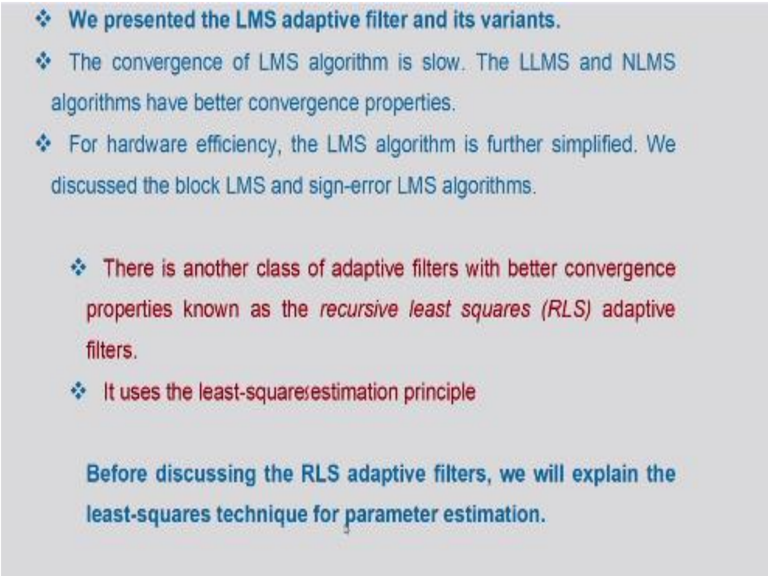
Statistical Signal Processing
Prof. Prabin Kumar Bora
Department of Electronics and Electrical Engineering
Indian Institute of Science – Guwahati

Lecture -30
Adaptive Filters 4 (Contd.)

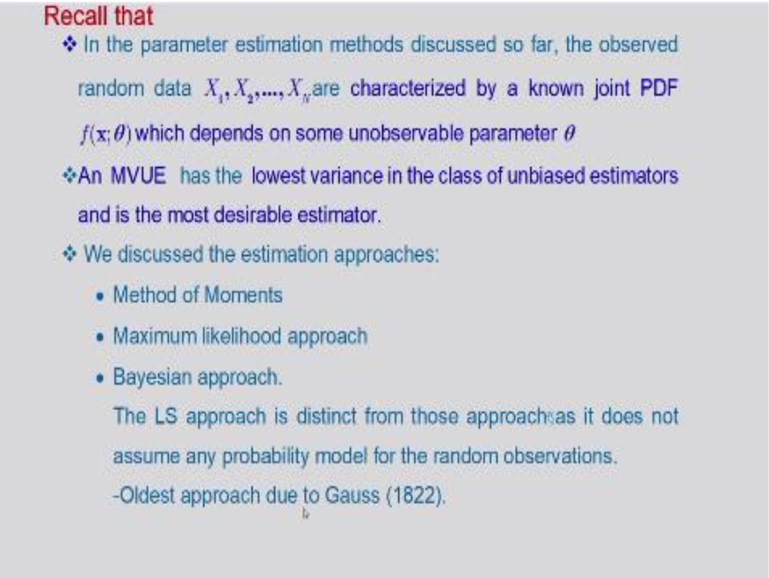
(Refer Slide Time: 00:38)



Hello students. Welcome to this lecture on Least Squares Estimator. We presented the LMS adaptive filter and its variants in the previous lectures. The convergence of LMS algorithm is slow. The leaky LMS and the normalize LMS algorithms have better convergence properties. For hardware efficiency the LMS algorithm is further simplified. We discussed the block LMS and sign-error LMS algorithms.

There is another class of adaptive filters with better convergence properties. This class is known as the recursive least squares adaptive filters. It uses the least squares estimation principle. Before discussing the RLS adaptive filters we will explain the least squares techniques for parameter estimation. Least squares is a parameter estimation technique we will first examine that.

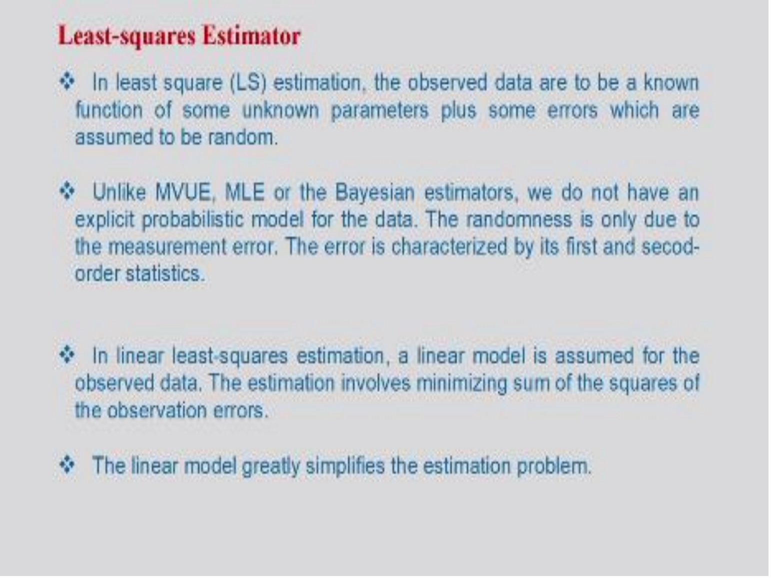
(Refer Slide Time: 02:03)



Recall that in the parameter estimation methods discuss so far the observed random data X1, X2 upto XN are characterized by known joint PDF f x theta or a known joint (()) (02:20) which depends on some unobservable parameter theta the parameter theta here is unknown and unobservable. An MVUE minimum variance unbiased estimator has the lowest variance has the lowest variance in the class of unbiased estimators and is the most desirable estimator.

We discussed about various approaches to find MVUE. We discussed the estimation approaches mentioned here like method of moments, maximum likelihood approach, Bayesian approach like minimum mean square error estimation, maximum a posterior probability estimation etcetera. The least squares approach is distinct from those approaches and it does not assume any probability model for the random observations. This is the oldest approach and due to Gauss in 1822.

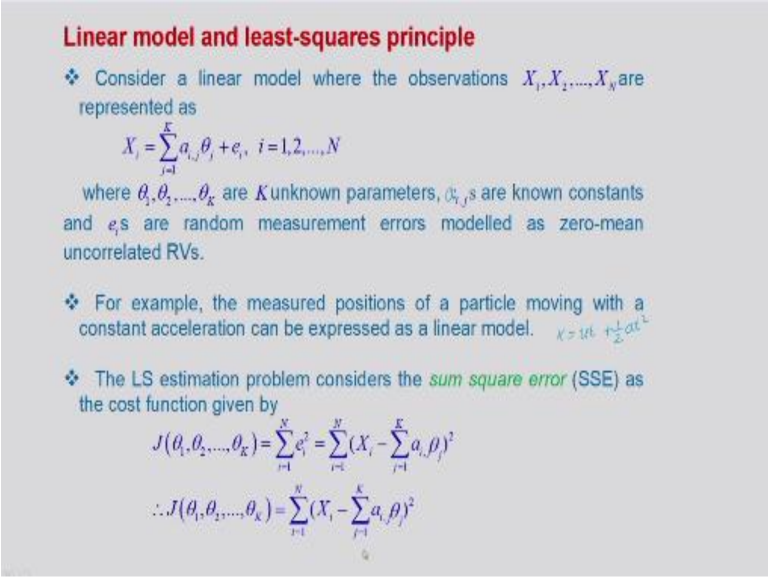
(Refer Slide Time: 03:30)



Let us discuss about least squares estimator. In LS estimation the observed data are to be a known function of some unknown parameters plus some errors which are assumed to be random plus some noise. Unlike MVUE MLE or Bayesian estimators we do not have an explicit probabilistic model this is the distinction. The randomness is only due to the measurement error. The error is characterized by its first.

And second order statistics that is important because only mean or second order statistics like variance, covariance etcetera are used. In linear least square estimation a linear model is assumed for the observed data. We will discuss about linear least squares estimation only. The estimation involves minimizing the sum of the squares of observation errors. So estimation involves therefore here estimation means minimizing the sum of the squares of the observation errors. The linear model greatly simplifies the estimation problem. Therefore, we have a linear model and the parameters are to be estimated.

(Refer Slide Time: 05:14)



Let us consider a linear model where the observation X1, X2 upto XN are represented as Xi that = summation aij theta j j going from 1 to K + ei where i = 1, 2 up to N this is the model of the signal. This is a linear model because the observed data is a linear function of the parameters plus some noise. Here theta 1, theta 2 upto theta K are unknown parameters aijs are known constants and eis are random measurement errors modeled as zero-mean uncorrelated random variable.

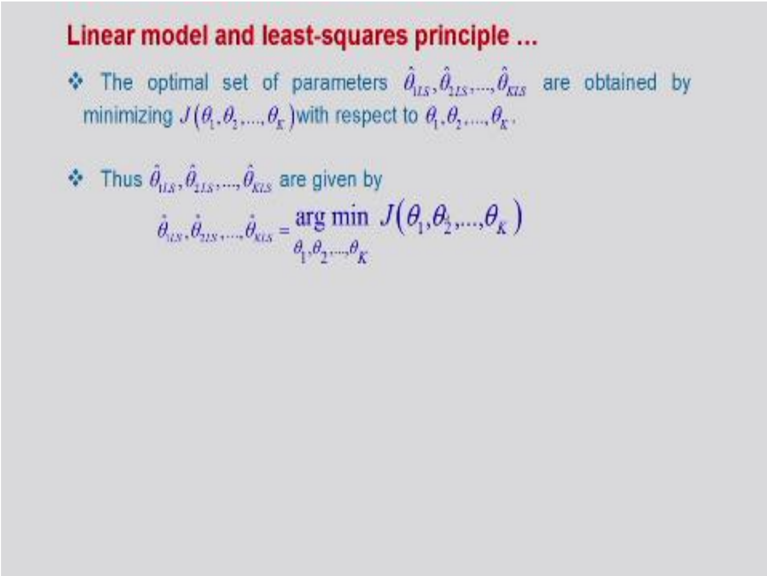
Therefore, the error is an zero-mean random variable and its covariance of any two errors is equal to zero. Here theta 1, theta 2 upto theta K are K unknown parameters aijs aijs are known constants and eis are random measurement errors modeled as zero-mean uncorrelated random variables. So eis are zero-mean and their covariance, covariance between any two samples is equal to zero.

For example, the measured position of a particle moving with a constant acceleration can be expressed as a linear model. Suppose position X = ut + ½ at square suppose this is the model then this u (()) (07:25) parameters therefore x is a linear combination of (()) (07:31). The LS estimation problem considers the sum square error. So this is SSE everywhere it as SSE so sum square error square of the errors.

And then sum up so that way this sum square error this is the cost function J theta 1, theta 2 up to theta K that is = summation ei square for all N observation. So that way summation ei square i going from 1 to N that = Xi – summation aij theta j j going from 1 to K whole square i going from 1 to N. So this is the because Xi is given like this therefore ei we can find out by subtracting this term from Xi.

So that way this cost function this sum square error and it is given by this so this is the cost function J theta 1, theta 2 upto theta K = summation Xi – summation aij theta j j going from 1 to K whole square i going from 1 to N. This is the cost function corresponding to LS estimation.

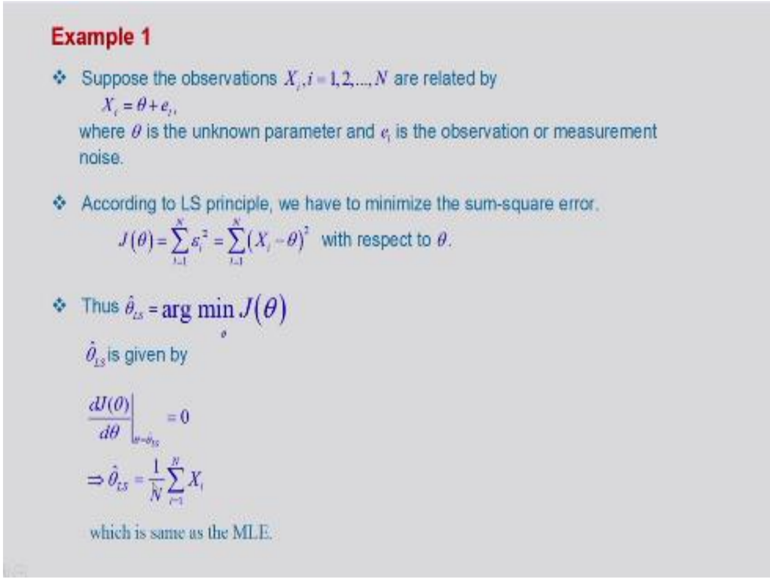
(Refer Slide Time: 09:04)



The optimal set of parameters that is theta 1 hat LS, theta 2 hat LS upto theta K hat LS are obtained by minimizing this cost function sum square error with respect to the parameters theta 1, theta 2 upto theta K. Thus, we can write that theta 1 hat LS, theta 2 hat LS upto theta K hat LS = arg min theta 1, theta 2 upto theta K of J theta 1, theta 2 upto theta K. So the parameters which minimizes this cost function (()) (09:51) LS estimators.

Since our cost function is a quadric cost function we can find out the LS estimators uniquely by applying the differentiation principle. So we will take the partial derivative to J with respect to all the parameters and set them to 0.

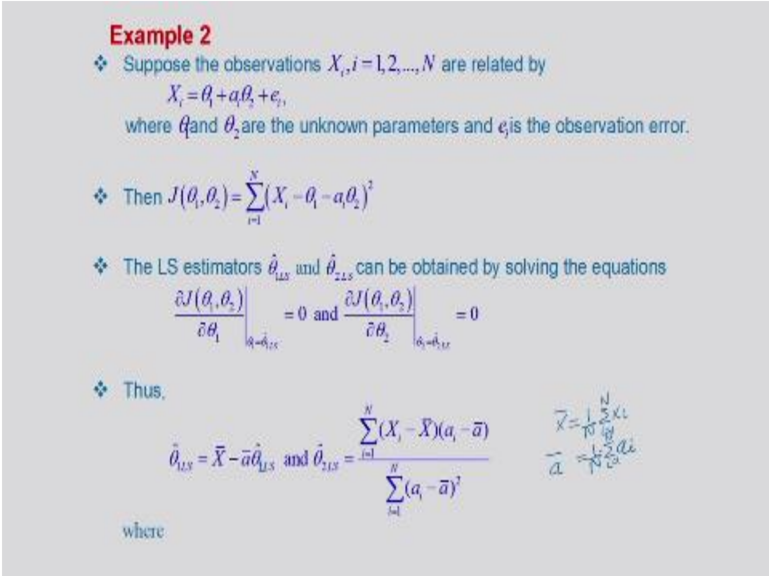
(Refer Slide Time: 10:20)



We will consider one example suppose the observations Xi are related by this simple model Xi = theta + ei where theta is the unknown parameter ei is the observation or measurement noise. Now this is the sum square error J theta = summation Xi – theta whole square i going from 1 to N we have to minimize this sum square error with respect to theta. Thus, theta hat LS = arg min theta of J theta.

So we have to minimize this J theta with respect to theta and it is given by del J theta del theta at theta hat LS = 0 and taking the derivative of J theta with respect to theta and setting it to 0 we get theta hat LS = 1/N into summation Xi i going from 1 to N. So this is theta hat LS which we obtain by differentiating this function and setting it to 0 and we observed that theta hat LS is same as the MLE because they are in that case also we found that theta hat MLE = 1/N summation Xi i going from 1 to N.

(Refer Slide Time: 12:06)

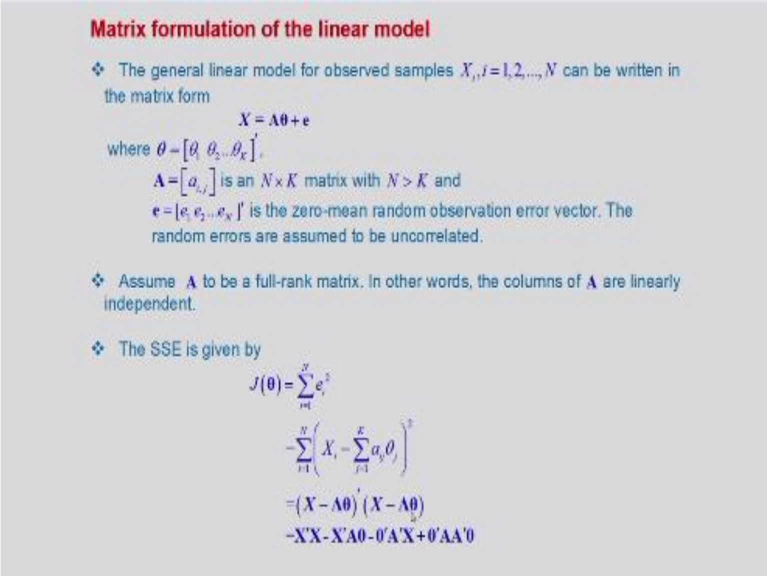


We will consider another example suppose the observations Xi i going from 1 to N are related by this relationship Xi = theta 1 + ai theta 2 + ei. So this is for i = 1, 2 upto N where theta 1 and theta 2 are unknown parameters and ei is the observation noise then the theta 1, theta 2 is given by this summation Xi this is the data – this model theta 1 – ai theta 2 whole square i going from 1 to N.

So LS estimators two terms we have to estimate theta 1 hat LS and theta 2 hat LS and they are obtained by taking the partial derivative with respect to theta 1 that is = 0 and taking the partial derivative with respect to theta 2 and that = 0 and ultimately we can take the partial derivative and solve these two equations. We can get theta 1 hat LS = X bar – a bar into theta 2 hat LS and theta 2 hat LS = summation Xi – X bar into ai – a bar i going from 1 to N/summation ai – a bar whole square i going from 1 to N.

So these are the expression for theta 1 hat LS and theta 2 hat LS where X bar and a bar are the sample mean. So that way X bar = summation Xi i = 1 to N/N and ai = summation ai i = 1 to N/N.

(Refer Slide Time: 14:23)



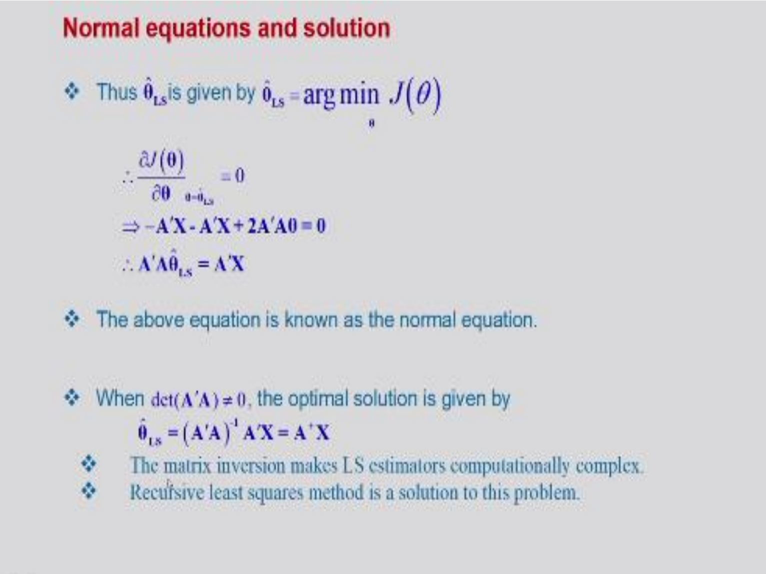
Now we will have a matrix formulation of the linear model. The general linear model of observed samples Xi i going from 1 to N can be written in the matrix form as X = A matrix times theta vector + e vector. This is the linear model in the matrix form. Here theta equal to the parameter vector comprising of theta 1, theta 2 upto theta K. So this is the theta vector similarly A matrix is the matrix of aij it is a N/K matrix N is the number of observation.

And K is the number of parameters and with the assuming that N > K. e is again corresponding to each observation we have one error therefore e is an N dimensional error vector comprising of e1, e2 upto eN ei are zero mean random variable. Further, the random errors are assumed to be uncorrelated. They are uncorrelated that is covariance of any two element will be = 0 and we also assume that they are of constant variance.

We assume A to be a full rank matrix. In other words, the columns of A are linearly independent. This is important for having the solution. The sum square error is given by J theta that = ei square i going from 1 to N and that = summation Xi – summation aij theta j j going from 1 to K whole square i going from 1 to N this is the cost function and this in matrix notation we can write X – A theta transpose into X – A theta.

So this summation we can write in the matrix form like this X – A theta transpose into X – A theta and if we expand this product we will get this as X transpose because transpose is here X transpose into X – X transpose into A theta now transpose of A theta is theta transpose A transpose. So that way theta transpose A transpose into X + this - - will be + theta transpose A A transpose into theta. So product of this transpose into A theta.

(Refer Slide Time: 17:36)



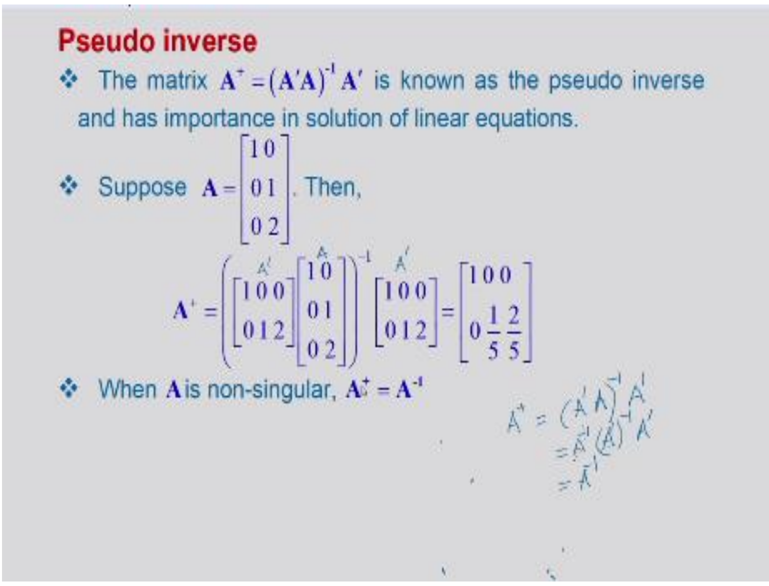
Now the estimator is given by theta hat LS = arg min theta of J theta and we can obtain by the relationship that is del J theta del theta hat theta hat LS = 0. Now this is a partial derivative with respect to theta vector that way we have to do the partial differentiation with respect to vector. So this we can carry out for this expression this is a constant term so this term will have zero partial derivative similarly corresponding to this term we will have A transpose X.

Similarly, for second term then this term also we will have the derivative A transpose X like that and the third term will be 2 times A transpose A into theta and that must be = 0. So this is the result we get after partial derivative with respect to theta vector. From this we get a equation because here A transpose X again A transpose X so it will be 2 A transpose X and there is 2 A transpose A theta.

So 2, 2 will get cancelled therefore we will get this relationship. A transpose A theta hat LS = A transpose X. This is the equation government LS estimation. So that means we have to pre multiply theta hat LS/A transpose A and right hand side is A transpose X this equation is known as the normal equation matrix form of normal equation. When determinant of A transpose A is not = 0 the optimal solution is given by theta hat LS = A transpose A whole inverse into A transpose into X and this we denote by A + into X.

So this is the least square solution and here this matrix inversion is there matrix inversion makes LS estimators computationally complex because when this matrix dimension is very high inversion is a big problem. Recursive least squares method is a solution to this problem which we will be discussing in the next lecture.

(Refer Slide Time: 20:46)

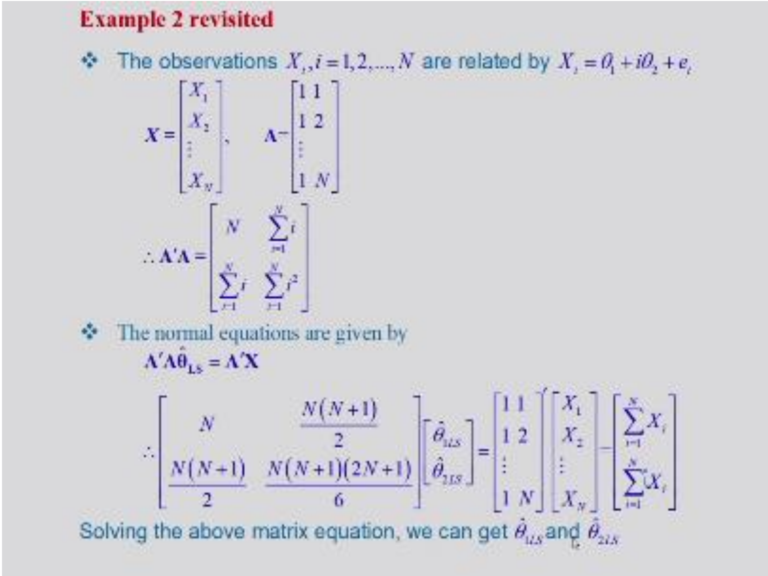


The quantity A + that = A transpose A whole inverse into A transpose is known as the pseudo inverse and has importance in solution of linear equation. In many cases of linear equations solution can be obtained in terms of pseudo inverse. Suppose A = this matrix, matrix comprising of 1 0 0 1 0 2. So 2 columns are there this 2 columns are independent and here pseudo inverse A that is A + will be = A transpose into A then whole inverse into A transpose so this is the expression for pseudo inverse.

So that way if I take the transpose of this I will get 1 0 0 0 1 2 and then A is this whole inverse multiplied by A transpose A transpose is 1 0 0 0 1 2. So if I carry out all this operations then we will get that pseudo inverse of A = 1 0 0 0 1/5 2/5. When A is non singular pseudo inverse of A will be simply A inverse itself because in that case A + will be = A transpose A inverse into A transpose.

And in that case now we can inverse like this so we can write this as A inverse then A transpose inverse into A transpose so this is identify matrix so we will have simply A inverse. So when A is non singular so pseudo inverse is A inverse itself.

(Refer Slide Time: 23:25)



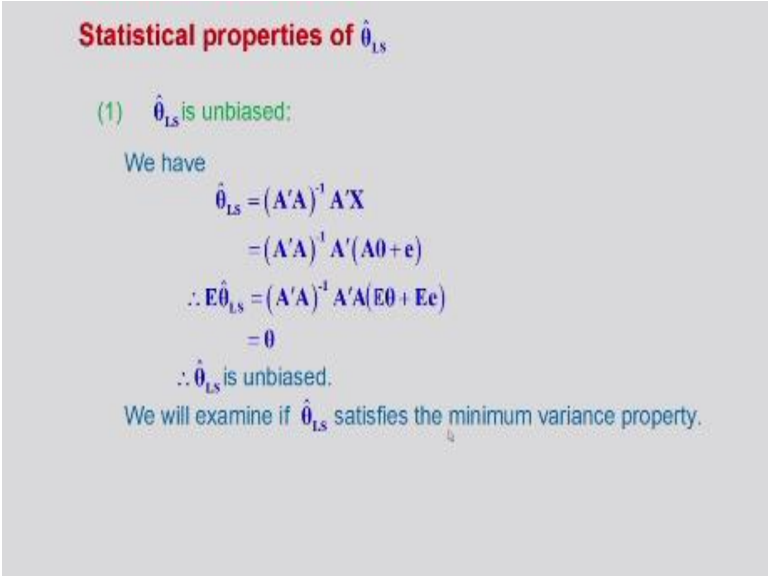
Let us consider example 2 again Xi are related by Xi = theta 1 + I theta 2 suppose ai = i + ei. So in that case X = this X factor is given by this A matrix is given by this and therefore A transpose A will get like this. First element will be N second element will be summation i i going from 1 to N and this element will be = summation i i going from i to N and this element will be summation i square i going from 1 to N.

This is the A transpose A so A is like this A transpose A will be like this. Therefore, we will have the normal equation now, normal equation is A transpose A multiplied by theta hat LS = A transpose X. This is the normal equation for least square estimation and we know now A transpose A this is given by this matrix into theta hat LS 2 component theta 1 hat LS, theta 2 hat LS.

And that must be = A transpose that is transpose of this matrix multiplied by X1, X2 upto XN this vector. So if we carry out this multiplication transpose of this matrix into this vector will get this right hand side will be = summation Xi i going from 1 to N and here it will be summation i Xi i going from 1 to N because this column has 1, 2 upto N therefore if I multiply transpose of this by this I will get this expression.

So that way we have now this set of expression it is (()) (25:34) matrix. Solving the above matrix equation, we can get theta 1 hat LS and theta 2 hat LS.

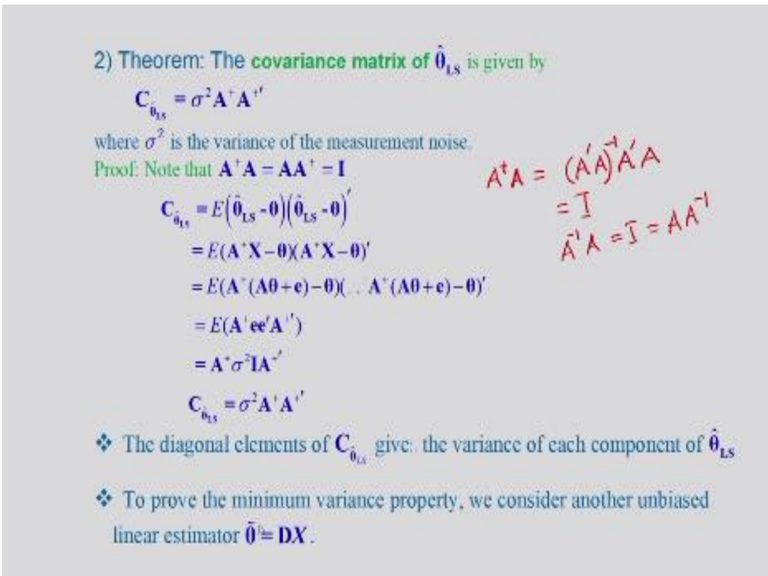
(Refer Slide Time: 25:47)



Let us see this statistical properties of theta hat LS. Number one is theta hat LS is unbiased. We have theta hat LS that is by definition A transpose A whole inverse into A transpose X that = A transpose whole inverse into A transpose now we know that X = A theta + error so that way we will have E of theta hat LS. Therefore, will be = A transpose A inverse A transpose into A into E of theta expected value of theta + E of e.

This part is 0 and this theta is a constant here so if I carry out this and this, this will be identity matrix and therefore we will get simply theta. So E of theta hat LS will be = theta. Therefore, theta hat LS is unbiased. So one important properly that least square estimator is an unbiased estimator. So this we proofed here. We will examine if theta hat LS satisfies the minimum variance property that is important property we have to examine.

(Refer Slide Time: 27:26)



We have proof that theta hat LS is an unbiased estimator. We will proof the theorem the covariance matrix of theta hat LS is given by C theta hat LS = sigma square A + into A + transpose where sigma square is the variance of the measurement noise and A + as you know it is this pseudo inverse of A. To proof this, first we note that A + A = A A + = I. So A + A that = A transpose A inverse A transpose that is A + into A.

So we see that this part is the inverse of this part therefore this will be simply = I matrix identity matrix and we can proof the other way A + into A will be also = I. This property is similar to the property of inverse that is A inverse into A A = I = AA inverse. So here this property is satisfied even by pseudo inverse. There A + A = AA + = I. Now let us see C theta hat LS by definition this is the expected value of theta hat LS – theta into theta hat LS – theta transpose.

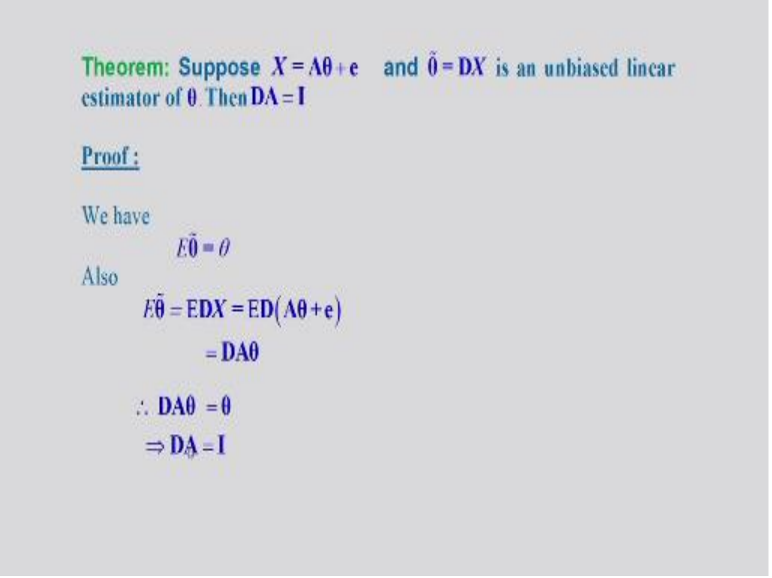
Now putting the value for theta hat LS that = A + X therefore this expression will be = A + X – theta into A + X – theta transpose. Now we know that X = A theta + E error vector. Therefore, this expression will be = E of A + into A theta + e – theta into A + then X = A theta + e – theta and whole transpose. So that way this expression can be written like this. Now we know that A + into A = I similarly here also A + A = I.

And therefore this first part will be simply = theta A + A into theta will be = theta – theta will get cancel. Similarly, here also A + A into theta itself this theta and this theta will get cancel. Therefore, this whole expression will be = E of A + ee transpose into A + transpose because here transpose is there so it will come to the right and then with transpose. Now we know that this error vector is uncorrelated with constant variance sigma square.

Therefore, the expected value of ee transpose will be = sigma square into I. Therefore, C theta hat LS will be = A + sigma square I into A + transpose and this we can write as because sigma square I if we multiply by I I will get the same matrix so it will be sigma square into A + into A + transpose. So this is the covariance matrix of theta hat LS. Now the diagonal elements of C theta hat LS give the variance of each component of theta hat LS.

So that is important now this is a covariance matrix and the diagonal elements are the variance. Therefore, the diagonal elements of C theta hat LS gives the variance of each component of theta hat LS. To proof the minimum variance property, we consider another unbiased linear estimator theta tilde = DX. Suppose this is an unbiased estimator and with respect to this unbiased estimator we will discuss the minimum variance property.

(Refer Slide Time: 33:03)

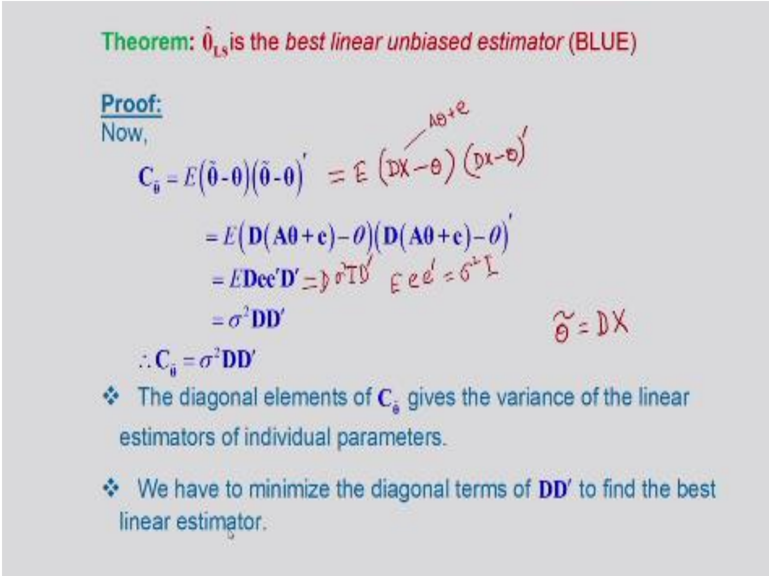


We will proof this result first. Suppose X = A theta + error vector this is the model for LS estimation and theta tilde = DX is then unbiased linear estimator of theta because it is a linear combination of X so therefore it is a linear estimator and it is assumed to be unbiased then D into A will be = I any linear estimator into D model matrix = I. We will proof the result we have the expected value of theta tilde = theta because theta tilde is an unbiased estimator.

Also E of theta tilde = E of DX by definition theta tilde = DX and now put X = A theta + e vector. Therefore, E of DX will be = E of D into A theta + e vector and we know that E of e vector will be = 0 vector. Therefore, we will be simply having DA into theta because theta is a constant quantity. So E of theta tilde will be = DA into theta, but we know that E of theta tilde = theta.

Therefore, this DA theta must be = theta which implies that this DA must be = I identity matrix. So that way we have proofed one important result that any linear unbiased estimator of theta will satisfy this property DA = I.

(Refer Slide Time: 35:09)



Now we will proof our final result theta hat LS is the best linear unbiased estimator. So this is the estimator with a minimum variance among the class of linear unbiased estimator. We will explain this again proof C theta that covariance matrix of theta tilde = expected value of theta tilde – theta into theta tilde – theta transpose by definition and theta tilde = D so this we can write as E of theta tilde = D into X – theta into DX – theta transpose.

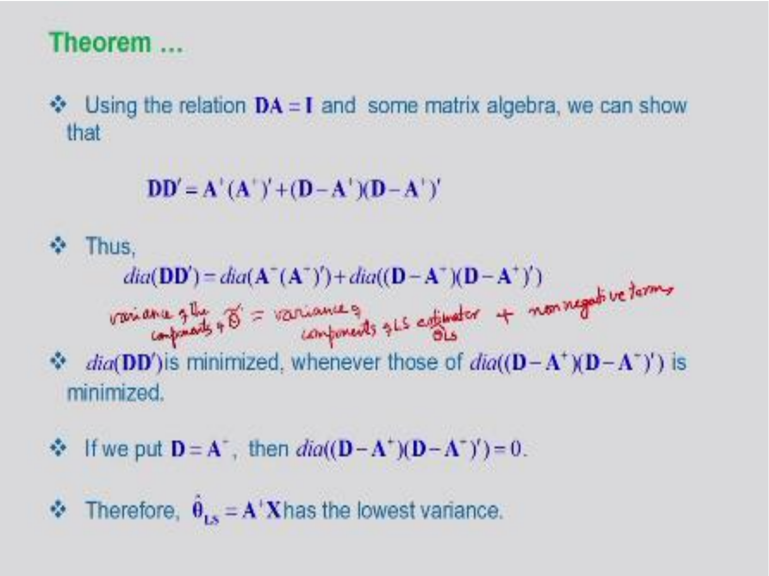
So we have written theta tilde = DX. Now we know that X = A theta + e vector. So therefore we will get this expression = E of expected value of D into A theta + error vector – theta into D into A theta + error vector – theta transpose. Now we will be using the previous identity D into A = I. Therefore, this part will be D into A will be = I therefore this will be simply theta and this theta and this theta will get cancel.

And therefore we will be left with D multiplied by e here. Similarly, from this side also since D into A = I therefore this theta and – theta will get cancel and we will have D into e transpose. So (()) (37:33) transpose will be given by e transpose into D transpose. So therefore this C of theta tilde = E of expected value of D into ee transpose into D transpose. Now we know that this ee transpose is a diagonal matrix we know expected value of ee transpose = sigma square times I matrix.

Therefore, here it will be simply D times sigma square I into D transpose and this will give me sigma square into D D transpose. Therefore, for any unbiased estimator theta tilde the covariance matrix C theta tilde is given be sigma square into D into D transpose and the diagonal elements of C theta tilde gives the variance of the linear estimators of individual parameters. Suppose DX theta tilde = DX and this is a vector and it component will have some variance and those variances will be given by the diagonal elements of the covariance matrix.

We have to minimize the diagonal terms of DD transpose to find the best linear estimator. We know that D is a linear estimator, but to be best linear estimator the variance should be minimum therefore and we know the variance is given by the diagonal elements of C theta tilde therefore we have to minimize the diagonal terms of DD transpose to find the best linear estimator.

(Refer Slide Time: 39:54)

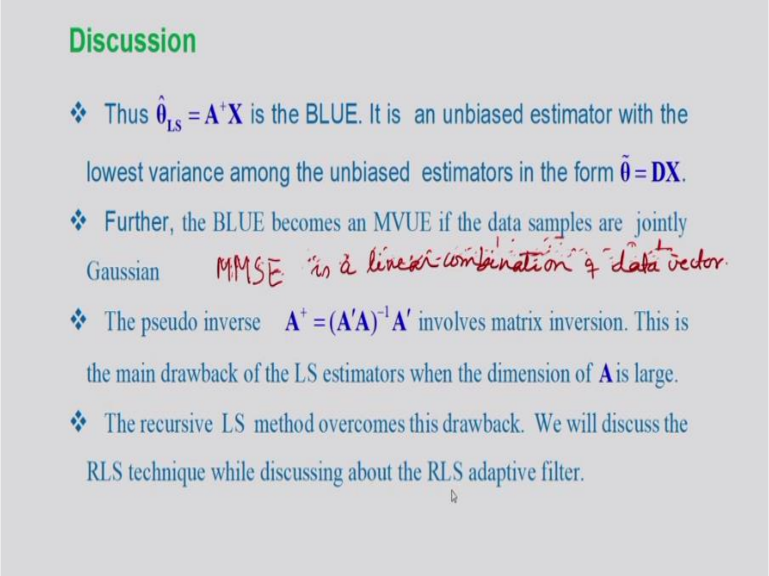


Now using the relation DA = I and some matrix algebra we can show that this is very important relationship DD transpose = A + into A + transpose + D – A + into D – A + transpose. So this is the covariance matrix of the least square estimators and this is some additional terms, but we know that this is a non negative quantity. Thus, the diagonal elements of DD transpose = diagonal elements of A + into A + transpose + the diagonal elements of this matrix D – A + into D – A + transpose.

So what does it means variance of the components of theta tilde = variance of components of LS estimators that is theta hat LS + some non negative terms. Therefore, this term will be minimize whenever this part is minimum variance of components of theta tilde will be minimum when this part the corresponding diagonal elements of D – A + D – A + transpose is minimum this is important observation.

That is diagonal of DD transpose is minimized whenever those of the diagonal elements of D – A + into D – A + transpose is minimize. If we put D = A + then the diagonal elements D – A + into D – A + transpose will be = 0. So this term will be lowest so when D = A + that means when DD lease square estimator A + D = A + then the variance will be minimum. Therefore, we conclude that theta hat LS = A + X that is pseudo inverse of A into X has the lowest variance. Thus, we have proofed that theta hat LS has the lowest variance among the unbiased linear estimators.

(Refer Slide Time: 43:25)



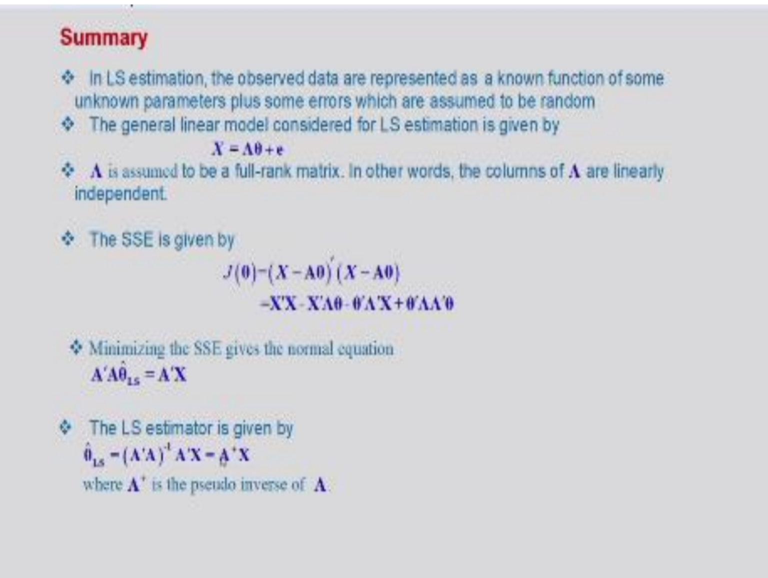
Therefore, theta hat LS that = A + X is the BLUE best linear unbiased estimator. It is an unbiased estimator with the lowest variance among the unbiased estimators in the form theta tilde = DX. We are considering estimators in this form only that is the linear estimators and with additional property that this estimator is unbiased. So this makes least square estimators very attractive.

It is the best linear unbiased estimator. Further, the BLUE becomes MVUE recall MVUE that is minimum variance unbiased estimator. So BLUE will become MVUE if the data samples are jointly Gaussian. Then MMSE minimum mean square error estimator is a linear combination of data vectors and here also this least square estimator is a linear combination of data vector.

Therefore, this estimator will become MMSE minimum mean square error estimator and further it is unbiased therefore it will be minimum variance unbiased estimator. A pseudo inverse A + = A transpose A inverse into A transpose involves matrix inversion. So this is the matrix inversion we have to do. This is the main drawback of the LS estimators when the dimension of A is large.

Because matrix inversion is computationally highly expensive. The recursive LS method overcome this drawback. There is a technique called recursive least squares estimation which overcomes this drawback. We will discuss the RLS technique while discussing about the RLS adaptive filter.

(Refer Slide Time: 46:00)

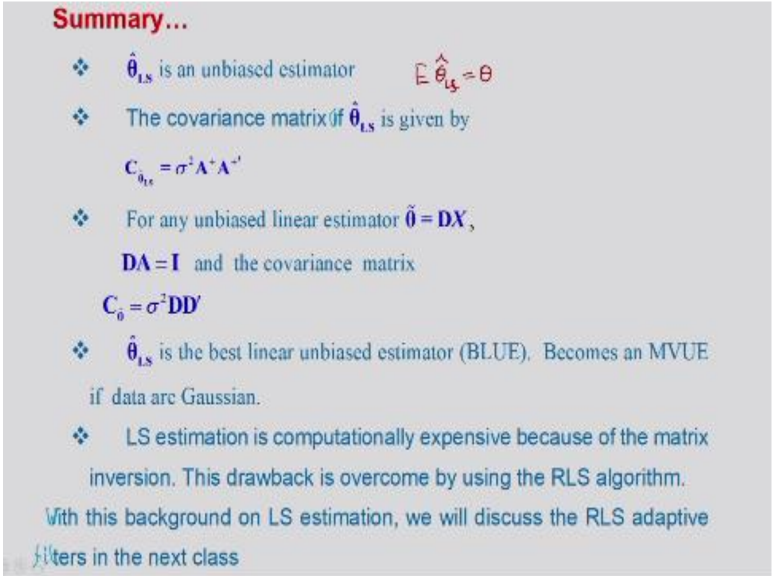


Let us summarize the lecture in LS estimation the observed data are represented as a known function some unknown parameters plus some errors which are assumed to be random. So only error is random otherwise it is a deterministic model. This general linear model considered for LS estimation is given by X = A theta + e vector. So this is the model so theta is the vector of unknown parameters A is a matrix and e is a error vector which is uncorrelated zero mean and with constant variance.

A is assumed to be a full-rank matrix in other words the columns of A are linearly independent. This is required to have the least square solution. Now here the cost function is sum square error and it is given by J theta = X – A theta transpose into X – A theta and this we expanded as X transpose X – X transpose A theta – theta transpose A transpose into A + theta transpose AA transpose into theta this is the sum square error cost function.

Minimizing the SSE gives the normal equation A transpose A times theta hat LS = A transpose into X this is the normal equation for LS estimation. Now the inverse of this matrix exists therefore the LS estimator is given by theta hat LS = A transpose A whole inverse into A transpose X and this quantity is known as the pseudo inverse that we denote by A + that is theta hat LS = A + into X where A + is the pseudo inverse of A.

(Refer Slide Time: 48:35)



We establish that theta hat LS is an unbiased estimator. So E of theta hat LS E of theta hat LS = theta. The covariance matrix of theta hat LS is given by C of theta hat LS = sigma square into A + into A + transpose where A + is the pseudo inverse and sigma square is the variance of the error. For any unbiased linear estimator theta tilde = DX we establish that DA = I I is the identity matrix and the covariance matrix C theta tilde is given by sigma square into DD transpose. This is also an important relation.

Next we establish using this expression we established that theta hat LS is the best linear unbiased estimator BLUE. The BLUE becomes an MVUE if data are jointly Gaussian. So that we establish that if random data vectors are jointly Gaussian then least square estimator is the MVUE because it is a BLUE and when data are jointly Gaussian the BLUE will become MVUE. LS estimation is computationally expensive because of the matrix inversion.

Because determining the pseudo inverse involves matrix inversion and matrix inversion is computationally expensive. This drawback is overcome by using the RLS algorithm recursive least square algorithm that we will be explaining in the next lecture. With this background on the LS estimation we will discuss the RLS adaptive filter in the next lecture. Thank you.