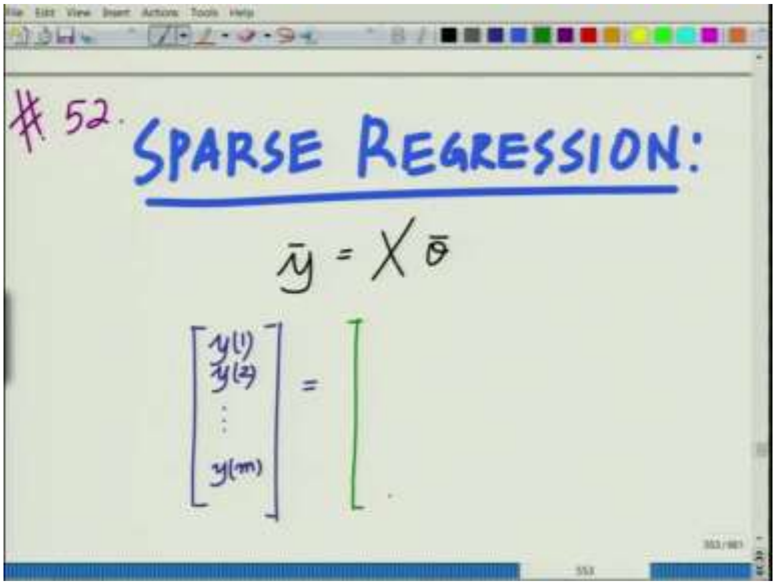
Applied Linear Algebra for Signal Processing, Data Analytics and Machine Learning
Professor Aaditya K Jagannatham
Department of Electrical Engineering
Indian Institute of Technology Kanpur
Lecture 52
Sparse Regression: Solution via the Orthogonal Matching Pursuit (OMP) Algorithm

Hello, welcome to another module in this massive open online course. So we are talking about sparse regression. Let us continue our discussion and continue to learn more about sparse regression.

(Refer Slide Time: 00:22)



52.

SPARSE REGRESSION:

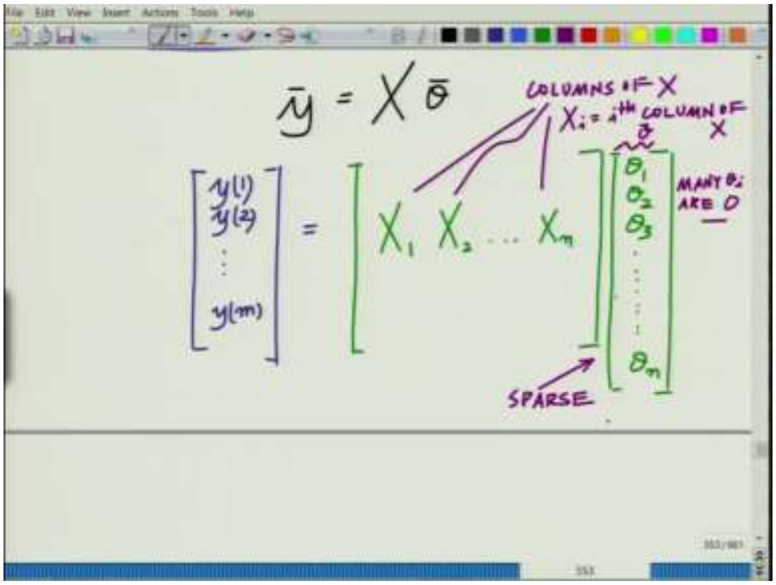
$$\bar{y} = X\bar{\theta}$$

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(m) \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

Specifically we have set up the model. Now we want to understand how to solve the sparse regression problem and remember we already looked at it, basically is a linear regression model for the output response using the fewest number of explanatory variables or regressors if possible.

Essentially that means that if you look at the weight vector theta bar, a large number of these, the regression vector, a large number of these regression coefficients theta i must be 0 and only very few must be non 0 which then we combine the corresponding explanatory variables. So we have seen previously the model can be formulated as the vector y bar equal to X theta bar where if you look at this y, this is m dimensional vector and this equals, let us write the matrix X in terms of its columns now.

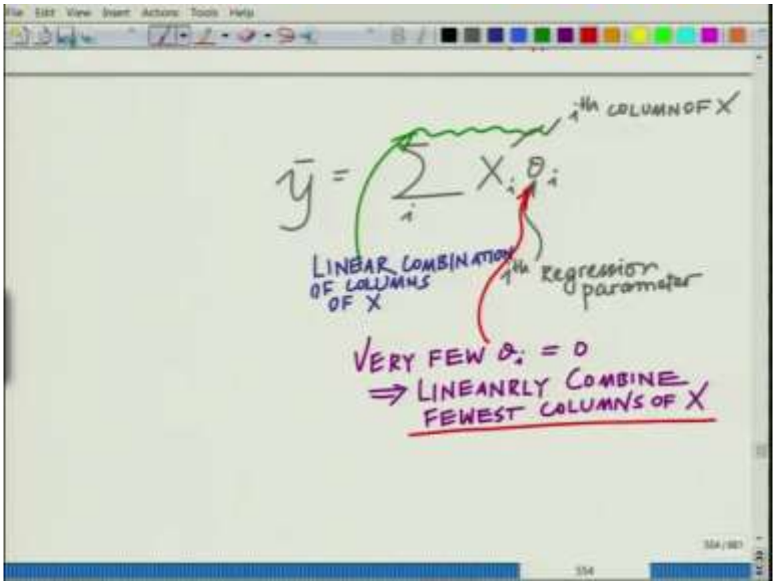
(Refer Slide Time: 01:38)


$\bar{y} = X\bar{\theta}$
COLUMNS OF X
X_i = i^{th} COLUMN OF X
$\bar{\theta}$
$y(1)$ $y(2)$ $\vdots$ $y(m)$
=
X_1 X_2 ... X_n
θ_1 θ_2 θ_3 $\vdots$ θ_n
MANY θ_i ARE 0
SPARSE

Let us say the first column is capital X 1 we have previously looked at the rows which are the training vectors now let us divide it in to the corresponding columns. So we have X 1, X 2, up to X n, these are the columns and then have the parameter vector. Remember this looks like a wide matrix that is essentially what we had said is the basis for the problem compressive sensing, plus, and this is essentially the model that we have, and what is X i these are the columns of X.

These are essentially the various explanatory variables, and then therefore for instance X i equals the i th column of the matrix X and this is of course, remember this is your regression parameter so this is your theta bar with many theta i equal to 0. So many theta i are simply 0. so this is essentially what we said is a sparse regression parameter vector.

(Refer Slide Time: 03:33)



$$\bar{y} = \underbrace{\sum_i X_i}_{\text{LINEAR COMBINATION OF COLUMNS OF X}} \theta_i$$

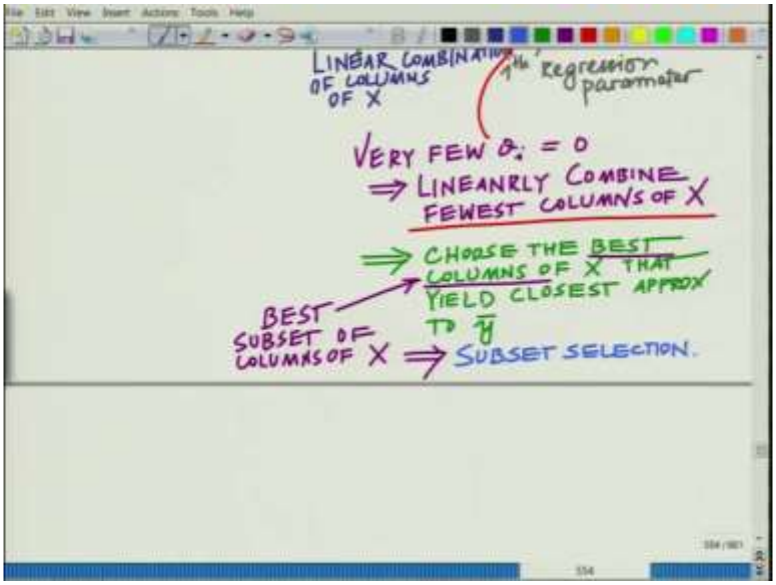
File Edit View Insert Actions Tools Help
i th COLUMN OF X
LINEAR COMBINATION OF COLUMNS OF X
i th Regression parameter
VERY FEW θ_i = 0
⇒ LINEANRLY COMBINE FEWEST COLUMNS OF X

Now if you look at this you will realize that what we are trying to find here is we are trying to express y bar as summation of i X i theta i this is the i th column of X and this is the i th regression parameter. Now very few, now what we have is, here very few theta i are, very few theta i equal to 0 implies linearly combining fewest columns of X right?

So essentially what we are doing is we are performing a linear combination of the columns of your matrix X, right? Your linearity combined in the regressors, right? Linearly combined in the columns of the matrix X to obtain the response vector y. So this is what it is doing.

Performing linear combination, this is linear combination of columns of X and therefore if we have very few theta i equal 0, so essentially we have very few theta i equal to 0, this implies you are linearly combining the fewest, you would like to linearly combine the fewest columns of X to obtain as good and the best approximation to y bar. That is the response vector. So this implies you have to choose the columns of X i carefully. The columns of X i which yields the best approximation to y bar.

(Refer Slide Time: 06:09)



LINEAR COMBINATION OF COLUMNS OF X

i^{th} regression parameter

VERY FEW $\theta_i = 0$

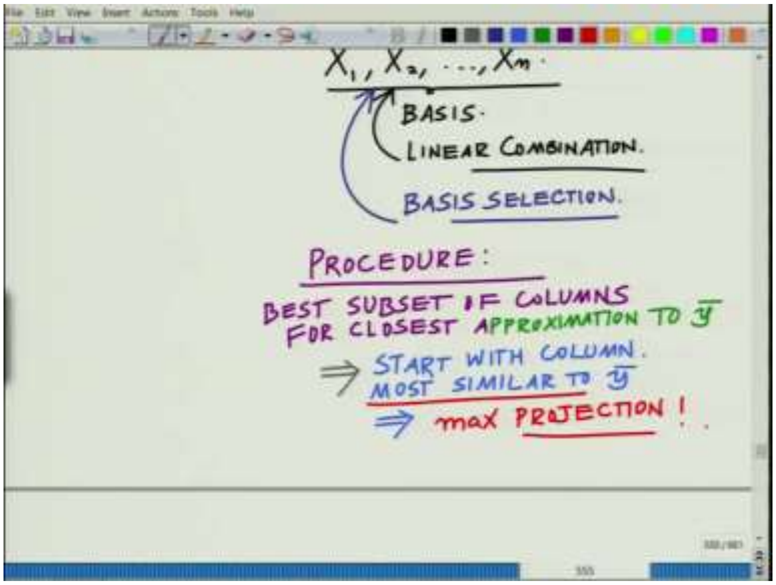
$\Rightarrow$ LINEANRLY COMBINE FEWEST COLUMNS OF X

$\Rightarrow$ CHOOSE THE BEST COLUMNS OF X THAT YIELD CLOSEST APPROX TO $\bar{y}$

BEST SUBSET OF COLUMNS OF X $\Rightarrow$ SUBSET SELECTION.

So if you want to use the fewest columns, right? So naturally that means choose the best columns of X that yield closest approximation to y bar. So we want to choose the best columns of X or rather we want to say, we want to choose the subset of columns, the best or we say we can choose the best subset. We want to choose the best subset of columns of X hence this is also termed as subset selection or basis selection

(Refer Slide Time: 07:27)



$X_1, X_2, \ldots, X_m$.

BASIS.

LINEAR COMBINATION.

BASIS SELECTION.

PROCEDURE:

BEST SUBSET OF COLUMNS FOR CLOSEST APPROXIMATION TO $\bar{y}$

$\Rightarrow$ START WITH COLUMN MOST SIMILAR TO $\bar{y}$

$\Rightarrow$ max PROJECTION !

You can also see that if we treat this as a space with basis as X 1, X 2, X n if we treat this as a basis since we are performing the linear combination. Since we are performing a linear combination of this essentially you can treat this as a space for which this set of columns on the basis and you are performing a linear combination.

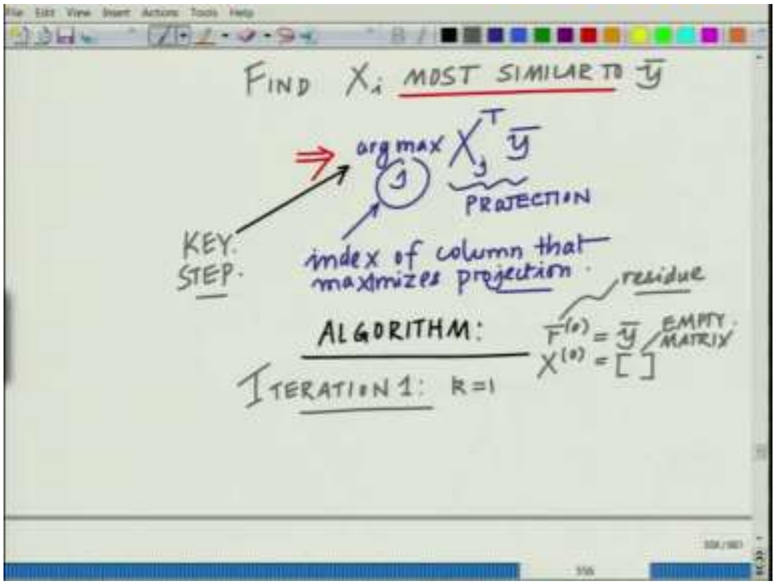
Now out of this you want to choose those vectors, those columns essentially which give you the best approximation to y bar. So you can think of this as also basis selection, subset selection. So this is also essentially becomes then a problem of basis selection. Now this can be performed as follows, so procedure will be as follows, so what is the procedure?

As we said we want to choose the best subset of columns for closest approximation to y bar. So now implies we start with the column. So if you want to choose the fewest number of columns of X which yield the best approximation to y bar start with the column that is the closest approximation to y bar to begin with.

Start with the column that is very similar to y bar. So you keep using columns essentially of X that are very close, very similar to y bar, then you will need to use fewer columns to approximate y bar, that essentially is the intuition. So start with the column that is most similar and how do we find this most similar column?

This implies essentially we find the projection, maximize the projection, this is always, that is we want to find two vectors, we want to find the sim vector which is the most similar to a given vector. We want to find the projection and using the projection the vector that has the largest projection is essentially the most similar vector. So as simple as that.

(Refer Slide Time: 11:04)



FIND X_i MOST SIMILAR TO $\bar{y}$

$$\Rightarrow \underset{j}{\arg\max}\ \underbrace{X_j^T \bar{y}}_{\text{PROJECTION}}$$

KEY STEP.

index of column that maximizes projection.

ALGORITHM:

residue

$r^{(0)} = \bar{y}$

$X^{(0)} = [\ ]$ EMPTY MATRIX

ITERATION 1: $k=1$

So this implies, so we want to find the column that is most find X i, most similar to y bar and this can be done very simply as follows. You have to choose essentially how to find the most similar vector, most similar to y bar, you have to choose, perform X j hermition or X j transpose if it is a linear vector, perform X j transpose y bar. This is essentially the projection.

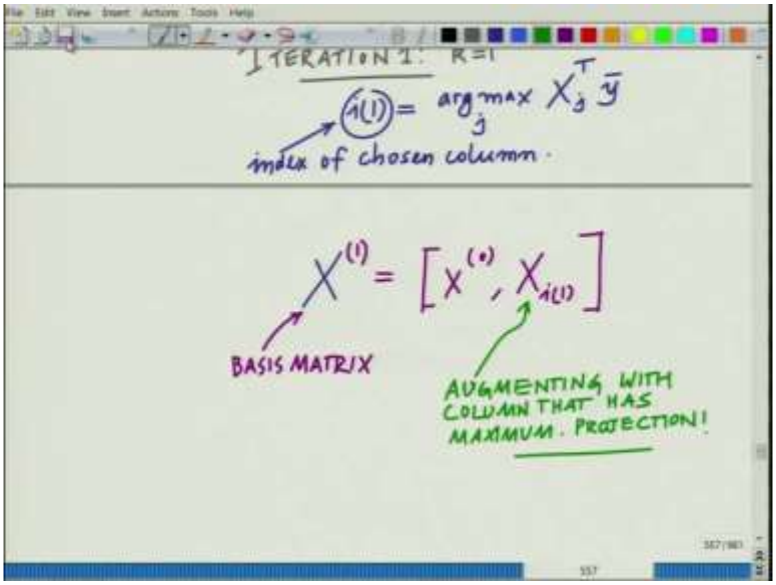
And you choose the j that maximizes this, so you choose the j that is column X j index of column that basically maximizes this projection. So let us term this, so let us look at this as an algorithm, so what is your algorithm? So this is the key step, if you look at this, this essentially is going to be the key step in this entire algorithm that is finding the projection to determine the column of X that is most similar to y bar and use that as a regressor, use that to build an approximation to y bar.

Use that as a probably explanatory variable to explain, which explains the response y bar, that is essential, which, so essentially you find the regressor or the set of regressors which are able to best explain y bar and what you are saying is these regressors have a very high projection over y bar or y bar has a very high projection over these regressors, essentially that is the central idea.

So let us now start with this residue. So let us term as iteration, so let us look at this as an iterative procedure and then it will be clear, so let us start with this as iteration 1 that is k equal to 1 and before we start we said the residue 0 is y bar and we also said this what is known as this basis matrix as the empty matrix.

So this is the residue or essentially what we can think of this as what is left to approximate in y bar. To begin with it is entire y bar itself. Progressively you will keep, as we keep getting approximations of y bar, we will keep subtracting those approximations to see what is the residue that is left in each iteration and this basis, remember we have to select the basis, the best set of columns of X i. So this basis matrix initially is an empty matrix. So this is initialized as in empty matrix.

(Refer Slide Time: 14:59)



ITERATION 1: k=1

$$\boxed{i(1)} = \arg\max_j X_j^T \bar{y}$$

index of chosen column.

$$X^{(1)} = [X^{(0)}, X_{i(1)}]$$

BASIS MATRIX

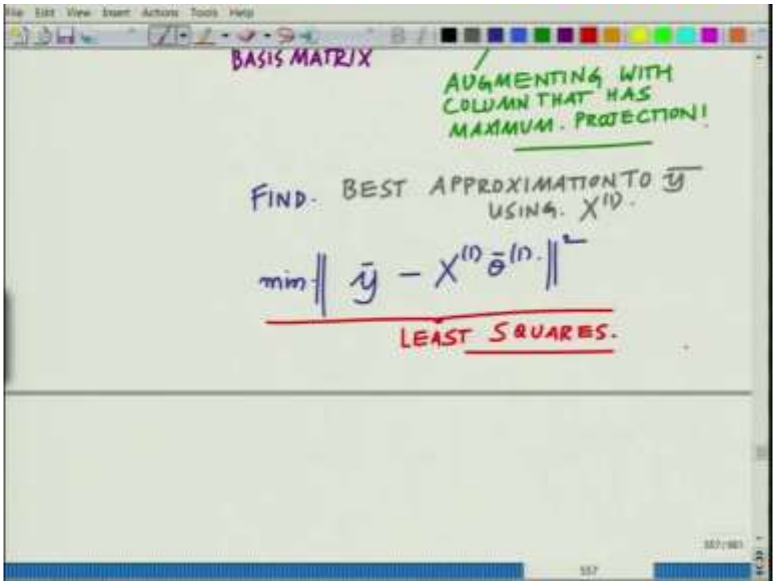
AUGMENTING WITH COLUMN THAT HAS MAXIMUM. PROJECTION!

And the step now therefore is to again what we have seen that is the index in the first iteration is to basically find the column that maximizes the projection over y bar X j transpose y bar and then you use this, so this is the index of the chosen column. And then you use, you augment the basis matrix in iteration 1, X 1 equals X 0 which is of course at this stage is an empty matrix, but this will keep getting full, times, and followed by this, now followed by this column which is your X i 1.

So augment the basis matrix. So what you are doing is this is your basis matrix. So what you are doing over here is essentially you are augmenting it with the column that has the maximum projection. This it he basis matrix that you find in the iteration. So this is we are augmenting, so essentially this is where your basis selection step is.

You are choosing the column that has the maximum projection on the residue what is left and then adding that in, so that you are putting that as your basis. That is essentially your best choice of the explanatory variable at this stage that explains y bar.

(Refer Slide Time: 0:17:29)



BASIS MATRIX

AUGMENTING WITH COLUMN THAT HAS MAXIMUM. PROJECTION!

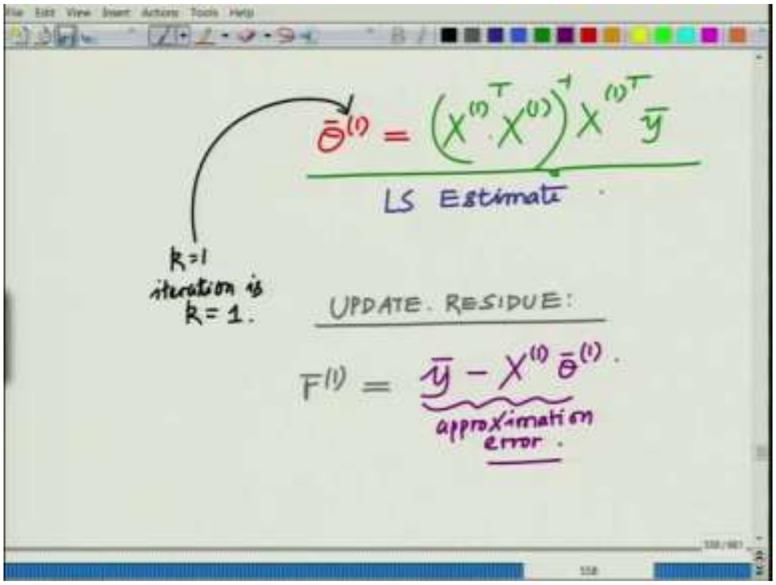
FIND. BEST APPROXIMATION TO $\bar{y}$ USING. $X^{(1)}$.

$$\min \| \bar{y} - X^{(1)} \bar{\theta}^{(1)} \|^2$$

LEAST SQUARES.

Now we have to find what is the parameter theta so now we find best approximation, so we are to yet find the regression coefficient. So find the best approximation to y bar using X 1, and this can be obtained by solving the least squares problem. Remember we want to find X 1, theta bar 1 at this stage and to find the best regression parameter vector at this stage you simply use the least square. So this is essentially, what is this? You minimize the least squares error. So this is essentially your least squares.

(Refer Slide Time: 18:42)



$$\bar{\theta}^{(1)} = \left(X^{(1)^T} X^{(1)}\right)^{-1} X^{(1)^T} \bar{y}$$

LS Estimate

k=1
iteration is
k = 1.

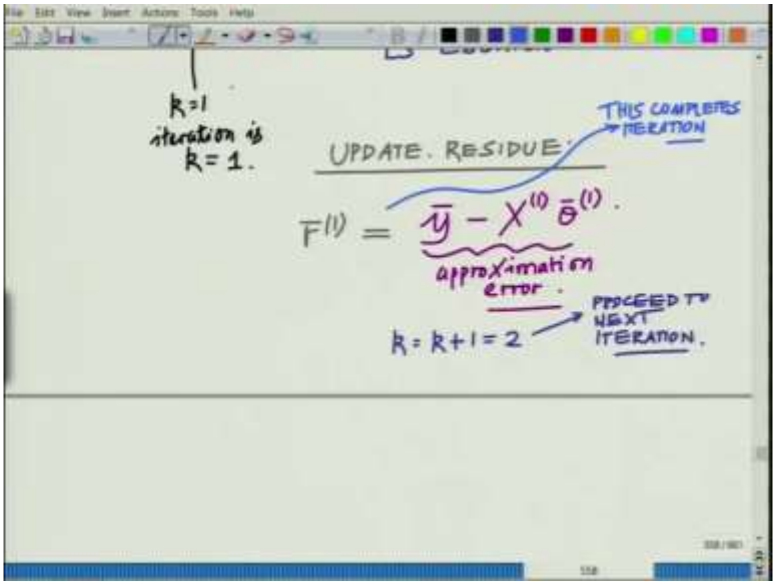
UPDATE. RESIDUE:

$$\bar{r}^{(1)} = \underbrace{\bar{y} - X^{(1)}\bar{\theta}^{(1)}}_{\text{approximation error}}.$$

And we know the solutions to this least squares problem that will simply be given as theta bar 1 that will be equal to X 1 transpose of this hermition of its complex times X 1 inverse X 1 transpose y bar. This is essentially your least squares estimate.

That is for, remember we are still talking, remember this bracket still shows that iteration is k equal to 1. This is the iteration, the iteration is still your k equal to 1. Now we update the residue. So it is the next, update the residue. How do we update the residue? So we have r bar in the first iteration equals y bar minus X that is the approximation error. This is the approximation error so update the residue.

(Refer Slide Time: 20:24)



k=1 iteration is k = 1.

UPDATE. RESIDUE:

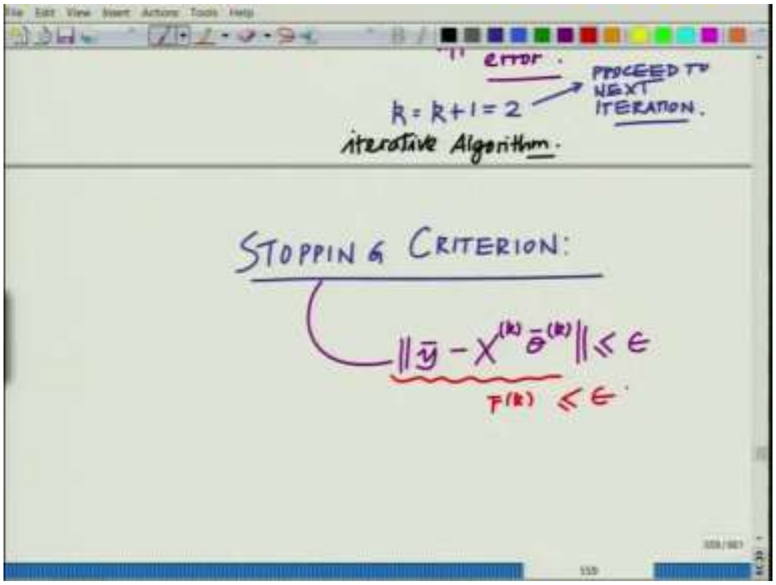
THIS COMPLETES ITERATION

$$F^{(1)} = \underbrace{\bar{y} - X^{(1)}\bar{\theta}^{(1)}}_{\text{approximation error}}.$$

$k = k+1 = 2 \rightarrow$ PROCEED TO NEXT ITERATION.

And now we proceed to the next iteration. Now you will k equal to k plus 1, proceed to next. So this essentially once you evaluate the residue this completes the iteration. So what you are doing, essentially you are projecting on the residue or you are projecting the residue in each column of X. Choosing that column which basically maximizes the projection, augmenting your basis matrix. Finding the best approximation to y bar. Subtracting the approximation, obtaining the residue. So in each iteration residue keeps progressively decreasing. You can stop either at two points. So now you have to choose a stopping criteria. Now let us talk about that.

(Refer Slide Time: 21:39)



error.

$k = k+1 = 2 \rightarrow$ PROCEED TO NEXT ITERATION.

iterative Algorithm.

STOPPING CRITERION:

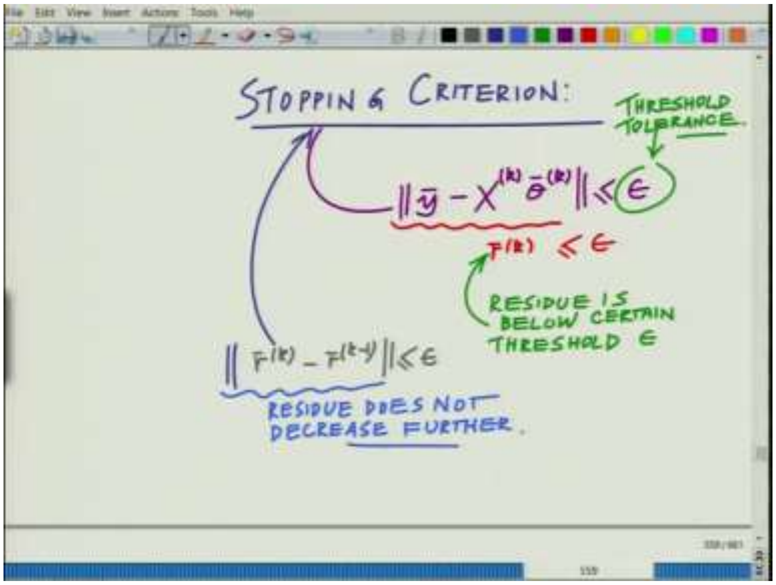
$$\underbrace{\|\bar{y} - X^{(k)}\bar{\theta}^{(k)}\|}_{r^{(k)}} \leqslant \epsilon$$

$$r^{(k)} \leqslant \epsilon$$

Now we have to choose a stopping criteria for this algorithm. So what is, what we call as the, because this is an iterative algorithm. How many iterations do you perform? So this is, we need to talk about, so since this is the iterative algorithm.

So this is, what is the stopping criteria? Stopping criteria either you have norm y bar minus at some iteration k X k theta bar k less than equal to epsilon, where epsilon is some suitably throws in, chosen threshold so let us say. So the error approximation error or this is what we call as the residue.

(Refer Slide Time: 22:40)



STOPPING CRITERION:

THRESHOLD TOLERANCE

$\|\bar{y} - X^{(k)}\bar{\theta}^{(k)}\| \leqslant \epsilon$

$\bar{r}^{(k)} \leqslant \epsilon$

RESIDUE IS BELOW CERTAIN THRESHOLD ϵ

$\|\bar{r}^{(k)} - \bar{r}^{(k-1)}\| \leqslant \epsilon$

RESIDUE DOES NOT DECREASE FURTHER.

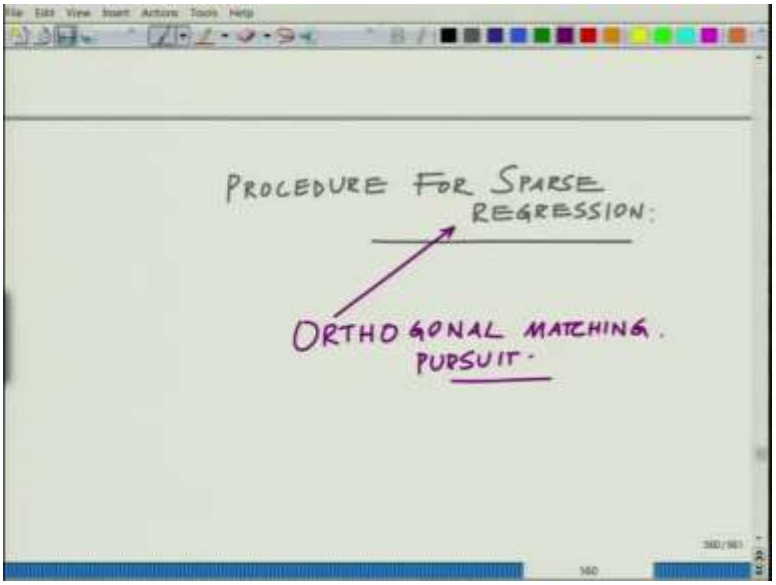
That is, the residue falls below a certain threshold. So approximation is good. Residue is below certain threshold. So this is essentially what we are terming as our threshold or you can also think of this as the tolerance of the algorithm. Tolerance of the approximation error. So if your approximation is good enough that is you have chosen a certain low value of epsilon and if your approximation is lower that let us say point 001 and your approximation error is less than this point 001 then you can term it as the algorithm.

Your approximation is good enough, because practically you are never going to get to an error that is 0, because of the noise and other aspects. There will always going to, we know that in practice there is always going to be an estimation error. So one has to be judicious in determining this threshold.

The other criterion is when you have a certain number of iterations and you say the residue stops decreasing that is if you look or stops decreasing significantly that is if you look at r bar k minus, that is residue does not decrease further. So what it means is essentially that when you are doing this approximation progressively residue is decreasing, decreasing, decreasing, after some point the residue is not decreasing significantly.

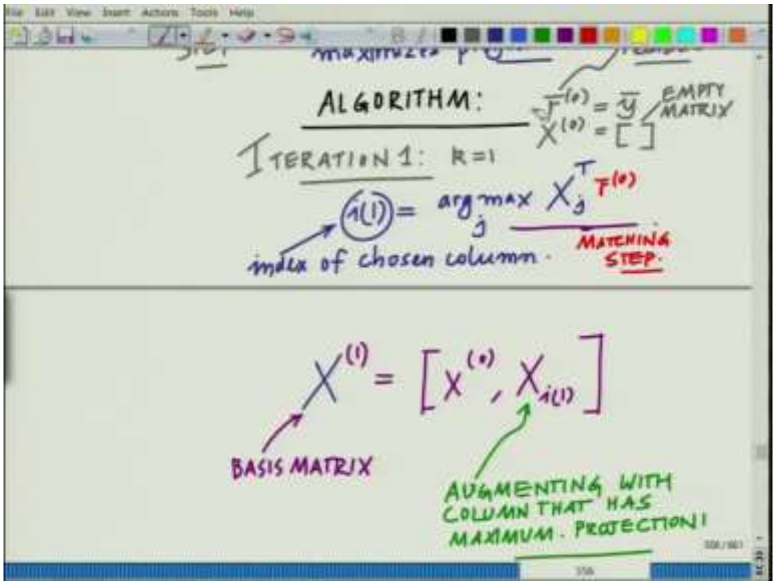
You are more or less left with same residue iteration after iteration which means there is no point adding mode regressors, adding no more non 0 weighting coefficients parameter, because that is not significantly helping the approximation, both are these are roughly the same essentially.

(Refer Slide Time: 25:04)



PROCEDURE FOR SPARSE REGRESSION:

ORTHOGONAL MATCHING PURSUIT.



maximizes

ALGORITHM:

$\bar{r}^{(0)} = \bar{y}$

$X^{(0)} = [\ ]$ EMPTY MATRIX

ITERATION 1: $k=1$

$\boxed{i(1)} = \underset{j}{\arg\max}\ \underline{X_j^T r^{(0)}}$

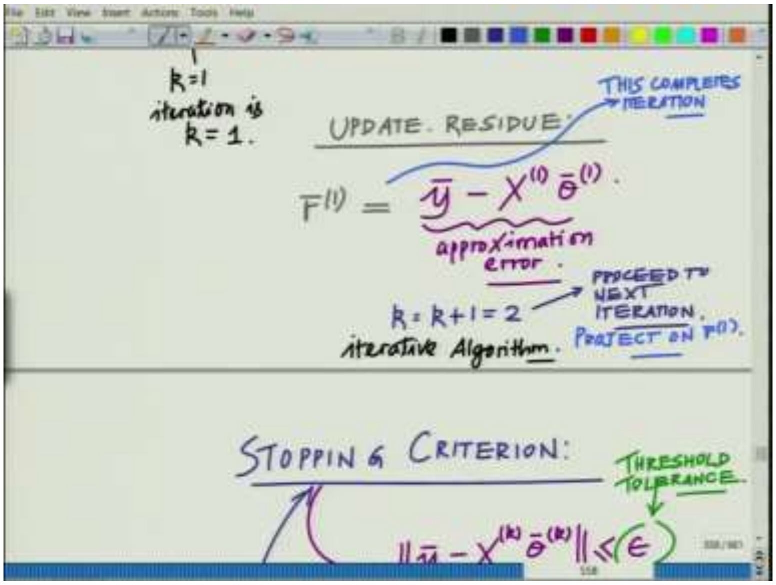
index of chosen column.

MATCHING STEP.

$X^{(1)} = [X^{(0)}, X_{i(1)}]$

BASIS MATRIX

AUGMENTING WITH COLUMN THAT HAS MAXIMUM. PROJECTION!



$k=1$ iteration is $k=1$.

UPDATE. RESIDUE:

THIS COMPLETES ITERATION

$$\bar{r}^{(1)} = \underbrace{\bar{y} - X^{(1)}\bar{\theta}^{(1)}}_{\text{approximation error}}.$$

$k = k+1 = 2$ → PROCEED TO NEXT ITERATION. PROJECT ON $\bar{r}^{(1)}$.

iterative Algorithm.

STOPPING CRITERION:

THRESHOLD TOLERANCE.

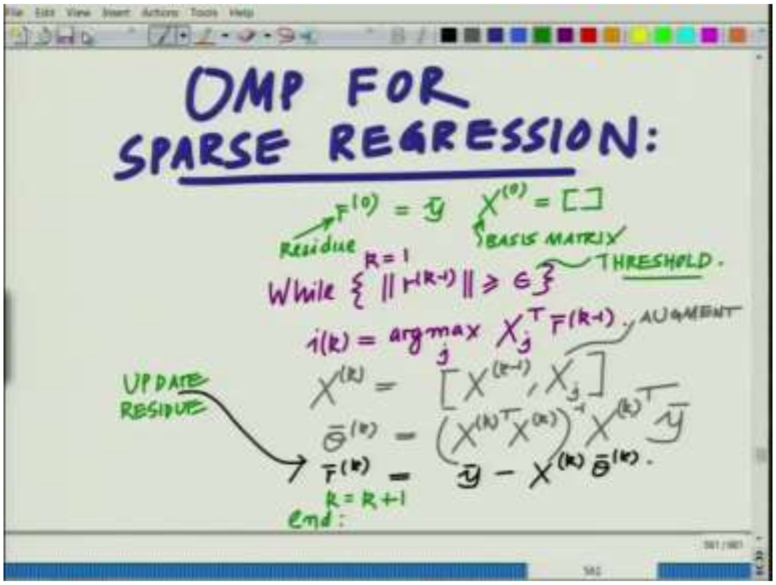
$$\|\bar{y} - X^{(k)}\bar{\theta}^{(k)}\| \leq \epsilon$$

So these are the two stopping criterion and this algorithm is known as the orthogonal matching pursuit. This algorithm, this procedure for sparse regression this is termed as what we have just talked about, this is termed as orthogonal matching pursuit. And in case you are wondering where the name comes from.

It simply comes from the fact that if you look at this step, the projection step where you are trying to find the best column, this is essentially what were termed as the matching step. This is the matching step, right? Essentially you are trying to match the columns of X with the residue. And, sorry this has to be here, r bar 0 or essentially r bar k minus 1. So we are projecting this on the residue from the previous iteration that is essentially what this is.

So this is X j transpose into r bar 0, which initially to begin with as we have noted r bar 0 equal to y and then it will be equal to the residues later. So next iteration, project on the residue, that is project on r bar, project each column r bar 1 in the next call, in then next iteration, you project on r bar 1. So to summarize this algorithm, let me just now, let me now summarize this algorithm, this OMP algorithm.

(Refer Slide Time: 27:19)



OMP FOR SPARSE REGRESSION:

$\bar{r}^{(0)} = \bar{y}$ $\quad X^{(0)} = [\ ]$

$k=1$

While $\{ \| \bar{r}^{(k-1)} \| \geq \epsilon \}$

$i(k) = \arg\max_{j} X_j^T \bar{r}^{(k-1)}$

$X^{(k)} = [X^{(k-1)}, X_j]$

$\bar{\theta}^{(k)} = (X^{(k)T} X^{(k)})^{-1} X^{(k)T} \bar{y}$

$\bar{r}^{(k)} = \bar{y} - X^{(k)} \bar{\theta}^{(k)}$.

$k = k+1$

end:

This is the OMP for sparse regression. If you think about this it is a very simple algorithm OMP. What you are essentially doing is you have your r bar 0, this is your r bar 0, iteration 0, this is what we are calling as y bar, then this matrix is the empty matrix. This is the residue, remember that this is your basis matrix.

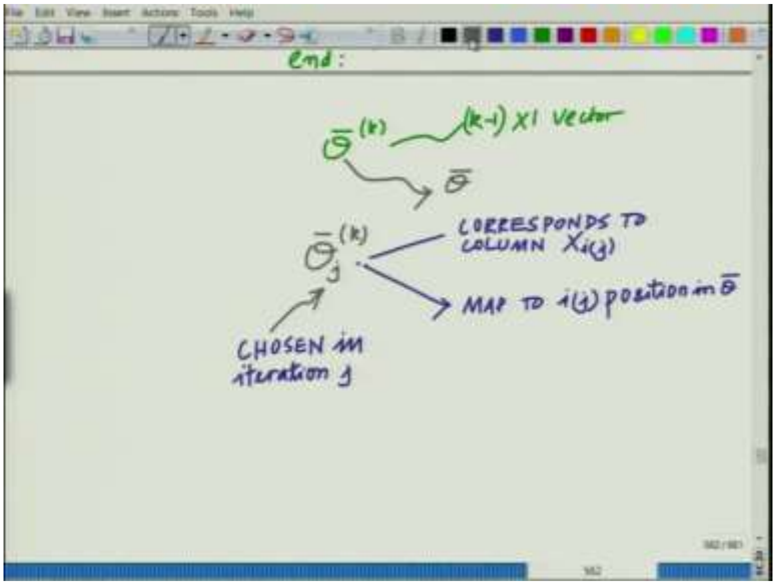
And then what is the next step? Let us say while, do this while, while your residue that is norm r bar, that is other thing. So said the iteration k equal to 1, iteration index, while the residue at k minus 1 in this case to start with it will be r bar 0 is greater than or equal to epsilon. Then form in this step i k equal to, this will be equal to argmax of over j X j transpose r bar of k minus 1, project on the previous residue, right?

Now X of j augment the basis matrix as X of k minus 1, X of j, so this is essentially your augmenting step. And then update the parameter X k transpose X k inverse, X k transpose y bar, so you update the parameter and then you update the residue y k minus theta bar k, update the residue. So this step is basically you update residue, and then of course you have k equal to update the iteration index and then you have the end.

So these are the steps of the OMP. This is a very basic as you can say, it is very, very simple, you simply have to choose the column that essential step is the matching. You have to choose the column that has the maximum projection on the residue, augment your basis matrix, find the best approximation to y bar, using the current basis, remove that approximation, find the residue, repeat it.

Until your residue is evidently high, that is basically you think, you can still achieve a better approximation of i bar, and that of course is an engineering, that comes with experience how to set that threshold, right? So this threshold, what is an acceptable threshold is in decision that I based on some kind of experience solving the problem, several times.

(Refer Slide Time: 31:38)

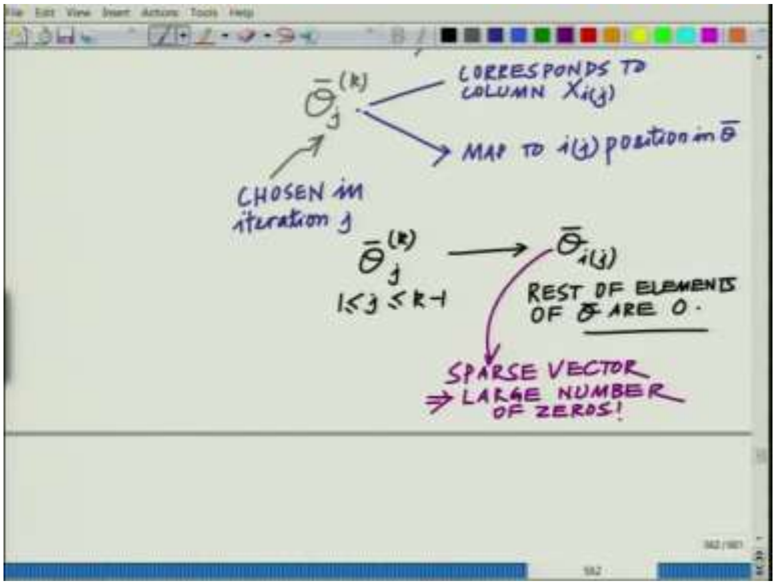


Now there is one last step that is mapping this, theta, still we have to have this theta bar k, remember at iteration k this will be at the, if you perform five, because k equal to k plus 1, this will be a k minus 1 cross 1 vector.

So how do you map this to the original element theta bar. So the j th element that is theta bar k of j will be mapped, remember this is chosen in the second iteration, j, that is, for instance theta bar k, first element is chosen in the first iteration, second element is chosen in the second iteration and so on.

So essentially this corresponds to, so this is chosen in, if you go back, take a look at the algorithm, chosen in iteration j. So this has to be mapped to the column corresponds to, so this corresponds to column X of i j, remember i j is the index chosen in the second iteration. Therefore map to i j position. So map to i j position in theta bar.

(Refer Slide Time: 33:40)



$\bar{\theta}_j^{(k)}$ — CORRESPONDS TO COLUMN $X_{i(j)}$

→ MAP TO $i(j)$ position in $\bar{\theta}$

CHOSEN in iteration j

$\bar{\theta}_j^{(k)} \longrightarrow \bar{\theta}_{i(j)}$

$1 \leq j \leq k-1$

REST OF ELEMENTS OF $\bar{\theta}$ ARE 0.

SPARSE VECTOR

⇒ LARGE NUMBER OF ZEROS!

So what we do is this j th element of this theta bar or the k th iteration, so theta bar k, the j the iteration for 1 less than equal to j less than equal to k minus 1, you map it to the original theta bar i j th element. Rest of the elements of theta bar are 0.

So that is essentially what gives, so you are mapping only, so at the end of this hopefully you have chosen only a few of the columns of x bar so those corresponding values in this theta bar k you map it, map them back to the corresponding, the indices, to the elements of the corresponding indices of the columns chosen in theta bar. Rest of the elements of theta bar are 0 that is what makes theta bar a sparse vector, and you will see there will still be a large number of vectors.

So this will be a sparse vector, implies, this implies you will have large number of 0's. So that is essentially the philosophy and the principle of sparse regression and we have seen the OMP algorithm that is the orthogonal matching pursuit for sparse regression and as I have shown you it is a fairly simple algorithm with significant implications because this is a path breaking radical field that is sparse signal recoveries, sparse estimation, compressive sensing, all of these essentially are latest techniques in signal processing, communication.

There are several applications. Signal processing, imaging, tomography and so on and so on. All right so let us stop here and in the next module we will look at a simple example to understand this algorithm better. Thank you very much.