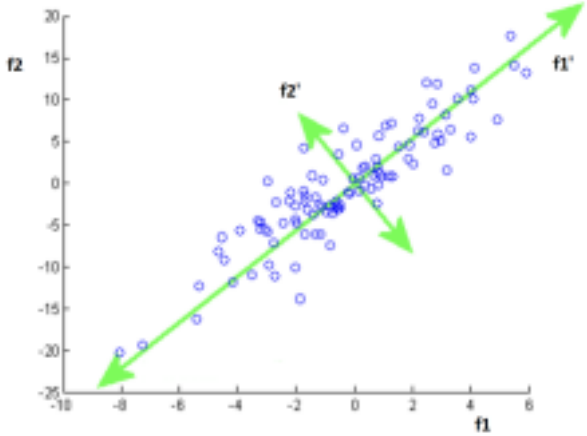
1. Principal Component Analysis (PCA)

- Dimensionality reduction techniques help to convert high dimensional data to fewer dimensions which can be then visualized using simpler plots or it can be used when we want to preserve the information of our feature columns.

- Principal component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets.
- PCA is an act of finding a new axis to represent the data so that a few principal components may contain the most information.
- The main objective here is whenever we are going from a higher dimension(d) to a lower dimension(d') we want to preserve those dimensions which have high variance (or high / information.)

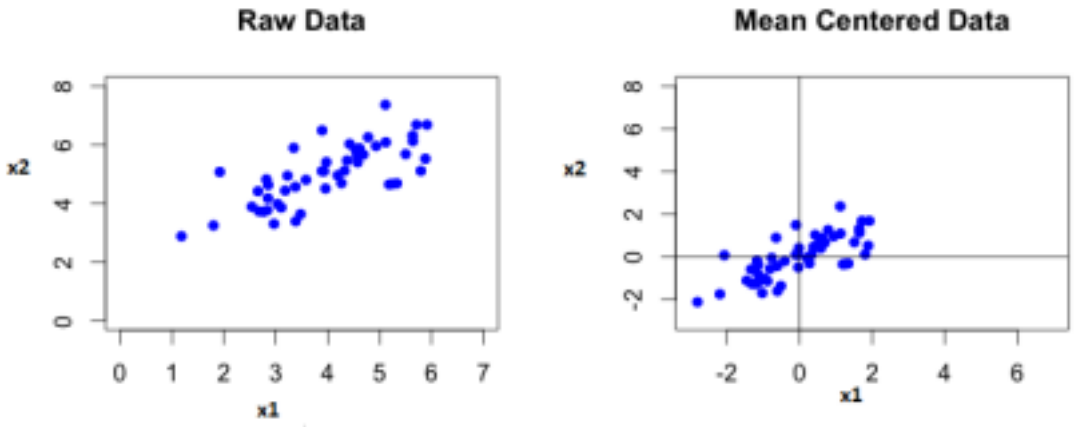
  For example: Imagine we have two features f1 and f2, and the data is spread as shown in the image below. We want to project the given 2-D data to 1-D.



- One thing we can observe here is that both axes have a good amount of variance. It will be difficult to decide which feature to drop.
- Therefore, we rotated the entire coordinate axes. It is shown in green.
- Let's call these new dimensions as $\mathbf{f_1}'$ and $\mathbf{f_2}'$. Now we can see that variability on $\mathbf{f_1}'$ is more than $\mathbf{f_2}'$ and we can now drop the $\mathbf{f_2}'$ axis.

Mathematics of PCA:

Step 1: Feature wise Centering: We compute the mean of each feature and subtract it from each value of that feature to center the mean of the data at origin.



Step 2: Standardize: It is done in order to keep all the features of our data in same scale. Best is to use standard scaling because it is less affected by the outliers.

Combining step1 and step 2, we get the value of our feature vector x as:

$$x' = \frac{x - x.mean()}{x.std()}$$

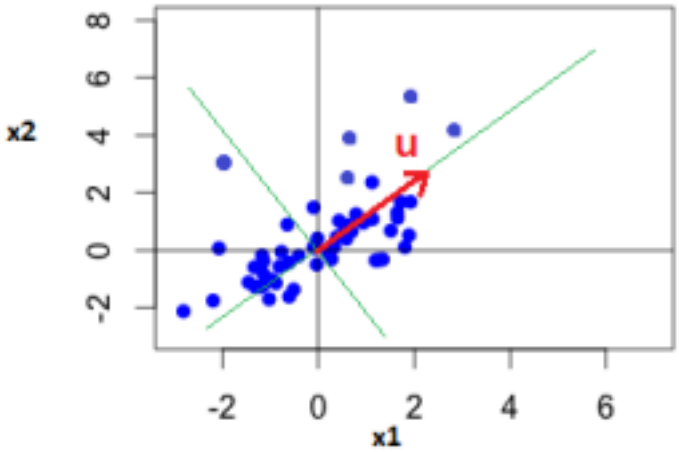
Step 3: Rotate the coordinate axes:

- There can be many angles of rotation possible, we'll consider a random axis and optimize it to get the best axis.

$\vec{u}$

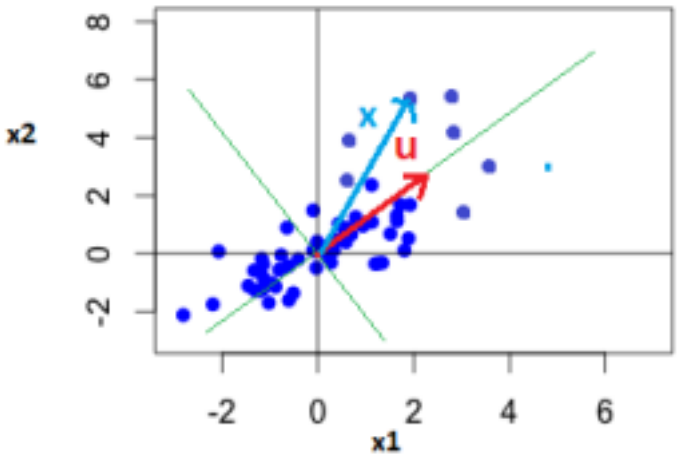
- Let's consider a unit vector in the direction of the new axis.

- Our optimization problem is finding unit vectors u such that the variance of all $\mathbf{x_i}'$ (features) are maximum.



- We can also think of it in terms of the length of projections.

- Consider a vector x in the space representing one of the points in our data.



- The projection of a vector x on vector u is given as: $\frac{\vec{u} \cdot \vec{x}}{\|\vec{u}\|}$

- The best u will be where the summation of the length of projections of all such

points(xi) on the vector u is maximum.

i.e. $$\max_{\vec{u}} \frac{1}{n} \sum_{i=1}^{n} \frac{\vec{u} . \vec{x_i}}{||\vec{u}||}$$

- However, the projections may be a negative value. Therefore, we take the magnitude of the numerator:

i.e. $$\max_{\vec{u}} \frac{1}{n} \sum_{i=1}^{n} \frac{|\vec{u} . \vec{x_i}|}{||\vec{u}||}$$

- Since the above objective function is not differentiable, we rewrite it as:

$$\max_{\vec{u}} \frac{1}{n} \sum_{i=1}^{n} \frac{(\vec{u} . \vec{x_i})^2}{||\vec{u}||}$$

- Now, since u is a unit vector, we introduce a constraint: $||\mathbf{u}|| = 1$

- Using Lagrange's multiplier, our loss function becomes:

$$max_{\vec{u},\lambda} \quad \frac{(X.\vec{u})^2}{n} + \lambda(||\vec{u}|| - 1)$$

We can optimize the above Lagrangian loss function using Gradient descent.

Optimization without Gradient descent:

- We know that, $\vec{u} . \vec{x_i} = \vec{x_i} . \vec{u}$

therefore, our new loss function is:

$$L = \frac{(X.\vec{u})^2}{n} + \lambda(||\vec{u}|| - 1)$$

- Now, since we can write the constraint $||\vec{u}|| = 1$ as $||\vec{u}||^2 = 1$, therefore,

$$L = \frac{(X.\vec{u})^2}{n} + \lambda(||\vec{u}||^2 - 1)$$

- Now we know that If a matrix A is an invertible square matrix, then:

$$A^2 = A^T.A = ||A||^2$$

Thus,

$$L = \frac{(X.\vec{u})^T(X.\vec{u})}{n} + \lambda((u^T.u) - 1)$$

- Using the identity, $(A.B)^T = B^T.A^T$

We get,

$$L = \frac{(u^T.X^T)(X.\vec{u})}{n} + \lambda((u^T.u) - 1)$$

Or

$$L = u^T \frac{X^T X}{n} u + \lambda(u^T u - 1)$$

Let $\frac{X^T X}{n} = V$

- Hence, $L = u^T V u + \lambda(u^T u - 1)$

V is the pairwise covariance matrix of all the features of our feature matrix X.

- Taking partial derivatives w.r.t μ and λ,

$$\frac{\partial L}{\partial u} = 2Vu + 2\lambda u \quad \text{and} \quad \frac{\partial L}{\partial \lambda} = u^T u - 1$$

- After putting both the partial derivatives equal to zero, we get,

$$Vu = \lambda u \quad \text{— (i)}$$

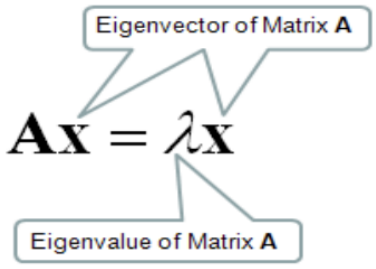
$$u^T u = 1$$

- By carefully observing the equation (i), we can conclude that u is the eigenvector and λ is the eigenvalue of matrix **V**.

Eigenvector and Eigenvalue:

- For any matrix A, there exists a vector **x** such that when this vector is multiplied with the matrix A, we get a new vector in the same direction having a different magnitude.

  - The vector **x** is called the Eigenvector and the length is called as

Eigenvalue. i.e. 

- There can be multiple eigenvectors, which are always orthogonal to

each other.

Conclusion:

- To find the best value of vector $\mathbf{u}$, which is the direction of maximum variance (or maximum information) and along which we should rotate our existing coordinates, we follow the below-given steps:

  Step 1: Find the covariance matrix of the given feature matrix X.

  Step 2: Then calculate the eigenvectors and eigenvalues of the covariance matrix. The eigenvector is the direction of best $\mathbf{u}$ and the eigenvalue is the importance of that vector.

- The eigenvector associated with the largest eigenvalue indicates the direction in which the data has the most variance.

- Therefore, we can select our principal components in the direction of the eigenvectors having large eigenvalues and drop the principal components having relatively small eigenvalues.

2. t-SNE (t-Distributed Stochastic Neighbor Embedding)