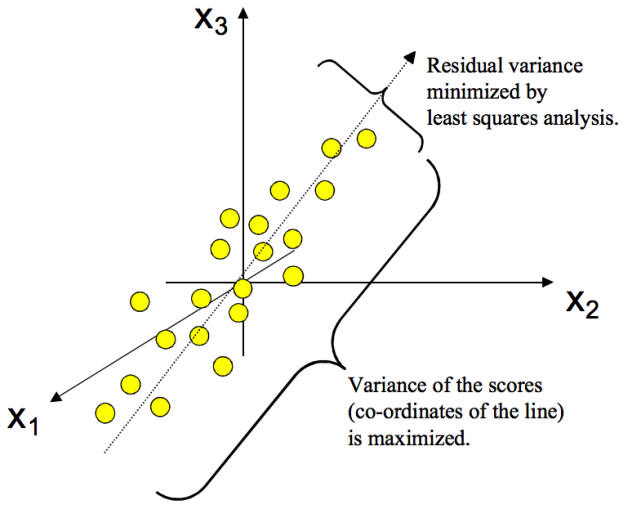
Principal Component Analysis (PCA)

Authors: Katelyn Vincent, Yiqun Tian, Yue Tian. (PDF)

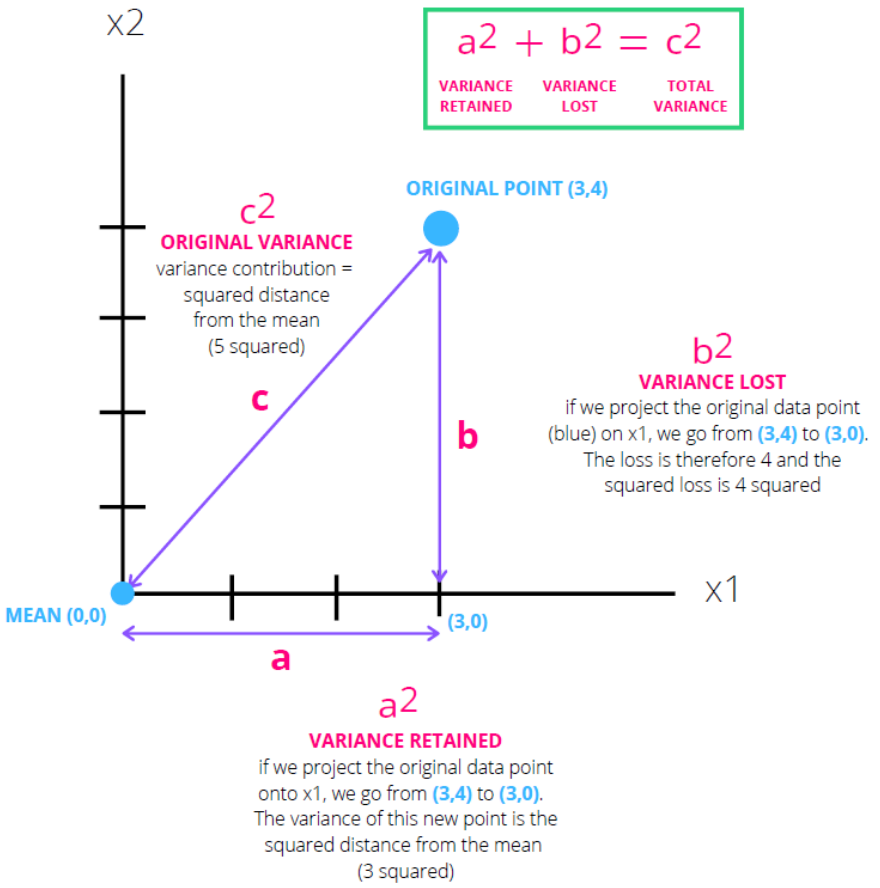
What is PCA?

In this lecture, we discussed feature extraction methods, focusing primarily on Principal Component Analysis (PCA). PCA is the simplest way of trying to project data into a low dimensional space. The goal is to reduce dimensionality while keeping as much variance (spread) as you can from the original data. Think about drawing a line and pressing on either side of that line until you squash all of the data points onto the line. This line is our principal component (PC), and we are 'projecting' all of the data points onto it. There are two ways of thinking about how to find the best line: we want to find the line that 1) maximizes the spread of data and 2) minimizes the sum of squared residuals.



We can have more than one PC, and each additional PC is orthogonal (at a right angle) to all of the other PCs. Sometimes what happens is that PC1 will tell you what the overall spread of the data is, and additional PCs (ex. PC2, PC3, PC4) will tell you how subpopulations differ from one another. Each PC

has an **eigenvalue** - a measure of how much variance is retained. PC1 will have the highest eigenvalue (eg 7.5), PC2 will have the second highest (eg. 2) and so on. In this example, we can say that the first two principal components capture 95% of the variance in the data (7.5 + 2). Typically, you'll want to pick enough principal components to cover 90-95% of the variance. If the first few eigenvalues are much higher than the others, PCA can capture the data using a much lower dimensional space. If all of the eigenvalues are roughly the same, it means the data falls the same way in every direction and there's no point in doing PCA.



When are the pros and cons of PCA, and when should you use it?

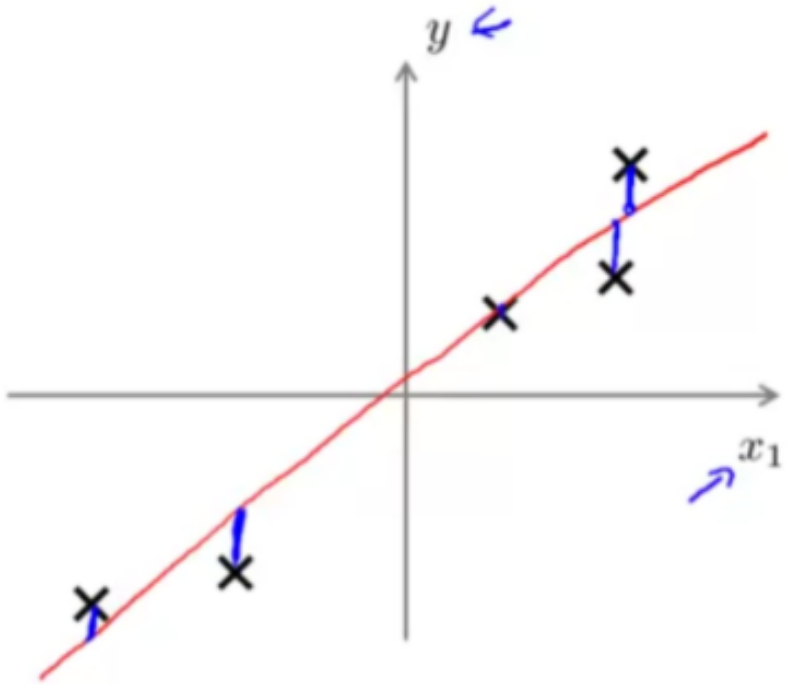
Pros: Removes correlated features, reduces overfitting, and improves performance One of the most obvious benefits of PCA is that it reduces dimensionality while retaining most of the information and variance in our original features. PCA is helpful when you need to reduce the number of features

for modeling, but it's not clear that there are individual variables you should remove. It is also helpful if you need to be sure that your features are independent of one another, since each of the principal components will be independent of one another. Because PCA reduces the number of features, it also helps to speed up training time.

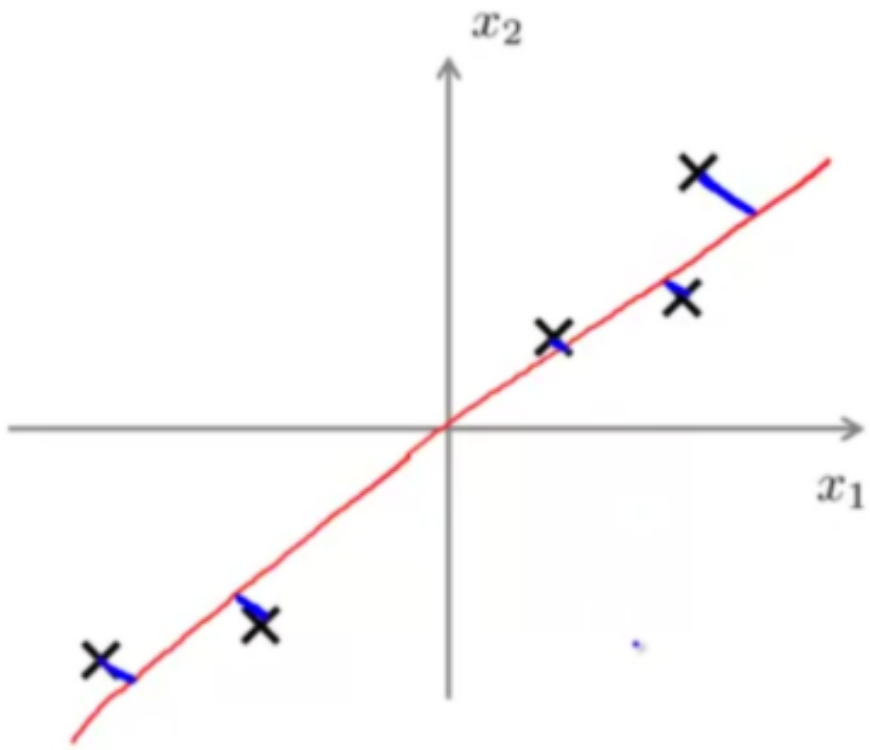
Cons: Variables are less interpretable, requires standardization, and possibility of information loss Because PCA combines information from multiple original features, the principal components are less interpretable than the original features. PCA is affected by scale, so it is also important to scale features before applying PCA, as well as convert categorical variables to numeric. Additionally, selecting too few principal components can result in information loss.

What is the difference between PCA and Linear Regression?

PCA is not linear regression. In fact, PCA and linear regression use totally different algorithms. For linear regression, it is trying to predict the value of Y given some info features of X and fit a straight line as to minimize the square error between the point and this straight line. However, for PCA there is no special variable Y that we are trying to predict. PCA minimizes the shortest orthogonal distance.



Linear Regression squared error



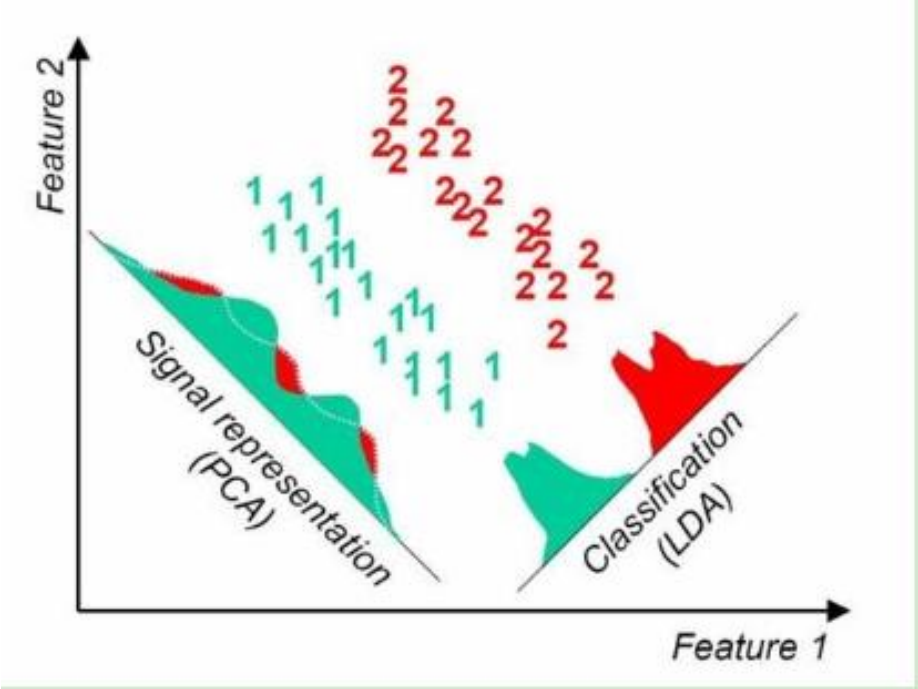
PCA - Projection Error

On the left graph, we can see linear regression calculates the square error as the vertical distance between true value and predicted value. On the right graph, PCA calculates the projected error as the shortest distance between true value and projected line.

Alternative for Classification - Linear Discriminant Analysis (LDA)

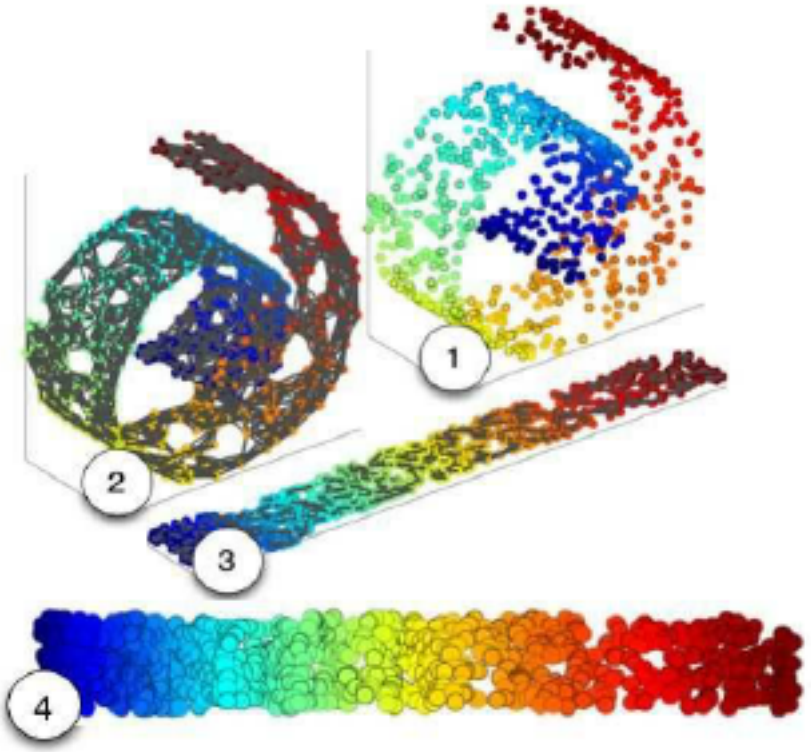
PCA does not use class information, so if you're trying to do classification then a better alternative might be Linear Discriminant Analysis (LDA). As the name implies, the technique is also linear but uses the class levels (unlike PCA). LDA can either be used for classification problems or as a dimensionality reduction technique in preprocessing. While Logistic regression is a classification algorithm traditionally limited to only two-class classification problems, if you have more than two classes then Linear Discriminant Analysis is the preferred

linear classification technique. LDA essentially plots your features, then creates a new axis that 1) maximizes the distance between means of the two classes and 2) minimizes the variation within each class. In simple terms, this newly generated axis increases the separation between the data points of the two classes.



Manifold Destiny

While this discussion has mostly focused on two-dimensional examples, there are many datasets where you can do a much better job of capturing the data if the projection space is curved (instead of flat). These surfaces are caled **manifolds**. In simple terms, an n-dimensional manifold is a space that locally looks like n-dimensional Euclidean space. You can think of it as putting a lot of n-dimensional shapes together to build a space. A simple example can be a flat map of the Earth. A map is a two-dimensional representation of the three-dimensional sphere, the Earth. Imagine the Earth has XYZ coordinates: locally, the z coordinate is barely changing, so the two-dimensional (XY coordinates) approximation is a good approximation about where a certain point is on the map. One of the most populat techniques to do nonlinear transformation mappings and find these manifolds is **t-sne**.



References

https://blog.umetrics.com/what-is-principal-component-analysis-pca-and-how-it-is-used

https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca/

https://sebastianraschka.com/Articles/2014_python_lda.html

https://www.programmersought.com/article/33174132390/

https://www.researchgate.net/figure/The-problem-of-manifold-learning-illustrated-for-N-800-data-points-sampled-from-a_fig1_201841023