

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

Lecture – 27
Bayesian Estimation

In this lecture we are going to be talking about the elements of Bayesian estimation. Bayesian estimation is the centerpiece of stochastic methods for data assimilation. We have already seen stochastically squares, we have already seen the notion of maximum likelihood, we did not elaborate on these techniques too much because it will take us too far from the main goal of dealing with data assimilation. Yet the fundamental principles of status statistical estimation are the foundations on which statistical methods for data assimilations are built that is why a good nodding under a good understanding of the fundamental principles of statistical estimation is very useful, and in my view is also necessary to be able to appreciate all the nuances relating to the techniques for dynamic stochastic dynamic data assimilation.

In the process we are going to start with the description of the Bayesian estimation schemes or Bayesian estimation methods. So, I am going to develop the Bayesian framework.

(Refer Slide Time: 01:40)

BAYESIAN FRAMEWORK

- $x \in \mathbb{R}^n$ – unknown, random – $p(x)$ is prior
- $z \in \mathbb{R}^m$ – observation about x – $p(z|x)$ – conditional distribution
- $\hat{x} = \Phi(z)$ be an estimate
 - Define error $\tilde{x} = x - \hat{x}$
- Cost function $c: \mathbb{R}^n \rightarrow \mathbb{R}$, $c(\tilde{x})$ is called the cost associated with error
- Properties:
 - $c(0) = 0$
 - $c(a) \leq c(b)$ if $\|a\| < \|b\|$]

$$f(x) = (z - Hx)^T (z - Hx)$$

\downarrow
 $\mathbb{R}^n \rightarrow \mathbb{R}$

Please remember Bayesian framework is an alternative to the Fishers framework. In the Bayesian framework let x be an unknown vector is random. This random unknown vector is supposed to follow a probability distribution called $p(x)$ which is known as the prior distribution. Prior distribution essentially is a summary of our understanding about the unknown so what is key here. I only know the probability distribution from which x is drawn I just do not know what is the actual realization of x that mother nature has picked. So, when you say x is unknown we do not know the actual value or the realization of x , but we know where the x come from the x comes from the probability distribution called prior distribution x is drawn from that probability distribution.

z again R^m is an observation about x z contains information about the chosen x . So, given x z has a conditional distribution which is probability of z given x . Let \hat{x} be equal to $\phi(z)$. So, ϕ is an estimator x as z is the data the estimator operates on the data give us an estimate we have already seen the notion of an estimator and an estimate. So, if \hat{x} is the estimate, if x is the value that mother nature has picked from prior distribution the difference between what the mother nature has picked and whatever estimate, estimate is something we create from the knowledge of the observation about x . More often they not \hat{x} and x may not be equal. So, there is an error \tilde{x} is the error in the estimate.

We are going to now associate a cost function c , c is a mapping c is a functional it maps R^n to R , the cost function is defined over the set of all errors, $c(\hat{x})$ is called the cost associated with the error. This is nothing new we have already talked about $f(x)$ within the context of static deterministic inverse problems. In that case we said z is given H of x is the model predicted observation z minus H of x is called the residue, residue error they are similar connotations. So, when I talked about z minus Hx inverse z minus Hx that is the essentially sum of squared errors that is a cost function, f is a cost function that maps R^n to R . So, $f(x)$ is simply sum of the squared error sum of the square of the residuals are errors.

Likewise in this case $c(\tilde{x})$ is the cost associated with the error \tilde{x} . We would like to impose some special conditions on this functional c . If I know z is equal to H of x and if I have $f(x)$ is equal to z minus H of x transpose times z minus H of x we already know it is quadratic. So, this $f(x)$ is quadratic we already know quadratic functions have unique minimum they have a convexity property. So, we did not have to worry too

much about further constraining f of x because f of x automatically by virtue of the definition in that case had all the quad properties, but in here we may not know right now what the form of the error is I am simply trying to define a function that captures the cost associated with the error therefore, to be consistent with what can be done we would like to impose conditions which are as follows, if the error is 0 the cost is 0 that makes lot of sense, much like in $f x$ if z is equal in H of x f of x is 0.

So, you can think of 0 as being the minimum of the cost then there is no error there should be no cost associated with it, there is no penalty associated with it. Secondly, if a is a vector b is a vector, if the norm of a is less than the norm of b c of a is less than c of b , norm of a less than norm of b means the length of the vector a is less than the length of the vector b . So, if a is one error, b is another error if the norm of a is less than norm of b means a is smaller the two errors; if a is smaller of the two errors the cost associated with a must be less than or equal to the cost associated with b what does it tell you it tells you some kind of a monotonicity property. The cost function increases with the length of the error vector which is also a very nice and a desirable property. So, you can see the cost function is such that it is 0 at the origin when you go away from the origin in the space that represents the error, when you go away the length of the error increases as the length of the error increases the cost function does not decrease it either increases or remains the same. So, that is the condition. So, you can see we are trying to already develop a bowl like shape for the cost function c .

(Refer Slide Time: 08:27)

EXAMPLES OF COST FUNCTION $C(\cdot)$

- Sum of squared error

$$c(\tilde{x}) = (\tilde{x} - \hat{x})^T (\tilde{x} - \hat{x}) = \tilde{x}^T \tilde{x} \quad \checkmark$$

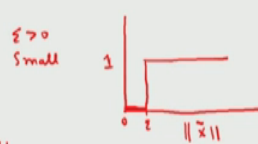
- Weighted sum of squared error

$$c(\tilde{x}) = \tilde{x}^T W \tilde{x} = (\tilde{x} - \hat{x})^T W (\tilde{x} - \hat{x}) \quad \checkmark$$

$$= \|\tilde{x} - \hat{x}\|_W^2$$

- Uniform cost

$$c(\tilde{x}) = \begin{cases} 0, & \text{if } \|\tilde{x}\| \leq \epsilon \\ 1, & \text{otherwise} \end{cases}$$



- Absolute error (x is a scalar)

$$c(\tilde{x}) = |(\tilde{x} - \hat{x})| \quad \checkmark$$

Why is that we are looking for such a bowl shape thing cost, in the optimization parlance cost we would like to be able to minimize. If we would like to be able to minimize the function should be endowed with the unique minimum that is why we would like to be able to require c to satisfy some simple condition that will guarantee the behavior of the function around 0, where 0 my estimate being equal to the unknown my estimate is equal to not the unknown; the estimate of the unknown is equal to the true value. So, here are some examples of cost function. So, x minus \hat{x} is the error x minus \hat{x} transpose x minus \hat{x} is also equal to $\tilde{x}^T \tilde{x}$. So, this is the quadratic function. So, this is called sum of squared error that is one way to be able to concoct the form of c . So, that is a simple quadratic form that we are already used to.

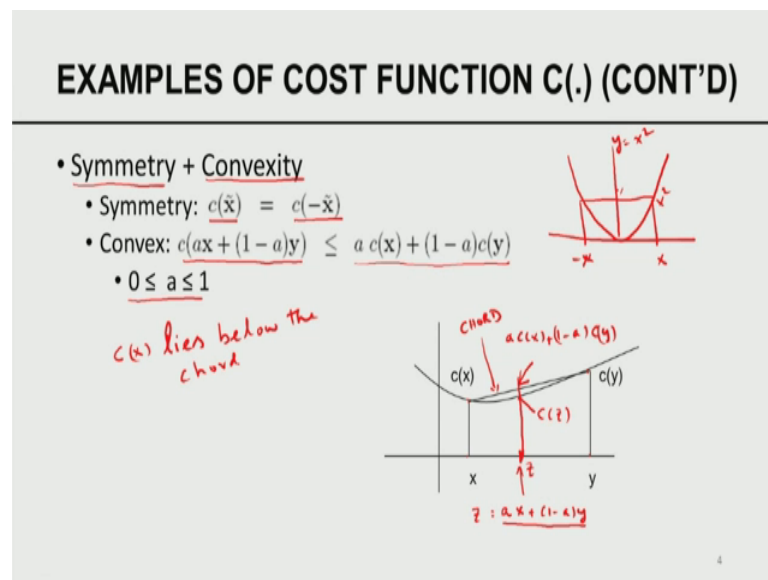
The next one is weighted sum of squared errors the first one there is no weight in the second one we are adding a weight W . So, this is simply weighted sum of squared errors which is sum, which is also called the energy norm the square of the energy norm of the error. So, we are very familiar with the first one as well as the second one.

There are couple of other ways in which one can design the cost function this must be \tilde{x} . So, if \tilde{x} is such that its norm is less than or equal to ϵ , ϵ is sum. So, ϵ is greater than 0 ϵ is pre specified and ϵ is small, given an ϵ greater than 0, but small if \tilde{x} is the error if the norm of the error is less than the chosen ϵ we will say the cost is 0. If the norm of the error is greater than ϵ we will say the cost is 1. So, how does this one look like functional pictorially? Let this be 0, let this be 0, so in a small neighborhood around the origin there are two ϵ . So, let me try to, let me try to I think up I want to refine this picture a little bit better sorry.

So, let us try to give a description of this. So, let us try to see let this be the norm of \tilde{x} , if the norm of \tilde{x} this let this be 0, let this be ϵ , if the norm of \tilde{x} lies between 0 and ϵ the cost is 0 otherwise it jumps to 1, it jumps to 1. So, this is how the cost function this is called uniform cost function. This is in the case where the error is a vector. The last one is suppose we are dealing with a scalar case x is a scalar \hat{x} is a scalar. So, \tilde{x} is a scalar. In this case one can think of what is called an absolute error, in this case this is simply the absolute value of \tilde{x} . So, the cost could be simply the absolute value of the scalar variable which happens to be the error in the estimate.

So, there are four, there are four different ways of picking the cost function each one of the way in which we pick has a different interpretation of the optimal estimate. All these things are used in decision making, two of them are already known to us quadratic the uniform cost, an absolute error cost are add on versions of the cost function.

(Refer Slide Time: 13:30)



So, what are the general forms of functions that will satisfy or that could be used as a cost function? One possibility is to require c symmetric, c symmetric and also c is convex. What is symmetry means? It is symmetric with respect to the vertical axis. So, for example, if you say x square x square is a symmetric function minus x and x have the same value, have the same value which is x square.

So, x square is a symmetric function therefore, if y is equal to x square there is an example of a symmetric function and that is exactly what this symmetry refers to the cost function for x tilde is the same as cost function for minus x tilde symmetry you can readily see. So, if x tilde is a vector if we change the sign of each component of that you get a point which is the reflection of that and c has the same value at these two points x tilde and x minus tilde.

Convexity, what does convexity means? The picture will help you your function c is said to be convex if I consider two points x and y it has a value $c(x)$ and $c(y)$, if I consider a constant a if I took a times x sorry if I consider a point if I consider a point in between x and y there exists an a in the interval 0 to 1 such that this point can be written as a times

x plus 1 minus a times 1 . This is called convex combination of the two endpoints. In fact, every point in the line segment from x to y can be obtained in this particular form by changing a in the interval 0 to 1 .

So, let us take let us draw a vertical line this is the point. So, let us call this point as z , z is the point which is equal to a x plus 1 minus a times y for sum a in the interval 0 to 1 . So, let us call this z this point the value of this point is c of z , but the value of this point is equal to a times c x plus 1 minus a times c y and what is that, it is it lies on the chord that joins the point c x to c y . So, this is point, this is the chord this is the point on the curve. So, what does this inequality says this inequality says? The value the function or intermediary point in between x and y is always less than or equal to the value along the chord; that means, the function lies below the chord, the function c x lies below the chord. Such a function is called convex functions. We have already alluded to convexity when we did optimization I am just trying to reinforce and remind you of some of the properties.

So, in general what are the desirable properties of cost function, we would expect the cost function to be symmetric what does it mean. The valley penalty or the costs for x and minus x are the same and the function c is naturally convex. Quadratic function especially the parabola is an example of a convex function convex functions generally have unique minimum therefore, some of these conditions essentially help you to be able to guarantee that the cost function is well defined it is the unique minimum. The guarantee of unique minimum is important because when we are trying to use the cost function in our estimation we would like to be able to get best estimates best in the sense of minimizing some associated cost function. So, we want to be able to choose our c appropriately so that we can make meaningful decisions when it comes to, when it comes to deciding algorithms for appropriate ways to estimate they are not.

(Refer Slide Time: 18:59)

STATEMENT OF THE PROBLEM

- Given $p(x)$, $p(z|x)$, z and $c(\cdot)$, goal is to minimize Bayes cost function:

$$B(\hat{x}) = E[c(\tilde{x})] = \int_{\mathcal{R}^m} \int_{\mathcal{R}^n} c(x - \hat{x}) p(x, z) dx dz$$

Since $p(x, z) = p(z|x)p(x) = p(x|z)p(z)$,

$$B(\hat{x}) = \int_{\mathcal{R}^m} B(\hat{x}|z) p(z) dz$$

where $B(\hat{x}|z) = \int_{\mathcal{R}^n} c(x - \hat{x}) p(x|z) dx$

- Since $p(z) \geq 0$, minimizing $B(\hat{x}|z)$, minimizes $B(\hat{x})$

Handwritten notes on the slide:
 - Red arrows point from $p(x)$ and $p(z|x)$ to $p(x, z)$ in the first equation, with the label "Joint distribution".
 - A red arrow points from $p(x, z)$ to $p(x|z)p(z)$ in the second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the tenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eleventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twelfth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirteenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fourteenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifteenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixteenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventeenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighteenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the nineteenth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twentieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the twenty-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirtieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the thirty-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fortieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the forty-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fiftieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the fifty-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixtieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the sixty-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the seventy-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eightieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the eighty-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninetieth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-first equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-second equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-third equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-fourth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-fifth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-sixth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-seventh equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-eighth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the ninety-ninth equation.
 - A red arrow points from $p(x|z)p(z)$ to the integral in the one hundredth equation.

So, now I am going to state the general Bayesian problem. We already know we are given the prior this is the conditional distribution, we are also given a cost function and the observation look at this now or what are all the given data. The prior, the conditional distribution the observation and the choice of the c function we already know c function can be chosen in one of many ways. I am going to now concoct what is called the base cost function $b(x \hat{z})$. Please remember $x \hat{z}$ is the estimate so this is the cost associated with estimate $x \hat{z}$ for Bayesian this is going to be equal to the expected value of c of $x \tilde{z}$, given $x \hat{z}$ there is an associated $x \tilde{z}$ the error associated with it. So, given $x \hat{z}$ I can compute $x \tilde{z}$ I have evaluated the cost we already know the estimate is a random variable $x \tilde{z}$ is a random variable therefore, c of $x \tilde{z}$ is also random I am considering e the expected value of the cost function. This expected value of the cost function can now be written explicitly by this integral.

Please remember c of $x \tilde{z}$ is equal to c of x minus $x \hat{z}$, $x \hat{z}$ depends on z depends on z observation x has its own prior information. So, there are two sources of randomness the prior one from the prior and from the observation. So, I would like to be able to integrate this cost function with respect to the joint distribution between x and z , x has prior z has a conditional distribution x is a random variable z is the random variable if I have two random variables are random vectors I can consider the marginal distribution as well as I can consider the joint distribution. Here this base cost function is simply expected value of the costs associated with the error in the estimation it is simply

given by this double integral one with respect to x another with respect to z where the probability is the joint distribution of x and z .

Please recall the joint distribution $p(x, z)$ can be written in this form using conditional probability, it can also be written in this form both ways are meaningful it is using these two we generally derive what is called the Bayes rule in elementary course on statistics. So, if I now replace the joint distribution by the product form the one of the product form I get a new formulation for the Bayes cost function. So, $b(x, \hat{x})$ now becomes if I substitute this in here I can absorb one of the integration in another quantity which is called Bayesian cost associated with the estimate given in the observation, please understand please recall the estimate is a function of the observation. So, I can rewrite this $p(z)$ as $p(x, z)$, $p(z)$ then I can associate this $p(x, z)$ with this. So, this integral with respect to x which is the internal integral I am going to denote it by $b(x, \hat{x} | z)$ and then if I plug this in here I get a new form I get a new form therefore, $b(x, \hat{x})$ has this particular form and this form is very useful and this is one of the forms that we are going to concentrate on.

Now, I am interested in minimizing. So, please remember minimize the Bayes bayes cost. So, I would like to be able to minimize $b(x, \hat{x})$, $b(x, \hat{x})$ from star is equal to the integral of $b(x, \hat{x})$ conditioner and z times $p(z)$, $p(z)$ in the region of interest is always greater than or equal to 0 it is that is a probability density function. So, if I want to be able to minimize $b(x, \hat{x})$ it is enough to be able to minimize $b(x, \hat{x} | z)$ because if I minimize this double star, I am taking your linear combinations, I am taking the weighted linear combinations of $b(x, \hat{x} | z)$ are given z with $b(z)$. So, if double star is minimum naturally it implies the minimum of star therefore, minimizing the Bayes cost function reduces to minimizing the conditional Bayes cost function conditioning on the fact that I have been given a set of observation z .

So, what is the basic idea here let me look back at it. Mother nature has picked x from the probability $p(x)$ I am going to observe and gain information about the chosen x . So, from the perspective of the estimator everything starts with z . So, everything given z is what we are going to be working at given z I am going to create an estimate \hat{x} , given \hat{x} I am going to have an error given the error I have I am going to have a Bayes cost function conditioned on the given observation which is given by double star. If I make double star the minimum that automatically minimizes the Bayes cost function $b(x, \hat{x})$

because p of z is 0 therefore, without loss of generality we could minimize the conditional Bayes cost and that is one of the important conclusion that comes out of this analysis.

(Refer Slide Time: 26:18)

SPECIAL CASES
A) BAYES LEAST SQUARES ESTIMATOR

- Define $\mu = E[\mathbf{x}|z] = \int_{\mathcal{R}^n} \mathbf{x} p(\mathbf{x}|z) d\mathbf{x}$
- μ is a function of the observations z
- Then choosing $c(\hat{\mathbf{x}}) = (\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{W}(\mathbf{x} - \hat{\mathbf{x}})$

$$\begin{aligned}
 B(\hat{\mathbf{x}}) &= E[c(\hat{\mathbf{x}})] \\
 &= E[(\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{W}(\mathbf{x} - \hat{\mathbf{x}})] \\
 &= E[(\mathbf{x} - \mu + \mu - \hat{\mathbf{x}})^T \mathbf{W}(\mathbf{x} - \mu + \mu - \hat{\mathbf{x}})] \\
 &= E[(\mathbf{x} - \mu)^T \mathbf{W}(\mathbf{x} - \mu)] + E[(\mu - \hat{\mathbf{x}})^T \mathbf{W}(\mu - \hat{\mathbf{x}})] \\
 &\quad + 2E[(\mathbf{x} - \mu)^T \mathbf{W}(\mu - \hat{\mathbf{x}})]
 \end{aligned}$$

Now, we are going to do with that as the background that is a very general principle. We are going to do what is called Bayesian least square estimators. Please look at this now. We are combining least squares with Bayes. What is the difference? In the statistical least squares that we saw in the previous talk we are simply given the observation using the observation, we are going to be able to estimate they are known we were not given any information about the unknown. So, that is what is called statistically squares. What is the difference between that statistically squares and basically squares? Within the Bayesian framework we always have two pieces of information one is the prior another is the new information coming from the observation given through the conditional distribution. So, I have two pieces of information. So, we are slightly richer within the Bayesian setup than compared to your simple statistically squares.

So, we are going to be able to revisit the design of least square estimation techniques within the context of the availability of prior information. So, all the notations, concepts, carry over I am now going to define some new quantity, \mathbf{x} is the unknown z is given to us from the previous slide we can see, we are all we are now interested in analyzing the problem conditioned on the given set of observation because that is what the basis for the

estimation techniques. So, given z I am going to talk about the conditional expectation of x given z that is an important quantity. I am going to denote that quantity by μ .

Now, please understand z is random. So, even though we are going to condition the expectation of x with respect to z conditional expectation is in general a random variable. So, μ is denoting the conditional expectation I want to remind you that μ in general is random. This μ as given by the conditional expectation is can be evaluated by the standard conditional expectation formula which is $\int x \cdot p(x \text{ given } z) dx$. So, μ is a function of the observation, observations are random. So, μ is random.

Now, I would like to be able to choose c of x to be the weighted sum of squares the weighted sum of squares. If I assume the way to be identity I get the simple sum of squares. So, it is a reasonably good way to start with this weighted version. Now b of \hat{x} , please remember from the previous slide is equal to expected value of c of \tilde{x} this is c of \tilde{x} sorry that must be c of \tilde{x} . So, c of \tilde{x} is x minus \hat{x} transpose W x minus \hat{x} .

Now, what do we do? There are two terms in here, there are two terms in here, so I add and subtract μ to this I add and subtract μ to this why; there is going to be a purpose for this they are going to get an important conclusion out of this. So, if we multiply and simplify I get one form like this another form like this and a third one yet another one is like this. So, I get sum of three expectations. It is a very simple idea, is a very simple idea. I am trying to compute the conditional expectation and I am going to compute the Bayesian cost I have inserted the conditional expectation into the Bayesian cost and I am going to see the role of conditional expectation in trying to minimize the Bayesian cost, that is the purpose.

So, let us consider one of the terms x minus μ transpose W μ minus \hat{x} let us look at this now, x minus μ transpose that is that is the term with the with the with the coefficient 2, x minus μ transpose W μ minus \hat{x} .

(Refer Slide Time: 31:21)

- Using iterated law of Conditional expectations

$$E[(\mathbf{x} - \mu)^T \mathbf{W}(\mu - \hat{\mathbf{x}})] = E\{E[(\mathbf{x} - \mu)^T \mathbf{W}(\mu - \hat{\mathbf{x}})|\mathbf{z}]\}$$

- But

$$E[(\mathbf{x} - \mu)^T \mathbf{W}(\mu - \hat{\mathbf{x}})|\mathbf{z}] = (\mu - \hat{\mathbf{x}})^T \mathbf{W} E[(\mathbf{x} - \mu)|\mathbf{z}]$$

$$= (\mu - \hat{\mathbf{x}})^T \mathbf{W} \{E(\mathbf{x}|\mathbf{z}) - \mu\}$$

$$= 0$$

Handwritten notes:
 $(\mathbf{x} - \mu)^T \mathbf{W}(\mu - \hat{\mathbf{x}}) = (\mu - \hat{\mathbf{x}})^T \mathbf{W}(\mathbf{x} - \mu)$
 \downarrow
 $\mathbf{x} = \mu$

$\hat{\mathbf{x}}$ is a function of the observation therefore, this expected, this expectation from basic probability theory sorry yeah; from basic probability theory it is very well known that this expectation can be expressed as iterated conditional expectations. So, what is the basic idea?

(Refer Slide Time: 31:27)

PROPERTIES OF BAYES LEAST SQUARES ESTIMATE

- $\hat{\mathbf{x}}_{MS}$ is unbiased: $E[\mathbf{x} - \hat{\mathbf{x}}_{MS}] = E\{E[\mathbf{x} - \hat{\mathbf{x}}_{MS}|\mathbf{z}]\}$
 $= E\{E[\mathbf{x}|\mathbf{z}] - \hat{\mathbf{x}}_{MS}\}$
 $= 0$
- $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}_{MS} \Rightarrow E(\tilde{\mathbf{x}}) = E[\mathbf{x} - \hat{\mathbf{x}}_{MS}] = 0$
 - Mean of the error is zero
- $B(\hat{\mathbf{x}}_{MS}|\mathbf{z}) = \int_{\mathcal{R}^n} (\hat{\mathbf{x}} - \hat{\mathbf{x}}_{MS})^T (\hat{\mathbf{x}} - \hat{\mathbf{x}}_{MS}) p(\mathbf{x}|\mathbf{z}) d\mathbf{x} \quad [W=I]$
 $= \text{total variance in the components of } \tilde{\mathbf{x}}$
 - Since $\hat{\mathbf{x}}_{MS}$ minimizes $B(\hat{\mathbf{x}}|\mathbf{z}) \Rightarrow \hat{\mathbf{x}}_{MS}$ also minimizes the variance in the estimate

First I could consider the conditional expectation of this quantity \mathbf{x} minus μ transpose \mathbf{W} μ minus \mathbf{x} transpose given \mathbf{z} if you calculate this conditional expectation it will becomes a function of \mathbf{z} and then you take another expectation with respect to the

distribution of z to be able to get rid of the randomness with respect to c as well. So, the inside one is given z the outside one is expectation with respect to z .

Now, let us look at the conditional expectation the inside factor. So, this is the conditional expectation which is inside. If z is given since \hat{x} is a function of z $\mu - \hat{x}$ is a known quantity given z . So, again the basic properties of conditional expectation wheels that I should be able to take this out of the expectation operator. So, it came here. W is also a non random quantity that also can be moved here therefore, by pulling these two quantities by pulling these two quantities, I can express this quantity is equal to this quantity. Please remember what is that we have used in here; $x - \mu$ transpose $W \mu - \hat{x}$ is also equal to $\mu - \hat{x}$ transpose $W x - \mu$ because W is a symmetric matrix and this quadratic form is a scalar, therefore, these two are essentially the same and we have used this particular fact in trying to go from there to answer to the right hand side.

Now, expectation is a linear operator expectation of a sum is the sum of the expectations. Given z μ is already known therefore, this expectation now can be written as expectation of x given z minus μ , but from the definition this is μ itself therefore, $\mu - \mu$ is 0 therefore, the cross term if you go back to the previous one this is the cross term this I would like to call this a cross term cross term is 0. If this cross term is 0 now I can further simplify, my Bayesian cost is essentially sum of only two terms, sum of only two terms.

(Refer Slide Time: 34:41)

$B(\hat{x}) = E[(x - \mu)^T W (x - \mu)] + E[(\mu - \hat{x})^T W (\mu - \hat{x})]$
 • The only control we have is the choice of \hat{x}
 • $B(\hat{x})$ is minimum when $\hat{x} = \mu = E(x|z) = \text{posterior mean}$ * $\mathbb{I} = \text{INDEPENDENT OF } \hat{x}$
 $\hat{x}_{MS} = E[x|z]$
 $= \int_{\mathcal{R}^n} x \left(\frac{p(z|x)p(x)}{p(z)} \right) dx$
 $= \frac{\int_{\mathcal{R}^n} x p(z|x)p(x) dx}{\int_{\mathcal{R}^n} p(z|x)p(x) dx}$

 $p(x) = \int_{\mathcal{R}^n} p(x, z) dz$
 $= \int_{\mathcal{R}^n} p(z|x)p(x) dz$

Now you see why we introduced the conditional mean into each of the expressions and what is the outcome of that mathematical simplification namely the third term with the coefficient two vanishes identically and that is one of the simplification that results from this manipulation.

Now, please understand I would like to be able to minimize b of x hat. To be able to minimize means I should have a free variable the only free variable I have is z is given to you x is unknown to you. So, what is the only chance you have; your only choice is the estimator ϕ . So, once you choose an estimator I got an estimate. So, the only control you have is to change the estimate. So, the choice the control is the choice of x hat.

Now, if you look at this expression the first term this is the second term; the first term does not have x hat as a part of it, first term independent is independent of x hat. If somebody does not depend on x hat I cannot change anything. The second term on the other hand depends on x hat the first term is a quadratic form expectation of the quadratic form and W is a symmetric positive definite matrix therefore, the quality form is always positive the first term is expectation of a positive quantity which is going to be greater than 0 unless x is equal to μ . The second term and we do not know whether x is equal to μ or not μ is my estimate μ is the conditional expectation of my estimate x is the unknown that mother nature has picked. So, it more often than not the first term will not be equal to 0.

Look at the second term μ is what I deliver it is the conditional expectation of \hat{x} , \hat{x} is my estimator. So, the second term is also a positive definite quadratic form the positive definite quantity form is 0 only when the vector is 0 and by changing \hat{x} I can affect only the second term. Therefore, by picking \hat{x} is equal to μ , \hat{x} is equal to μ is called the posterior mean; why \hat{x} is equal to μ is called for it is something that is done after observations are given that is posterior therefore, it is called conditional mean if you want to call a conditional mean. So, if I choose the second term is 0, I cannot change the first term therefore, $b \hat{x}$ is minimum when x is when \hat{x} is equal to μ .

So that means, what is the optimum estimate the optimum estimate \hat{x} is μ . What is μ ? μ is the conditional expectation. So, conditional expectation is an optimal estimate within the Bayesian framework. Why this is called least squares? The function c we have chosen let us go back to this is a least square Bayesed function is the sum of squared errors. So, this is the sum of squared errors. So, the least square comes in here. How Bayesian comes in here? They are posterior I am sorry we have prior and conditional expectation. So, we have combined everything in a beautiful way namely mean square quadratic cost given prior, given posterior we have combine all of them together to come to an important conclusion that the Bayes cost function is minimum when the estimate is equal to the condition of main.

The estimate is equal to condition of main, therefore, we are now going to give you a special estimator which is called \tilde{x}_{MS} , \tilde{x}_{MS} is called the means Bayesian mean square estimate it is like \hat{x}_{LS} for the least square estimate. So, \tilde{x}_{MS} is equal to the conditional expectation of x with respect to z . And what is the expression for this conditional expectation? It is the integral over R^n of the expected value of x with respect to the conditional distribution of x with respect to z . The conditional expectation of x with respect to z we already know from Bayes theorem this is equal to $p(x|z)$ and $p(x|z)$ by simple application of the Bayes rule is given by $p(z|x)$ times $p(x)$ divided by $p(z)$. So, this is the characterization of the optimal estimate within the context of the Bayesian framework, I have prior, I have conditional distribution, I have the mean.

Now, p what is $p(z)$? $p(z)$ is the probability density of observation you can readily see, $p(z)$ is the integral is the $p(z)$ is the marginal density with respect to d of x this is

\mathbb{R}^n . So, this is the joint density I am trying to integrate the joint density with respect to x to get $p(z)$. This can also be written as integral over \mathbb{R}^n $p(z)$ of x given times p of x d of x . So, this is the expression for p of z and that expression is used in this denominator. So, by combining this we get the important formula for the optimal Bayesian estimate, optimal least square Bayesian estimate and that is the structure of the optimal estimator that comes out of this analysis.

(Refer Slide Time: 41:49)

PROPERTIES OF BAYES LEAST SQUARES ESTIMATE

- \hat{x}_{MS} is unbiased: $E[x - \hat{x}_{MS}] = E\{E[x - \hat{x}_{MS}|z]\}$
 $= E\{E[x|z] - \hat{x}_{MS}\}$
 $= 0$
- $\tilde{x} = x - \hat{x}_{MS} \Rightarrow E(\tilde{x}) = E[x - \hat{x}_{MS}] = 0$
- * Mean of the error is zero
- $B(\hat{x}_{MS}|z) = \int_{\mathbb{R}^n} (\tilde{x} - \hat{x}_{MS})^T (\tilde{x} - \hat{x}_{MS}) p(x|z) dx$ [W=I]
= total variance in the components of \tilde{x}
- Since \hat{x}_{MS} minimizes $B(\hat{x}|z) \Rightarrow \hat{x}_{MS}$ also minimizes the variance in the estimate

Now, we are look at this now again least squares comes in a beautiful way. So, we are going to be able to do the least squares either within the Bayesian concept where that is two pieces of information prior and the observation or in official like situation where there is no prior. I only have only one thing coming from the observation all I can do is to be able to extract the juice out of the observation and we have already seen both in the case of deterministic least squares also in the case of statistical least squares now we are redoing the same least squares within the framework of Bayesian analysis.

So, having established the structure the optimal Bayesian estimate using the least square criterion, I am now going to discuss some of the properties. This is the Bayesian estimate they are going to the first climb is this estimate is unbiased. So, what is the main? E of x minus, x is unknown \hat{x}_{MS} is the estimate of x from the definition of unbiased knows from the earlier discussion expected value of x minus \hat{x}_{MS} must be equal to 0 to verify it is equal to 0 again I use the notion of repeated conditional expectation that

is a very beautiful mathematical trick that one could use we have already used this in the derivation of the Bayesian structure of the Bayesian estimate in the in the previous slides.

So, I can express this expectation as the iterated expectation with respect to the conditional expectation. I can $E[x | z]$ of the function of z , and expect this expect the conditional expectation is again a linear operator. So, I can express this function as the difference of the conditional expectation of x with z and $E[x]$, but if a little reflection reveals this is also equal to the same as that and the difference is 0; that means, the Bayesian estimate is automatically automatically unbiased. \tilde{x} is equal to x minus $E[x]$ therefore, $E[\tilde{x}]$ is equal to $E[x] - E[x]$ minus the mean square is 0; that means, mean of the error of the estimate is 0.

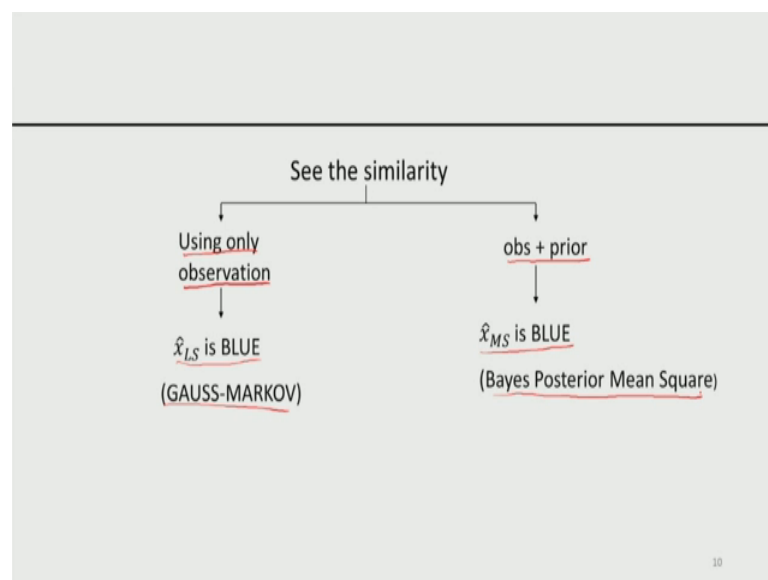
Now, let us come back and revisit the Bayesian cost. So, $E[\tilde{x}^2 | z]$ given z is equal to that is a Bayes cost which we have already given is the conditional Bayesian cost given the observation. So, going to the previous slide, I would like to be able to, I would like to be able to see what this one is this is given by \tilde{x}^2 minus x MS. So, what is that x MS is the optimal estimate \hat{x} . So, this is the, I think there should be that should be x that should be x sorry that should be \hat{x} . So, x minus \hat{x} m s; that means, the error in the Bayesian estimate this is the sum of the squared error in the Bayesian estimate. I am trying to integrate this with the conditional density the posterior density with respect to x I have assumed W is I , but, what is this? This is the random variable minus the mean transpose times random variable minus the mean. So, that is the total variance in the components of \tilde{x} therefore, the Bayesian cost the conditional Bayesian cost function is equal to the total variance and by virtue of our property of \hat{x} MS, it readily follows it readily follows that the Bayesian estimates minimizes the Bayesian cost.

The conditional Bayesian cost the conditional Bayesian cost is essentially variance and it minimize the variance. So, this is also minimum variance estimation, the minimum variance estimation. So, this would be a beautiful interpretation of what the Bayesian estimate is all about and some of the properties associated with this Bayesian estimate. I hope I hope the concept is clear very very nicely. Please also remember unbiasedness when the estimator is unbiased minimum squared error is equal to is equal to the minimum variance which we have already seen that is one of the reasons why we require unbiasedness we have alluded to when we discussed unbiasedness. So, because it is unbiased I am the right hand side relates to the conditional variance, conditional variance

given the observation the left hand side is the Bayesian cost function given the observation and by virtue of the fact that the Bayesian estimate minimizes the Bayes, the conditional Bayes cost function.

So, we already know the left hand side is minimum because \hat{x}_{MS} is a minimizer of the Bayes cost function the conditional Bayes cost function. The right hand side has an interpretation of variance that is why we can associate important properties with this class of Bayesian estimation namely the optimal estimate is equal to a posterior conditional mean, it also minimizes the total variance in the estimate, total variance in the estimate.

(Refer Slide Time: 48:19)



So, I would like to now bring out the similarity. Until now we used only observations using only observations in which case we said \hat{x}_{LS} is a blue and we had Gauss-Markov theorem. So, Gauss-Markov theorem refers to optimality of the least squares estimate when there is only one observation. Now, when there is only observation there is no prior. When there is observation and prior within the Bayesian context \hat{x}_{MS} is also a blue, \hat{x}_{MS} is called Bayes posterior mean. So, \hat{x}_{LS} is the blue using only observation \hat{x}_{MS} is a blue using observation and prior. So, you can see the least squares coming in both sides one with another without the use of prior. So, you can see the parallelism between the arguments in the Bayesian context as well as in the non Bayesian way of estimation all within the context of the squares.

So, thus far we have talked about the Bayesian context with the least square cost function. Please remember you could have considered the Bayesian cost function with any choice of cost function and we have already given four choices for the cost function the first to relate to the least squares, the third one uniform cost is something else and the fourth one the absolute cost is something else. So, you can readily see if I changed the form of the cost function the form of the estimate also will correspondingly change; that means, we will get a variety of different types of Bayesian estimates one for each possible choice of the cost function that goes to show the richness of the Bayesian formulation.

(Refer Slide Time: 50:32)

EXAMPLE 16.2.1

- $z = x + v$ $v \sim N(0, \sigma_v^2)$ $x \sim N(m_x, \sigma_x^2)$, x and v uncorrelated
 $E(xv) = 0$
 $\Rightarrow z \sim N(m_x, \sigma^2)$ where $\sigma^2 = \sigma_x^2 + \sigma_v^2$
- Compute $p(x|z)$
 - Recall $p(z|x) = N(x, \sigma_v^2)$, $p(x) = N(m_x, \sigma_x^2)$

$$\begin{aligned}
 p(x|z) &= \frac{p(z|x)p(x)}{p(z)} \\
 &= \frac{N(x, \sigma_v^2) N(m_x, \sigma_x^2)}{N(m_x, \sigma^2)} \\
 &= \beta \exp\left\{-\frac{1}{2}\left[\frac{(z-x)^2}{\sigma_v^2} + \frac{(x-m_x)^2}{\sigma_x^2} - \frac{(z-m_x)^2}{\sigma^2}\right]\right\}
 \end{aligned}$$

$z = x + v$
 $p(z|x) \sim N(x, \sigma_v^2)$

Of course even in the case when there is no prior we could have considered very many different types of cost functions. Since we are trying to estimate, since we are trying to seek optimal estimates we are chose the least square criteria because least square criteria has very nice properties with respect to with respect to minimum existence of unique minimum and so on.

So, I am going to quickly illustrate some of these using simple example. Let z be equal to x plus v let us z be equal to x plus v because x is an unknown v is the noise, v has 0 mean sigma square v as the variance x is a random variable its mean is m of x and sigma square x is the variance of x we are assuming x and v are uncorrelated x and v are not in correlated I am sorry it must be uncorrelated. That means, x v transpose must be equal to

0. So, that must be a v transpose yeah because I am going to get the, in this case I do not have to worry about the transfer sorry because I am assuming x and v are essentially scalar variables that is correct; $x^T v$ the expectation of the product must be 0 that is correct.

Now, z is the sum of two random variables therefore, x and v are uncorrelated, I know x and v are both normal the sum of two normals sum of two uncorrelated normals is also a normal variable. So, z is a normal variable whose distribution is given by this where σ_z^2 is equal to sum of the variances. So, if you add two random variables not only the mean changes, but also the variance changes. Here the mean of the sum is the sum of the means the variance of the sum is the sum of the variances this happens because in the simple case z is the sum of x and v x and v are both Gaussian and x and v are uncorrelated. If they are all correlated the expression for the variance will be different we are considering only the simplest possible case. So, what is that we have? We have all the information now, so I would like to be able to compute the posterior using Bayes rule.

Please recall the conditional distribution is given by, so if you are given x z I would like you to realize this x is z is equal to x plus v . So, if x is given x is fixed z is random because of v and v has 0 mean therefore, conditional distribution of z given x is normal with x is the mean and σ_v^2 as the variance. So, that is the formula that is given in here. The prior x is already given to us with a mean m_x and σ_x^2 as the variance therefore, the posterior using Bayes rule is given by this ratio. I am now going to substitute $p(x|z)$ $p(x)$ and $p(z)$ we already know everything. So, we can substitute the form of the normal functions we know the functional form of normal random variables if you substitute for each of the normal random variable as the normal density and simplify you get the following expression which is β times β is a constant that depends on π and variances, one can explicitly compute that and they and the exponent of the exponential is given by this sum of three terms. I would like you to be able to substitute and do the simplification and I think is a very good exercise.

(Refer Slide Time: 55:15)

EXAMPLE CONTINUED

- Simplifying the term in square brackets:

$$\frac{(z-x)^2}{\sigma_v^2} + \frac{(x-m_x)^2}{\sigma_x^2} - \frac{(z-m_x)^2}{\sigma^2} = x^2 \left[\frac{1}{\sigma_v^2} + \frac{1}{\sigma_x^2} \right] - 2x \left[\frac{z}{\sigma_v^2} + \frac{m_x}{\sigma_x^2} \right] + \left[\frac{z^2}{\sigma_v^2} + \frac{m_x^2}{\sigma_x^2} - \frac{(z-m_x)^2}{\sigma^2} \right] \rightarrow (1)$$

- We need to compute it as a **perfect square**

- Define

$$\frac{1}{\sigma_e^2} = \frac{1}{\sigma_v^2} + \frac{1}{\sigma_x^2} = \frac{\sigma_v^2 + \sigma_x^2}{\sigma_v^2 \sigma_x^2} \rightarrow (2)$$

and

$$\frac{\hat{x}_{MS}}{\sigma_e^2} = \frac{z}{\sigma_v^2} + \frac{m_x}{\sigma_x^2} \rightarrow (3)$$

Now, if you look at the term in the, under the bracket that is the term that term can be rewritten this is the method of perfecting the square I can express this as x squared times this two x times this that is a constant times that. So, this becomes this by simple simplification up by routine simplification.

Now I have an x squared term I have a 2 x term I would like to be able to express it as x minus sum whole square plus sum constant. So, I can rewrite this by method of perfecting the square to that end we are going to define a new quantity called sigma e square I am defining one over sigma e squared to be in other words the information this is the reciprocal of the variance that is the information, this is the reciprocal the variance of the noise, this is the reciprocal of the variance of the prior. So, the sum of the reciprocal is the reciprocal of sigma square e which can be rewritten I like this. I can also concoct a quantity called X hat MS divided by sigma square e which is given by this quantity using 2 and 3, using 2 and 3,

(Refer Slide Time: 56:55)

EXAMPLE CONTINUED

- R.H.S. (1) becomes

$$\frac{1}{\sigma_e^2} [x^2 - 2x\hat{x}_{MS} + \hat{x}_{MS}^2] = \frac{1}{\sigma_e^2} (x - \hat{x}_{MS})^2 \rightarrow (4)$$

- $\therefore p(x|z) = \frac{1}{\sigma_e} \exp\left[-\frac{1}{2} \frac{(x - \hat{x}_{MS})^2}{\sigma_e^2}\right]$
- \therefore Posterior mean is

$$\hat{x}_{MS} = \left(\frac{\sigma_e^2}{\sigma_x^2}\right)m_x + \left(\frac{\sigma_x^2}{\sigma_z^2}\right)z \rightarrow (5)$$

$$= \left(\frac{\sigma_e^2}{\sigma_x^2 + \sigma_z^2}\right)m_x + \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_z^2}\right)z$$

$$= \alpha m_x + (1 - \alpha)z \rightarrow (6)$$

$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$
 $m = \text{mean}$
 $\sigma^2 = \text{var}$

we can express the right hand side of 1. Please remember, 1 has a right hand side with three terms by using a perfect square and using the definitions in 2 and 3 the right hand side of 1 now the right hand side of 1; that is the important thing. The right hand side of 1 becomes this quantity this quantity can be expressed as simply by 1 over sigma square e x minus X hat MS, X hat MS has been defined in the previous page; sigma square e has been defined in 2. So, 2 and 3 gives you the basic definitions. So, with that in place I can indeed rewrite p x the posterior density the posterior density p x z p of x given z as this particular form where alpha is a constant, as alpha is a constant.

Alpha is a constant that can again be explicitly expressed, but in this particular case we already know this is the density function and alpha can be expressed in explicit form using the data that we have used earlier. We would like to leave the explicit computation, so the constant beta alpha as an exercise. So, if you now look at this expression it follows that that X MS, X hat MS is the mean please go back what is the standard normal distribution p of x is equal to 1 over square root of 2 pi sigma exponential minus x minus m squared divided by 2 sigma square.

In this case m is the mean and sigma square is the variance of the random variable x. So, if I use that analogy and use it here it readily follows the posterior mean is given by this, the posterior mean is given by this and this formula from the previous page can be written like this. Sigma square e by expressions in 2 and 3 can be written like this. Sigma

square v by, sigma square v by, sigma square x plus sigma square v sigma square x plus sigma square x plus sigma square v , I would like to remind this α I think must be beta sorry this α must be must be the same as beta. I do not want to confuse that that α must be the same as that. In here we are introducing an α which is this ratio and this one is one minus α . So, what does this tell you? This essentially tells you the following. The best Bayesian least square estimate is the convex combination between the prior mean and the observation.

α times m_x plus 1 minus α times z , so what does it mean? The m_x is a point here, z is the point here any point along this line has this particular form that is called the convex combination therefore, the Bayes least square estimate is a convex combination the prior mean and the observation that is an important conclusion that comes out of this analysis. Yes I have not talked about the algebraic simplifications, I think it is a very good exercise to be able to go through this algebraic simplifications.

(Refer Slide Time: 60:56)

EXAMPLE CONTINUED

$\hat{x}_{MS} = m_x + \underbrace{\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2} \right)}_{\text{Kalman gain}} \underbrace{[z - m_x]}_{\text{innovations}}$ ← Kalman-filter form

\therefore If $\sigma_x^2 \gg \sigma_v^2 \Rightarrow$ observation has a larger weight
 If $\sigma_x^2 \ll \sigma_v^2 \Rightarrow$ prior has a larger weight

$\hat{x}_{MS} = a m_x + (1-a)z$
 $= m_x + a m_x - m_x + (1-a)z$
 $= m_x + (1-a)[z - m_x] \rightarrow (7)$

$a = \frac{\sigma_v^2}{\sigma_x^2 + \sigma_v^2} = \alpha$

Adaptive nature!

I can rewrite the expression in 6, again as the Bayesian estimate \hat{x}_{MS} the Bayesian \hat{x}_{MS} is equal to m_x which is the prior mean plus z minus m_x times this quantity you can readily see that quantity is sigma square x by sigma square x plus sigma square v is it comes in here. In the previous slide we call it 1 minus α we simply now call it a gain term. So, the Bayesian estimate is equal to the Bayesian estimate is a very beautiful structure and interpretation, the Bayesian estimate is equal to

the prior plus z minus m of x , m of x is the mean z is the new information. So, m of x ; z minus m of x gives me what is called innovation. Innovation is the information in excess of what I already knew. So, the new estimate is equal to prior plus a constant times the innovation.

This form is the form, that underlie the well known Kalman filter. So, you can readily see the form of a Kalman filter coming in here. So, if I did not have any new information I would have the second term, I would my best estimate is the prior mean, but in addition to the prior mean if I get the new information I am going to update my belief to get the posterior mean. So, the posterior mean which is the Bayesian optimal estimate is equal to the mean plus a correction term, the correction term has a structure product of product of a gain term and the innovation and that form is a very standard form which is a very standard form. So, you can think of this in the form of a Kalman filter equation.

This form has also an adaptive property. What is the adaptive property? If σ^2_x is greater than σ^2_v if σ^2_x is much much greater than σ^2_v what does it mean observations are more reliable than the prior. If the prior variance is very large compared to the observation variance observations are more reliable than the prior therefore, in the previous equation 6 observations will be given more weight.

On the other hand if the observations are less reliable than the prior for example, σ^2_v is much much larger than σ^2_x ; that means, prior is much more reliable than the observation then the prior gets larger weight. So, that is a beautiful adaptivity property in the formula that is given in 6, also given in the Kalman filter form and that is called the adaptive property, that is what is called the adaptive property and that is called the adaptive property. So, by substituting this is equal to well this is the 7, is another way of rewriting the same thing I believe we should a is a constant. So, in this case a is equal to a is equal to the previous thing that comes in here which is σ^2_v a is equal to σ^2_v by σ^2_x plus σ^2_v a that is the definition of a which we have also called α . So, x MS can be written as a times m x plus one minus a time z which can be written as m x plus this plus this and that and that can be written like this. So, you can see these two forms are essentially the same form.

So, this is also called the Kalman filter representation, Kalman filter like representation k f form. So, that is the basic idea behind the Bayesian estimation.

(Refer Slide Time: 65:32)

VECTOR CASE: $Z = HX + V$

- v : $E(v) = 0$ $E(vv^T) = \Sigma_v$ $v \sim N(0, \Sigma_v)$
- x : $E(x) = m_x$ $\text{cov}(x) = \Sigma_x$ $x \sim N(m_x, \Sigma_x)$
 - $\Rightarrow Hx \sim N(Hm_x, H\Sigma_x H^T)$
 - Note: v, x are uncorrelated
- z is normal
- $E(z) = Hm_x \rightarrow (8)$
- $\text{cov}(z) = E[(z - Hm_x)(z - Hm_x)^T]$

$$= E[(Hx + v - Hm_x)(Hx + v - Hm_x)^T]$$

$$= E[H(x - m_x) + v][H(x - m_x) + v]^T$$

$$= E[H(x - m_x)(x - m_x)^T]H^T + E[vv^T]$$

$$= H\Sigma_x H^T + \Sigma_v \rightarrow (9)$$

$$\begin{aligned} \text{cov}(Hx) &= E[(Hx)(Hx)^T] \\ &= E[Hxx^T H^T] \\ &= H E[xx^T] H^T \\ &= H \Sigma_x H^T \end{aligned}$$

We continue our illustration of the Bayesian estimate, Bayesian least square estimates. For the vector case we have already seen the properties of the Bayesian estimate for the simple linear case, this is an extension to a vector case. So, Z is equal to H of X plus V , Z is a vector in R^m , X is a vector in R^n , all the properties we have been utilizing all the, all these properties of matrices vectors we have been utilizing all along. v such that E of v is 0; that means, its mean is 0; v is such that its covariance is Σ_v . Please remember we have utilized sometimes we are now going to utilize Σ_v . So, v is normal with 0 mean and Σ_v as the covariance, x is the unknown, x has a prior distribution expected value of x with respect to the prior distribution is m_x , covariance of x is Σ_x .

So, x as a normal distribution the prior distribution is m_x the vector Σ_x is a matrix. Therefore, H of x is the deterministic function of x . So, H of s random because x is random H of x has a distribution which is H times m_x and its covariance changes $H \Sigma_x H^T$ that is a very simple exercise in probability theory. So, covariance of H of x is equal to expected value of H of x times H of x transpose this is equal to expected value of $Hx, x^T H^T$ that is equal to H times expected value of $xx^T H^T$. So, that is equal to $H \Sigma_x H^T$. So, that is the formula

that comes out of it that is why this occurs. We are also assuming x and v are uncorrelated. z is the sum of H times x plus v . x is normal, H of x is normal, v is normal, x and v are uncorrelated. So, z is normal. The mean of z is equal to H times m of x . The covariance of z is given by this outer product. If you do the simplification as given in here we get the formula in 9 which is $H \Sigma_x H^T + \Sigma_v$. So, that is the covariance of z .

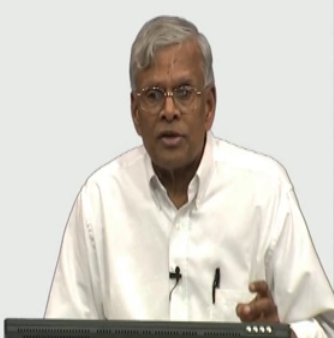
I think it is worth remembering that z gets a randomness from two different directions one through H another through v therefore, and H and v are uncorrelated. Therefore, the covariance of z is the sum of the covariances one coming from x through H , H of x another through the additive part of v therefore, the total covariance of z is given by 9 it is a sum of these two terms that is an important thing to realize. That again comes from basic considerations of probability calculations.

(Refer Slide Time: 69:33)

VECTOR CASE

- $E[z|x] = E[Hx+v|x]$
 $= Hx + E[v|x]$
 $= Hx \quad \rightarrow (10)$
- $\text{cov}(z|x) = E(vv^T) = \Sigma_v \rightarrow (11)$
- $p(z|x) = N(Hx, \Sigma_v) \rightarrow (12)$

$E[v|x], E(v) = 0$



So, E of x of, E of the conditional expectation of z with respect to x is equal to condition expectation of H of x plus v given x . Since x is given H of x is already known. So, H of x comes out of the expectation operator. We are left with the condition expectation of v with the respect to x , x and v are uncorrelated therefore, E of v given x is equal to E of v which is equal to 0 therefore, conditional expectation z with respect to x is H of x .

The covariance of z given x is Σ_v because from z you must subtract H of x that is why the conditional expectation, the conditional expectation, conditional expectation of z

with respect to $H^T x$, H of x the conditional covariance of z with respect to conditional covariance of z given x is equal to Σ_v . Therefore, I have the conditional mean, I have the conditional covariance if I have a conditional mean as a conditional covariance I have a conditional distribution because the conditional distribution is also normal. So, this is the conditional distribution of z given x . So, that is the distribution of the observation condition on the fact x has already been chosen by mother nature even though I do not know the value of x this is the conditional distribution of the observation given by 12.

So, now I would like to be able to do a posterior analysis.

(Refer Slide Time: 71:31)

POSTERIOR ANALYSIS

- $p(x|z) = p(z|x)p(x)/p(z)$ - BAYES

$$= \frac{N(Hx, \Sigma_v) N(m_x, \Sigma_x)}{N(Hm_x, \Sigma)}$$

$$= \alpha \exp \left[-\frac{1}{2} (z-Hx)^T \Sigma_v^{-1} (z-Hx) - \frac{1}{2} (x-m_x)^T \Sigma_x^{-1} (x-m_x) + \frac{1}{2} (z-Hm_x)^T \Sigma^{-1} (z-Hm_x) \right] \rightarrow (13)$$
- Consider the exponent:

$$x^T [H^T \Sigma_v^{-1} H + \Sigma_x^{-1}] x - 2 [H^T \Sigma_v^{-1} z + \Sigma_x^{-1} m_x]^T x + z^T \Sigma_v^{-1} z + m_x^T \Sigma_x^{-1} m_x - (z-Hm_x)^T \Sigma^{-1} (z-Hm_x) \rightarrow (14)$$

$$\equiv (x - \hat{x}_{MS})^T \Sigma_e^{-1} (x - \hat{x}_{MS}) \rightarrow (15)$$

$$\Rightarrow \Sigma_e^{-1} = (H^T \Sigma_v^{-1} H + \Sigma_x^{-1}) \rightarrow (16) \quad \leftarrow \text{ITS COVARIANCE}$$

$$\hat{x}_{MS} = \Sigma_e [H^T \Sigma_v^{-1} z + \Sigma_x^{-1} m_x] \rightarrow (17) \quad \leftarrow \text{ESTIMATE}$$

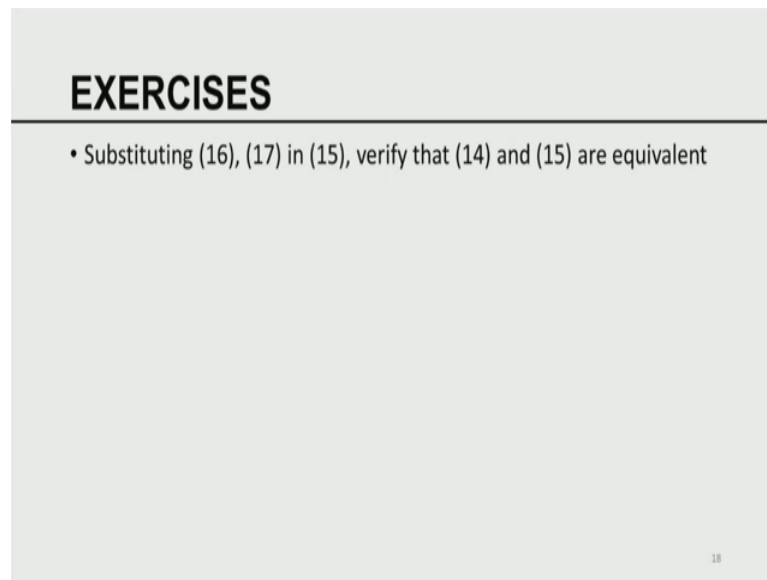
Posterior analysis this is the posterior distribution the posterior distribution by invoke into the Bayes rule. So, this is essentially a statement of Bayes rule. So, each of them are normal distributions. So, you remember that we this is exactly what we did in the scalar case as well it is the ratio of the numerator the product of two normal distribution denominator yeah another normal distribution. So, ratio of normal distributions at this can be again expressed as constant times this complicated expression. Even though looks complicated arithmetically is easy to simplify this now consider the exponent term again we are going parallel to what we did in the scalar case in the scalar case everything where scalar quantities here we simply need to consider matrix vector quantities matrix vector quantities.

So, we are considering the exponent $x^T H^T \Sigma^{-1} H x + x^T \Sigma^{-1} \mu + \frac{1}{2} \mu^T \Sigma^{-1} \mu$. So, the exponent can be written like this. Look at this now, this term is quadratic in x , this term is linear in x , the other terms do not have any x . So, if I have a quadratic term a linear term at a constant term what is the basic idea here, you try to use the method of perfecting the square. So, you add and subtract a constant so that I can extract a square root after that. The same principle that we did in the scalar case, but the algebra is little bit more involved in the vector case therefore, the exponent can be simplified the exponent in three can be simplified as 14, 14 after doing the completion of the perfect square becomes identically equal to this, where Σ_e^{-1} is given by this and μ_{MS} is given by this.

So, you can verify by substituting 16 and 17 and 15, 14 and 15 are equal. Yes that is that is the government of algebra involved in here and since our aim is to be able to indicate all the major steps we are going to leave the algebra for the reader to verify. I think it is absolutely essential for anyone who wants to understand these derivations thoroughly must go through the details of all the simplifications. So, with this we have now derived an expression for the best Bayes in estimate, I am sorry this is the best Bayes in; this is the best Bayes in estimate and this is its covariance.

Again that activity property that we talked about in the case of scalar case also applies in here, but the interpretation is a little bit more complicated because the fact there are matrices and the operator H comes into play, but in principles all the conclusions the adaptivity with respect to which weight is more which weight is less. For example, there are two pieces of information the posterior mean is going to be a linear function of the prior mean and the new information that comes from the innovation and how these two terms are weighted relatively that depends on the relative values of the covariance matrices for the prior and the observation the covariance of the conditional distribution of z given x . And the similarities is very obvious and I would definitely like the reader to be able to compare scalar expressions with vector x expressions and identify which term becomes which term corresponds to what term in the scalar case and the corresponding term the vector case I think it will be very beneficial for anyone everyone to do that.

(Refer Slide Time: 76:20)



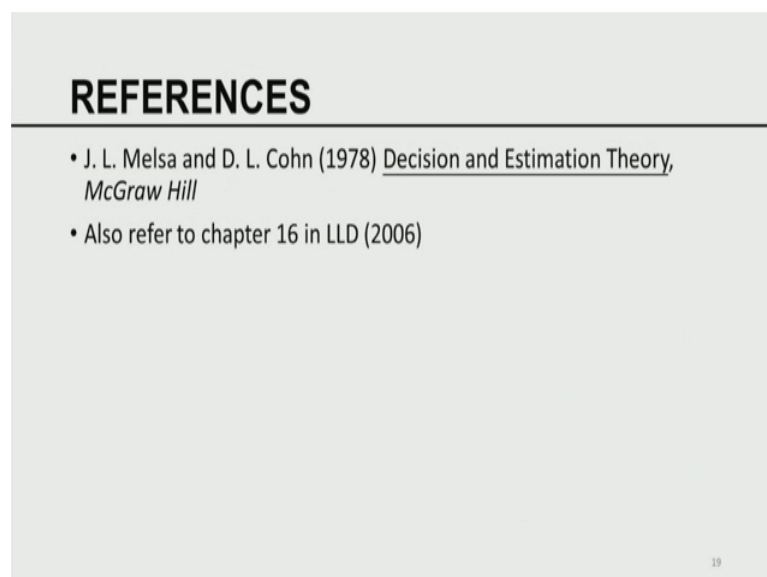
EXERCISES

- Substituting (16), (17) in (15), verify that (14) and (15) are equivalent

18

So, with that we come to the end of the discussion of the Bayesian methods. Now we can see it is this Bayesian method that is going to be the basis for stochastic aspects of data simulation. What is an exercise? The exercise is substituting 16 17 and 15, verify 14 and 15 are equivalent and again I have already mentioned to this is relate, this relates to the principle of perfecting the square in the matrix vector notation algebraically it is non-trivial please do that.

(Refer Slide Time: 76:58)



REFERENCES

- J. L. Melsa and D. L. Cohn (1978) Decision and Estimation Theory, McGraw Hill
- Also refer to chapter 16 in LLD (2006)

19

Again a reference for this is by Melsa and Cohn and Decision Estimation Theory, McGraw Hill. Also you can refer to our chapter 16 Lakshmi (Refer Time: 77:08), Lewis Lakshmi (Refer Time: 77:10) 2006. With that we conclude the discussion, an elementary discussion of the Bayesian least squares estimation.

Thank you.