

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

Lecture – 25
Statistical Least Squares

In the previous module 6.1, we listed several of the desirable properties of estimates, estimators these are extremely fundamental properties one should be, one should always look for whenever they are trying to do statistical estimation. Because if you show your result to a statistician they will immediately challenge you by asking this question is it unbiased, how efficient is this, is this a consistent estimate and so you have to be aware of these potential question that an educated person who is well versed in statistics could ask because if you project data assimilation as one of estimation of unknowns you cannot escape analyzing properties of estimates that one has to consider as a fundamental base.

So, in this module we are going to turn to a particular method of estimation least square estimation. We have already talked about deterministically squares so far, now we are going to be talking about statistical least squares. So, the whole lectures are centered around least square methods. Until now we talked about deterministically square principles now because we considered static deterministic problem, dynamic deterministic problem everything was deterministic up until now, now we have ventured into the world of stochastic or statistical estimation with respect to model being stochastic with respect to observations are always stochastic. So, this theory essentially calls for some nerding understanding of the fundamental underpinnings of stochastic or statistical estimation.

One of the work hours of the statistical estimation methodology is again statistical least squares.

(Refer Slide Time: 02:22)

STATISTICAL LEAST SQUARE ESTIMATE

- Given $z = Hx + v$, $z \in \mathbb{R}^m$, $x \in \mathbb{R}^n$
 - $E(v) = 0$
 - $E(vv^T) = R$ - known, S.P.D. $\Rightarrow R^{-1}$ information matrix exists
 - x and v are NOT correlated
- Define residual $r(x) = z - Hx$
- Weight sum of squared residuals

$$\begin{aligned} f(x) &= \frac{1}{2} r^T(x) R^{-1} r(x) = \frac{1}{2} \|r(x)\|_{R^{-1}}^2 \\ &= \frac{1}{2} (z - Hx)^T R^{-1} (z - Hx) \end{aligned}$$

So, we are going to not develop the grand theory of statistical least squares, but we are going to illustrate the fundamentals of statistical least squares using simple examples that are relevant to data assimilation. Again let us start with let us z be an observation let us z be a linear function of the unknown x v be a noise, v be such that its mean is 0, its covariance is R is known R is SPD. R is a covariance matrix in statistics the inverse of the covariance matrix is called the information matrix. For example; let us consider the following let sigma square be the variance. If the sigma square is large what does it mean? Their under variable can find itself in a large domain; that means, the uncertainty is large. When sigma square is small means what, the variance is very small if the variance is very small there is a very small variation that the underlying random variable encounters. So, we know pretty much the values much more confidently than when the variance is large.

So, the inverse of the variance if sigma square is large 1 over sigma square is very small, a sigma square is small one over sigma square is large therefore, the inverse of variance is call information, they are inversely related to each other one is the reciprocal of the other. So, larger variance means less information and smaller variance means more information the covariance matrix inverse of a covariance matrix we simply do not say inverse of a covariance matrix another name for it is information matrix. So, if R is SPD information matrix exists. x is an unknown x could be random for example, as we have already said, the unknown could itself be a random variable for example, a climate

variable has a natural variability so climate variables are random variables. I am going to make observations of this underlying climate variables v is the observation noise. So, x is a random variable v is another add on noise coming from observations. So, there are two components in here I am going to assume that the observation noise and the underlying intrinsic variation x are uncorrelated that is a very standard assumption and it is also very valid assumption.

The climate does not change because I am trying to measure the nature. So, v depends on the instruments I use, but the climate itself has an underlying variability for example, el nino comes once in several years in some form of a cycle while we cannot pin down the exact period of the cycle we know roughly what happens in here. So, we know the overall behavior become pin down the values of period. So, the period with which el nino occurs is the random variable. So, el nino phenomena has a natural variation a el nino varies naturally the associated climate also varies naturally. So, x is unknown, x is endowed with a natural variability, x is a random variable, v is an additional randomness that I am introducing into by virtue of measuring about our gaining information about x . So, when you have multiple random variables in statistics one of the things that we have to worry about what is the relation between these two random phenomena simplest possible assumption is also very valid assumption x and v are uncorrelated.

Now, define the residual. This, sounds very familiar sound very familiar to us that is what we have deal in deterministic least squares. So, I can now consider the weighted sum of squared residuals the weighted sum of squared residual is f of x is equal to $r^T x$, R is the noise covariance which is also called the energy norm square of $r^T x$ where the weight is R inverse. The explicit expression is given by z minus $H x$ transpose R inverse z minus $H x$ I have stuck in a factor half, I have already argued that many of the arguments do not change the only convenience for half is that it simplifies a little bit of an algebra.

(Refer Slide Time: 07:14)

STATISTICAL LEAST SQUARE ESTIMATE (CONT'D)

- $\nabla f(x) = (H^T R^{-1} H)x - H^T R^{-1} z \rightarrow (1)$
- $\nabla^2 f(x) = H^T R^{-1} H \rightarrow (2)$
- $\hat{x}_{LS} = (H^T R^{-1} H)^{-1} H^T R^{-1} z \rightarrow (3)$

We can compute the gradient of that we already know we have, we should be already masters of this we may have by now we have seen this term several times over. So, this is the expression for the variance, this is the expression for the Hessian if I set the variance to 0 and solve the equation 1 I get the estimate least square estimate like this.

If I assume z to be deterministic \hat{x}_{LS} the least square estimate it is also be deterministic, but in our case z is random z is random because x is random z is random, z inherits randomness from two sources, the two sources are uncorrelated therefore, \hat{x}_{LS} is random. So, this estimate is a random vector it has its own underlying sampling distribution the estimator, the estimate has a distribution. So, I would like to be able to ask myself the following question under what condition this least square estimate is unbiased and so on. We would like to be able to analyze the properties of this estimate.

(Refer Slide Time: 08:26)

OBSERVATIONS

- Unbiasedness

$$\hat{x}_{LS} = (H^T R^{-1} H)^{-1} H^T R^{-1} z \quad \text{and } Z = Hx + V$$

$$= x + (H^T R^{-1} H)^{-1} H^T R^{-1} v$$

$$E(\hat{x}_{LS}) = x + (H^T R^{-1} H)^{-1} H^T R^{-1} E(v)$$

$$= x.$$
- \hat{x}_{LS} is unbiased
- Covariance of the estimate

$$COV(\hat{x}_{LS}) = E[(\hat{x}_{LS} - x)(\hat{x}_{LS} - x)^T]$$

$$= (H^T R^{-1} H)^{-1} H^T R^{-1} E(vv^T) R^{-1} H (H^T R^{-1} H)^{-1}$$

$$= (H^T R^{-1} H)^{-1} \underbrace{R^{-1} R R^{-1}}_{= R^{-1}}$$

$$= [\nabla^2 f(x)]^{-1}.$$

So, first question is this least square estimate is this estimate unbiased. So, I am going to discuss unbiasedness. Please from the previous slide this is the expression for the estimate, but please remember z is equal to $Hx + V$. So, if I substitute this z in here and simplify simple algebra leads you to this. So, if I took the expected value of the left hand side is this, right hand side is this, the expected value of v is 0. So, the expected value of the estimate is equal to the unknown and hence the estimators are unbiased. So, what does it mean? The least square estimate is unbiased in this case.

I would like to be able to compute the covariance of the estimates the covariance of the estimate z of x LS being a vector. The covariance consists of the outer product, this is the column that is a row we have to consider a matrix expected value of the matrix, expected value of the matrix is a expected value of every element of the matrix, but in this particular case from this equation I do know the expression for \hat{x}_{LS} . So, expression for \hat{x}_{LS} is equal to x plus this. So, using this relation in here as well as in here we can readily see these two factors relate to this I am going to talk about the expected value is it is very easy for me to draw these things and tell you, but I would like you I would like to emphasize that you should do all the verify all these algebra that is very fundamental.

Now look at this now, this term is not random, this term is not random, this term consists of the covariance of the noise vector and we all know the covariance of the noise vector is

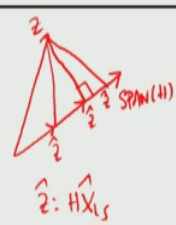
R. This R will can get cancelled with one R inverse. Once it gets cancelled with the R inverse the other term also gets cancelled living behind H transpose R inverse H inverse. That is the expression for the least square least covariance of the square estimate.

Now, I would like to go back to the previous slide. If you look at equation 2, the Hessian I want to emphasize the Hessian of the f of x is equal to H transpose R inverse H now let us come back in here the covariance of the estimate is H transpose R inverse H inverse. So, you can readily see the covariance of the least square estimate sorry the covariance of the least square estimate is simply inverse of the Hessian of the objective function it is a beautiful property its one of the one of the most one of the most beautiful property that relates the gradient the Hessian the estimates, the variance of the estimate and so on. Therefore, what is the conclusion the conclusion is the least square estimate in this setup is unbiased, I can also compute its variance, the variance is related to the inverse of the Hessian of the objective function that the least square estimate tries to minimize.

(Refer Slide Time: 11:53)

OBSERVATIONS (CONT'D)

- Relation to Projection

$$\begin{aligned}\hat{z} &= H\hat{x}_{LS} \\ &= H(H^T R^{-1} H)^{-1} H^T R^{-1} z \\ &= Pz \\ P &= H(H^T R^{-1} H)^{-1} H^T R^{-1}\end{aligned}$$

- Idempotent: $P^2 = P$
- Non-symmetric: $P \neq P^T$
- $\Rightarrow P$ is an oblique projection
- Note: when $R^{-1} = I$, $P = P^T$ is symmetric (orthogonal projection)

Please remember we have already talked about the relation between projections and least square estimates. So, relation to projections I would like to bring to our attention. So, what is z hat? So, let we can think of it like this now, this is the span of H we have already talked about it in our deterministic methods, this is z if I project this this is the orthogonal projection, if I project this that is an oblique projection. So, I in the case of orthogonal projections we had in the case of oblique projection just be this, I could also

get an oblique projection like this there is only one orthogonal projection all the other projections are oblique. Oblique projection occurs when there is a weighted least squares orthogonal projection comes in only when you do ordinary squares. These are our weighted least squares so we are going to get oblique projections in here. Therefore, \hat{z} whichever projections you have \hat{z} is always is equal to H of x LS, x LS the formula for that is already known you substitute that formula this is the expression for the projection of z onto the span of H . I am going to concoct a matrix that matrix is the projection matrix P which is given by this expression.

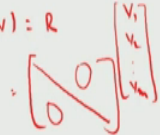
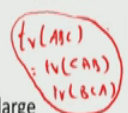
We have already seen this in the case of deterministic case I am trying to redo them within the context of statistic least squares to see the relation between least squares whether it is statistical or deterministic. It can be verified this matrix P is idempotent what does it mean P square is equal to P please verify that. This matrix is not symmetric P is equal not equal to P , P transpose please again verify that. So, any idempotent matrix that is not symmetric gives rise to an oblique projection that is the basic theory we have not proved that theory we are not proved that theorem, I would like to if you are a person who wants to know more and more of fundamentals what is the fundamental result in here, any orthogonal projection matrix must be idempotent and symmetric, any oblique projection matrix as an operator must be idempotent, but not symmetric.

What is the intrinsic relation between projection operators and matrices with these properties these are produced in an advanced book on matrix analysis, matrix theory? There are number of books we have alluded to especially the book by Horn and Johnson is one of my favorites it has very nice proofs of these basic facts from operator theory as well as projection theory. When R inverses I P is equal to P transpose it becomes symmetric it becomes orthogonal projection. So, this is something we have already come across within the context of deterministic I wanted simply inform ourselves of the fact that the same properties also carry over to the statistical control path.

(Refer Slide Time: 15:13)

OBSERVATIONS (CONT'D)

- Uncorrelated noise: $R = \sigma^2 I$
 $\hat{x}_{LS} = (H^T H)^{-1} H^T z$
 $COV(\hat{x}_{LS}) = \sigma^2 (H^T H)^{-1}$
- $H^T H$ is symmetric
- $\Rightarrow (H^T H)Q = Q\Lambda$ – Eigen decomposition of $H^T H$
 $(H^T H) = Q\Lambda Q^T$, $Q^T Q = Q Q^T = I$, $\Lambda = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$
- $(H^T H)^{-1} = Q\Lambda^{-1}Q^T$
- $\Rightarrow \text{tr}[COV(\hat{x}_{LS})] = \text{tr}[\sigma^2 (H^T H)^{-1}] = \sigma^2 \text{tr}[Q\Lambda^{-1}Q^T] = \sigma^2 \text{tr}[Q^T Q \Lambda^{-1}]$
 $= \sigma^2 \text{tr}[\Lambda^{-1}] = \sigma^2 \sum_{i=1}^n \frac{1}{\lambda_i}$
- If $H^T H$ is nearly singular, then λ_i is close to 0. $\Rightarrow COV(\hat{x}_{LS})$ is large

$COV(v) = R$

 $COV(v_i, v_j) = 0$ if $i \neq j$
 $Var(v_i) = \sigma^2$
 $R = \sigma^2 I$


Now, I am going to make some further simplification. Until now we assumed the covariance of v , the covariance of v is equal to R . In general R is a speedy, so R is a symmetric matrix in general it can have nonzero of diagonal elements now I am going to come to a very special case I am going to consider the noise component v are uncorrelated. So, what does it mean? v has v_1, v_2, \dots, v_m ; the covariance of v_i with respect to v_j if it equal to 0 for i not equal to j . What does it mean? All the elements here are 0 all the elements under here are 0.

Further I am going to assume the variance of each of the, is equal to sigma square; that means, they all have the common variants. In this case my R becomes sigma square times I , sigma square times I is an identity matrix multiplied by sigma square and the means it is the matrix where off diagonal elements are 0 all the diagonal elements are same. And this makes sense why is that? I am going to measure observations at different places and collate them into a vector. The observational errors of different places when I am measuring from as they are uncorrelated I am using the same instrument at every price so they have common variants. So, that is the physical import of this assumption. So, in this case R is equal to sigma square I .

In this case my expression for x_{LS} if I substitute this becomes this. Now, please remember now in this case there is no sigma square there is no R because it magically goes away that is that is the nature of the expressions in here. The covariance of this is

simply the inverse of the Hessian which is given by this. $H^T H$ you remember that that is the Hessian that is the Gramian, $H^T H$ is symmetric if I do an eigen decomposition of $H^T H$, I get this Q is the matrix of columns of the eigenvectors.

$H^T H$ is equal to $Q \Lambda Q^T$, $Q^T Q$ of I is I , Λ is the diagonal element with n eigenvalues. I know I am going little faster, but I am sure we have come across this several times. So, I do not want to overly repeat what we have already done, we have come across these things in context of the discussion of decomposition and the same ideas come over here because this is the Gramian. So, given this, if $H^T H$ is equal to this its inverse is equal to the inverse of the right hand side we already know inverse of the product is the product of the inverses. In the case of Q , Q is an orthogonal matrix transpose is the inverse. I am combining all those results to get this. This is a very special formula for the $H^T H$ inverse, why am I interested in the $H^T H$ inverse because $H^T H$ inverse is a matrix that relates to the covariance or the yeah the least square estimate least square estimate.

Now, if I consider a covariance matrix the diagonal elements of the covariance matrix are all variances of the different components. So, I am now going to compute the trace of the covariance of the estimate. Please remember we already talked to the trace the trace of a matrix is equal to sum of the diagonal elements, when the matrix is the covariance matrix diagonal elements are all related to variance. So, trace of the covariance matrix gives you the total variance in the components of the estimate. So, trace of the covariance of \hat{x}_{LS} is equal to the total variance in all the elements of the estimated vector \hat{x}_{LS} . Now, we already know that is equal to σ^2 this. This is a constant that multiplies by trace. So, that comes out from the definition of trace.

So, now I am going to utilize this formula to substitute for this. So, this becomes this, there are lots of matrix algebra in here and we also have seen trace of ABC is equal to trace of CAB is equal to trace of BCA . The trace is invariant under the circular shift of the product, so because of that I get this. $Q Q^T$ is I , therefore, this quantity becomes a trace of Λ^{-1} trace of Λ^{-1} is simply sum of $1/\lambda_i$ times σ^2 beautiful, beautiful. So, that the total variance in the estimate is simply σ^2 times sum of the reciprocals is their eigenvalues of the gramian $H^T H$. It is a absolutely beautiful result.

So, what does it tell you? If $H^T H$ is nearly singular one of the eigenvalues is going to be close to 0, when one of the eigenvalues is going to be close to 0, one over the smallest eigenvalue that is exploded on your face; that means, the covariance estimate is large. So, what does it tell you? You remember the condition number we talked about is something similar to that. When the condition number goes large, the condition number please remember is the ratio of the largest eigenvalue to the smallest eigenvalue is the smallest eigen value while remaining positive is very close to 0 that is going to explode on your face that is exactly what is happening in here, that is exactly what is happening in here.

The condition number the gramian $H^T H$ is very large another interpretation in the context of statistical least squares is that, the covariance of the least square estimate can explode on your face if the matrix H is ill conditioned or nearly ill conditioned that is that is exactly the conclusion coming from the last line of the slide.

(Refer Slide Time: 21:58)

OBSERVATIONS (CONT'D)

- Estimation of σ^2 : Let $R = \sigma^2 I$
- Define the residue e (error in the estimate)

$$\begin{aligned}
 e = z - \hat{z} &= z - H\hat{x}_{LS} \\
 &= (I - P)z \\
 &= (I - P)(Hx + v) \quad [(I - P)H = (H - PH) = H - H = 0] \\
 &= (I - P)v,
 \end{aligned}$$

$E(e) = E[(I - P)v] = (I - P)E(v) = 0.$

$R: \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$
 \uparrow

Now I am going to talk about, I can do all the estimates I am assuming my covariance, noise covariance is sigma square I until now I assume my I know my sigma square what happens if sigma square is not known. So, I could estimate that. Statisticians are very clever people they will relax every possible assumption what if that is known how did not we how to estimate this, what if that is not known how to estimate this. So, they will talk about every possible combinations of knowns and unknowns in the context of

estimation problem as we have seen in the previous exercise. So, the whole question is this can I use this framework to be able to estimate the observation noise covariance.

Let me post the problem like this we have already seen you use satellite observations, we generally do not know what is the covariance of satellite observations. So, is there a way to formulate your estimation problems such that using which I can at least hope to estimate the variance of satellite observations or radar observations; that is one way to think of this problem as a motivational fashion. So, I would like to be able to estimate I would like to be able to estimate σ^2 σ^2 is related or by this.

Now, to be able to do this estimation I am going to go back to my residual, residue e which is there on the estimate e is equal to I sometimes I call it R sometime I call it e depending on the context I do not think that should bother you. I can call the same quantity by different names at different places depending on convenience, I hope it does not throw you out of the board. So, z minus H of LS, this is the model counterpart this is actual observation the difference is the error. I can express this error as I minus P times z where P is the projection matrix which we have already talked about in the last slide.

I can substitute for z as Hx plus v and if I do the multiplication it turns out that this error is equal to I minus P times v and why this is I minus P times v because of the algebra that is given in here. So, because of this algebra it turns out that this product while it should have 4 terms essentially the product of I minus P of the Hx becomes 0 because of the argument in here I am going to leave it to you to enjoy the argument. So, the expected value of this error is 0; that means, this error is unbiased it is the covariance of this error I am interested in estimating.

(Refer Slide Time: 24:58)

OBSERVATIONS (CONT'D)

- Estimation of σ^2 (cont'd)

$$\begin{aligned}
 E(e^T e) &= E[v^T (I - P)(I - P)v] \\
 &= E[v^T (I - P)v] \quad ((I - P) \text{ is idempotent}) \\
 &= E[\text{tr}(v^T (I - P)v)] \quad (\text{tr}(a) = a \text{ for scalar } a) \\
 &= E[\text{tr}(v v^T (I - P))] \quad (\text{tr}(ABC) = \text{tr}(CBA)) \\
 &= \sigma^2 \text{tr}(I - P) \quad P \in \mathbb{R}^{m \times m}, H^T H \in \mathbb{R}^{n \times n} \\
 &= \sigma^2 [\text{tr}(I) - \text{tr}(P)] \quad [\text{tr}(P) = \text{tr}[H(H^T H)^{-1} H^T] = \text{tr}[(H^T H)(H^T H)^{-1}] = \text{tr}(I)] \\
 &= \sigma^2(m - n) \quad (m > n)
 \end{aligned}$$

$e e^T$
 $e^T e \rightarrow$
 $E(v v^T) = \sigma^2 I$
- $\hat{\sigma}^2 = \frac{e^T e}{(m - n)}$ is an unbiased estimate of σ^2

Therefore I want to be able to estimate sigma square sorry, I want to be able to estimate sigma square in order to able in order to estimate sigma square E is the error vector e transpose e , what is that? That is the sum of the square errors. I want you to remember two things now $E e$ transpose is e matrix because E is a column vector, but e transpose e is a scalar. I want you to distinguish two things $E e$ transpose e is simply sum of squared errors.

So, expected value of the sum of square errors is given by this because we have already derived expression for e in the previous slide e is equal to I minus P times v . I plug in that value in here I am sorry, I plug in that value in here sorry yeah right this is the right place; I plug in that value in here. I minus P , P is the projection, projection operators idempotent. So, I minus P times I minus P is I minus P , I want you to verify I minus P is idempotent because P is idempotent. v transpose I minus P v is a scalar therefore, a scalar is a quadratic form in v that is equal to its own trace because trace of a scalar is itself that is a mathematical trickery that we bring in to make the analysis more meaningful and H , the trace of a product by rule by the cyclic rule that trace is essentially given by this, this is essentially.

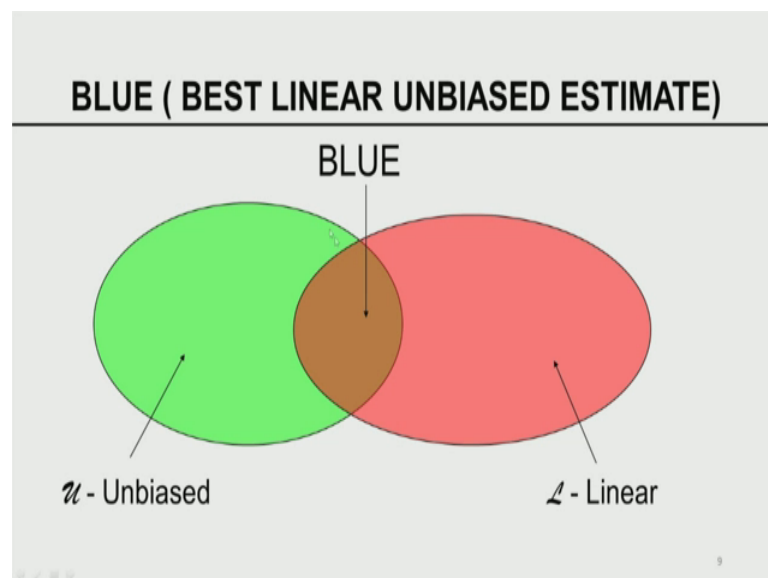
So, I can now ah trace of $v v$ transpose is sigma of I want you to remember E of the $v v$ transpose is equal to said I sigma square i . So, if you use that particular property this becomes this which is essentially a trace of sigma square times trace of I minus P trace of

a sum is the sum of the traces you can see. So, many facts are used trace of I is m because I am concerned with the m by m matrix, trace of P trace of the projection matrix is n that is that is proved in here I have attached the proof for every one of these things therefore, sigmas. So, the expected value of sum of squared errors is sigma square times m minus n . I am concerned with an over determined system where m is greater than n that is the underlying assumption we have make we have made all through.

Therefore, what could be in the structure of estimator for sigma square, this is the resulting structure for the estimator. So, E is the error. Please remember E is the error or the residual in the estimate. So, E is computable $E^T E$ is a scalar that is computable you divide that by m minus n that hence to be an unbiased estimate that tends to, that is a very good unbiased estimate for sigma square.

So, look at this now using this framework we not only can assume sigma square is known and estimate x which is unknown. After having estimated x we can also use the same framework to be able to come to estimating sigma square. So, we can kill 2 birds in one stroke, we can kill 2 birds in one stroke. It is a very powerful device, it is a very powerful device. I hope you do go over this very leisurely and enjoy the moment to understand the beauty and the structure the arguments and the power of this example.

(Refer Slide Time: 28:57)



So, now I am going to summarize everything we have seen. So, I am talking about the world of estimators U is the world of unbiased estimate that is given in green, L is the

world of linear estimates. Some estimates are linear some estimates are non-linear, some estimates are biased some estimates are unbiased. So, outside the green is biased outside the red or is non-linear. The intersection of unbiased and linear is called linear unbiased among all the linear unbiased estimate LUE, L for linear U are unbiased E for estimate. I am interested in the best linear unbiased estimate. Best in what sense, best in the sense of minimum variance a that is what Kalman Filter Rahsaan, Kalman Filter makes estimations which are blue, Kalman filter estimates are blue that is the reason why we use them repeatedly.

And what gives us the power to use Kalman filters and make predictions because it is a very beautiful underlying property called blue, blue is the best linear unbiased estimate we have talked about unbiasedness, we have talked about linear estimate, we have talked about also best relation efficiency linear relate the structure of the estimator unbiased is also relate to the property of the estimate. So, far we have seen via an example properties of least square statistical least square estimate a reader can very easily recognize the parallelism between deterministic least squares and statistical least squares, we not only are able to estimate the unknown we are able to show that is unbiased we are able to compute its variance. We are also able to test the properties of the resulting estimate against the standards unbiasedness and consistency and being able to estimate the unknown covariance and so on. That gave rise to the notion of class of all unbiased estimators, the class of all linear estimators.

Why we are going to be concerned with linear estimators because computationally, linear estimators are easy to compute non-linear estimators requires more time. If I can get very good results with linear estimators why not unbiasedness is very fundamental to any and every estimator whenever wherever possible. So, I am interested in the intersection of these two properties namely linear unbiased estimates LUE. Once you confine yourself to LUE you also want to be able to next talk about the other dimension namely variance of the estimate. If the total variance of the estimate can be minimized then it is going to be optimal in some sense such optimal estimates are call best linear unbiased estimate. So, we are interested in blue. Blue is a key characterization of estimations and as we will see later in when we do the Kalman filtering this notion of blue plays a very fundamental role in the definition of filtering equations known as carbon filter equations.

(Refer Slide Time: 32:31)

GAUSS-MARKOV THEOREM: OPTIMALITY OF LEAST SQUARES VERSION II

- Let x be the unknown being estimated, $x \in \mathbb{R}^n$
- Pick μ and define $\Phi(x) = \mu^T x$, $\mu \in \mathbb{R}^n$
- Consider the problem of estimating $\Phi(x)$
- We are seeking a linear unbiased estimate for $\Phi(x)$
- z is the data and let $a^T z$ be an estimator for $\Phi(x)$, $z \in \mathbb{R}^m$, $a \in \mathbb{R}^m$

$$E[a^T z] = E[a^T (Hx + v)] = a^T H E[x] + a^T E[v] = a^T H x$$

Now, I am going to bring to your attention a very fundamental theorem called Gauss-Markov theorem. Gauss-Markov theorem relates to an inherent optimality of the least square estimates. Thus far we talked about some of the routine properties this Gauss-Markov theorem takes the least square estimates one step further and puts it on a pedestal. So, what does Gauss-Markov theorem says? Among all the linear unbiased estimate then the least square estimate is the best. So, it tries to attest the importance of the least square estimate within the framework of estimation theory. So, that is what is call optimality of least square estimates.

Least square estimates are optimal in a very natural sense the discovery of this natural property of least squares is call Gauss-Markov theorem. Gauss lived in 1700s, Markov lived in early 1900s. Markov is a Russian mathematician, Gauss is a famous German mathematician. So, in their honor I believe it is call Gauss-Markov theorem. So, let us try to formulate the basic aspects of the time of the Gauss-Markov theorem.

Let x be an unknown to be estimated. Let us pick a μ a vector a vector of the same size as x and let us concoct a function ϕ of x which is μ transpose x . μ is a vector of x , x is a variable. So, μ transpose x is a linear function with an inner product. So, ϕ of x is the linear function. So, this is a linear estimator ϕ of x is a linear function is a linear estimator. Now, what is that we want to do? Even though μ is known even though μ is something I picked, x is not known, components of x is not known. So, μ of μ

transpose x is essentially sum of μ_i, x_i . So, instead of estimating x_i I would like to be able to estimate this linear combinations of the components of x which is some of μ_i, x_i . So, the problem is to consider estimating ϕ of x . I hope that is clear now.

Ultimately I am interested in estimating x , but I am considering a special case of estimating ϕ of x . So, you can think of ϕ of x as a functional. So, ϕ is a mapping from \mathbb{R}^n to \mathbb{R} , ϕ is a function to which if I give x it gives me $\mu^T x$ and ϕ is defined by μ . So, that is the idea here. So, what is that we are seeking? We are seeking a linear unbiased estimate for ϕ of x .

Now, how is this related to estimating x ? Suppose, now, you can see μ is a vector which is $\mu_1, \mu_2, \dots, \mu_n$, suppose if I set μ_1 is 1 all the other μ s are 0 $\mu^T x$ becomes x_1 , suppose I pick μ_2 as one everybody else is 0 I pick x_2 and so on. So, if I know how to estimate $\mu^T x$, I know how to estimate every component of x if I can estimate every component of x I can estimate x therefore, the problem of estimating x is embedded in this problem of estimating a functional of x . So, without loss of generality let us consider the problem of estimating a linear functional of x . So, we all know linear function functional, functional is a mapping from a vectors from \mathbb{R}^n to \mathbb{R} and so on.

So, let z , so we talked about the structure of what we want to do. Let Z be the data or the observation that has information about x . Let $a^T z$ be the estimator for ϕ of x . So, ϕ of x is the so, what is that now we want to do? We do not want to we do not want. So, there is x that is ϕ of x we want to be able to estimate ϕ of x to estimate ϕ of x I need an estimator, estimator is a function of the observation. So, I am going to concoct that $a^T z$ where z is the observation be an estimator for ϕ of x . ϕ of x is a scalar x is a vector let a be another vector, I would like to be able to consider $a^T x$ to be a potential candidate for estimating the value of ϕ of x . I hope that is all clear now. a is \mathbb{R}^m , Z is \mathbb{R}^m , μ is \mathbb{R}^n , x is \mathbb{R}^n . So, you can see all the relations all the players in this game.

If $a^T z$ is my estimator I am sorry is the structure of my estimate I can compute the expected value my Z has a standard formula Z is the linear function of x . So, $a^T H x + v$ I am substituting for Z in here the standard one, expectation of the sum is the sum of the expectations, expectation of v is 0, expectation of E of x if the

unbiased estimate is x therefore, $E(a^T z)$ is a transpose Hx . So, that comes very very nicely that comes very nicely.

(Refer Slide Time: 39:18)

GAUSS-MARKOV THEOREM (CONT'D)

- \Rightarrow 1) $a^T z$ is unbiased only if
 - $\Phi(x) = \mu^T x = E(a^T z) = a^T Hx$
 - $\Rightarrow \mu^T = a^T H$ or $H^T a = \mu$
- 2) Since it is unbiased, M.S.E. = variance
 - $\text{var}(a^T z) = E[a^T z - E(a^T z)]^2$

$$= E[a^T (Hx + v) - a^T Hx]^2$$

$$= E[a^T v]^2$$

$$= a^T E(vv^T) a$$

$$= a^T R a$$

$\left\{ \begin{aligned} (a^T v)^2 &= (a^T v)(a^T v) \\ &= (a^T v)(v^T a) \\ &= a^T (vv^T) a \end{aligned} \right.$

Now, $a^T z$ is an unbiased estimate only if $\Phi(x)$ is equal to $\mu^T x$ is equal to expected value of $a^T z$ is equal to $a^T Hx$. Please remember for the unbiased estimate it must be equal to the value I am seeking therefore, if I were to relate these two quantities that essentially tell you μ^T must be equal to $a^T H$ or $H^T a = \mu$. So, that provides the relation between the a vector that I use in the estimator and the μ vector I use in the definition of the functional to be estimated. So, the a and μ cannot be 2 distinct they must be related through H , H is a function that relates the state to the observation I hope that is clear.

Since, it is unbiased estimator we already know the mean square value is equal to variance. Please remember in one of the earlier lecture we have already said that if the estimate is unbiased means square error is equal to the variance. So, minimizing the mean square error is equal to minimizing the variance. Therefore, I am not going to compute the variance of the estimate $a^T z$. The variance of the estimate is equal to expected value of the error square. I am again going to substitute for z which is given by this.

We already know $a^T z$ expected value is $a^T Hx$. So, if you simplify this you get this a is a constant, v is the is a random vector, $a^T z$ so we can write like

this a transpose v square is equal to a transpose v transpose I am sorry a transpose one second; times a times a transpose v here a transpose v times v transpose a because the symmetry the inner product or symmetric therefore, this can be written as a transpose v v transpose a. If I took the expectations I get what is this. So, to go from here to here this is the simple algebra. So, like this I would like you to do all the basic algebra to see which steps come from the previous one why and how and the variance of v is R. So, this is a transpose Ra, a transpose Ra, R is a matrix a transpose Ra is a quadratic form. So, a transpose is this, this is the matrix that is the vector. So, this is a scalar. So, the variance of a scalar is a scalar everything matches. So, I have computed an expression for the variance of the estimate which is a transpose a, the variance of the estimators a transpose Ra.

Yes, I know I am going a little faster, but what I am trying to tell is nothing unknown to you except for the algebra. So, I would like you to be able to go through these algebra and unconvinced yourself in my regular class where I teach a student I also do not go over this algebra in the class because it is a part of the folklore of the course where you need to struggle learn to do many of the matrix manipulations and these exercises these simplifications are very educative I hope you will pursue that.

(Refer Slide Time: 43:16)

GAUSS-MARKOV THEOREM (CONT'D)

- Seek to minimize $a^T R a$ when $H^T a = \mu$

$L(a, \lambda) = a^T R a - \lambda^T (H^T a - \mu)$, Lagrangian, $\lambda \in \mathbb{R}^n$
 $\nabla_a L(a, \lambda) = 2 R a - H \lambda = 0$
 $\nabla_\lambda L(a, \lambda) = H^T a - \mu = 0$
 $\therefore a = \frac{1}{2} (R^{-1} H \lambda)$, $H^T a = \mu$
 $\frac{1}{2} (H^T R^{-1} H) \lambda = \mu$
 $\lambda = 2 [H^T R^{-1} H]^{-1} \mu$
 $a = R^{-1} H (H^T R^{-1} H)^{-1} \mu$ (*)

Handwritten notes and diagram:
 $z = Hx + v$
 $\phi(x) = x^T x$
 $a^T a$
 $\text{Min. w.r.t. } a$
 $\text{LINEAR UNBIASED BLUE}$
 $H^T a = \mu$

Now, what is that, what do you want to do? We want to be able to get the best estimate, best estimate is in the sense it minimizes the variance of the estimate. So, what is the

estimate? The estimate is $\mu^T x$, I am sorry the estimate is a transpose z . What is the functional being estimated? $\mu^T x$. What is the variance of the estimate? The variance of the estimate we have already seen is given by a transpose Ra .

Now, I would like to minimize this variance with respect to a , a is the variable you remember a is a vector I picked to make to design my estimator which is a transpose z z is given a is something I picked now I am going to take responsibility in how to pick a what are the conditions for the choice of a that is where we are coming to. So, I need to be able to minimize this quadratic function a with this quadratic function with respect to a , but a is not a free variable. Please remember the unbiasedness requires a and μ to be related $H^T a$ must be μ therefore, I am not interested in minimizing a for any a I am interested in minimizing $a^T R a$ under the condition that $H^T a$ is μ . So, this is the constraint, this is an equality constraint. So, I am interested in minimizing something with equality constraints, now how many times we have seen this constraints minimization the equality constraint what is the rule Lagrangian multiplier.

Now, you know the importance of multivariate calculus optimization theory. So, formulate this as a Lagrangian problem, a is a free variable with respect to which I am going to do the minimization λ is a vector with respect to which I am going to build my constraint into my Lagrangian function. So, this is the Lagrangian function. Compute the gradient of the Lagrangian function with respect to a , compute the gradient of the Lagrangian function of λ equate them to 0, solve them simultaneously the solution process is given here ultimately the minimizing a is given by a is equal to $R^{-1} H (H^T R^{-1} H)^{-1} \mu$, wow. Look at this now. I want to go back I know there is lot of lot of side steps there is involved, z is equal to $Hx + v$, I do not want to estimate x by I want to estimate a functional of x which is $\mu^T x$.

To estimate this I picked an estimator to be a transpose z . This estimator is unbiased, this estimator has a variance which is a transpose Ra I am interest. So, this estimator is linear I am sorry this estimator is linear it is also unbiased therefore, it is already LUE. I want to be able to introduce a B to that I want to make it a blue to make it a blue I have to minimize this with respect to a , but the unbiasedness requires a and μ to be a related $H^T a$ must be equal to μ therefore, this is the constraint this is the function to be minimized. So, I combine the two, I minimize this with this as a constraint using

Lagrangian multiplier technique with a little bit of algebra and simplification I have now found a formula for the optimal a .

What is this value? The a that I you should use in my estimator is not coming out of the blue sky, but it is going to be decided by this structure which is given by star. Now, let us look at the structure star it depends on R the covariance matrix of the noise it depends on H which is the linear map between the model space in the observation space. It also depends on μ , what is μ ? μ is the coefficient of the functional that we originally started that we originally started with. So, we have solved the problem. If I pick my a to be this and use that a in my a transpose z the resulting estimator is not only is not only is not only linear it is unbiased it is also minimum variance it is also minimum variance.

(Refer Slide Time: 48:04)

GAUSS-MARKOV THEOREM (CONT'D)

- \therefore Linear, unbiased minimum variance estimate of $\Phi(x) = \mu^T x$ is

$$a^T z = \mu^T (H^T R^{-1} H)^{-1} H^T R^{-1} z = \mu^T \hat{X}_{LS}$$

Least squares estimate
- \therefore If $\mu = (1, 0, \dots, 0)^T$
- \Rightarrow is the best estimate of x_1 and so on.

13

So, linear unbiased minimum variance estimate of ϕ of x μ transpose x is equal to is given by this expression. So, you can readily see this is the least square estimate, the H transpose R inverse H inverse H transpose R inverse z that is the least square estimate. So, it is μ transpose times the least square estimate. So, μ transpose z sorry μ transpose z that is the structure of that is the structure of the estimate that comes out.

So, if I pick μ to these $1 \ 1 \ 0 \ 0 \ 0$ I get the best estimate of x_1 , if I picked $0 \ 1 \ 0 \ 0 \ 0$ I pick x_2 and so on. So, I can by this formulation can estimate any component of x , I can estimate all components of x each of the component of x will be, it will be I can estimate

use m and the properties of the estimate is blue. So, that is the basic idea of, that is the basic idea of the setup, that is the basic idea of the set up.

(Refer Slide Time: 49:30)

NOTE

- If v is $N(0, R)$, then \hat{X}_{LS} is the best among all estimators - Rao-Blackwell Theorem.
- If v is not Gaussian, there exists non-linear estimates whose variance is smaller than linear estimate.

14

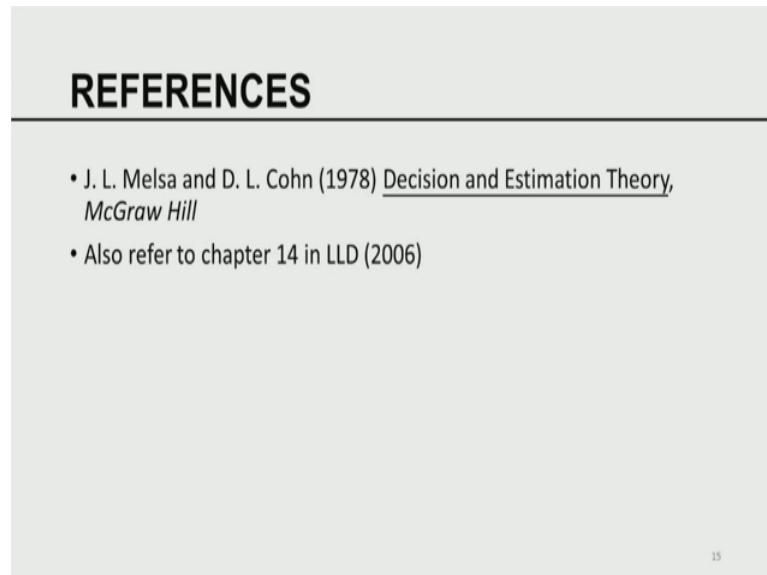
So, I would like to take you through some of the slides. So, in the Gauss-Markov theorem we have we have established linearity, we have establish unbiasedness we have also now established it being the best, being the best.

Now, I am going to further embellish this with some other claims some extensions if the noise. So, until I good until now I only assume the noise v has a mean 0 and the covariance R , we did not explicitly invoke to the Gaussian nature of v . So, until now all the analysis depends on the fact that v has a mean 0 and the covariance R it need not this will be Gaussian. So, in addition if somebody tells you v is indeed Gaussian as in this case there is a further extension of this Gauss-Markov theorem and that Gauss-Markov theorem is the extension is called Rao-Blackwell theorem.

The Rao-Blackwell theorem essentially tells you the least square estimate with the Gaussian noise is the best among all estimates, linear and non-linear. If v is not Gaussian there exists there exists non-linear estimates whose variance is smaller than the linear estimate these are all deeper results from statistical estimation theory, these results can be gleaned from the book by Dr Rao, C R Rao we referred to in the previous talk. So, I believe I had given you, a reasonably good picture of the properties of statistical least

squares as supposed to deterministic least squares that we saw within the context of, within the context of deterministic static estimation theory.

(Refer Slide Time: 51:15)



My favorite book on this topic is again Melsa and Cohn is the same book that I had already referred to earlier. Decision and Estimation Theory by McGraw Hill is a small little book no more than 200 pages, but its beautifully written it has an engineering flavor and whenever I have trouble with these I always fall back on Melsa and Cohn or Melsa and stage in Melsa these are the two books and we also refer to many of these things in our chapter 14 of our book on data assimilation. So, with these two you should be able to get a rather complete picture of the basic elements and properties of statistical least square estimation and its underlying properties.

Thank you.