Machine Learning for Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 13
Principal Component Analysis and Regression Applications in Agriculture (Contd.)

Welcome friends to this 13th lecture of NPTEL online certification course of Machine Learning for Soil and Crop Management. And this is actually the third lecture of week three. And in this week, we are talking about the principal component analysis, as well as the regression applications in agriculture.

Now, in the previous two lectures, we have covered some of the important aspects that is principal component analysis as well as we have covered the principal component regression, what is principal component regression. You all remember that principal component analysis is a dimensionality reduction technique and it helps in both unsupervised as well as supervised algorithms.

Specifically in case of classification, in case of regression it acts as a supervised method and sometimes it is unsupervised method because, here we are trying to extract the information, in unsupervised method we are trying to extract the information from the feature space without considering the target variable.

So, we have discussed what is principal component analysis, how to compute principal component analysis. Apart from that, we have also seen the different types of plots like score plot, then, we have seen the loading plots and then biplot, we have seen how to select the principal components, what is the basis of selection of principal components using the screeplot what are the eigen values, what are the eigen vectors, why principal component analysis, why principal components are uncorrelated to each other.

So, we have discussed in details. Remember that dimensional data reduction is must when there are multicollinearity and to reduce the overfitting. So, we have also discussed in that line we have also discussed the principal component regression. Remember principal component regression is computing the principal components and then regressing the target variable using those principal components as inputs.

So, we have also seen some application of principal component analysis in agriculture, specifically focusing on soil and crop and we have also seen some of the applications of principal component regression in soil. So, today we are going to start partial least squares regression. Partial least squares regression is a widely accepted method or widely accepted cammo metric method for measurement for predicting several soil properties.

And partial least squares regression has a special application or I would say it is widely used in case of soil spectral characterization studies. For developing spectral prediction model, for soil properties, scientists have widely used the partial least squares regression model followed by other models. So, we are going to discuss what is partial least squares regression model, how it differs from principal component regression.

(Refer Slide Time: 4:08)

CONCEPTS COVERED

- Partial Least Squares Regression
- Mathematical expression of PLSR
- PLSR- soil and plant applications

So, these are the important concepts which we are going to discuss today. One is partial least squares regression and their features. Secondly, we are going to discuss the mathematical expression of PLSR, and finally, we are going to talk about the PLSR application for soil and plant size.

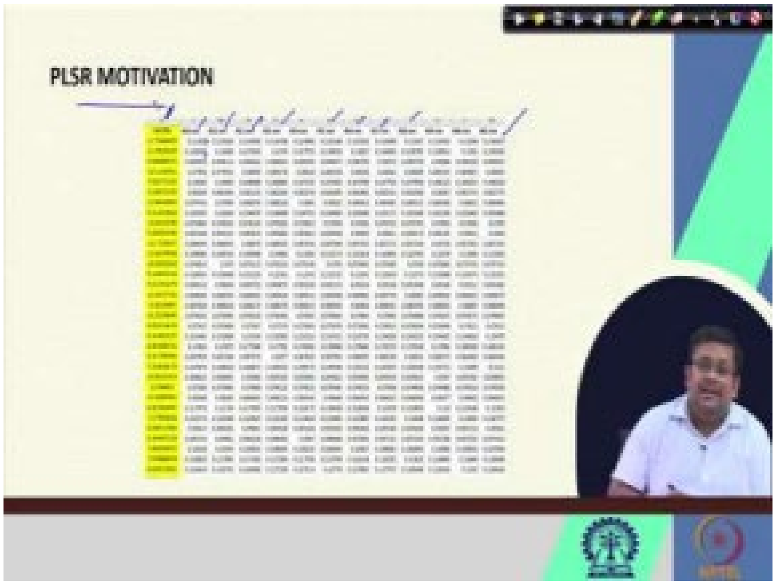
(Refer Slide Time: 4:33)

KEYWORDS

- PLSR
- Latent factor
- PLSR loading plot
- Soil heavy metal prediction
- Algal pigment prediction

So, some of the keywords we are going to see, we are going to encounter in this lecture are PLSR, then latent factor, then PLSR loading plot, then soil heavy metal prediction and also algal pigment prediction. So, let us see what is partial least squares regression, or what is the motivation of PLSR.

(Refer Slide Time: 4:57)



PLSR MOTIVATION

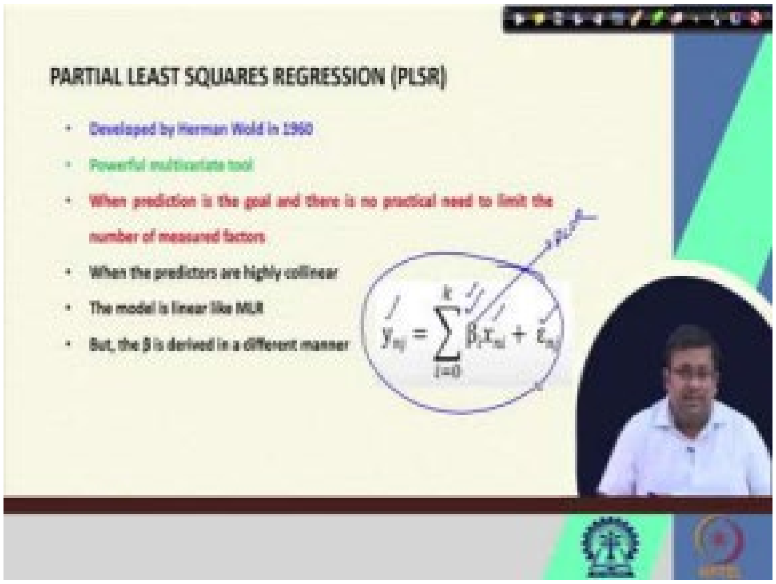
Now, you have already seen the snippet of a spectral data set in our previous lectures and you know that a soil property here in this snippet we can see this is our target soil property that is loss on ignition. And you can see that starting from 450, 451 nanometer, 452 nanometer, 453 nanometer up to 461 nanometer, we are having the spectral data set, these are the reflectance values we are getting.

Now, the one thing is for sure from this snippet that when we are using 1000s and 1000s of this type of spectral reflectance variable, these are the wavelengths and these are the spectral reflectance values as you can see for each sample for each wavelength. Now when we add these variables, some of these variables are correlated to each other.

And when they are, when there are multiple variables, which are correlated to each other, then we termed that as multicollinearity. So, when there is a multicollinearity and when there are a huge number of variables in a model, there is always chance of getting overfitting in the model. What is overfitting, we have already discussed in our previous lectures.

Now, so to reduce the overfitting or to deal with the multi collinearity in any data set, specifically in a regression problem, we use the PLSR method. Now, let us see some of the important features of PLSR.

(Refer Slide Time: 7:11)



PARTIAL LEAST SQUARES REGRESSION (PLSR)

- Developed by Herman Wold in 1960
- Powerful multivariate tool
- When prediction is the goal and there is no practical need to limit the number of measured factors
- When the predictors are highly collinear
- The model is linear like MLR
- But, the β is derived in a different manner

$$y_{nj} = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_{nj}$$

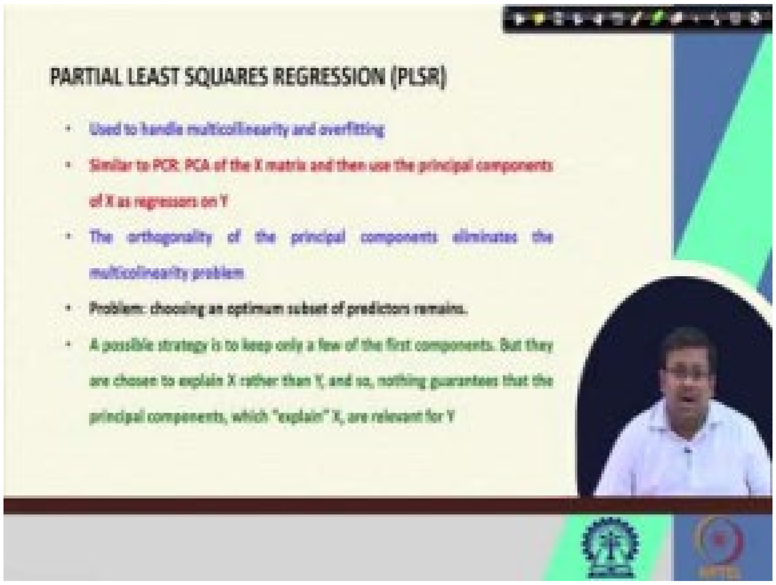
So, this PLSR regression was first developed by Herman Wold in 1960 and it is a powerful multivariate tool. And when in any multivariate problem our target is or our goal is to predict the outcome and there is no practical limit for, there is no practical need to limit the number of measured factors, then we always go with the partial least squares regression. In short, we call it PLSR regression.

And also when the predictors are highly collinear to each other, then also we can use this partial least squares regression. So, this model is ultimately linear model and it is a parametric model. What is parametric model we have discussed, what are the assumptions of parametric model we have already discussed.

Now, the model is a linear model and it takes the same linear model function, it assumes the same linear prediction model function as we are seeing here where the y is our target, whereas, betas are the coefficients and x is the predictor variable and these are the residuals. So, you can see the basic formula PLSR is also pretty much similar with that of multiple linear regression.

However, the beta, what is the difference between a multiple linear regression and a partial least squares regression, because, you see, the calculation of beta in case of partial least squares regression is different than that of multiple linear regression. So, the way we derive the beta coefficient is different in partial least squares regression as compared to the multiple linear regression. We are going to discuss that how in details in our subsequent slides.

(Refer Slide Time: 9:37)



PARTIAL LEAST SQUARES REGRESSION (PLSR)

- Used to handle multicollinearity and overfitting
- Similar to PCR: PCA of the X matrix and then use the principal components of X as regressors on Y
- The orthogonality of the principal components eliminates the multicollinearity problem
- Problem: choosing an optimum subset of predictors remains.
- A possible strategy is to keep only a few of the first components. But they are chosen to explain X rather than Y, and so, nothing guarantees that the principal components, which "explain" X, are relevant for Y

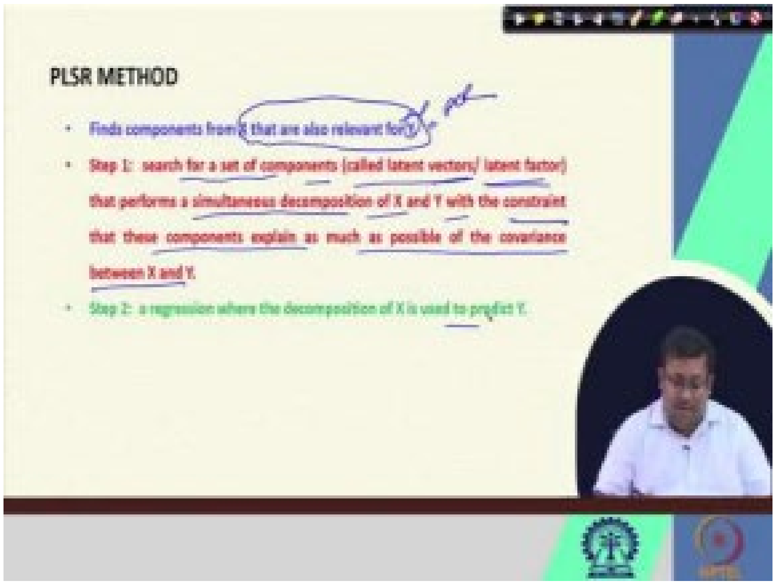
So, what are the other major features of partial least squares regression? The partial least squares regression generally used to handle the multi collinearity and overfitting as I have already mentioned. Now, it is similar to principal component regression because in case of principal component regression, what do we do, we generally calculate if there is an x matrix of predictors we generally calculate, we generally go with the principal component analyses for those x metrics, and then we calculate the principal components and then subsequently we regress our target based on those principal component.

Now, the orthogonality in the principal component, when we calculate the principal component the orthogonality feature in the principal components helps to eliminate the multi collinearity problem, but there is a problem in principal component regression. Now, choosing the optimum subset of predictor remains always a problem in case of principal component regression.

Now, what is the possible strategy for that, how we can identify the optimum subset of predictors when we are also considering our target variable? So, a possible strategy is to keep only the few of the first components, but they are chosen to explain the x rather than y. So, nothing guarantees that the principal components which explain X are relevant to Y.

Because, whenever we consider the principal component regression, we keep only the first principal, important principal components, we have already seen, based on the screeplot, we can keep some of the first 5 or first 10 or first 15 principal component based on their Egon values or based on their cumulative variance explaining power.

Now, when we are doing that, we are not considering the variation of our target variables. So, we do not have any idea whether the subset of the principal components which we are selecting will guarantee the proper explanation of our target variable Y that cannot be guaranteed, in case of principal component regression, because we are not considering the target variable while calculating or while selecting the important principal components. That is the point where partial least squares regression comes into picture.
(Refer Slide Time: 12:54)



PLSR METHOD

- Finds components from X that are also relevant for Y.
- Step 1: search for a set of components (called latent vectors/ latent factor) that performs a simultaneous decomposition of X and Y with the constraint that these components explain as much as possible of the covariance between X and Y.
- Step 2: a regression where the decomposition of X is used to predict Y.

So, how partial least squares regression can address that drawback or shortcoming of principal component regression? In case of partial least squares regression, first we find this method or PLSR regression method first finds the components from X that are also relevant to Y. And this was missing in case of PCR, because we are not considering the Y while calculating the PCR.

Although, Y is there as a target variable, but the principal calculation and selection of the principal components were not affected by the variation of Y. So, here PLSR try to rectify it and find the components from the X that are also relevant to Y. So, there are generally two steps which, there are mainly two steps in partial least squares regression, first of all these algorithms search for a, search for a set of components, we call them latent factors and latent vectors or latent factors. Sometimes we call them latent variables also.

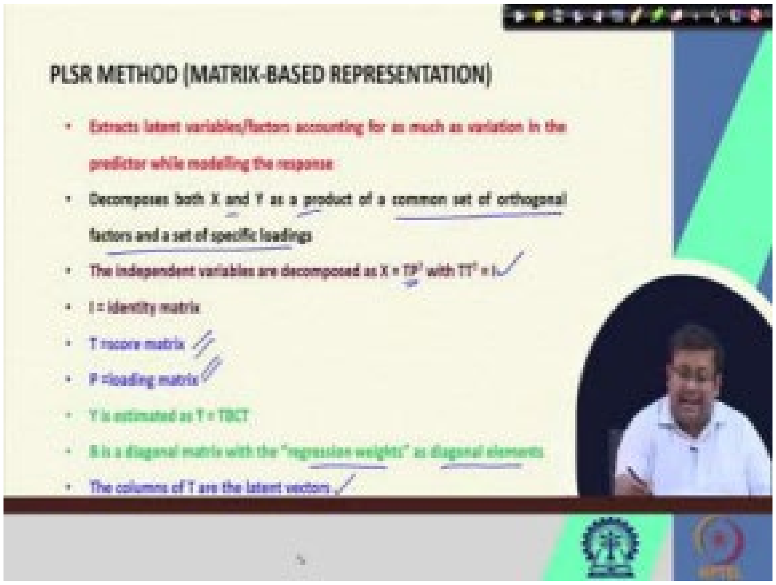
So, these terminologies will be used interchangeably, sometime I will be using latent factor, sometime I will be using latent vectors, sometime I will be using latent variables, all these are synonymous. So, so, first these algorithms search for a set of components that performs a simultaneous decomposition of both X and Y with the constraint that these components explain as much as possible of the covariance between X and Y.

So, the idea is you search for a set of components which we call the PLSR latent factors and that performs simultaneous decomposition of both the target variable as well as the matrix of our input variable with the constant that these components explained as much as possible of the covariance between X and Y. So, here the explaining power of the selected partial least squares regression latent variables ensure that they explain the direction or selected or computed direction among Y that explains the variation of Y maximum.

In other words, we try to extract the latent factors that explains the variation of Y, that maximum, if there are, if there are latent factors, PLSR latent factors, these PLSR latent factors will explain the variation of Y to the maximum. So, then step two is a regression where decomposition of X is used to predict the Y. Now, it is simple ordinary least squares regression.

Once we calculate the latent factors or latent variables, PLSR latent variables, the next step is a simple linear regression. Again, I am telling you, the latent vectors or latent variables are selected in a fashion that they will explain the maximum variance in Y.

(Refer Slide Time: 16:55)



PLSR METHOD (MATRIX-BASED REPRESENTATION)

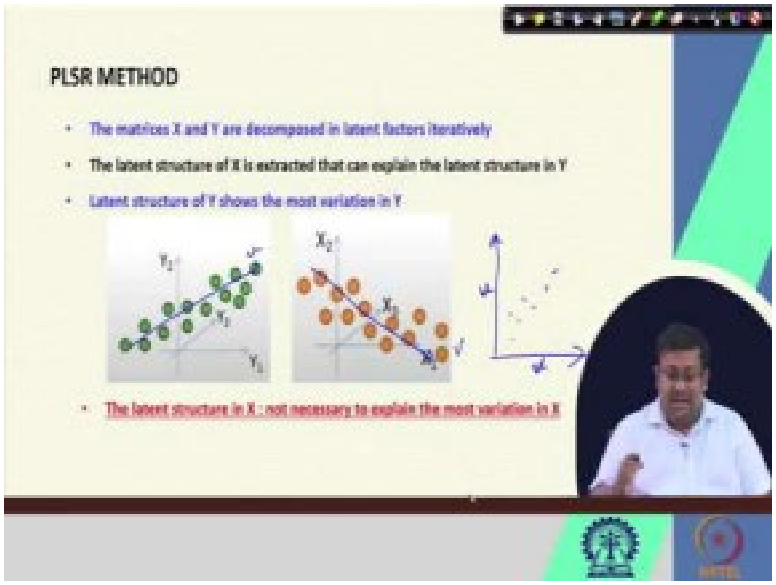
- Extracts latent variables/factors accounting for as much as variation in the predictor while modelling the response
- Decomposes both X and Y as a product of a common set of orthogonal factors and a set of specific loadings
- The independent variables are decomposed as $X = TP^T$ with $TT^T = I$
- I = identity matrix
- T =score matrix
- P =loading matrix
- Y is estimated as $Y = TBC^T$
- B is a diagonal matrix with the "regression weights" as diagonal elements
- The columns of T are the latent vectors

So, if we consider the matrix based representation, I have not included the too much mathematical details here, but I will try to explain the step by step process here. So, if we, those who already know matrix algebra, they will understand that these steps, you can see the first step is to extract the latent variables and factors accounting for as much as variation in the predictor while modeling the response.

So, how it starts? It starts with decomposing both X and Y as a product of common set of orthogonal factors and a set of specific loadings. So, you can see, the independent variables are decomposed as, suppose this is a independent variable, matrix is X and it is decomposed that T transpose of P, T multiplied by transpose of P, where TT transpose this is an identity matrix, you know that, where T is the score matrix and P is the loading metrics.

What is loading you already know. And simultaneously Y is estimated as Y hat equal to TBCT. So, here B is a diagonal matrix with the regression weights as diagonal elements. So, the column of T are the latent vectors. So, this is how we calculate the PLSR step by step, we decompose both the X and also we decompose both also the Y feature space, Y or target variable.

(Refer Slide Time: 18:54)



PLSR METHOD

- The matrices X and Y are decomposed in latent factors iteratively
- The latent structure of X is extracted that can explain the latent structure in Y
- Latent structure of Y shows the most variation in Y
- The latent structure in X : not necessary to explain the most variation in X

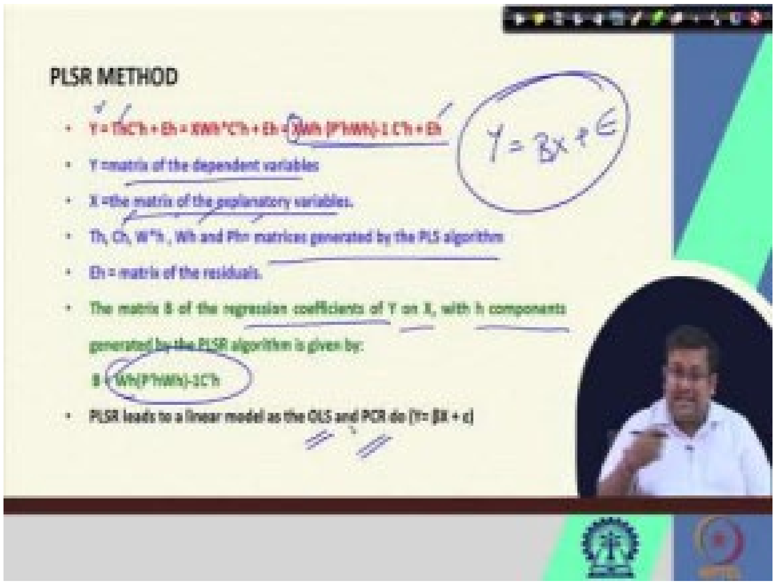
So, in other words, in the PLSR method, the matrix X and Y are decomposed in latent factors iteratively. So, the latent structure of X is extracted that can explain the latent structure of Y. So, we extract the latent structure of both X and latent structure of Y and then, we try to extract the latent factors of X in such a fashion that can explain the latent factors of latent structure of Y.

So, here you can see that there are three dimension Y1, Y2 and Y3 and we are getting a vector, suppose, this is u. So, the Latent, we are getting the latent structure represented by a vector of u, vector denoted by u and suppose we have also got a direction called v which is explaining the vector u maximally.

So, now, we will try to develop a regression between, then after we extract the two vectors then we go for the further regression. So, here we regress u on v and then or v. So, if we do a regression then it will be called a partial least squares regression. So, the latent structure is X remember that the latent structure of Y shows the most variation in Y.

And this is guaranteed that this shows the most variation in Y, however, since extraction of the latent structure of X is dependent on the latent structure of Y, we cannot guarantee that this direction or this vector v will explain the most variation in X. So, this is how most variation in x. So, this is the difference between a principal component regression and a partial least squares regression.

(Refer Slide Time: 21:57)



PLSR METHOD

- Y = ThC'h + Eh = XWh*C'h + Eh = XWh (P'hWh)-1 C'h + Eh
- Y =matrix of the dependent variables
- X =the matrix of the explanatory variables.
- Th, Ch, W*h , Wh and Ph matrices generated by the PLS algorithm
- Eh = matrix of the residuals.
- The matrix B of the regression coefficients of Y on X, with h components generated by the PLSR algorithm is given by:
  B = Wh(P'hWh)-1C'h
- PLSR leads to a linear model as the OLS and PCR do (Y= βX + ε)

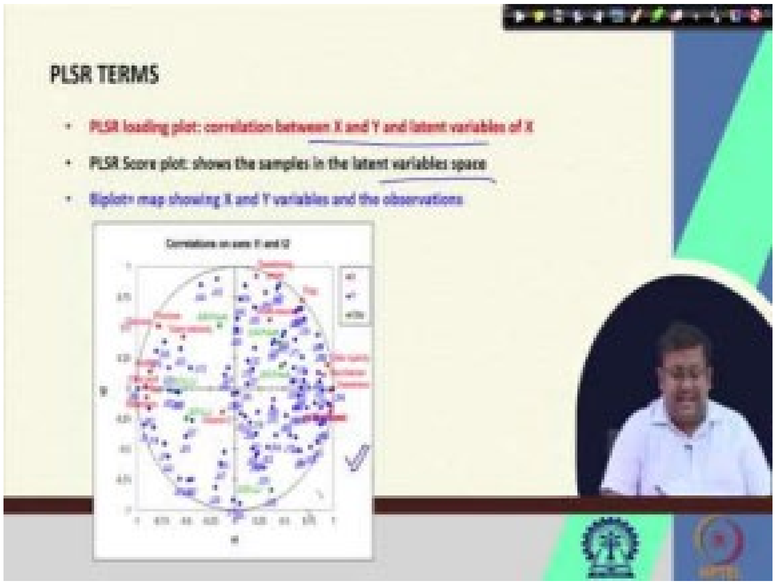
So, let us move ahead and see other ways of representing the PLSR method. Here you can, you see that this is Y which can be represented at that XWh Ph Wh inverse C dash h plus Eh whereas, this is the decomposition of the whole Y and it is an extended representation of Y here, these beta coefficient is this Wh P dash h Wh inverse C dash h.

So, here Y is the matrix of the dependent variables, X is, so, these X is a matrix of the explanatory variable and then this all other variables which are mentioned here like Th, then Ch then W star h then Wh Ph these are the metrics generated by the PLSR algorithm and here Eh is the matrix of the residuals.

So, the matrix B of the regression coefficients or beta of Y on X with h components generated by PLSR algorithm can be calculated using this formula. So, here you can see ultimately we are getting Y equal to beta X plus the residuals. So, here we can see that ultimately our model is linear.

So, PLSR leads to a linear model as the OLS and PCA do. So, in case of ordinary least squares in case of principal component regression we know that both the models are linear model. So, similarly, in case of PLSR also, ultimately the model is linear model.

(Refer Slide Time: 23:51)



PLSR TERMS

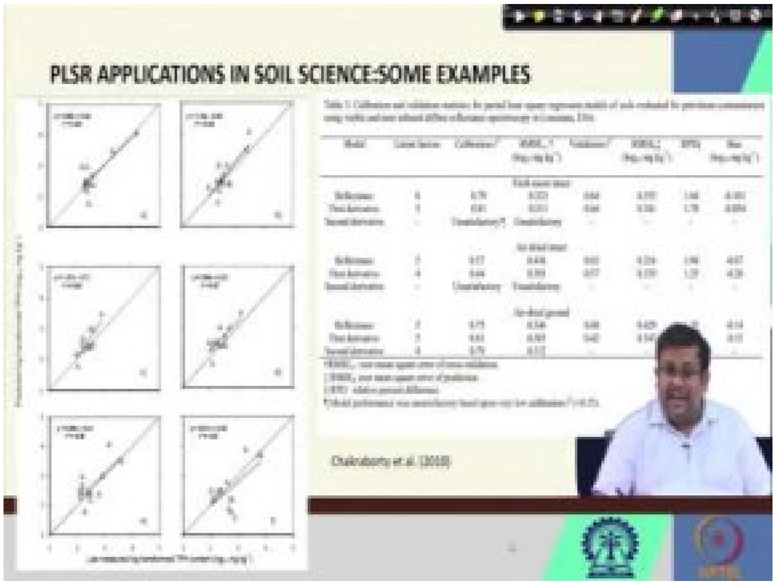
- PLSR loading plot: correlation between X and Y and latent variables of X
- PLSR Score plot: shows the samples in the latent variables space
- Biplot= map showing X and Y variables and the observations

So, there are certain PLSR terms, for example, the PLSR loading plot for example, which shows the or PLSR sometime we call it PLSR correlation loading plot, which is the, which shows the correlation between X and Y and latent variables of X.

So, here you can see an example of PLSR loading plots. Also there are a PLSR score plot which shows the samples in the latent variable space and biplot, biplot is, this is a biplot actually, this is the biplot which is a map showing X and Y variables and that observations.

So, these are the terms which matches well with the principal component analysis. So, here you can see these red dots are the X variables and Y dots are the Y variables and the observations are marked using these green dots. So, this is the example of biplot where we are seeing both X and Y variables as well as the other observations.

(Refer Slide Time: 25:06)



PLSR APPLICATIONS IN SOIL SCIENCE:SOME EXAMPLES

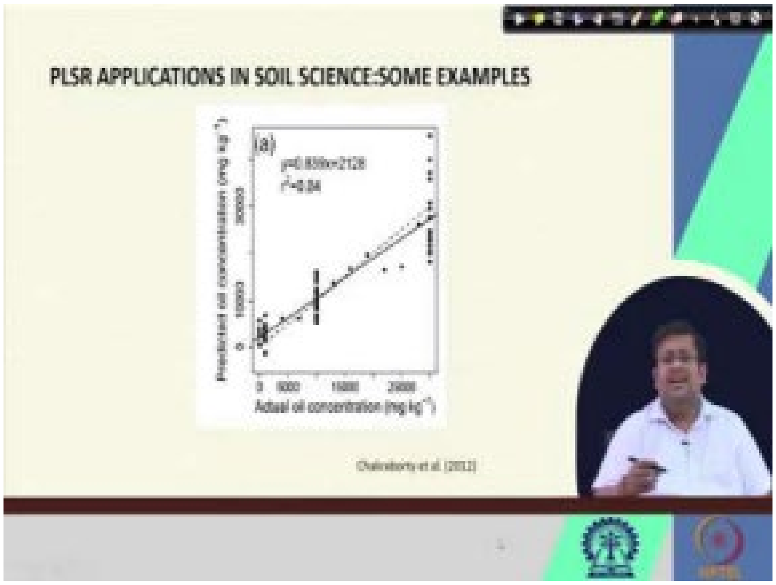
Model	Latent factors	Calibration r²	RMSE_CV (log₁₀ mg kg⁻¹)	Validation r²	RMSE_P (log₁₀ mg kg⁻¹)	RPD	Bias (log₁₀ mg kg⁻¹)
			Field-moist intact				
Reflectance	6	0.79	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
First derivative	5	0.81	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Second derivative	–	Unsatisfactory¶	Unsatisfactory	–	–	–	–
			Air-dried intact				
Reflectance	5	0.77	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
First derivative	4	0.84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Second derivative	–	Unsatisfactory	Unsatisfactory	–	–	–	–
			Air-dried ground				
Reflectance	5	0.75	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
First derivative	5	0.83	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Second derivative	4	0.76	[illegible]	–	–	–	–

Chakraborty et al. (2010)

Now, let us see some of the applications of PLSR in soil science application. So, this is one application from our research group, which where we have tried to predict the petroleum contaminated soils, in our previous lecture, we have talked about the petroleum contaminated soils. So, which we tried to predict the petroleum contamination in the soil using the partial least squares regression, using the spectral data from the diffuse reflectance spectroscopy.

So, you can see here an example where the partial least squares regression was used to predict the total petroleum hydrocarbon which is an indicator of petroleum contamination in the soil. So, we have predicted the total petroleum hydrocarbon using spectral data as input variables to predict the TPH through partial least squares regression and these are some of the model statistics from different conditions. So, that shows the application of partial least squares regression for soil environmental studies.

(Refer Slide Time: 26:26)



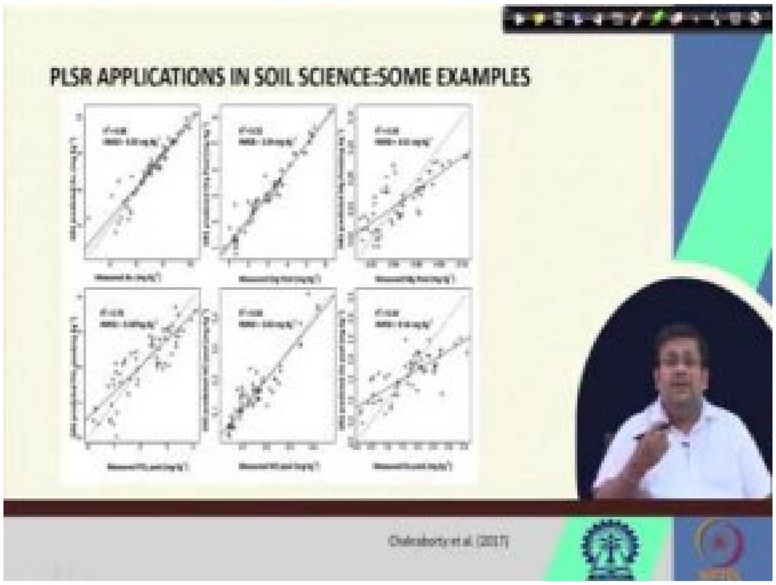
PLSR APPLICATIONS IN SOIL SCIENCE:SOME EXAMPLES

(a)
y=0.839x+2128
r^2=0.84
Predicted oil concentration (mg kg^{-1})
0
10000
30000
0
5000
15000
25000
Actual oil concentration (mg kg^{-1})

Chakraborty et al. (2012)

Also in another research also we tried to develop another partial least squares regression model to predict the soil petroleum hydrocarbon contamination and we got a very good R squared values of 0.84 while predicting the pair, the oil concentration in the soil.

(Refer Slide Time: 26:47)



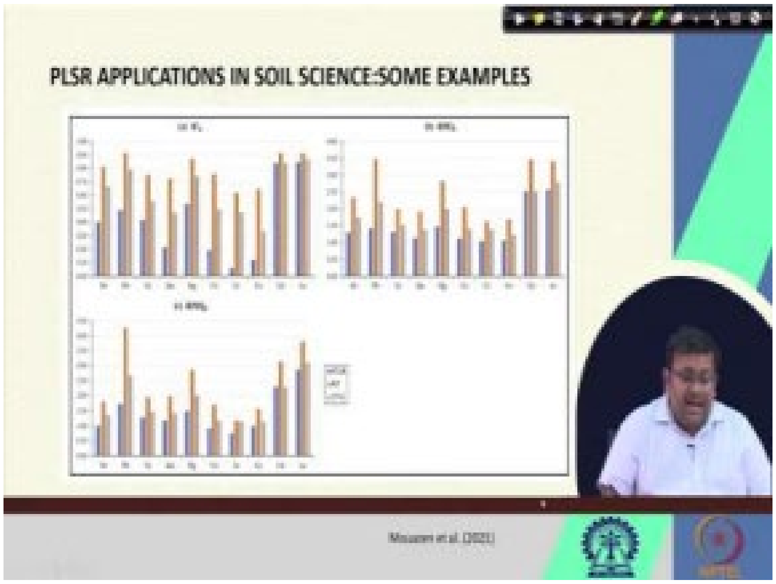
PLSR APPLICATIONS IN SOIL SCIENCE:SOME EXAMPLES

Chakraborty et al. (2017)

Similarly, in another research we have utilize this partial least squares regression for predicting the different pools of soil arsenic. Now, soil arsenic it is an important heavy metal, which has different deleterious effects for human. So, we tried to predict the soil arsenic content using the diffuse reflectance spectroscopy or spectral method using the partial least squares regression and not only that total arsenic we also tried to measure the fraction with which the soil arsenic are present.

So, in soil arsenic is present in different forms in conjunction with different pools. So, we try to predict those pools using partial least squares regression, whereas, our predictors were the spectral variables. So, we have seen that our method is useful for predicting. So, partial least squares regression has produced very good R square for predicting the total arsenic as well as the organic matter associated arsenic and also the HCL pool of arsenic and you can see the phosphate pool of arsenic is also being measured with the satisfactory accuracy.

(Refer Slide Time: 28:26)



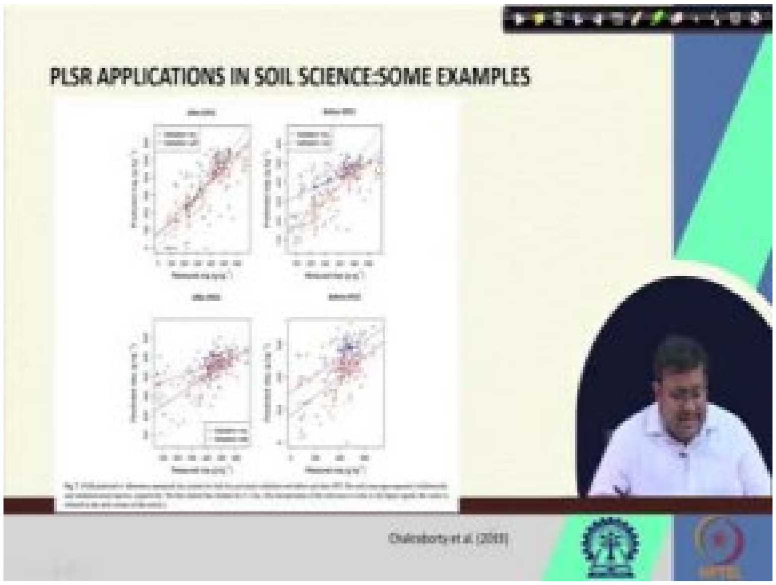
PLSR APPLICATIONS IN SOIL SCIENCE:SOME EXAMPLES

Mouazen et al. (2021)

So, this shows the application of partial least squares regression in different spectroscopic based soil characterization studies. Here, there is another recent application where Mouazen et al has tried to use the partial least squares regression for predicting the different heavy metals and you can see, they have compared three different models, partial least squares regression model, random forest model and support vector machine, these three models they have compared for predicting different heavy metals in the soil.

You can see the R squared values they have compared the R squared values they had compared the RPD values which is another important metrics called, the RPD is another important metrics which is used to denote the model accuracy. And we will going to do, we are going to discuss more about RPD while we will be discussing the spectroscopic method and also RPIQ, which is another important indicator, so, we can see that they have tried to use the partial least squares in along with other models to compare their performance for predicting different soil heavy matters.

(Refer Slide Time: 29:44)



PLSR APPLICATIONS IN SOIL SCIENCE:SOME EXAMPLES

Chakraborty et al. (2013)

Also, here you can see that we have also used the partial least squares regression model for predicting the soil clay content. Soil clay content so, numerous studies have utilized partial least squares regression for predicting the soil clay content using the spectral methods, not only the spectral method, but other methods also they have used.

(Refer Slide Time: 30:10)

So, one example let us, one or two examples let us see from the plant research. So, one research recently was published which is predicting the grape sap flow in a greenhouse. So, remember that understanding the variation which is published in 2021 by Peng et al. So, they conclude that the understanding the variation in a sap flow rates and the environmental factors that influence sap flow is important for exploring grape water consumption, patterns and developing reasonable greenhouse irrigation schedule.

So, they have established three irrigation levels for this study, one is adequate irrigation W1, moderate deficit irrigation W2, and deficit irrigation W3 and grape sap flow estimation models were constructed using the partial least squares and random forest algorithms, we are going to talk about random forest in our upcoming classes. And the simulation accuracy of the stability of these models were also evaluated.

(Refer Slide Time: 31:15)

So, these are the model stats for this random forest and partial least squares regression. So, this shows the application of these types of models for predicting different growth stages of the crop based on certain predictive variables.

(Refer Slide Time: 31:32)

We can see another application here, where these partial least squares application in plant research. So, here the partial least squares regression was applied for rapid assessment of algal biomass and pigment contents. So, here you can see that this PLSR model was used to predict
this is the biomass and this is chlorophyll A content and this is the chlorophyll B content and so on, so forth. Different other parameters from two different algal species.

So, using the partial least squares regression you can see that it has a wide spread application in both soil science and plant science applications. And this is one of the most important chemo metric model. Why we call it chemo metric model? Because this type of model is utilized for measuring the chemistry of any metrics or any substance through regression.

So, these type of models or PLSR model have been widely used in different domains of agriculture, specifically in soil and crop and so people are also nowadays trying to improve the PLSR model and by selecting some of the variables which are called PLSR VIP methods, so for improving the model prediction accuracy.

(Refer Slide Time: 33:12)

So, these are the references for this lecture. I hope that you have got a basic understanding of what is PLSR and what is the basic principal of PLSR and how it differs from the other methods like principal component regression and we have seen couple of examples of PLSR application in soil and crop.

So, guys I hope that you have got some new information in this lecture. So, let us wrap up this lecture and let us meet in our next lecture to start from here and then we will go and discuss other models also. Thank you very much.