

Machine Learning for Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 12

Principal Component Analysis and Regression Applications in Agriculture (Contd.)

Welcome friends to this lecture 12 of NPTEL Online Certification Course of Machine Learning for Soil and crop Management. We are in week 3 and this is the second lecture of week 3. And in this week we are continuing principal component analysis and regression applications in agriculture. In our first lecture of week 3, we have already discussed what is principal component analysis.

Just to summarize, principal component analysis is a dimensionality reduction approach, sometime it can be both supervised and unsupervised. In case of classification regression problem, it is supervised; sometimes it is unsupervised, when you want to see the clustering among the samples. And principal component analysis as I told you it is a dimensionality reduction technique, where the correlated variables...

It generally applies for high dimensional data, where multiple features are correlated to each other, or in other words when there is a multicollinearity available. So, when there is a multicollinearity, that also can involve that also can induce over fitting. So, to remove this problem, we generally go with the, we generally go with the principal component analysis.

Principal component analysis deals with the calculation of principal components, which are the linear combination of the original variables in the feature space. And we linearly combine them together in a single principal component and this principal components are ordered and named based on their relative predictive power. So, principal component 1 always show higher predictive power followed by principal component 2 and so on so forth.

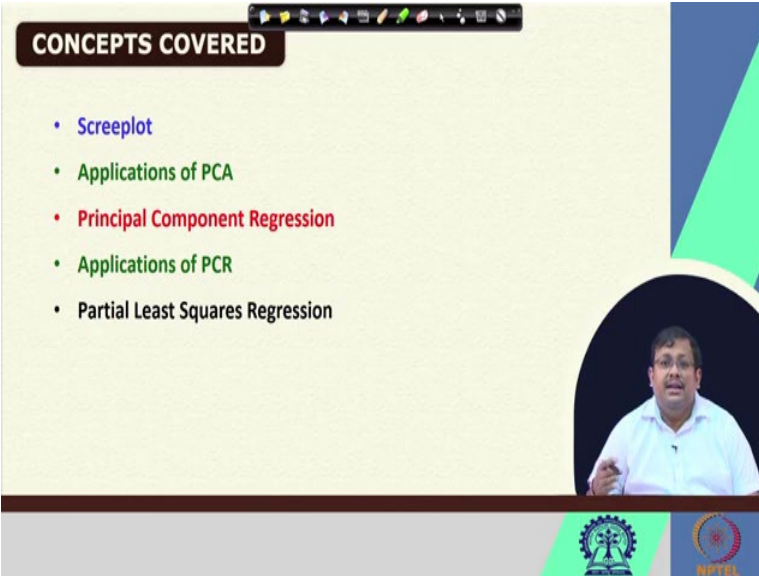
So, with if there is a n number of features in a data set, we can actually calculate n number of principal components. And then we can selectively remove some principal components, which are not showing very good explaining power for the data set, in that way we can reduce the dimension without neglecting the importance of relatively weak predictor because while calculating the principal components, we capture their importance also.

We capture the importance is the linear combination; we capture the information coming from both the important features and also some unimportant features also. So, it does not matter if we selectively remove some of the principal components, we can still capture their contribution of the, we can still capture the contribution of the unimportant variables.

So, how to calculate the principal components? What are the eigen values and eigen vectors? We have discussed based on the matrix algebra, although there is no scope of mathematically explaining all these terms at this point of time because our major focus is to show the application of this technique for soil and crop problems.

So, today in this lecture, we will start from here and I will discuss how to select the important principal components for subsequent analysis. What are the considerations? And then we will discuss what is principal component regression? And we will see some examples of application of principal component analysis and then principal component regression. We will also see some example of principal component biplot.

(Refer Slide Time: 4:43)



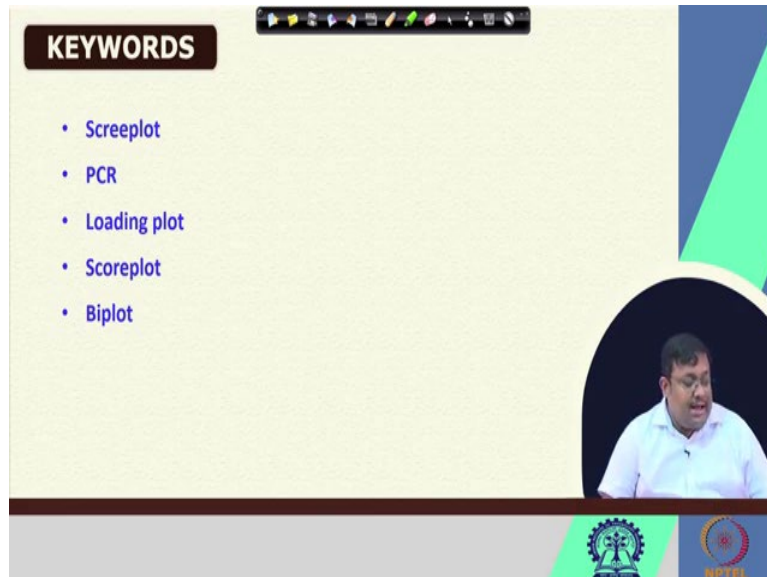
The image shows a presentation slide with a title bar at the top that reads "CONCEPTS COVERED". Below the title, there is a bulleted list of topics: "Screplot", "Applications of PCA", "Principal Component Regression", "Applications of PCR", and "Partial Least Squares Regression". The text "Principal Component Regression" is highlighted in red. In the bottom right corner of the slide, there is a circular video inset showing a man in a white shirt speaking. At the bottom of the slide, there are two logos: the Indian Institute of Technology (IIT) logo on the left and the NPTEL logo on the right.

So, these are the concepts, which we are going to cover in this lecture. We are going to talk about the screeplot, we are going to talk about the application of PCA. Then we are going to talk about the application of principal component regression, or PCR. We are also going to talk about the

applications of PCR. And then partial least squares regression, if time permits, but that will we are going to mainly discuss up to applications of PCR.

And briefly we will be discussing what is partial least square segregation. However, partial least squares regression will be discuss in details in our upcoming lectures.

(Refer Slide Time: 5:27)



Now, these are the important keywords for this lecture. We are going to discuss screeplot, we are going to discuss PCR. We are going to discuss what is loading plot? We are going to discuss what is score plot? And also we are going to discuss the biplot.

(Refer Slide Time: 5:46)

HOW MANY PCs TO KEEP

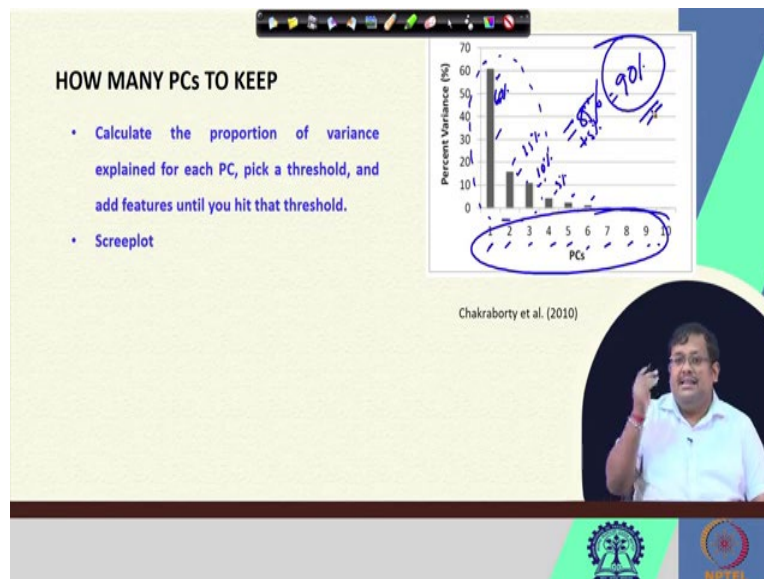
- For n original dimensions, correlation matrix is $n \times n$, and has up to n eigenvectors. So n PCs.
- Where does dimensionality reduction come from?

Now, when we do the principal component, suppose there are n number of variables and we are doing the principal component analysis and calculating the n number of principal components. The most important question comes to our mind is ok, now what; how many principal components we should keep for subsequent analysis?

How many of them are really important? So, for n number of principal components, correlation matrix is of course n into n and this will give us n number of eigen vectors. And n number of eigen vectors means, will be having n number of principal components.

So, how does this dimensional reduction come from? I told you briefly in our previous lecture, that we can selectively remove some of the unimportant principal component. But how we know that which principal component is unimportant? To identify the unimportant and important principal component, we generally use screeplot.

(Refer Slide Time: 6:56)



So, here you can see this is one of the way, this is the example of screeplot. But before discussing the screeplot let us see the approach, the approach is calculate how to how to develop the screeplot? So, first of all we calculate the proportion of the variance explained for each principal components and then we pick a threshold and add features until you hit that threshold.

So, generally for example this example you can see this is a called a screeplot and here we have used the spectral data to calculate the principal components. And we have calculate 10 principal component, principal component 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. And simultaneously we have calculate their percent variation, percent variance, or proportion of variance they can explain in the data set.

So, we have seen of course as the name suggests, principal component 1 will show the highest variance followed by principal component 2, principal component 3, principal component 4, principal component 5, and up to principal component 10 but how to select, how many of them are important? Now, it generally varies from one application to another application.

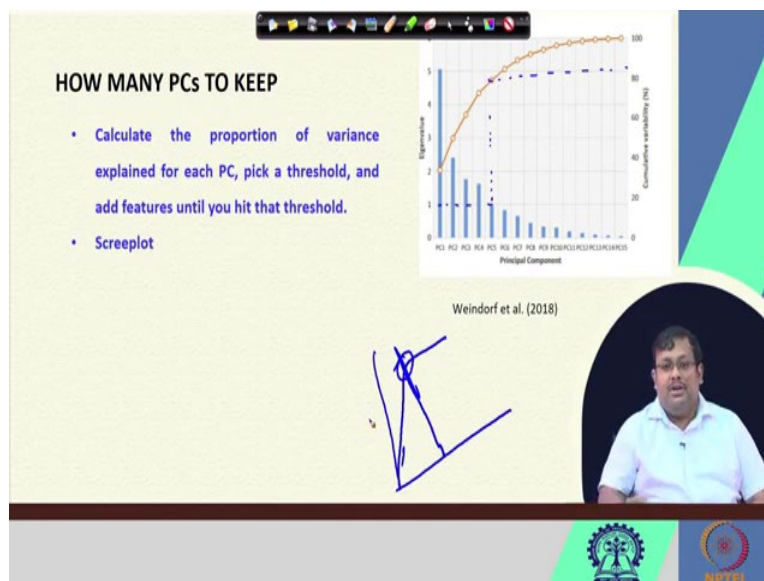
Suppose you have that this principal component 1 is showing almost 60 percent of the variance. And this principal component 2 is showing 15 percent of variance, principal component 3 is showing 10 percent of the variance. So, cumulatively 60 plus 15 plus 10, that means we are getting 85 percent variability, just from this first 3 principal components.

So, if we fix a threshold of including the principal components up to 85 percent, then only we select these three principal components for subsequent analysis. And we remove all the seven remaining principal component analysis principal components. In other words you can also see here the number 4 principal component showing almost 5 percent.

So, if we add it up it will give us 90 percent variability. So, you can see instead of using all 10 principal components, if we use only first 4, then only we can capture 90 percent of the principal 90 percent of the variance in the data set. So, 90 percent of the variance in the data set, we can we can capture by only these 4 principal components.

So, it depends on our threshold, we set up a threshold and then we add the feature, or add the principal until we get to that point of 90 percent principal component. We can see here that up to 4, we are getting this 90 percent and then we will stop. So, this is called scree plot. Sometime the screen plot you will also see are shown based on the eigen values also.

(Refer Slide Time: 10:35)



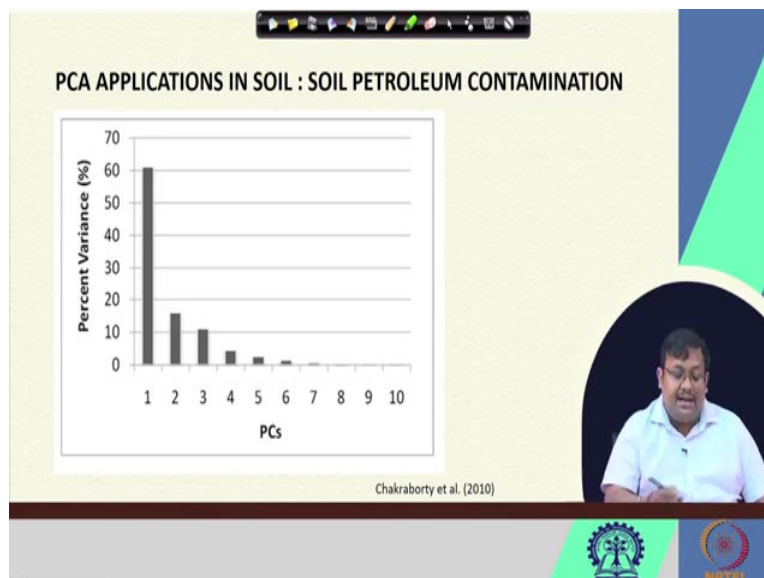
So, another approach for calculating here you can see here another example is given we they have calculated the eigen values and they have ordered the eigen values of course, it is another application. And here you can see for principal component 1, they have calculated the eigen values and they have plot the eigen values. And at the simultaneously you have they have plot the cumulative variability shown by these principal components.

So, if we fix any threshold, suppose we are fixing any threshold here, in terms of cumulative variability. Then we can see up to 5 principal component, they are showing the 80 percent variability in the data set. So, we will keep 5 principal components and will remove the rest of the principal components. Other way people also use the principal components with the eigen values greater than 1.

So, the eigen values, which are greater than 1, sometime also researcher used to keep only those principal components because that shows the more important features in the principal component space. So, these are known as the screeplot and these are the ways through which we can keep the principal component. In other words sometime also people use the elbowing, or elbow in this cumulative variability.

Suppose the cumulative variability in the principal component goes like this and suppose and then it reaches a plateau, then people always go with the principal components up to this point also. It is called the elbow point. So, they can select the principal components up to the elbow point. So, there are different ways through which you can calculate, or you can select the principal components for subsequent analysis.

(Refer Slide Time: 12:36)



So, this is an example of principal component analysis screeplot, you can see here if we see that first two principal components. So, first two principal components collectively explain almost 80

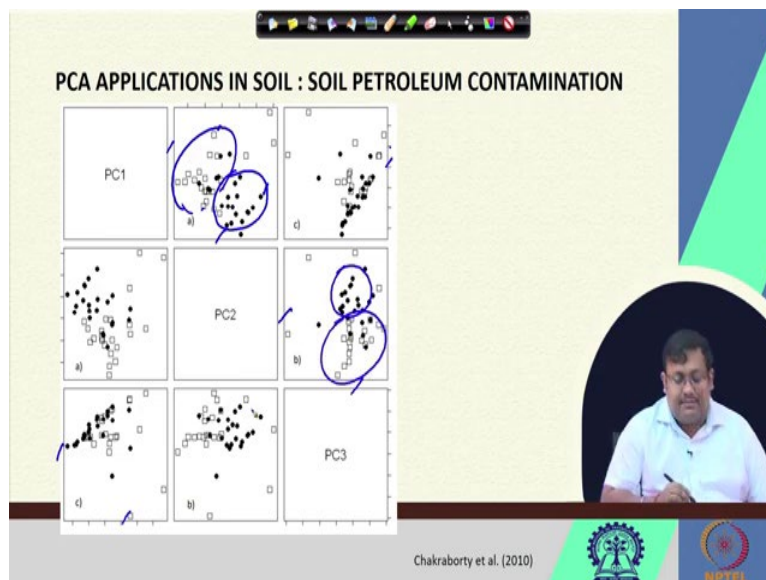
percent of the variation. So, in the subsequent PCA analysis, we can we should we should focus only based on the first 2 , or first 3, if we combine first 3 we are getting around 85 percent variation.

So, if we combine if we if we plot our data into this principal component space, we can use up to first 3 principal components. So, this is a one example of soil petroleum contamination. Now, I am showing some soil application and crop application. This data was gathered from the principal component analysis of a from the spectral data, which we have gathered using a spectro radiometer.

The spectro radiometer and its principal and how it collects the data we will get we will discuss in our, in coming weeks. But at this point of time remember that, there are thousands of spectral features starting from 350 to 2500 nanometer wavelength, at each 10 nanometer interval 350 nanometer, 360 nanometer, 370 nanometer, up to 2500 nanometer.

So, using these spectral features, we have calculated this principal components 10 principal components and from this 10 principal components, we have seen that up to 3 principal components our data up to 3 principal components we can explain up to 85 percent of the data variability.

(Refer Slide Time: 14:35)

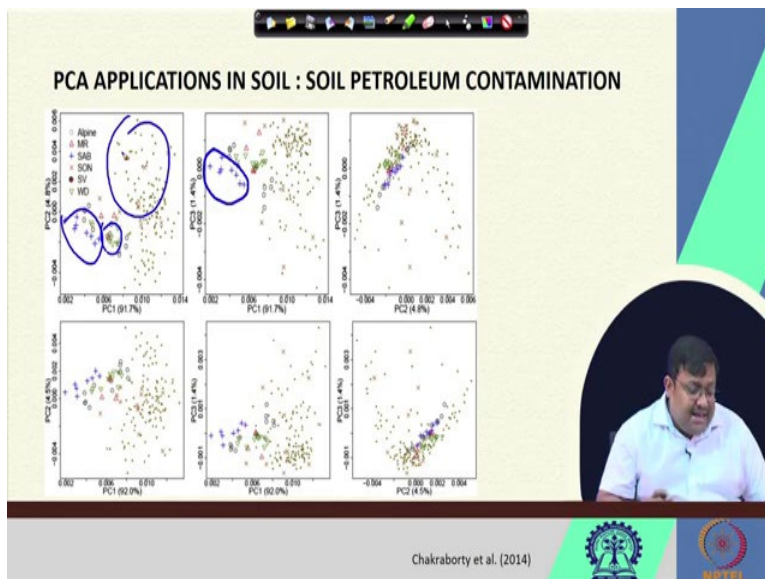


And then you can use this, this score plot and you can see here these are the PCA score plot. In the PCA score plot, these you can clearly see these solid points are contaminated samples and these hollow points are non-contaminated samples. So, using this is PC 1 versus PC 2 plot, this is PC 2 versus PC 3 plot, and this is also PC 1 versus PC 3 plot. So, also this is PC 1 versus PC 3 plot.

So, you can see that using this type of principal component, we can capture, or we can cluster different samples. So, here you can see we can clearly classify the samples, or cluster the samples, which are contaminated and which are non-contaminated. So, this is the application of principal components.

And for this for seeing this we do not have to we do not have to plot up to principal component 10, principal component 1 and 2 will be sufficient and then principal component 2 versus principal component 3 are also showing the clustering features as you can see here, the cluster they are clustering the data based on the similarity in their spectral feature.

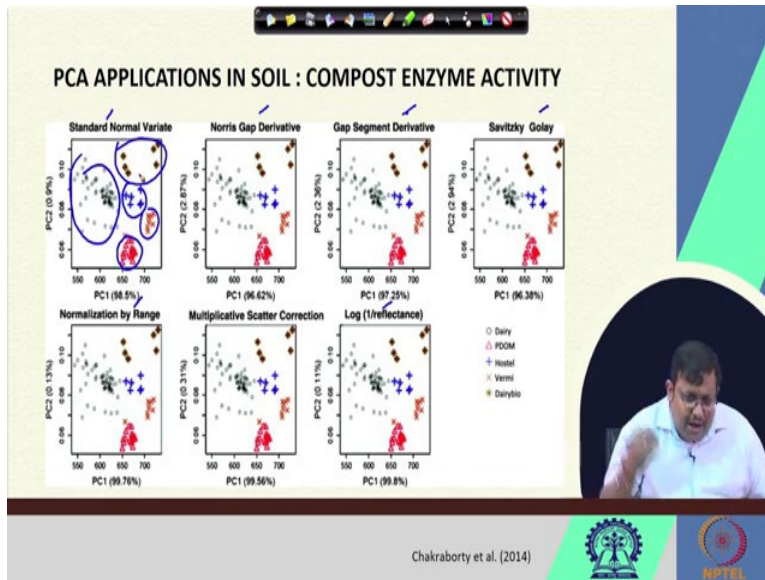
(Refer Slide Time: 16:03)



You can see here another example for soil petroleum contamination, published in 2014. And you can see here also the data set another data set is , they are showing the clustering among the data collected from different, collected from different locations. So, here you can see one cluster, here you can see one cluster, of course there are some mix mixing between the cluster samples. But

you can clearly see some cluster here. So, this is how you can identify the clustering among the data set using the principal component 1.

(Refer Slide Time: 16:53)



Another example we have used for identifying the compost enzyme activity, using the spectral method. So, here we have calculated, we have taken the spectra of different types of compost, five to six different types of compost. And then we have plotted their principal components.

So, here you can see for different types of spectral preprocessing. These are different types of spectral preprocessing; you can see here normal, very standard normal variate, Norris Gap Derivative, Gap Segment Derivative, Savitsky Golay, normalization by range, multiplicative scatter correction. And then log 1 by reflectance.

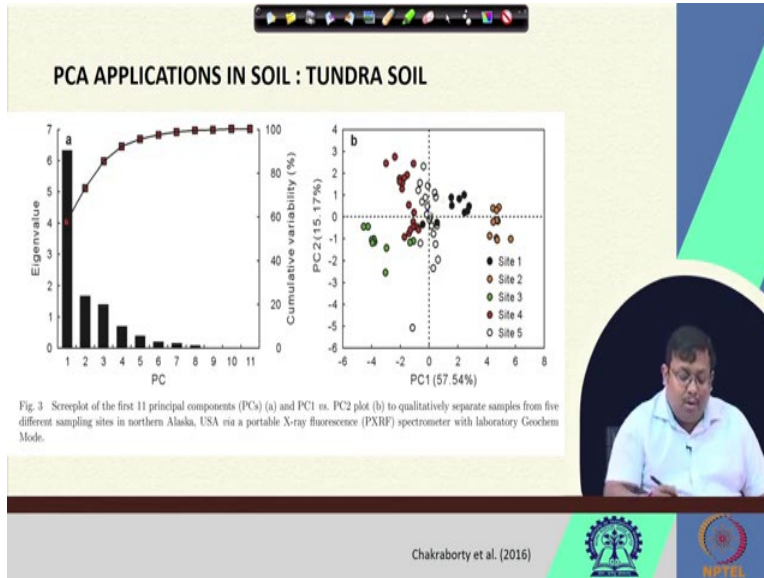
So, we will discuss this spectral processing in our upcoming weeks, but remember that once we do this kind of spectral processing and then we do the principal component analysis, we can get this type of clustering. So, you can see clearly identify, the clusters of sample based on the similarity of their spectral features. At this similarity of the spectral features also identity also indicate the similarity among their property.

So, that means these compost samples, these compost samples are different than this compost cluster, this is also different than this compost cluster, this is also different than this compost

cluster, this is also different than this compost cluster. And they might be varying from each other based on certain properties.

So, PCA gives us a qualitative identification of separation, or possible clustering among the data set. So, this is one example of based on the compost enzymatic activity.

(Refer Slide Time: 18:49)

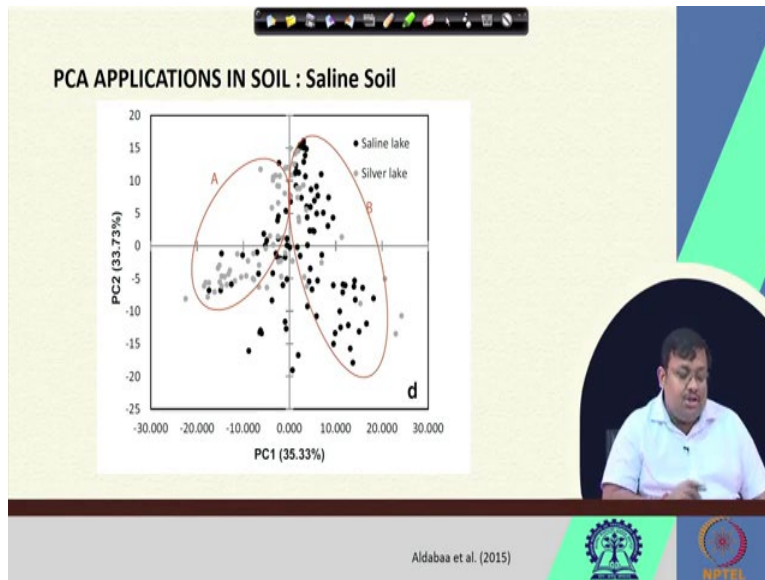


Another example you can see here, this is the PC application in tundra soil. So, we have collected soil from 5 different sites and based on the principal component, we have first produced this screeplot. And this screeplot from this screeplot you can see by combining this PC 1, PC 2, cumulatively showing around 72 percent variability.

And this is the principal component 1 versus principal component 2 score plot, which showing the differences and clustering among different samples coming from different slides, different sites. So, here this PC 1 versus PC 2 plot, qualitatively separate samples from 5 different sampling sites in northern Alaska.

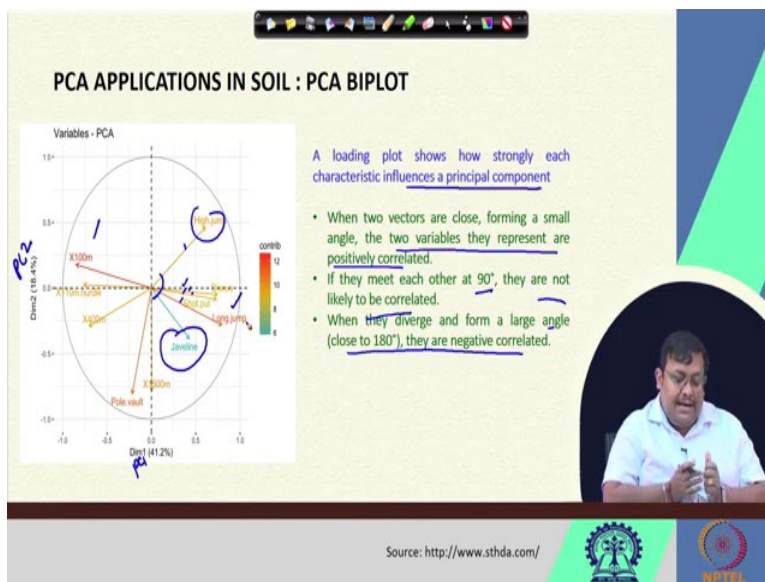
And these are actually calculated based on the elemental results collected using the portable X-ray fluorescent spectrometer, we are going to also discuss that in our coming lectures.

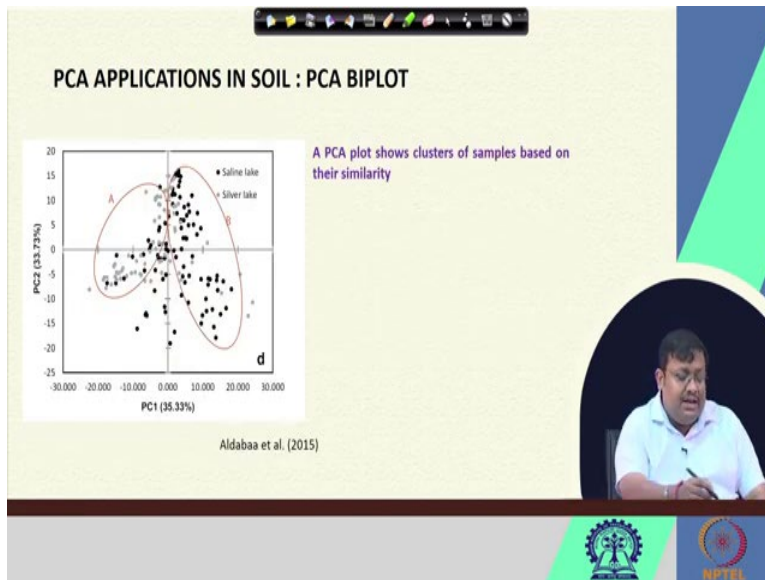
(Refer Slide Time: 19:53)



We have seen that principal component analysis can be useful for separating soils coming from two different types of lakes. And showing variance in the salinity, you can see here this principal component 1 versus principal component 2. So, we can see some clustering pattern and also that shows the sample heterogeneity among the two different types of areas, or sampling site, and also shows the similarity among those samples coming from the same site. So, you can see here, this is how we apply the principal component analysis.

(Refer Slide Time: 20:39)





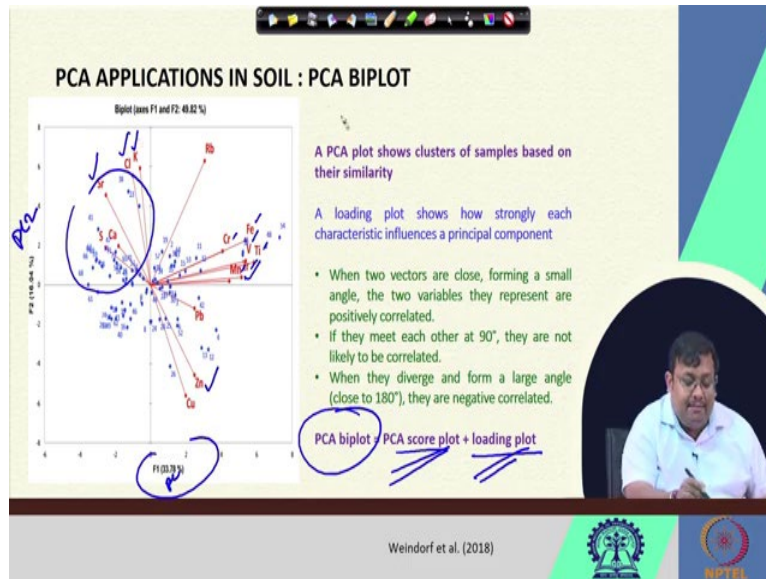
Now, the next question comes what is principal component biplot? So, principal component score plot you have already seen this is called the principal component score plot, where we calculate the principal component score for individual data point and plot them into principal component space. Now, what is a loading plot? Loading plot shows how strongly each characteristic influence a principal component.

So, suppose this is a principal component 1, that is denoted by PC dimension 1, and this is PC 2. And you can see these are the features, or the variables. Now, when two variables are vectors are close forming a small angle, then the two variables they represent are positively correlated. So, here you can see this is a vector called discuss and here short part, long jump, they are very much close and forming a small angle so they are these 3 are positively correlated.

So, if they meet each other at 90 degree they are not likely to be correlated because we have already seen that when there is a orthogonal to each other there is no correlation. So, one example could be this javelin and high jump they are almost 90 degree they are making a 90 degree angle so they are not correlated to each other.

However, when they diverge and form a large angle which is close to 180 they are negatively correlated. So, here you can see this long jump and 100 meter, 100 meter run they are almost 180 degree far apart and they are negatively correlated. So, from the loading plot we can see the relative influence or relative closeness among different features based on their angle.

(Refer Slide Time: 22:59)



So, what is a biplot? This term comes very frequently PCA biplot is, so we know that a PCA plot generally shows the cluster of sample based on their simplicity and we know that the loading plot shows how strongly each characteristic will influence a principal component. Now, the combination of principal component score plot and loading plot is known as principal component biplot.

So, here you can see this is an example of principal component biplot, it is principal component 1 and this is principal component 2 and these blue dots are the principal component scores of several compost samples and here you can see the elements which are reported by the PXRF or portable XRF we have produced there and these vectors, the vectors of those elements which are features which you have used for calculating this principal components.

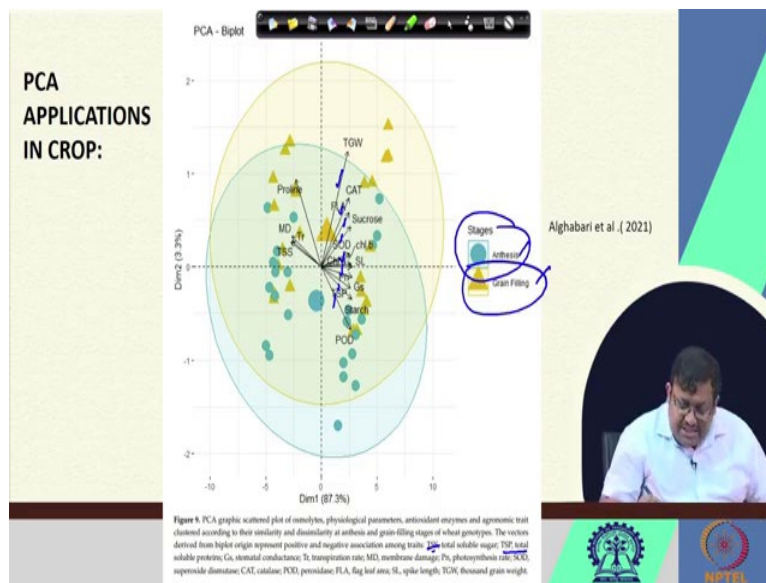
So, here you can see these are the loading so we can see iron, vanadium, titanium, zirconium, manganese, chromium they are highly correlated to each other. Whereas, manganese and potassium are non-correlated or chlorine is non-correlated to each other and we also can see that zinc and strontium are inversely related. So, we also can see what are the features which are important for these samples from this PCA biplot.

So, in the PCA biplot not only we can see the cluster, we can see the variables which are important for those samples and also we can identify the correlation among the samples so from

here we can see that for calculating the principal components this copper for this samples, copper, zinc and lead are important features.

Whereas, for these samples, for these samples strontium, calcium, sulphur, chloride and potassium are important features. So, this is how we get the idea of the distribution of the samples as well as their, as well as their loading from the PCA biplot.

(Refer Slide Time: 25:27)



So, here one example of principal component analysis in crop. Here you can see that this PCA graphics scatter plot of osmolytes and physiological parameters and antioxidant enzyme and agronomic trait clustered according to the similarity, so clustered according to their similarity and dissimilarity at anthesis and grain filling stage. So, there are two stage, anthesis stage and grain filling stage and they cluster together of wheat genotypes.

So, from this, so the vectors derived from the biplot origin represent positive and negative association among the traits. So, you can see these are different types of traits based on which we have clustered, they have clustered these, they have defined these two clusters, so the vectors derived from the biplot origin represents a positive and negative association just I have discussed in our previous slide.

We can also see which are the features, which are positively correlated and which are negatively correlated. So, we can see the TSS generally is the total soluble sugar, TSP is the total soluble

protein, then GS is the stomatal conductance, Tr is the transpiration rate, MD is the membrane damage, PN is the photosynthetic rays, SOD is the superoxide dismutase enzyme, CAT is a catalyst enzyme, POD is a peroxidized enzyme, FL is the flag leaf area, SL is the spike length and TGW is a thousand grain weight.

So, these are the feature based on which we have clustered the sample into two category, one is the, we have clustered the wheat genotypes into grain filling stage as well as the anthesis stage. So, this is how, this is the very good application of PCA in crop. So, now you have understand why we use the PCA.

(Refer Slide Time: 27:28)

PRINCIPAL COMPONENT REGRESSION (PCR)

- Principal Components Regression is a technique for analyzing multiple regression data that suffer from multicollinearity
- When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value
- Based on PCA
- Regression where PCs are independent variables
- One typically uses only a subset of all the PCs for regression, making PCR a kind of regularized procedure and also a type of shrinkage estimator

The slide includes a video inset of a man in a white shirt speaking. There are also some hand-drawn annotations in blue ink, including a box around the text 'Regression where PCs are independent variables' and a line pointing to the text 'One typically uses only a subset of all the PCs for regression, making PCR a kind of regularized procedure and also a type of shrinkage estimator'. The slide also features a logo for NPTEL at the bottom right.

Now, next comes the principal component regression. Principal component regression is a technique for analyzing the multiple regression data that suffers from multicollinearity when multicollinearity occurs least square estimations are unbiased but their variants are large so they may not be, may be far away from the true value. So, this principal component regression generally it is very simple.

So, first we do the principal component analysis on to the independent variable matrix, matrix of the independent variables and then the regression, in the second step we do the regression where PCs are independent variables. So, instead of incorporating the original variables as an

independent variable we first calculate the principal components and select the important principal components and incorporate them in the model as independent data set.

So, one typically uses only a subset of all the pieces for regression making PCR a kind of regularized procedure and also a type of shrinking estimator. So, this is how the principal component regression is done.

(Refer Slide Time: 28:42)

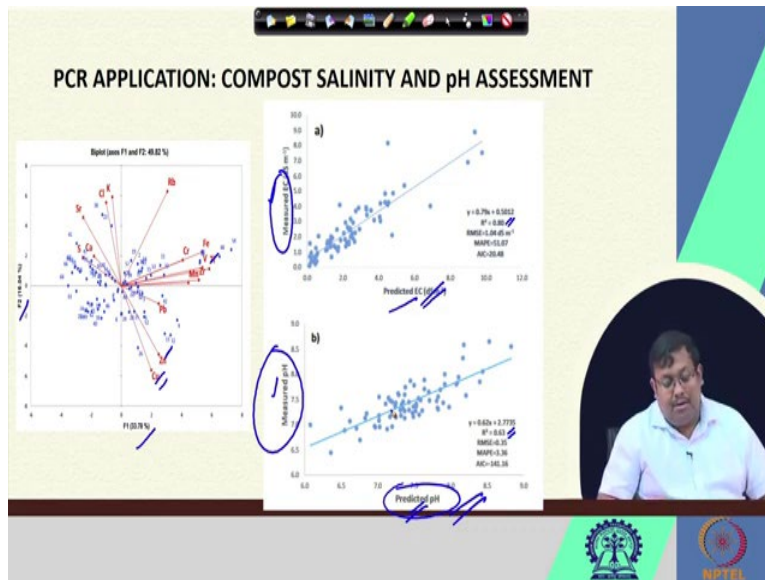
PRINCIPAL COMPONENT REGRESSION (PCR)

- The PCR approach involves constructing the first M principal components, and then using these components as the predictors in an OLS linear regression model
- The key idea is that often a small number of PCs suffice to explain most of the variability in the data, as well as the relationship with the response
- We assume that the directions in which X_1, \dots, X_p show the most variation are the directions that are associated with Y
- When performing PCR, predictors should be standardized prior to generating the principal components

The PCR approach involves constructing the first M principal components and then using this component as the predictor in the ordinary least square regression. So, principal component is similar to ordinary least square regression, only the exception is here the independent variables are the principal components. So, the key idea that often a small number of PCs suffice to explain most of the variability in the data as well as the relationship among the response.

So, we assume that the direction in which this X_1, X_2, \dots, X_p show the most variation are the direction that are associated with the Y or dependent variable. So, when performing the PCR predictors should be standardized prior to generating the principal components.

(Refer Slide Time: 29:33)



So, it is one example of principal component regression analysis for measuring for compost salinity and pH prediction. So, I have already showed you this biplot in one of my previous slide you can see here, here this is principal component 1 and principal component 2, we can see the importance of the features and this is the scores, these are the scores.

So, after selecting the important principal components we have predicted the compost electrical conductivity or salinity and compost pH and we can see that and this is the measured pH versus predicted pH, measured pH versus predicted pH and both these cases we can see in case of electrical conductivity we are getting very high R square values or coefficient of determination, in case of measurement of pH also we are getting moderate prediction accuracy.

So, that shows one application of principal component regression. Here, remember the regression is done where measured electrical conductivity of the compost samples are target whereas the predicted values are calculated based on the principal components which are calculated from the input variables. And similar in the measure pH relationship versus predicted pH relationship also.

So, the pH are predicted when ECs are predicted based on the principal component analysis and then we are showing this prediction plot. So, guys I hope that you have now a very good idea

about the principal component analysis, their application and biplot, their application and also the principal component regression and their application in soil and crop sectors.

(Refer Slide Time: 31:36)

REFERENCES

- Aldabaa, A.A.A., D.C. Weindorf, S. Chakraborty, A. Sharma, and B. Li. 2015. Combination of proximal and remote sensing methods for rapid soil salinity quantification. *Geoderma*, 239-240:34-46. doi: <http://dx.doi.org/10.1016/j.geoderma.2014.09.011>.
- Alghabari, F.; Shah, Z.H.; Elfeel, A.A.; Alyami, J.H. Biochemical and Physiological Responses of Thermotolerant Wheat Genotypes for Agronomic Yield under Heat Stress during Reproductive Stages. *Agronomy* 2021, 11, 2080. <https://doi.org/10.3390/agronomy11102080>
- Chakraborty, S., D.C. Weindorf, C.L.S. Morgan, Y. Ge, J. Galbraith, B. Li, and C.S. Kahlon. 2010. Rapid identification of oil contaminated soils using visible near-infrared diffuse reflectance spectroscopy. *Journal of Environment Quality* 39(4): 1378-1387.
- Chakraborty, S., B.S. Das, N. Ali, B. Li, M.C. Sarathjith, K. Majumder, and D.P. Ray. 2014. Rapid estimation of compost enzymatic activity by spectral analysis method combined with machine learning. *Waste Management* 34: 623-631.
- Chakraborty, S., D.C. Weindorf, B. Li, N. Ali, K. Majumder, and D.P. Ray. 2014. Analysis of petroleum contaminated soils by spectral modeling and pure response profile recovery of n-hexane. *Environmental Pollution* 190: 10-18.
- Weindorf, D.C., S. Chakraborty, B. Li, S. Deb, A. Singh, and N.Y. Kusi. 2018. Compost salinity assessment via Portable X-ray fluorescence (PXRF) spectrometry. *Waste Management* 78:158-163.

Now let us, these are the references and let us wrap up our lecture here and in our next lecture we will start from the partial least squares regression, remember partial least squares regression is another modification of principal component regression and the calculation of the factors we call them latent factors or PLSR latent factors are somewhat different than calculating the principal components.

So, partial least squares regression is more robust algorithm than principal component regression and partial least squares regression is widely used as a machine learning tool for soil and crop application. So, we are going to learn about partial least squares regression in our next lectures. And also, we are going to learn other type of regression, multivariate regressions also and their application in crop and soil. So, thank you, let us meet in our next lecture.