REMOTE SENSING FOR NATURAL HAZARD STUDIES

Course Instructor: Dr. Rishikesh Bharti
Associate Professor
Department of Civil Engineering
Indian Institute of Technology Guwahati
North Guwahati, Guwahati, Assam 781 039, India
e-mail: rbharti@iitg.ac.in
Website: https://fac.iitg.ac.in/rbharti/

Lec 8b: Remote Sensing Data Analysis-I Part B

Hello everyone, welcome back to the second part of Lecture 8. So, we will continue this remote sensing data analysis using principal component analysis. So, before we start the principal component analysis and its application in remote sensing, let us understand the problems that we are facing in remote sensing data. When we talk about the remote sensing data, we have multiple bands. Captured by sensors in different wavelengths. So, what happens is that remote sensing images are high in dimension and difficult to analyze; that means when we talk about the dimension, we are referring to 1, 2, 3, 4, 5, 6, and 7 bands.

So, these 7 bands have been used to capture or represent the information on how it behaves in different wavelengths. So, it can happen that some of the features, let us say this is an example of vegetation. So, some of the features are appearing multiple times in different images, let us say I have a sensor that is sensitive in these regions, So, if I refer to one particular object, let us say this is vegetation. So, this is the second band, the third band, and the fourth band, So, in this case, if you consider this, it is the reflectance of vegetation, which is high in NIR.

So, for this reason, you will not get any additional or unique information about this vegetation for this target. This is one of the examples. So, if I have 3 bands, I will have the similar information. So, what will this similar information do? If you plot these three bands together, or let's say two by three. So, you will get the highly correlated information.

This is the positive correlation; if this is like this, then it will be a negative correlation, but even in both cases, both data sets are highly correlated. So, this is a big problem with remote sensing images because we try to gather a larger number of bands. using different wavelength ranges. So, there is a chance that a similar object will be captured in the image and it will have similar information in all 2, 3, 4, or 5 bands, So, even though it is very easy in multispectral to segregate this information, when we talk about hyperspectral it is a big challenge because the number of bands is in the order of 200, 300 and the bands are very narrow and contiguous in nature. So, the correlation there will be much greater than your multispectral data.

So, in such a situation, what do we have to do? We have to segregate, or we have to identify the unique information that is available in 200 or 300. The number of bands is. So, that is why it is written that remote sensing images are high in dimension and difficult to analyze. Acquired images are correlated with similar information in two or more bands to each other. In general, bands acquired in different wavelengths are correlated.

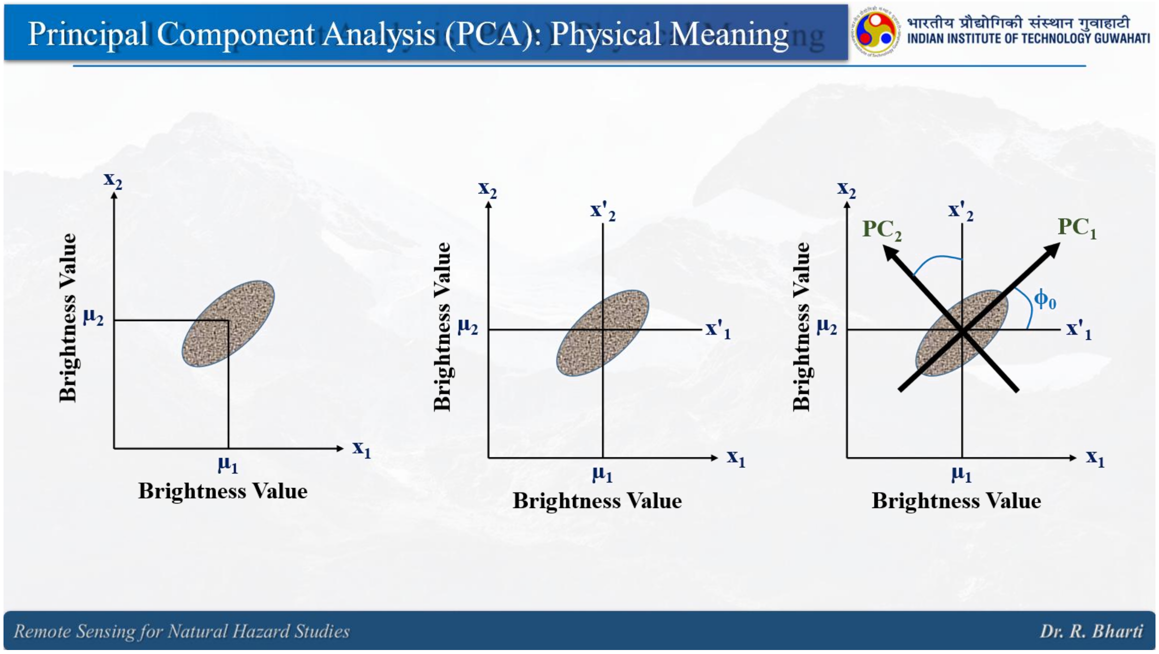
It is complicated to extract information about an object; not all data sets are independent of each other. Now, let us take another example: here you have 400 to 2500 nanometer. Now, one band, which we are generating from the first image from that sensor, is generated using 450 to 550 nm. Now, another one is using 1000 to 1200 nanometer. Another one is using 2300 to 2500 nanometer, and let us consider vegetation. So, here, if you see, this is also capturing the green band. So, green is from 500 to 600 nanometer. So, blue, green, and red remember that 400 to 500 nanometer is blue, 500 to 600 is green, and 600 to 700 nanometer is red wavelength. So, this red wavelength is somewhere around here. Now, part of this green is coming into this right now. So, if I redraw this, it will be like this: in this 3-band, all the reflected portions of this vegetation are captured. So, if I draw any one first versus second or second versus third, because of this correlation, this will come as a positive correlation, but if I have a band here, let us say this is the second one; now this will become the third, and this will become the fourth. In this case, what happens this second will not be correlated with 1, 2, or 4; Why? Because this is having different information about the target. So, in this situation, band 2 is not correlated with bands 1, 3, and 4, but if I compare bands 1, 3, and 4, they are comparable and have similar information. So, that is why we are saying that all the datasets are not independent, regardless of what we generate through remote sensing.

So, we try to make this highly correlated data uncorrelated. So, if this is the data point, we will not have any correlation, So, it says it is not dependent on the x or y axis, So, if this is the second band and this is the third band, in this case, these two are not related, but if the data sets are like this. So, we will see that we will have a high correlation. So, we try to make this highly correlated information in such a way that it becomes uncorrelated data. So, for that, what we have to do is identify the inherent dimensionality of the data, which means the unique information that is present in the data will be identified, and we will have a new set of images that will be used to identify the features.

So, the first problem is data redundancy, and the second is the noise in the data. So, these two will be a big problem or challenge associated with your remote sensing data. So, we try to use this principal component analysis, and we will be able to segregate the noise and identify unique information in a few bands or maybe an equal number of bands, depending on your application that we will try to understand in this particular lecture. Principal component analysis is a statistical method that uses orthogonal transformations.

Orthogonal means perpendicular, right, to convert a set of datasets into a set of values of linearly uncorrelated variables.

So, I hope you understand. That remote sensing data captured in different wavelengths may be highly correlated because similar objects have similar responses in different wavelength regions, The data reduction through correlation that is possible in general bands acquired in different wavelengths is correlated. To highlight or reduce the data redundancy, the number of output principal component images will be less than or equal to the number of input bands. Highlights specific features using its responses in different bands. Now, there are a few terms that are coming up again and again.



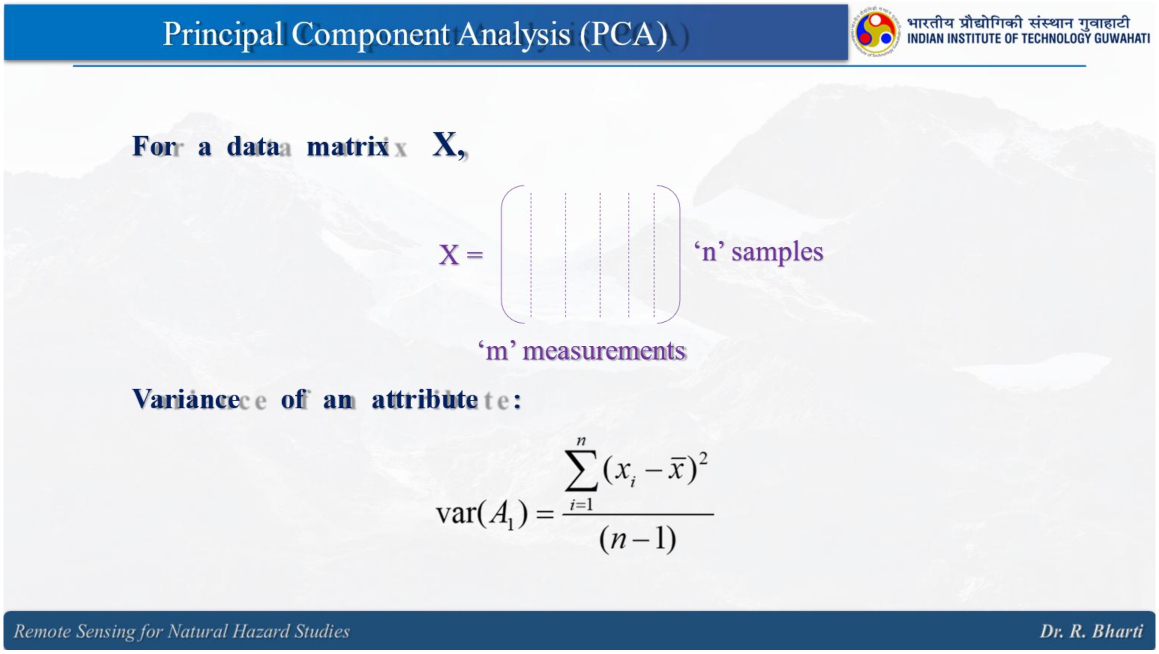
Principal Component Analysis (PCA): Physical Meaning

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Remote Sensing for Natural Hazard Studies

Dr. R. Bharti

So, I have intentionally put them to help you understand that these are very, very important concepts. So, you need to understand this data redundancy, why it is happening, and what the role of principal component analysis is here to reduce the data. And we should have the uncorrelated images, and with these images, we can have the output, which can be equal to the number of bands of the input, or it can be less than the input number of bands, it can be used to highlight a specific feature that we will try to understand with feature-oriented principal component analysis. So, before that, let us first understand this. This is the first case where we have this highly correlated data, and then what we do is try to shift the origin from here to here because this is an ellipse.



Principal Component Analysis (PCA)

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

For a data matrix X,

X =

'n' samples

'm' measurements

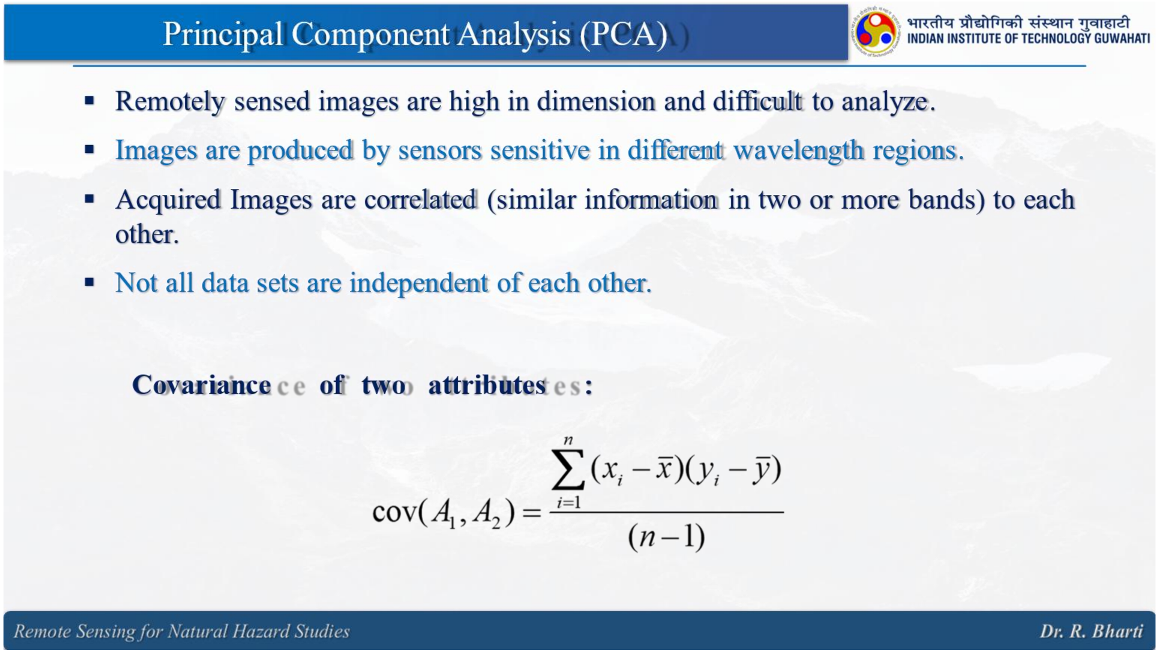
Variance of an attribute:

$$\mathrm{var}(A_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

Remote Sensing for Natural Hazard Studies

Dr. R. Bharti

Here you will have the semi-major and semi-minor axes, and this axis will be shifted here with a certain rotation. What we will do is try to rotate, and we will make it uncorrelated, So, how do we do that mathematically, that we will try to understand. So, for a data matrix x, which is nothing but m by n, it is a set of digital numbers, So, the size of x is m by n, and the variance of an attribute can be calculated using this formula. This is the normal variance; how do we calculate the variance for two given parameters, So, remotely sensed images are high in dimension and are difficult to analyze. Images are produced by sensors sensitive to different wavelength regions. Images acquired are correlated, and not all datasets are independent of each other.



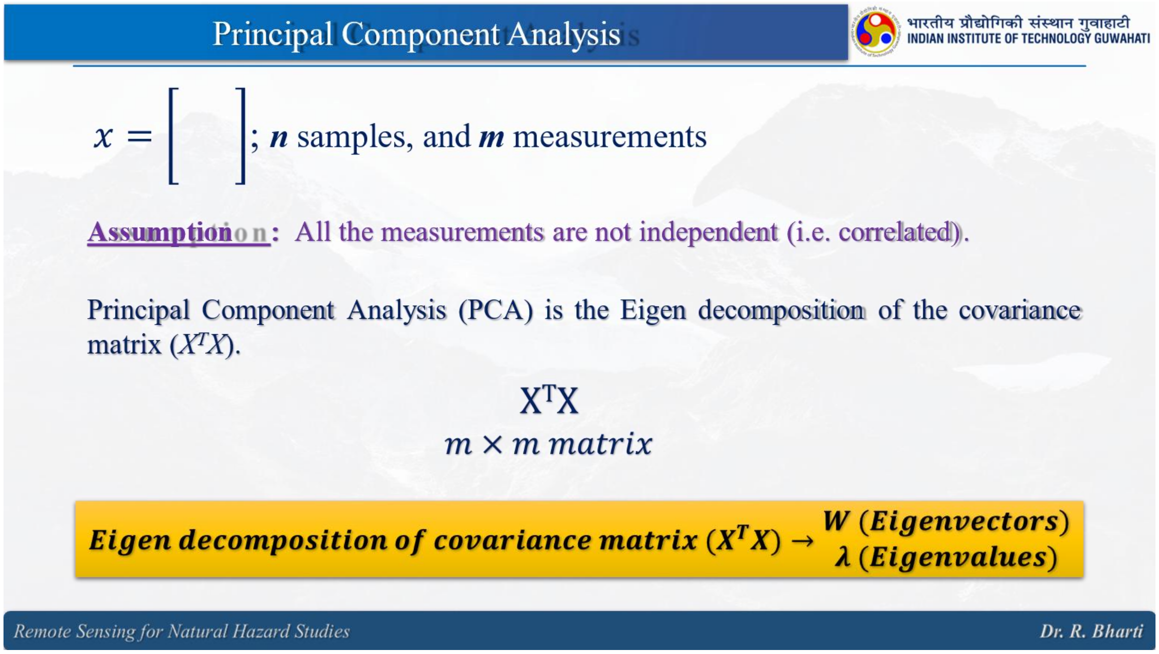
Principal Component Analysis (PCA)

- Remotely sensed images are high in dimension and difficult to analyze.
- Images are produced by sensors sensitive in different wavelength regions.
- Acquired Images are correlated (similar information in two or more bands) to each other.
- Not all data sets are independent of each other.

Covariance of two attributes:

$$\mathrm{cov}(A_1, A_2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

After variance, we have to calculate the covariance correctly. So, the covariance of two attributes can be calculated using this formula, and now we have variance and covariance, Now, if I have this x matrix which has n samples and m measurements. So, here the size will become m times n, and what you assume here is that all the measurements are not independent, So, we are assuming that our measurements are highly correlated. Principal component analysis is the eigen decomposition of the covariance matrix, which is nothing but x transpose x. So, this is x transpose x that we calculated in the previous slide. So, here we are going to do the eigen decomposition. So, this x transpose x, so the size will become m by m.



Principal Component Analysis

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

$x = \begin{bmatrix} \quad \end{bmatrix}$; $\boldsymbol{n}$ samples, and $\boldsymbol{m}$ measurements

Assumption: All the measurements are not independent (i.e. correlated).

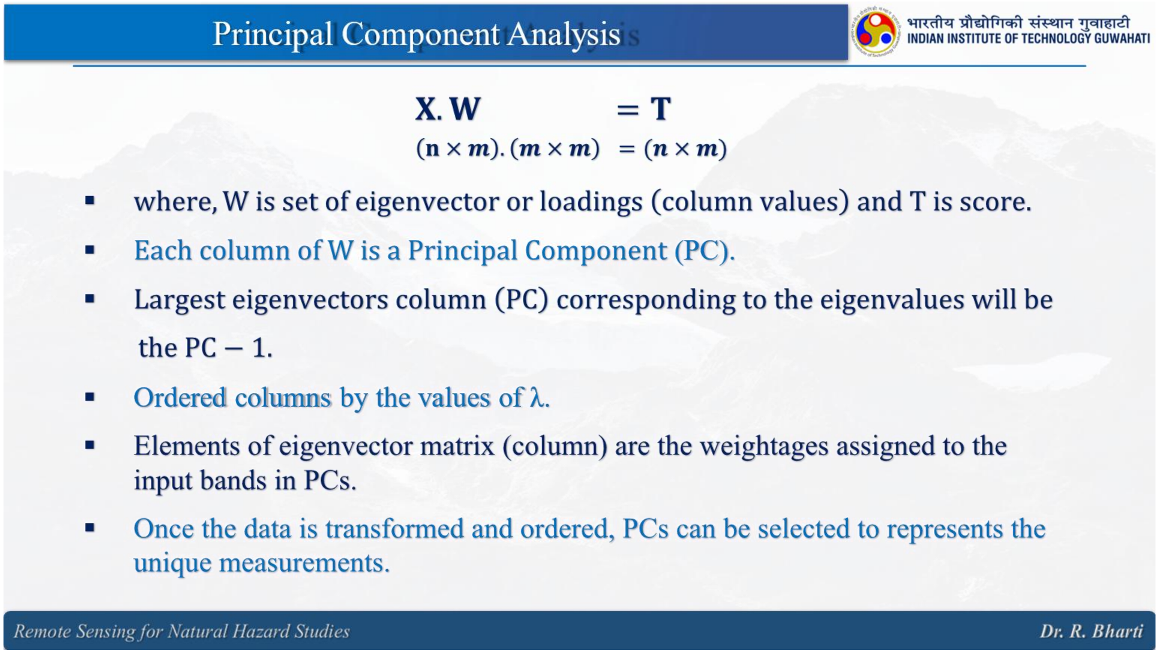
Principal Component Analysis (PCA) is the Eigen decomposition of the covariance matrix (X^TX).

$$\mathrm{X^TX}$$
$$m \times m \; matrix$$

$$\boldsymbol{Eigen\ decomposition\ of\ covariance\ matrix\ (X^TX)} \rightarrow \begin{matrix} \boldsymbol{W\ (Eigenvectors)} \\ \boldsymbol{\lambda\ (Eigenvalues)} \end{matrix}$$

Remote Sensing for Natural Hazard Studies *Dr. R. Bharti*

So, the eigen decomposition will have two components. So, we will have eigenvectors and then we will have eigenvalues, we will be referring to these eigenvectors and eigenvalues to generate our principal component images,



Principal Component Analysis

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

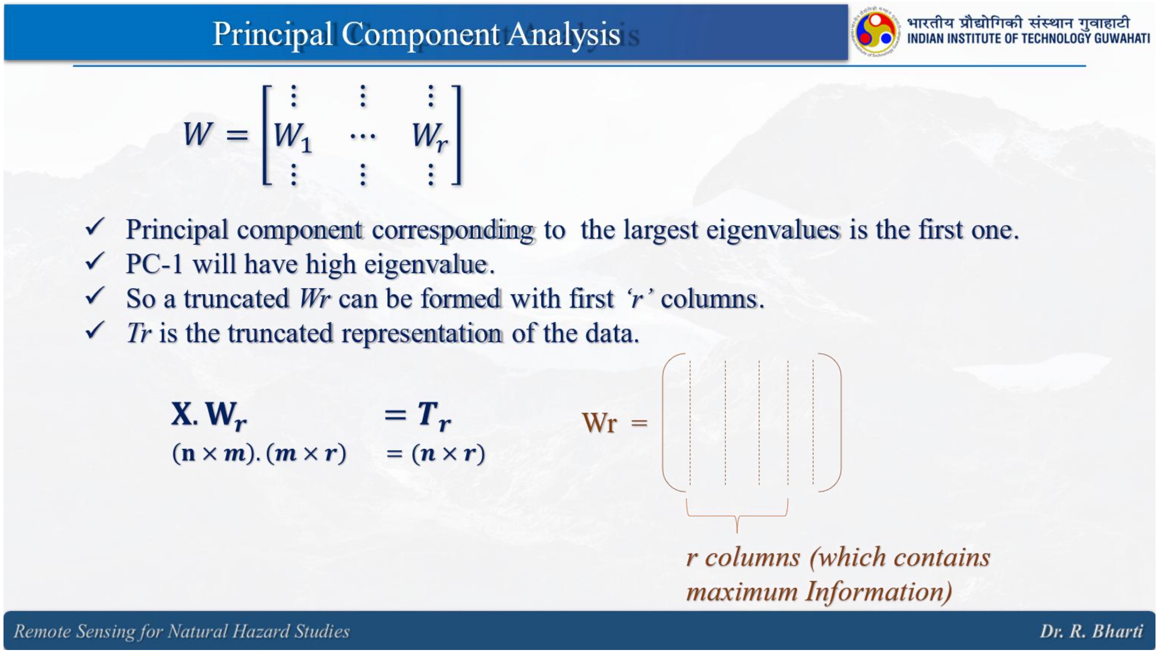
$$\mathbf{X.W} \qquad\qquad = \mathbf{T}$$
$$(\mathbf{n} \times \boldsymbol{m}).(\boldsymbol{m} \times \boldsymbol{m}) \;\; = (\boldsymbol{n} \times \boldsymbol{m})$$

- where, W is set of eigenvector or loadings (column values) and T is score.
- Each column of W is a Principal Component (PC).
- Largest eigenvectors column (PC) corresponding to the eigenvalues will be the PC – 1.
- Ordered columns by the values of λ.
- Elements of eigenvector matrix (column) are the weightages assigned to the input bands in PCs.
- Once the data is transformed and ordered, PCs can be selected to represents the unique measurements.

Remote Sensing for Natural Hazard Studies *Dr. R. Bharti*

So, here you can see this X. W=T. So, W is the eigenvector or loading in remote sensing; many times, you will refer to it or you will notice that instead of eigenvector we are calling it loading, And T is the score. Each column of W is a principal component. Now, we will see how the matrix is changing. So, this W is basically your eigenvector. So, the eigenvector column in the calculation will be your first principal component.

The largest eigenvector column corresponding to the eigenvalue will be the first PC, or PC1.



Principal Component Analysis

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

$$W = \begin{bmatrix} \vdots & \vdots & \vdots \\ W_1 & \cdots & W_r \\ \vdots & \vdots & \vdots \end{bmatrix}$$

- ✓ Principal component corresponding to the largest eigenvalues is the first one.
- ✓ PC-1 will have high eigenvalue.
- ✓ So a truncated *Wr* can be formed with first *'r'* columns.
- ✓ *Tr* is the truncated representation of the data.

$$\mathbf{X}.\mathbf{W}_r = \boldsymbol{T}_r$$

$$(\mathbf{n} \times \boldsymbol{m}).(\boldsymbol{m} \times \boldsymbol{r}) = (\boldsymbol{n} \times \boldsymbol{r})$$

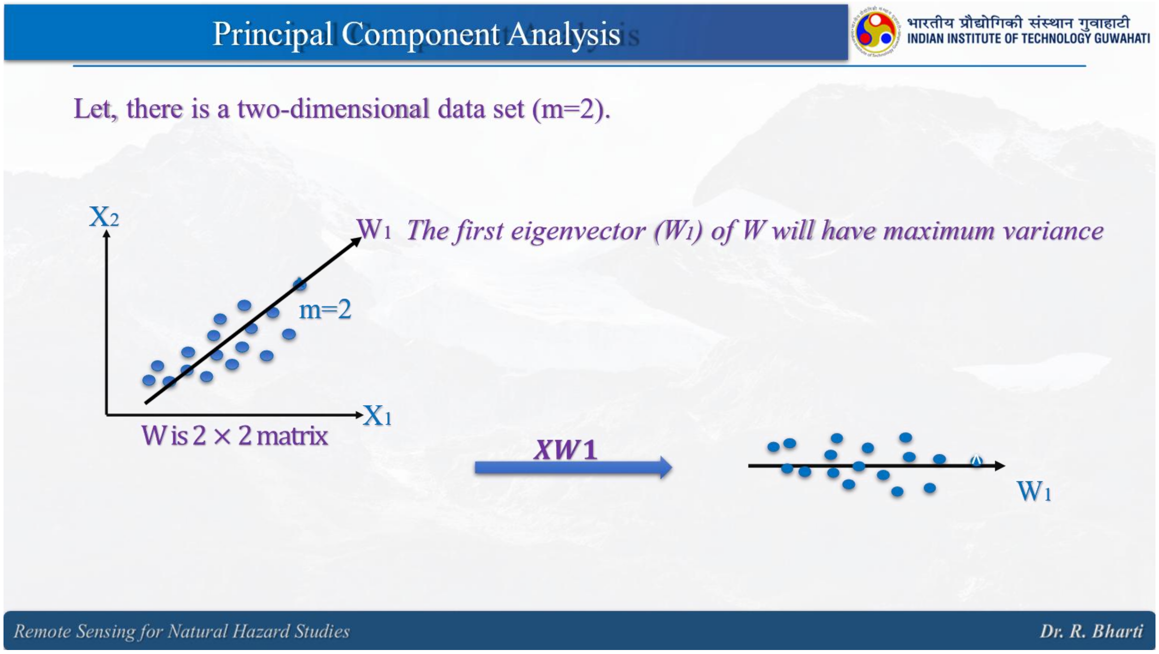
Wr =

r columns (which contains maximum Information)

Remote Sensing for Natural Hazard Studies *Dr. R. Bharti*

After calculating this, you have to order the column by the value of the eigenvalues because a high eigenvalue means you have more information in that particular column. Elements of the eigenvector matrix columns are weights assigned to the input bands of the PCs. Once the data are transformed and ordered, PCs can be selected to represent the unique measurement. We will try to understand this in the table. So, before that, when we are having this W, here W1 is your PC1, PC2, PC3, and so on, like that PCR. So, the principal component corresponding to the largest eigenvalue is the first one. So, here you have the highest eigenvalues; PC1 will have the eigenvalue. So, that truncated WR can be formed when the first R column. So, as I discussed, this output can be controlled, and here the output can be less than the number of bands provided in the calculation.

So, this WR can be introduced. So, R can be less than or equal to the number of bands. So, depending on that r and after you have sorted this table with these eigenvalues. So, the W1 has the highest eigenvalue. So, like the first, second, third, and fourth, the first "r" will be considered, and the rest will be discarded. So, the R column, which contains maximum information, will be in your output data, like this.

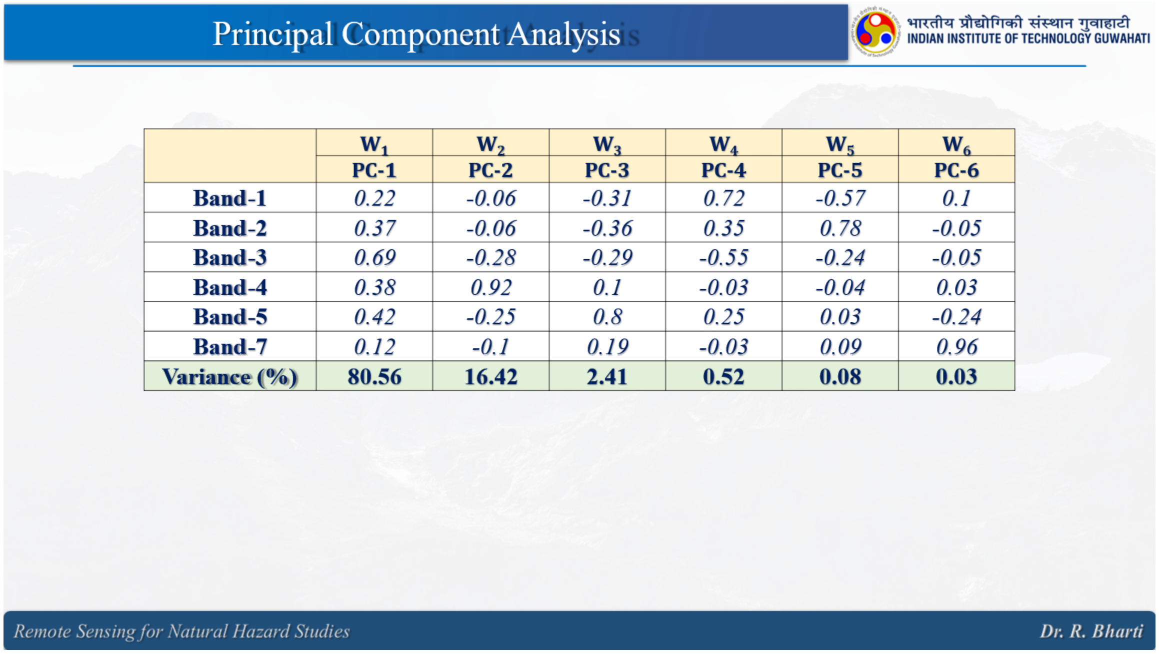


Principal Component Analysis

Let, there is a two-dimensional data set (m=2).

The first eigenvector (W_1) of W will have maximum variance

There is a two-dimensional data set where m is equal to 2. So, this data is highly correlated, So, the first eigenvector wi of w will have maximum variance; that means lambda, here, lambda refers to variance, and after the principal component analysis, this w1 will become less correlated or uncorrelated. Finally, we will have this kind of table where you can see all the bands listed here. Now here you have w1, w2, w3, w4, w5 and w6. So, this is your output and this is your input. So, we have used the Landsat data. So, the first, second, third, fourth, fifth, and seventh bands of Landsat 7 that were used here, and this is the W1; this is the eigenvector and this is the variance. So, these are eigenvalues. So, you see this table has been sorted based on this variance. So, the first column, which refers to this eigenvector W1, here you have this PC-1, So, the first element of this table is a loading. So, this is also an eigenvector, and earlier I referred to the eigenvector; we will be referring to it as the loading, which means it is the weightage for band number 1. So, when we talk about this 0.22, this is the weightage for band number 1, this is the weightage for band number 2, this is the weightage for band number 3, and like that, we will have the weightage for all the bands.



Principal Component Analysis

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

	W_1 PC-1	W_2 PC-2	W_3 PC-3	W_4 PC-4	W_5 PC-5	W_6 PC-6
Band-1	*0.22*	*-0.06*	*-0.31*	*0.72*	*-0.57*	*0.1*
Band-2	*0.37*	*-0.06*	*-0.36*	*0.35*	*0.78*	*-0.05*
Band-3	*0.69*	*-0.28*	*-0.29*	*-0.55*	*-0.24*	*-0.05*
Band-4	*0.38*	*0.92*	*0.1*	*-0.03*	*-0.04*	*0.03*
Band-5	*0.42*	*-0.25*	*0.8*	*0.25*	*0.03*	*-0.24*
Band-7	*0.12*	*-0.1*	*0.19*	*-0.03*	*0.09*	*0.96*
Variance (%)	**80.56**	**16.42**	**2.41**	**0.52**	**0.08**	**0.03**

Remote Sensing for Natural Hazard Studies *Dr. R. Bharti*

So, if I have to produce principal component 1, what will the expression be? So, 0.22 with band 1 plus 0.37 multiplied by band 2 plus 0.69 multiplied by band 3 plus 0.38 multiplied by band 4 plus 0.42 multiplied by band 5 plus 0.12 multiplied by band 7. So, if we do this calculation, we will have the first principal component right. So, this is how the output is generated using principal component analysis, So, this is the input band; now we are taking the same Landsat 7 data. So, the band one is basically your blue band; here you see this information is like this: this is green wavelength.

You see how images are different when they are captured in different wavelengths, So, this is red earlier was green. Then you have NIR; now you see in NIR the brighter pixel. Remember the vegetation, So, here you will find brighter pixels that are from the vegetation; this is SWIR. This is again SWIR, and this is the false color composite. I hope you remember how we generate the false color composite.

So, this is just a mapping, we have three guns in our display system, and then we are trying to put different combinations here, and then we are generating the color composite. So, this false color composite is used in that combination to display this particular image. So, here all the red colors are basically your vegetation. Finally, after the principal component analysis, we have this table, and this table will provide you with all the necessary information to generate the principal components because you have the inputs. Now, this is the output: principal component 1, principal component 2, principal component 3, principal component 4.

So, here we have truncated it; let us say we do not have these 2 columns. So, do you remember the wr? So, here we have given that wr is equal to 4. So, the first four columns will be used to generate your output, So, in this case, we have all these principal components. So, this PC 6 that means, here you have 6 output and 6 input that means, both are equal.

So, we have not introduced any thresholds here. Depending on your application, you have to select which particular principal component is suitable or will provide you with the information. So, let us say one example: this is PC6. So, all these weightages are used here. and this corresponding bands and then PC6 is generated. Now, we will talk about feature-oriented principal component analysis because we have understood how to decorrelate the data and how to identify the inherent dimensionality of the data, and then we try to identify features or the target with this method, which is called feature-oriented principal component analysis; it is also known as the Crosta technique, So, we will discuss that now.

To understand this feature-oriented principal component analysis, let us take the example of the previous slide; this is the table. Now, we will consider this Landsat 7 ETM+ image, which is used to generate this particular table. Now, here this particular figure is used to explain or demonstrate how gypsum spectra behave between 400 and 2500 nanometers. Now, you see it has several absorption features at different wavelengths. So, these are the characteristic wavelengths, if you remember.

So, if I have to identify gypsum using this principal component analysis, then we have to refer to this feature-oriented principal component analysis, which is known as the Crosta technique. Now, this gypsum and these, you can see; I hope this is clear. So, here are the wavelengths that are used to generate the first band, second band, third band, fourth band, fifth band, and seventh band, So, now here is the distribution of this gypsum. So, here 1st, 2nd, 3rd, and 4th, all these bands are having only the reflected energies, but if you see this band 5, it has the peak that is being captured, and fortunately, Landsat 7's fifth band is capturing this particular phenomenon, we would have been very lucky if we had gotten this particular absorption feature in one of the images, but that is not available. So, now we will rely on the other information that is captured in band number 7, I hope this is clear.

So, now we have two different bands that are of interest to us when we talk about gypsum. So, this is Band 7 and this is Band 5. Now we will go back to this particular table where we have the loadings correct. So, these are the loadings or the weightage, or this is the eigenvector value, So, this eigenvector we will refer to, and we will try to find which of the principal components has a very contrasting loading of band 5 and band 7, So, if I refer to this band W1 or the PC1, the bands 1, 2, 3, 4, 5, and 7 all have almost similar loadings; band 3 has the maximum, but they all have positive loadings, which means the positive weight is correct. When we talk about principal component 2, here band 3 has the lowest value, but band 5 also has the lowest.

So, our interest is to find out the negative or positive loading of band 7 and band 5, where all other bands have different combinations so that these two can be easily identified. So, now we will go back to this table, and we will refer to this third column, which is PC3. Here, you also see that the values are in the same range, these three are negative, but these three are positive. So, it is very difficult to differentiate between this band 4, band 5, and band 7. Then band 4 or the PC4 also has a similar nature, but when we come to this PC6, here you see that band 1 has a weight of 0.1, band 2 has a weight of minus 0.05, and band 3 has a weight of minus 0.05, which means these 3 bands are correct. Then band 4 has 0.03. This is also fine, but here, band 7 is having a very high loading, which is for this particular absorption feature.

Now, for band 5, it is minus 0.24. That means it has the maximum negative loading for band 5 and the maximum positive loading for band 7. So, we have a combination where we have these two absorption and reflectance peaks, and if we can use them together, then we will be able to separate this gypsum pixel from the other pixels in the image, So, here is what we will do. We will go with this band 5, band 7, and the characteristic absorption features of this gypsum are 1700, 2210, and 2440 nanometer, So, here we have used this 2440. So, here we have used this particular absorption feature so that we can identify it. So, we can generate this principal component. Now my question is, since we have got this band 7 very high loading. So, this high loading means the pixels will be bright or dark? Since we are having high loading, this will be bright, and this is minus 0.24. Now see this clearly: when you have band 7, which indicates very high loading, and band 7 is the absorption feature of this gypsum, So, this is a dark pixel in the image because this value is not available; this is absorbed, and all other values will be high. So, in this particular image, this absorption feature will be represented by the darker pixel, and when you have a high loading of this darker pixel in the output of the principal component, which is PC 6, all the pixels that are dark will represent your gypsum.

Similarly, here you have used the different, or here we also have the positive loading for this band 5. Since we have the negative loading for band 5, which is 0.24, and here we have this 0.96, this combination will provide you with a darker pixel in principal component 6. So, in your output, you will be able to easily identify gypsum pixels using this feature-oriented principal component analysis. With feature-oriented principal component analysis, I will end Lecture 8 and will continue this course with Lecture 9 in the subsequent or upcoming videos.

Thank you.