihood function another is called the conditional distribution. That is the fundamental difference. And this difference is an enormous difference that led Fisher to be able to concoct a new class of methods called maximum likelihood method.

So, Fisher in 1920 Fishers principle given z what is the value of x that will minimize what is the value of z. I think there is maximize, maximize sorry it is the maximum likelihood I have said it correctly that is maximize the probability of observing the sampled z given x that is the basic idea here .

(Refer Slide Time: 06:58)



THE ML METHOD

- $z = h(x) + v$
- $E(v) = 0,\ E(x^T v) = 0,\ E(vv^T) = \cancel{0}\ \Sigma$
- Let $v \sim N(0, \Sigma) \Rightarrow p(z|x) \sim N(h(x), \Sigma)$
- By def. $L(\hat{\mathbf{x}}_{ML}|\mathbf{z}) \geq L(\hat{\mathbf{x}}|\mathbf{z})$
- $\Rightarrow ln\ L(\hat{\mathbf{x}}_{ML}|\mathbf{z}) \geq ln\ L(\hat{\mathbf{x}}|\mathbf{z})$
- A necessary condition is:

$$\nabla_{\mathbf{x}}[\underbrace{ln\ L(\mathbf{x}|\mathbf{z})}_{:\ln p(z|x)}] = \frac{1}{L(\mathbf{x}|\mathbf{z})}\nabla_{\mathbf{x}}\ L(\mathbf{x}|\mathbf{z}) = 0$$

for any other estimate

So, the maximum likelihood method. So, we that is the underlying principle of Fishers strategy maximum likelihood ML method, without loss of generality I can start with the with the with a non-linear observation with a non-linear case let z be equal to h x plus v, v has this property means 0 x and v uncorrelated the I am sorry this must be R, I will correct that, v is mean 0 x and v are uncorrelated the covariance of v is R. I am sorry in this particular case I is not R, I am going to change this I am sorry in here I am assuming this is sigma. Therefore, v is normal with 0 and sigma as the covariance. Therefore, you h of x is deterministic v is random, z is random. If v has a distribution 0 mean and variance covariance sigma z has a distribution whose mean is the h of x and covariance sigma, that is the conditional distribution. So, one of the basic canons of Fisher's theory is that I should know the conditional distribution in its exact form.

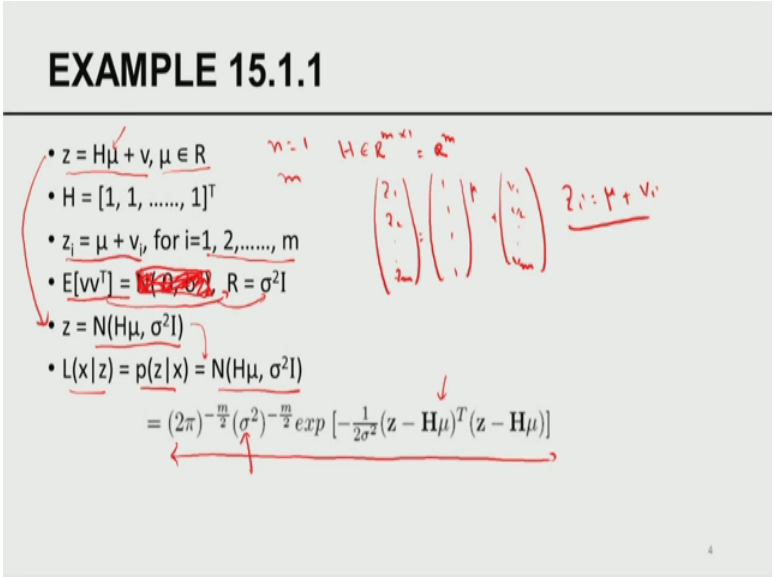
So, given x I should know the distribution of observation z a are conditioned on x. So, this is the conditional distribution that Fishers method rests on. So, we need to have that distribution up.

So, this distribution has x and x is unknown so what is that we are looking for. Now please remember I can relate p to L. So, looking at L as the likelihood what am I looking for let x hat be any estimate let x hat ML be the maximum likelihood estimate and how do I define the maximum likelihood estimate. Maximum likelihood estimate is an estimate that maximizes the likelihood of observing z given that estimate compared to any other estimate. So, the likelihood of the probability the probability observing. So, what does the left hand side say. The probability of observing a sample z when you set the parameter to be x hat ML is larger than the probability of observing z for any other estimate x hat, among all the estimate the maximum likelihood estimate gives you the most probable value of the unknown based on which you will observe what is being observed.

So, this inequality essentially underlies the definition of maximum likelihood estimation technique. In other words I am interested in a x hat ML that satisfies this property please remember L and p the conditional distributions are related as we have seen in the previous slide. If L must be greater than this logarithm is an increasing function. So, if I took the logarithm of both sides the inequality must be preserved. So, the natural logarithm of the likelihood of the left hand side must be greater than equal to the natural logarithm of the likelihood on the right hand side.

I am not going to derive this Roth's book gives a beautiful definition the book by Melsa and Cohn gives a very good very good derivation of this a necessary condition for this to happen is that the gradient of the log of the likelihood function, please understand the log of the likelihood function is essentially p of z of x. So, this is also equal to log of p z of x. Log of p z of x I have to compute the derivative with respect to x that derivative is given by 1 over L times the gradient with respect to x that must be 0 that comes from the maximization property of the likelihood function. This necessary condition is extremely simple to be able to look at why this is necessary for this inequality holder good.

(Refer Slide Time: 11:39)



EXAMPLE 15.1.1

- $z = H\mu + v, \mu \in R$
- $H = [1, 1, \ldots\ldots, 1]^T$
- $z_i = \mu + v_i$, for i=1, 2,......, m
- $E[vv^T] =$ ~~$N(0, \sigma^2)$~~, $R = \sigma^2 I$
- $z = N(H\mu, \sigma^2 I)$
- $L(x|z) = p(z|x) = N(H\mu, \sigma^2 I)$

$$= (2\pi)^{-\frac{m}{2}}(\sigma^2)^{-\frac{m}{2}} exp\,[-\frac{1}{2\sigma^2}(z - H\mu)^T(z - H\mu)]$$

$n = 1$ $\quad H \in R^{m \times 1} = R^m$

m

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \qquad z_i = \mu + v_i$$

4

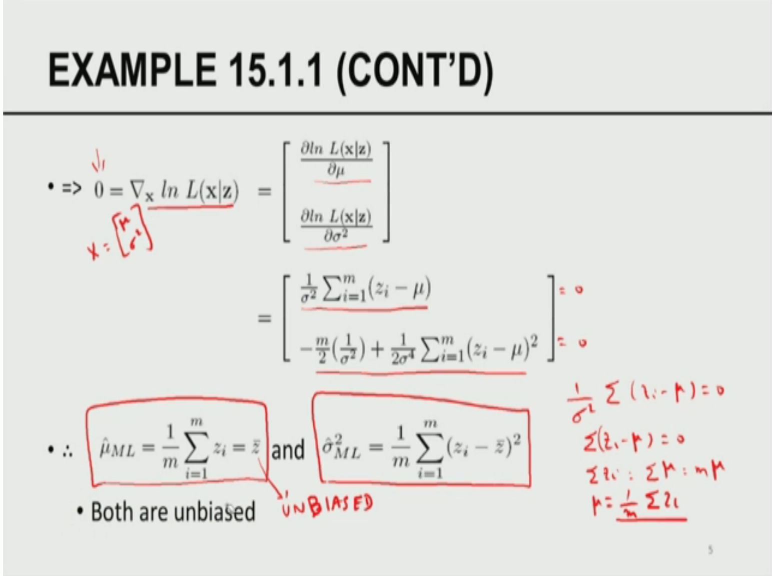
Now, I am going to illustrate this fundamental principle using a very simple example. So, let us pretend I want to be able to estimate an unknown mu, mu is a constant, but mu is not observable z is observable z is equal to h times mu plus v. In this case I am I am assuming mu is not even a vector, mu could be just a real number. So, I am assuming n is equal to 1, I have m observations. So, H is m by 1 it is simply a vector which is simply a vector, I am assuming h to be all ones therefore, I have z 1, z 2, z m is equal to 1 1 1 1 times mu plus v 1, v 2, v m. So, each z i is equal to mu plus vi that is the observation there are m such observation. So, z i is equal to mu plus v i for i is equal to 1 to m.

I am going to assume my v is such that my v, I think that is I should have this is not right, this is equal to R equal to this. So, we go from here to here to here, the covariance of v is equal to R which is equal to sigma square i. So, H mu is a constant v is a random vector. So, if I add a constant to a random vector its essentially shifts the mean. So, the distribution of z is given by normal with mu h mu sigma square i. So, everything is right, but this should not be here I am sorry I will correct this later, I hope that is clear now.

So, the likelihood function is given by this, this likelihood function is given by this. So, I know the functional form the functional form of the like dual function is normal with H mu is a mu sigma square I as the variance. This is the explicit form of the function please understand the variable to be estimated is not x s we call it mu because it is an unknown constant. So, this is a multivariate Gaussian distribution, this is the expression for the multivariate Gaussian distribution, this is the function which is this is called the likelihood function. When considered as if given mu when consider the function of z is called a conditional distribution function given z consider the function of mu is called the likelihood function.

So, there are two variables mu and z, mu and z. So, whether you are going to consider this a conditional distribution or a likelihood function the maximum likelihood estimate tries to find the optimal value for mu optimal in the sense of trying to maximize this distribution, this likelihood function. I hope that is clear now.

(Refer Slide Time: 14:52)



EXAMPLE 15.1.1 (CONT'D)

- => $\overset{\Downarrow}{0} = \nabla_{\mathbf{x}} \underline{ln\ L(\mathbf{x}|\mathbf{z})} = \begin{bmatrix} \underline{\frac{\partial ln\ L(\mathbf{x}|\mathbf{z})}{\partial \mu}} \\ \underline{\frac{\partial ln\ L(\mathbf{x}|\mathbf{z})}{\partial \sigma^2}} \end{bmatrix}$

$x = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$

$$= \begin{bmatrix} \underline{\frac{1}{\sigma^2}\sum_{i=1}^{m}(z_i - \mu)} \\ \underline{-\frac{m}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{2\sigma^4}\sum_{i=1}^{m}(z_i - \mu)^2} \end{bmatrix} \begin{matrix} = 0 \\ = 0 \end{matrix}$$

- ∴ $\boxed{\hat{\mu}_{ML} = \frac{1}{m}\sum_{i=1}^{m} z_i = \bar{z}}$ and $\boxed{\hat{\sigma}^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(z_i - \bar{z})^2}$
  - Both are unbiased

UNBIASED

$\frac{1}{\sigma^2}\sum(z_i - \mu) = 0$

$\sum(z_i - \mu) = 0$

$\sum z_i = \sum \mu = m\mu$

$\underline{\mu = \frac{1}{m}\sum z_i}$

5

So, I can compute the derivative. So, let us go back now. So, given this I can solve two problems one is to be able to estimate mu sorry, one is to be able to estimate mu I can also formulate this problem as one of estimating sigma square. Please go back mu is known and the noise it covariance is not known. So, there are a number of estimation problem associated with it. So, if you are interested in estimating mu you consider the derivative of the likelihood function with respect to mu whichever variable you are interested in estimating you have to, we are interested in maximizing the likelihood with respect to that particular parameter. So, if you are interested in estimating mu you have to make the log of the likelihood function maximum with respect to mu; that means that derivative of the log of the likelihood function with respect to mu must be 0 at the maximum.

If you are trying to estimate sigma square again the same principle you have to compute the derivative of the max derivative of the likelihood function of the sigma square at the maximum, I am sorry at the maximum the derivative must be 0 standard principles and optimization. So, you can see as early as 1920 he has mixed several ideas condition the conditional distribution interpreters likelihood function, maximizing likelihood function, maximization as an optimization problem. So, you can see the role of optimization embedded in estimation theory least squares least, best maximum likelihood, maximum, best. So, optimization theory and estimation theory are inseparable every estimation that every estimation problem.

We are going to solve we are going to solve it as an optimization problem that is why if data simulation is estimation the estimation is posed as an optimization problem you can see the intrinsic interest in optimization intrinsic role played by optimization in data simulation. So, I could compute, I am killing two birds in one stroke I am computing, so I am assuming the unknowns are to be estimated x are equal to a vector mu and sigma square. So, I am computing gradiant with respect to these two values. Please remember L is a scalar function if I am going to differentiate a scalar function with respect to vector variable the gradient is a vector the vector has two components given the expression for the likelihood function as given at the bottom of page 4, I could compute these derivative explicitly the derivatives.

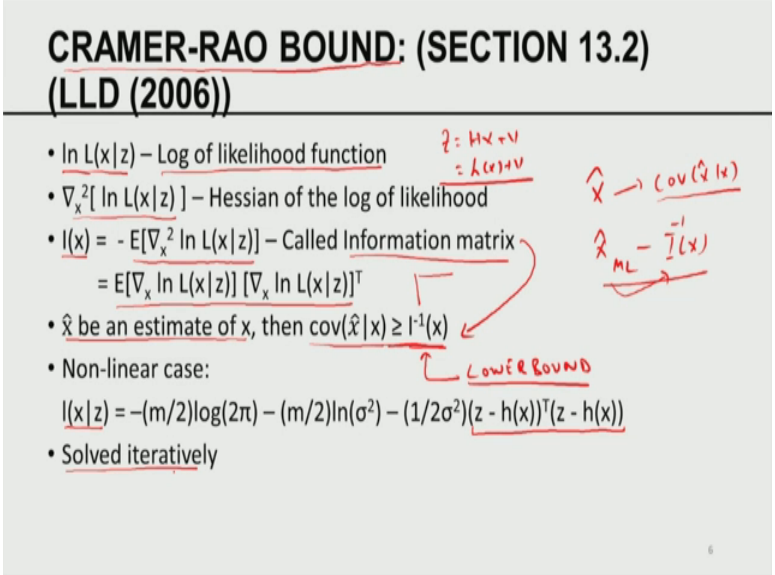
The derivatives the derivative of L with respect to mu is given by this, the derivative of the derivative of the log of L with respect to mu is given by this, the derivative of the log of L with respect to sigma square is given by this. These are interesting exercise I would like you, I would strongly urge you to do this exercise. Now please understand at the maximum I am these derivative must vanish so it must be equal to 0; that means, this component the first component must be 0, the second component must be 0, the first component being 0 gives raise to a function form of the estimator. Please look at this now if this must be 0; that means, that m.

So, let us look at this now, 1 over sigma square summation z i minus mu must be equal to 0, this is a fraction a fraction is 0 only when the numerator is 0. Therefore, if the numerator is to be 0 summation z i minus mu must be equal to 0 this essentially tells you summation z i must be equal to summation mu the summation is over i, i running from 1 to m. So, this is equal to m times mu therefore, mu must be equal to 1 over m times summation z i. What is that? That the average value.

Tada, we have now rediscovered a formula that we already knew what is that we knew from least square estimation when we did a statistical estimation theory average is the best least square estimate, average is also best in the sense of maximum likelihood. So, mu hat ML is the maximum likelihood estimate that is also z hat. So, average has very beautiful property of being optimal simultaneously from the least square sense from the maximum likelihood sense.

Again can I can equating the second term to 0 and simplifying you can readily get an estimate for the variance. So, sigma square hat ML is the estimate for the maximum likelihood estimate for the variance this is the expression for the variance and I think this is this expression is not correct. This is not back, this is unbiased, this is unbiased and I think I am sorry I should be able to erase this, that is right.

(Refer Slide Time: 21:05)



CRAMER-RAO BOUND: (SECTION 13.2) (LLD (2006))

- $\ln L(x|z)$ – Log of likelihood function
- $\nabla_x^2[\ \ln L(x|z)\ ]$ – Hessian of the log of likelihood
- $I(x) = -E[\nabla_x^2 \ln L(x|z)]$ – Called Information matrix

  $= E[\nabla_x \ln L(x|z)]\,[\nabla_x \ln L(x|z)]^T$
- $\hat{x}$ be an estimate of x, then $\mathrm{cov}(\hat{x}|x) \geq I^{-1}(x)$
- Non-linear case:

  $l(x|z) = -(m/2)\log(2\pi) - (m/2)\ln(\sigma^2) - (1/2\sigma^2)(z - h(x))^T(z - h(x))$
- Solved iteratively

$z = Hx + v$

$= h(x) + v$

$\hat{x} \rightarrow \mathrm{cov}(\hat{x}|x)$

$\hat{x}_{ML} - I^{-1}(x)$

LOWER BOUND

6

So, both the estimates I had given you here. Now I would like to talk about another related property. So, you that is what is called Cramer-Rao Bound in the sense of optimality intrinsic optimality of the maximum likelihood estimate. This is the likelihood function this is the log of the likelihood function, so the likelihood function you know what the Hessian of the log of the likelihood function with respect to x, we can compute that Hessian exist. The yellow facts is called the information matrix the information matrix is essentially negative of the expected value of the Hessian of the log of the likelihood, it can be shown that this information matrix is also equal to the outer product. Now look at this, now L the log of the likelihood the gradient with respect to this, that is a vector the transpose is the outer product, the expected value the outer product, this is the expected value the matrix.

So, what is the theory here? The theory here is that the outer product matrix the outer product matrix and the Hessian matrix are related. So, what is the fundamental result? Again there is a ton of theory goes with it, but I want to tell you to expose you to some of the existing results, x hat be an estimate of x then the covariance. So, if x hat is any other estimate if I have a maximum likelihood estimate information matrix, what is the information matrix? Information matrix is the reciprocal of the covariance matrix. So, inverse of the invariant of information matrix on the right hand side covariance of estimate of any estimate. So, what does this inequality says? This essentially tells you the covariance of the maximum likelihood estimate is always less than or equal to covariance of any other estimate that is what this inequality is, this inequality is very fundamental. So, what does it mean? I have two estimates one is x hat, another is x hat ML; the I L I am sorry L inverse this I x is the information matrix please understand information matrix is the inverse of the covariance matrix. So, I inverse x is the covariance of the estimate of this.

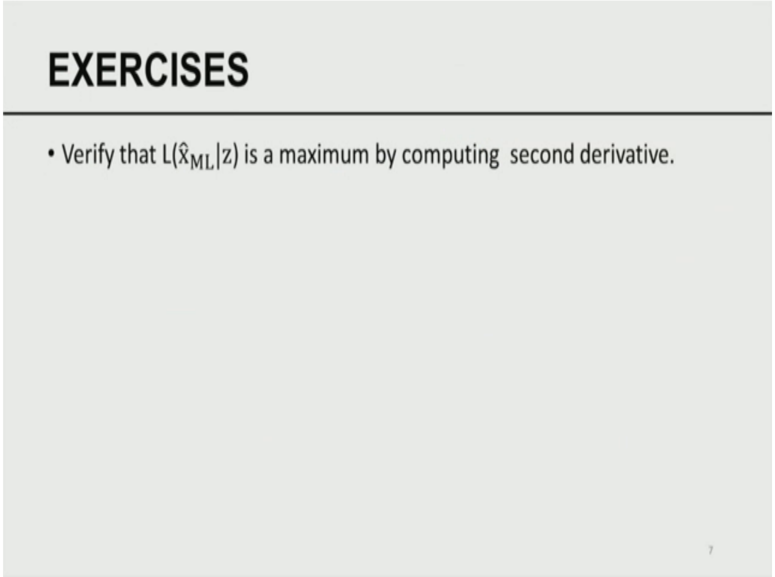
The covariance of the estimate of this is covariance of x hat given x. So, this is the covariance of any other estimate, this is the covariance of maximal likelihood estimate. One of the fundamental results is that the covariance of any other estimate is greater than or equal to the covariance of the covariance of the maximal likelihood estimate this inequality is called Cramer-Rao inequality or Cramer-Rao bound. What is the bound? The estimate the covariance of the maximum likelihood estimate is the lower bound on the covariance of any other estimate. So, this is the lower bound. That is the least value and this bound is attained by the maximal likelihood estimate. So, what does this mean? Maximum likelihood estimate gives the best estimate in the sense the covariance of estimate resulting from the maximal likelihood estimation is the smallest among the possible values that estimate, the covariance estimate can take, covariance estimate can take.

So, when we are dealing with when we are dealing with linear functions of the observation z is equal to H of x plus v when you are dealing with non-linear functions of observation h is equal to this. The linear observations are continuing or computationally a simpler in the case of a non-linear observations we can see the non-linear function comes in here in this case the log of the likelihood function, the log of the likelihood function is a non-linear function maximizing this is not easy, we cannot get explicit expressions easily and we cannot we may not be able to solve for the 0 of the gradient of these. So, we may have to find the maximum in the non-linear case only iteratively.

So, so log of the likelihood function computing the derivative of the log of the likelihood function, equating the derivatives of 0, solving the resulting equation, the solution of the resulting equation gives rise to the optimal maximum likelihood estimates, all these processes are simpler in this context when there is linear observation. All these processes are little bit complex in the case of non-linear observations. In the non-linear observations all the methodology the methodology still holds good except that the solution process have to be applied only numerically iteratively, it gives rise to iterative optimization.

Of course, we have already provided methods for iterative optimization, namely gradient method, conjugate grading method. So, we can use one of the very well known techniques that we already covered to do the maximization of this log likelihood function. Therefore, the theory applies to both linear as well as non-linear functions of the state. We would like to end this talk by asking you to do a homework problem of trying to compute the Hessian of the log likelihood function and computing the derivative.

(Refer Slide Time: 27:30)

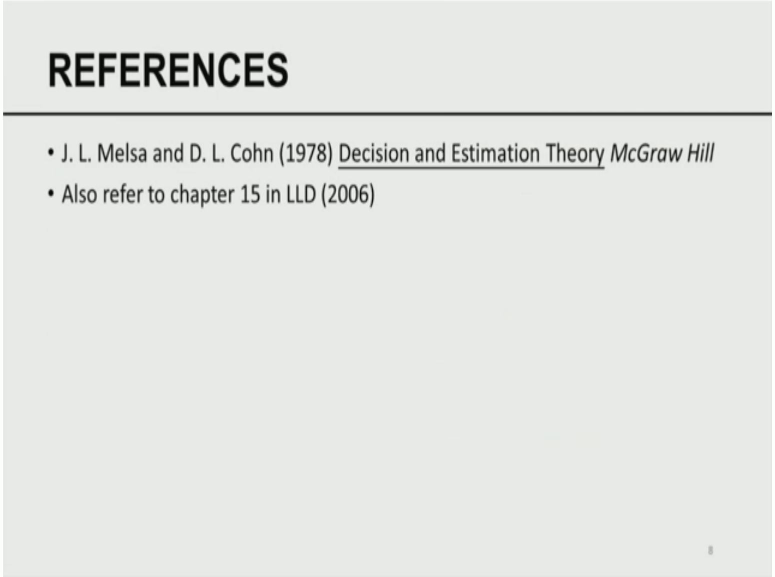


EXERCISES

- Verify that $L(\hat{x}_{ML}|z)$ is a maximum by computing second derivative.

So, computing the first and second derivative of the log likelihood function is a homework problem, is a homework problem.

(Refer Slide Time: 27:47)



REFERENCES

- J. L. Melsa and D. L. Cohn (1978) Decision and Estimation Theory *McGraw Hill*
- Also refer to chapter 15 in LLD (2006)

And again my favorite coverage of this is in Melsa and Cohn, 78. We also cover this in chapter 15.

This the cover a broad and a quick overview of maximum likelihood estimates. There are very few papers in the data assimilation literature relating to maximum likelihood, I will not say there is none there are a couple of them. They are done within the context of

Kalman filter and under relation to other problems. Even though in our illustration we will not invoke the maximum likelihood estimation we are largely going to be concerned with least squares, it is better to know what are the things out there and what are the alternate ways of thinking about the problems. That the reason I am trying to introduce to you some of these techniques, so that it will open our windows and our eyes to other related areas in estimation theory.

Thank you.