Dimensionality Reduction

WHAT?

Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

WHY?

1. The higher the number of features, the harder it gets to visualize the training set and then work on it.
2. Removes redundant features and noise. Sometimes, most of these features are correlated, and hence redundant.
3. Avoiding possible overfitting.
4. Less misleading data means model accuracy improves.
5. Less dimensions mean less computing. Less data means that algorithms train faster.
6. Less data means less storage space required.
7. Less dimensions allow usage of algorithms unfit for a large number of dimensions

HOW?

1. Feature selection: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem.
    - Advantages: simplicity and maintaining interpretability of your variables.
    - Disadvantage: missing out on whatever the dropped variables could contribute to our model.
2. Feature extraction: This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions. Say we have ten independent variables. In feature extraction, we create ten “new” independent variables, where each “new” independent variable is a combination of each of the ten “old” independent variables. However, we create these new independent variables in a specific way and order these new variables by how well they predict our dependent variable.
    - Advantages: keeping the most valuable parts of our old variables. each of the “new” variables after PCA are all independent of one another.
    - Disadvantage: variables become not interpretable

PCA

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Assumptions

Assumption #1: You have **multiple variables** that should be measured at the **continuous level** (although **ordinal variables** are very frequently used).

Assumption #2: There needs to be a **linear relationship between all variables**.

Assumption #3: You should have **sampling adequacy**, which simply means that for PCA to produce a reliable result, large enough sample sizes are required.

Assumption #4: Your data should be **suitable for data reduction**.

Assumption #5: There should be **no significant outliers**.

Step 1: **Standardization**

$$z = \frac{value - mean}{standard\ deviation}$$

- Avoiding larger ranges variables to dominate over those with small ranges

Step 2: **Covariance Matrix computation**

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

- covariance matrix of Z = Z^TZ
- identify correlations between variables
- Positive: correlated; Negative:Inversely correlated
- covariance is commutative (Cov(a,b)=Cov(b,a)). So, symmetric

Step 3: **Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components**

Eigenvectors & Eigenvalues: an eigenvector or characteristic vector of a linear transformation is a non-zero vector that changes by only a scalar factor when that linear transformation is applied to it. T is a linear transformation, v is an eigenvector of T if T(v) is a scalar multiple of v.

$T(\mathbf{v}) = \lambda\mathbf{v},$ where λ is a scalar, known as the **eigenvalue**

The linear transformation T can be represented as a square matrix A, and the vector v by a column vector, rendering the above mapping as a matrix multiplication on the left-hand side and a scaling of the column vector on the right-hand side in the equation

$A\mathbf{v} = \lambda\mathbf{v}.$

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.

$A = Z^TZ$ (covariance matrix of Z)

The **eigenvectors of the Covariance matrix** are actually the directions of the axes where there is the most variance(most information) and that we call Principal Components. And **eigenvalues** are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.

Step 4: **Feature vector**

By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance.

The **feature vector** is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. if we choose to keep only ***p*** eigenvectors (components) out of ***n***, the final data set will have only ***p*** dimensions.

Last step: **Recast the data along the principal components axes**

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

***FinalDataSet*^{T}(I believe)**

Or Z* = ZP* Z*: FinalDataset P*: FeatureVector

Example:

- ***Z* : m*n (m samples, n features)**
- ***Z*^{T}: n*m**
- **A = *Z*^{T}*Z (n*n)***
- ***Z*^{T}*Z * P = DP***
- ***P (n*n) eigenvectors***
- ***D (n*n) diagonal are eigenvalues, others zeros***
- ***Keep p eigenvectors, FeatureVector (n*p)***
- ***FinalDataset* = (n*p)^T * (m*n)^T= (p*n)*(n*m) = p*m**
- ***FinalDataset*^{T} (m*p) (m samples, p new features/principal components)**

How many features to keep?

Method 1: We arbitrarily select a number of principal components to include.
Method 2: Calculate the proportion of variance explained ($\lambda_1/(\lambda_1 + \lambda_2 + \ldots + \lambda n)$) for each feature, pick a threshold, and add features until you hit that threshold.
Method 3: This is closely related to Method 2. Calculate the proportion of variance explained for each feature, sort features by proportion of variance explained and plot the cumulative proportion of variance explained as you keep more features (a scree plot).

https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
https://towardsdatascience.com/a-step-by-step-explanation-of-principal-component-analysis-b836fb9c97e2