Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week - 11

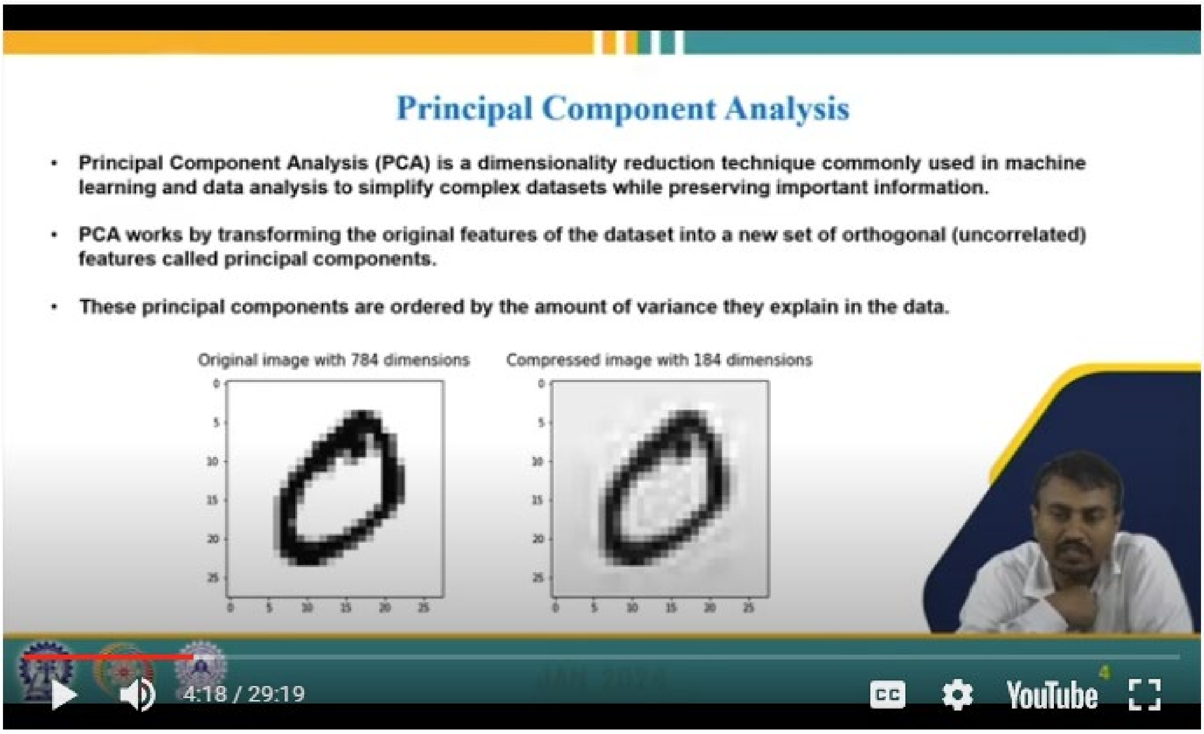
Lecture 54: Principal Component Analysis (PCA)

Welcome back to my course on mine automation and data analytics. Today, we will discuss another unsupervised machine learning approach: principal component analysis. Let me brief you a few minutes before we manipulate today's PCA. PCA is a machine learning approach that every day we face and use. Our activity in social media is relatively high compared to the last decade. If you notice that when we share an image with our peers and friends if you manually observe the data size, the file size of the image you have just shared as a picture is around 3.

4 MB, and when it has been transferred shared, it is around 200 kilobytes. So, a slight reduction is made in general. These media platforms need to do this because, for example, now it is Holi time, so people share crores of pictures with their families or friends. So, it would be tough to manage the load on the network that will be created due to the sharing of this small amount of data.

So, data compression techniques are instrumental nowadays when we handle extensive data, such as big data, in a network. Similarly, in the mining example, when we are building parallel mines and connected mines, most of the operations are monitored through sensors, transforming the data. So, a massive amount of data regularly comes through the network from the mine site to the control station or headquarters. So, to maintain this connection and manage the system effectively, a data compression technique is required. So, PCA might be a good option in those cases.

So, let us focus on this method. Principal component analysis. So, in today's lecture, we will cover principal component analysis and its algorithm, then determine the number of dimensions in the PCA, the assumptions, its pros and cons, and finally, we conclude this lecture with the applications of this particular method. So, in one sentence, PCA is a dimension reduction technique. This dimension reduction technique will be used for different applications based on the necessity, based on the purpose more so for efficient data management, more so for efficient data management because that high degree of data, the dimension might not be required all the time and to be stored in a space that also a big challenge. So, it reduces the dimension and is a machine learning approach. The fundamental assumption is that while reducing the dimension it reserves, it preserves the fundamental features, the most essential features in the data, in the information.



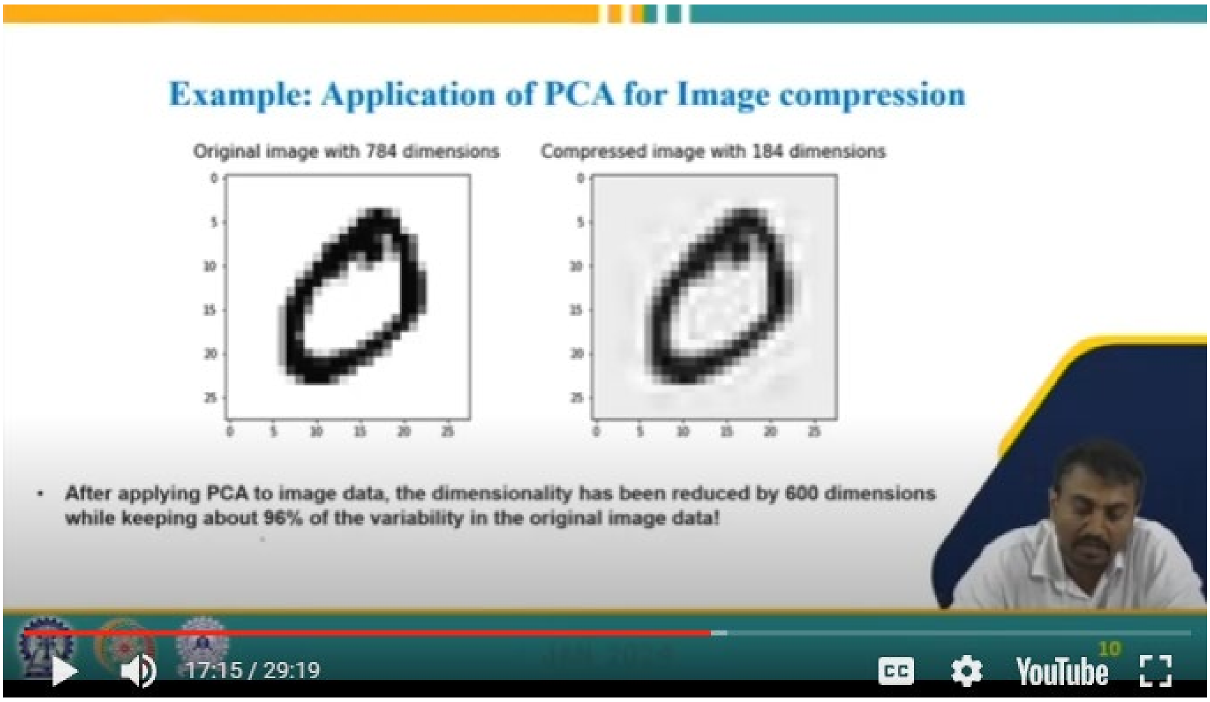
So, in that way, it does not compromise with the principal feature of the data. So, it preserves the principal feature, the principal patterns, and the principal behavior of the data, and it works on that principle. So, it works by transforming the original feature of the data set into a new set of orthogonal features called principal components. Let's discuss how it works with the 2D example. So, for example, this is the data space, and we found that the data are spatially located like this, okay, for example.

So this is the data set, okay? This is variable 1, this is variable 2. These variables are spatially interrelated in 2D space, okay. Now, if we start with an imaginary line, okay, and this imaginary line is rotated like that again along the origin, we want to find out the projected distance and maximize this distance square. And what we will find finally, we may see this line, okay, that has the highest d1 square plus d2 square plus d3 square and so on up to d6, here 6 number of points, here 7 number of points, d7.

Similarly, we want to maximize the distance of the projected points to this new line. So this particular line is okay, where we have the maximum distance we want to maximize. Now, if I finally rotate this line along this particular axis, and we have already found this is the maximum, then that line alignment is PC1, okay. Perpendicular to this line is the PC2 for this data space. Why are we doing this exercise? We are doing this based on the amount of variance in the data, okay?

Along this line, it has been found that it best fits this particular line. Now, after fitting this specific line, I operate the data. Now, the data will be located. So what will we find? We will then see that this small amount of PC1 distance, how much is the PC2 distance? So if its ratio is four to 1, PC1 has much more weightage in the data set, okay, or the variable 1. So what will happen? It shows that in this data set, the data, variable 1, has maximum influence over the pattern of the data set.

This is the meaning of this kind of transformation. By this principle, mathematically, we can handle a large amount of data set and find suitable orthogonal combinations of the data set. By that, we are achieving a sound reduction in the data size. So, this is the original image with 784 dimensions. Now, it is reduced to 184 dimensions; 600 reduction is already there, but principally, it looks like a 0, but it also looks like a 0. PCA, or this application, is a good option for handwritten exercises and script learning.



So, how does it work? Let us summarize that in a few sentences. So, PCA typically starts with centering the data by subtracting the mean of each feature from the corresponding feature values. This ensures that the data is 0, which means okay, a prerequisite for the PCA. Then, if we have n number of points, data points, and d number of features, we represent the data set into n into d matrix in X, capital X, where each row corresponds to a data point and each column corresponds to a feature. So, the mean of each feature j is denoted by mu j.

$$X_c = X - \mu$$

So, the center data matrix Xc is computed as Xc, equal to X minus mu. So, mu is the vector containing the mean of each feature: step number 2, covariance matrix. So now we compute the covariance matrix of the center data. Center data means the mu is already removed, or the mu is minus.

So, the covariance between the two features j and k is given by covariance Xj. Xk is equal to 1 divided by n minus one summation over i equals 1 to n, Xij minus Xj bar into Xik minus Xk bar. So Xij and Xik are the ith sample features of j and k, respectively, and Xj and Xk bar are the sample means of features j and k, respectively, okay? So, the covariance matrix is the d into d matrix where element jk represents the covariance between the feature j and k. So the covariance is nothing but one by n minus one into Xc transpose into Xc: step number 3, eigen decomposition.

$$\text{cov}(X_j, X_k) = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

$$\Sigma = \frac{1}{n-1} X_c^T X_c$$

So once we get the covariance matrix one divided by n minus one into Xc transpose into Xc, we perform an eigendecomposition on it to find the eigen vectors and eigen values. The eigenvectors represent the direction that is the principal component of maximum variance in the data, and the corresponding eigenvalues represent the amount of variance explained by each eigenvector. So let lambda 1, lambda 2 and lambda d are the eigen values of let V1, V2 and Vd are the corresponding eigen vector of the covariance. So, these eigenvectors are the principal components of the data. Then, step 4 is choosing the principal components.

The principal components are ranked in descending order of their corresponding eigenvalues, and typically, we retain only the top k principal components that capture the most variance in the data. This reduces the dimensionality of the data set from d to k—finally, the projection. So, we project the center data onto the subspace spanned by the selected principal components. This is done by multiplying the center data matrix Xc by the top k eigenvectors V matrix, which forms a d into the k matrix. So Y is equal to Xc into V.

So Y is the projected data matrix with dimension n into k. So here, we have summarized the mathematical process of PCA while preserving the maximum variance as much as possible. So, this is the representation of the projection onto the subspace. Here, you can see that this data is especially in 2D space. It is plotted, and it has been found that the data distribution, the data spread along the C1, is maximum.

So this is PC1. The perpendicular to this is PC2. So what is happening? We are finally reducing the data size with this compression technique. It indicates that along this value, the data variance is maximum along this direction. Using the data might help use this data further. So, the data we have shown earlier shows that 784 dimensions were reduced to 184.

So, 600 dimension reduction is done while we are keeping 96% of the variability in the original image data. So, it is efficient for this method to be applied in real-life applications, particularly for dimension reductions. Dimension reduction is an essential tool when handling large amounts of real-time data. So, let us understand how to determine the number of dimensions. So selecting the number of dimensions that are the principal components to retain the PCA is a crucial step as it directly impacts the amount of information preserved in the reduced dimensional representation of the data.

So let us discuss that. The first one is the variance explained. So, one common approach is to examine the cumulative proportion of variance explained by each principal component of PC. So, we can plot the cumulative explained variance against the number of components and choose the number of elements that capture a significant portion of the total variance. A standard threshold is to retain components that cumulatively explain around 75 to 95 percent of the variance, depending on the specific requirements of our analysis. The scree plot plots the explained variance against the component number, known as the scree plot.

So, the point at which the explained variance sharply decreases the elbow of the point indicates the number of components beyond which adding more components does not significantly increase the explained variance. So, we can choose the number of elements corresponding to this elbow point. So, this is the number of components for which we keep the 95 percent cut-off threshold for this cumulative variance. So here, we keep the 95 percent threshold value up to 9 components we consider for this particular data. The assumption of this specific method is linearity.

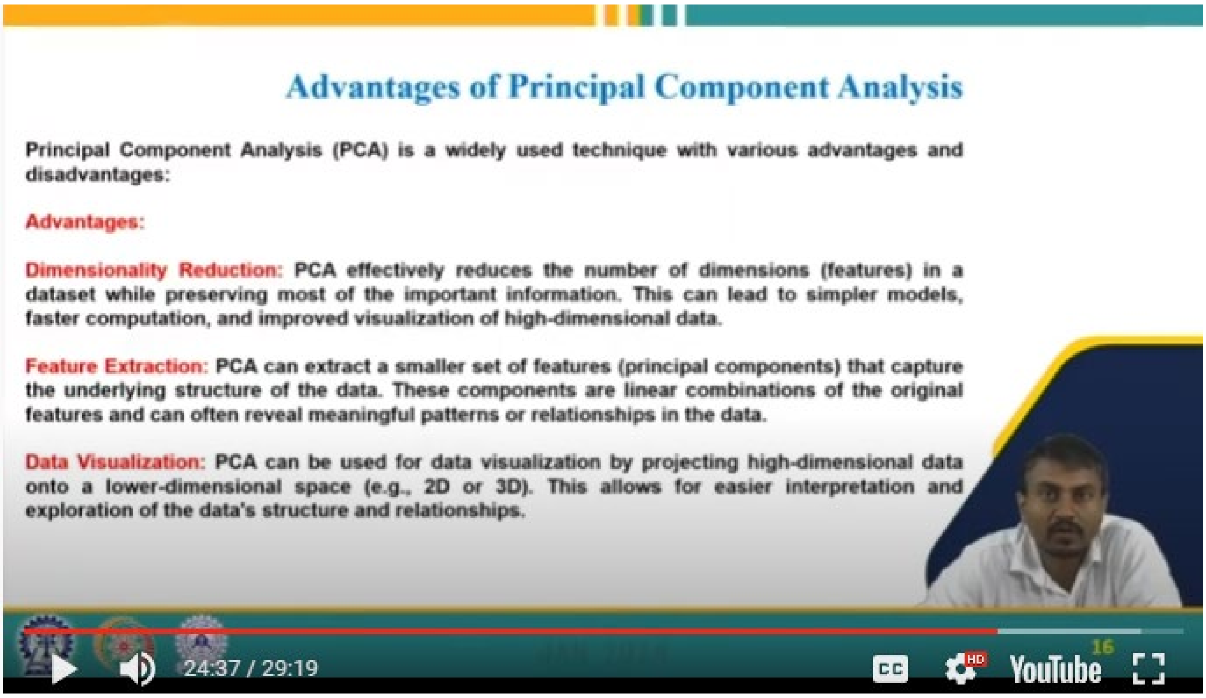
PCA assumes the underlying structure of the data is linear. So, it seeks to find the linear combinations of the original features that capture most of the variance in the data. PCA may not be the most suitable technique if the relationship between variables is highly nonlinear. So what is variance direction? PCA assumes that the direction in the feature space where the data varies the most, the direction with the data's most considerable variance, is the most important. It seeks to identify these directions, known as principal components, and reduce the dimensionality of the data by projecting it onto them.

Orthogonality. PCA assumes that the components are orthogonal to each other. So this means that they are linearly independent and do not contain redundant information. The orthogonality property simplifies interpretation and facilitates dimensionality reduction. Homoscedasticity. PCA assumes that the variance of the data is constant along each principal component direction.

In other words, the spread of the data points remains consistent across different directions in the feature space. If the variance varies significantly along specific directions, PCA may not accurately capture the underlying structure of the data. Mean centered data. PCA typically assumes that the data has been mean-centered before performing the analysis Xc-μ. So this means that the mean of each feature is subtracted from the corresponding feature value.

So, mean centering helps remove any bias in the data and ensures that the first principal component captures the direction of maximum variance. Normality. While PCA does not strictly record the data to be normally distributed, it can be sensitive to extreme outliers and non-normality. In some cases, transforming the data to approximate normality through log transformations may improve the performance of these PCAs. So, we have to remember that these assumptions are generally in the context of the PCA and that only some may be accurate for some cases.

So, we have to understand these assumptions and the purpose of the assumption. Based on that, we must apply the PCA to a corresponding data set. So sometimes, for handling outliers and non-linear data, a variant of the PCA might be used, or robust PCA might be used to handle this kind of situation so that it can handle the complex data as well. The advantages of these methods. Another benefit is that it effectively reduces the dimension in the data while preserving most of the vital information.



So this can lead to simpler models, faster computation, and improved visualization of high dimensional data—linear extraction. PCA can extract a smaller set of features and principal components that capture the underlying structure of the data. These components are a linear combination of the original features and can often reveal meaningful patterns or relationships in the data. Data visualization. The PCA can be used for data visualization by projecting high-dimensional data onto a lower-dimensional space in 2D or 3D format.

So this allows for a more straightforward interpretation and exploration of the data structure and relationship—noise reduction. PCA tends to reduce the impact of noise and irrelevant features by focusing on the direction of maximum variance in the data. Removing noise can lead to improved model performance and generalization.

Orthogonality. The PCA obtained are orthogonal to each other. So, it simplifies the interpretation and facilitates further analysis. The disadvantages. Loss of interpretability. So, understanding those data is critical.

So, in those cases, PCA might not be the method to apply because when it reduces the data dimension, the principal component may only sometimes help interpret the data further for further analysis. Assumption of linearity. PCA assumes that the underlying structure of the data is linear. If the relationship between the variables is highly nonlinear, PCA may not capture the underlying structure effectively.

Sensitive to scale. As we have seen, it is on the d1 square d2 square, which measures the maximum distance and projected distance from the origin of the data points. It may capture or perform with bias with large-scale and small-scale combinations. So, it is always essential to standardize the data before applying PCA to ensure that all features contribute equally to the analysis—information loss. PCA aims to retain as much variance as possible in the reduced dimensional space.

However, in some cases, reducing the dimensionality too much can result in significant information loss, leading to a loss of predictive power or essential pattern in the data—outlier sensitivity. PCA is very sensitive to outliers, as outliers significantly affect the estimation of the covariance matrix and the resulting principal components. So, a robust PCA technique or other outlier detection methods may be necessary to mitigate this issue. For nonlinear relationships, PCA captures the relationship between linear relationships between variables, but it may not effectively capture nonlinear relationships.

Techniques such as kernel PCA may be more appropriate when underlying relationships are highly nonlinear. So, in summary, PCA is a potent tool for dimensionality reduction, feature extraction, and data visualization. Still, it is essential to consider its limitations and potential drawbacks when applying this to real-world examples. So we can use this method in connected mines, mining 4.0, geotechnical monitoring, supply chain optimization, resource and reserve estimations, market analysis and pricing optimization, safety risk assessment, and accident prevention in different mining areas, as well as in general applications day. We have already shown you that social media platforms are one of the most used methods for reducing data dimensions.

So these are the references, and let me conclude in a few sentences. We have covered what PCA is, its algorithm, how to determine the number of dimensions and the assumptions, advantages, and disadvantages, and finally, its applications. Thank you.