Data Analysis & Visualization

LVC 1: Data Exploration and Visualization

Data collection is the process of collecting data from diverse sources. It is the primary and the most important step for exploratory data analysis and any possible research to be conducted. Data can be collected from diverse sources, for example: survey data, data from the web, real-life monitoring data, etc. These sources of data can be classified as follows:

The objective of data collection is to ensure that a information-rich and reliable data is collected in the process. Good quality data will lead to a healthy analysis and provide reliable insights post the exploration and analysis.

The data collection process needs to be **unbiased** and **well-randomized**, which makes the data, and by extension the insights and decisions taken post-analysis of the data, reliable.

- **Randomization**: It is an experimental process where a certain sample is selected at random from the given population. The benefit of using randomization is that it avoids the **selection bias** (where some groups are underrepresented in the sample) and the **accidental bias** (the unwanted intrusion of a certain variable in the experiment being performed). This also helps in avoiding the **confounding effect**.
- **Confounding effect:** The confounding effect occurs when apart from the independent features, the outcome is affected by some other variables as well.

Once the collected data is ensured to be fair, reliable, and trustworthy, it becomes a requirement to ensure that the sample data is a good representative of the population of which it is a part. To do this, a certain type of test (namely the **Hypothesis testing**) is performed in **inferential statistics**. Let us understand it with an example of the **Health Insurance Plan (HIP) Study.**

Health Insurance Plan (HIP) Study:

- In the United States of America, one of the common malignancies among women is breast cancer. The **Mammography case study** is done to understand the effect of **offering mammography** on a patient susceptible to breast cancer.
- **Mammography:** It is the medical process of creating the **mammogram** of a woman. A **mammogram** is an **X-ray** picture of the breast of a woman. It is taken to detect whether the concerned patient is prone to breast cancer or not.
- Using **randomization,** a sample of 62,000 women who are in the age group 40-64 is taken for testing. It is considered a **large-scale study** due to the size of the sample taken.
- It is the **first large-scale** randomized, controlled experiment on mammography performed in the year 1960.
- The study is about whether **offering mammography** will reduce the death rate due to breast cancer. It is conducted only with women who are **offered** mammography, because someone can not be forced to take the test. This makes the process more random, unbiased, feasible, and hence more trustworthy.
- The sample is divided into two groups namely the **control group** and the **treatment group**.
    - **Control group**: It contains 31,000 women from the sample. These are the women who were **not offered mammography**. For this group, the death rate is 2.0 (2 patients out of 1000 are dying due to breast cancer in this group).
    - **Treatment group:** It also contains 31,000 people from the sample. These are the women who are **offered mammography**. The treatment group is further divided into two groups, namely the screened group and the refused group.
        - **Screened group**: Women who have gone through the screening, are in this group.
        - **Refused group**: Women who refused to go through the screening but they were offered the test, are in this group.

Before proceeding further, let us understand some basic definitions from Hypothesis testing that are required during this process.

Hypothesis testing:

- **Population:** A population is the collection of objects or persons that hold some **common feature** among them.
- **Sample:** A sample is a subset of the population that is collected for some analytical purpose. It is expected/assumed that the sample is a **good representative** of the population.
- **Hypothesis:** A hypothesis is a **claim** made about the population of a quantity. For example, offering mammography affects the death rate due to breast cancer among women, is a hypothesis. The hypotheses are generally of two types:
    - **Null Hypothesis -** It is a hypothesis that claims that there is **no effect** or there is no difference between the presence and absence of a characteristic. It is denoted as Ho. In the current scenario, the null hypothesis is the claim that offering mammography will not reduce the death rate of women due to breast cancer.

      Ho = Offering mammography does not affect the death rate of women due to breast cancer
    - **Alternate Hypothesis:** It is a claim about the population that shows the **dependency** between cause and effect. It is always against the null hypothesis and is denoted by Ha. In the current scenario, the alternate hypothesis will be “offering mammography will reduce the death rate of women due to breast cancer”.

      Ha: Offering mammography will reduce the death rate of women due to breast cancer
- **Hypothesis Testing:** It is the process of validating the hypothesis made about the population by conducting a test. In hypothesis testing, it is estimated whether the

evidence collected against the null hypothesis is enough to reject it or not. To do this, different methodologies are applied based on different test conditions. The test statistic is calculated using the data acquired against the null hypothesis. If the evidence against the null hypothesis is strong enough to reject the null hypothesis, then Ho is rejected, else it will not be rejected.

- **Significance Level:** Significance level is considered as the probability of rejecting the null hypothesis (Ho), when it is true. It is denoted by alpha. If alpha is 5%, then it means that there is a 5% risk of concluding there is a difference, while in fact, no difference was there in actuality.

- **Test Statistic:** It is a value calculated using the evidence against the null hypothesis that will differentiate between the null and the alternate hypothesis.

- **p-value** - It is the summation of probabilities of all the events that are **equally likely** or **rarer** than the observed data. It can also be interpreted as the probability of occurrence of the alternate hypothesis given the null hypothesis is true. If the p-value is less than the significance level, then the null hypothesis is rejected. If the p-value is greater than the significance level, then it is understood that there is no strong evidence against the null hypothesis and hence we fail to reject it.

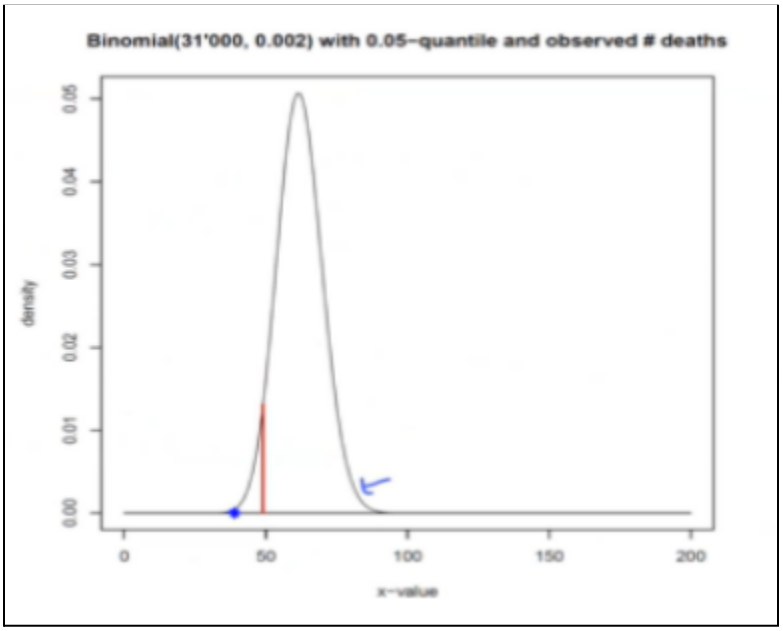
Now that we understand these basic terminologies, let's get back to the HIP study.

	Table 1. HIP data. Group sizes (rounded), deaths in 5 years of follow ups, and death rates per 1000 women randomized.					
		Group size	**Breast Cancer**		**All other**	
			No	**Rate**	**No.**	**Rate**
Treatment	**Screened**	20,200	23	1.1	428	21
	Refused	10,800	16	1.5	409	38
	Total	31,000	39	1.3	837	27
Control		31,000	63	2	879	28

The above table shows the distribution of women in the treatment and the control group. The death rate per 1000 women is taken under consideration for 5 years' data. In the treatment group, those

who are screened, have a death rate of 1.1 while those who refused, have a death rate of 1.5. As a total, in the treatment group, the death rate is 1.3. This indicates that the death rate is lowest among those who are screened, in comparison to the refused group and the control group. In the control group, the corresponding death rate is 2, which is the highest. Now, we need to test whether this difference is significant or not.

- In the control group (where no test is offered) the death rate is 0.002, i.e., (63/31000). So, the **null hypothesis** is that offering mammography will not reduce this death rate due to breast cancer.
    - $H_o : P_i = 0.002$
- The **alternate hypothesis** is that offering mammography will reduce the death rate.
    - $H_a : P_i < 0.002$
- It is a one-tailed test where the comparison has to be done only on one extreme of the test statistic.
- The binomial model with a significance level of 0.05 is given as follows:
    - T = The number of deaths under Ho
    - T ~ Binomial (31000, 0.002)



- In the above figure, the red line shows the cut-off region for a significance level of 0.05, while the blue dot represents the actual test statistic. For the HIP study, the number of deaths in the treatment group is 39 while the critical value (the number of deaths corresponding to the significance level) is 50. Corresponding to these numbers the p-value for this study is as follows:

  p-value = 0.0012, while alpha = 0.05

  So as per the binomial test, the p-value < alpha and hence the null hypothesis will be rejected. The evidence found against the null hypothesis is strong enough to reject it.

  Conclusion: Offering mammography **reduces the death rate** of women due to breast cancer in the United States.

Apart from the healthcare industry, hypothesis testing has found its extensive use in a diverse range of industries in the modern era. It has enough utilization in every sector. Let us take a few examples to understand where it is applied in various industries.

Hypothesis testing application outside of healthcare:

- **Finance:** In the finance sector, hypothesis testing can be applied to decide whether an investment made is going to give a satisfactory return or not.
- **Business:** While being in business, hypothesis testing can be used to test whether making a certain decision will cause the business to grow or not.
- **Manufacturing:** In the manufacturing sector, it can be used for quality management - whether a new process / technique is likely to reduce the number of defective products or not.
- **Advertisement:** Here, it can be used to test whether a certain advertisement will increase the sales (or number of clicks on a page, etc.) or not.

There are multiple dimensions of studies done on cancer over the entire world. An extensive amount of study is found over the effect of food items consumed on the possibility of someone getting cancer. Below are some interesting studies done in this field:

Example research findings:

- Here different food items are tested with the risk of cancer. Some food items are found to affect the chances of cancer while others do not. The p-value corresponding to each food item is calculated. Below are the details:
    - Intake of **tomato sauce** (p = 0.001), **tomatoes** (p=0.03), **pizza** (p=0.05). They are food items that reduce the risk of prostate cancer.
    - **Tomato juice** (p=0.67), **cooked spinach** (p=0.51), and many other vegetables were found not to be significant as shown by a higher p-value.

The above conclusion is made against a significance level of 5%.

While doing hypothesis testing over a certain sample and population sometimes, it becomes important to **protect the outcome of the test**. There are places where it is required to be strict with the test while in some other conditions being lenient is also good enough. It is especially important when the test has to be conducted multiple times. In such cases, to reduce the acceptable count of errors made, the protection levels are used. Let us understand different protection levels that are applied in real-life scenarios.

Different Protection Levels:

To protect the hypothesis testing, we need to identify the error rate among the following two.

- **FWER:** It is named as **family-wise error rate**. It is the **probability of at least one false significant result.**
  It is majorly used by the Food and Drug Administration (FDA) and is considered to be the most stringent protection level because it allows very little error during the test. Majorly, it is deployed where stringent protection is needed to apply.
- **FDR** - It is named as **false discovery rate**. It is defined as the **expected fraction of false significant results among all significant results.** It is less stringent than the FWER.

Computing these protection levels is the next task in the process to give protection to the test statistic. There are multiple ways of making corrections and getting the values of FWER and FDR. Below are some of them:

a) **Bonferroni correction:**

i) It rejects the H_o when $m \times p \leq \alpha$

Where m is the number of tests performed.

ii) Bonferroni correction implies that $FWER \leq \alpha$

b) **Holm-Bonferroni correction:**

i) For all the m tests, the corresponding p-values are sorted as follows:

$$p(i) \leq ..... \leq p(m)$$

ii) Ho is rejected when $(m - i + 1) \times p(i) \leq \alpha$

iii) Holm-Bonferroni correction implies $FWER \leq \alpha$

iv) This method is more powerful than the Bonferroni correction because it takes under correction the probability outcome in the i^{th} trial.

The above two methods are used to get the value of the family-wise error rate. Now, to get the false discovery rate, we can use the below method:

c) **Benjamini-Hochberg correction**:

i) For all the m tests, the corresponding p-values are sorted as follows:

$$p(i) \leq ..... \leq p(m)$$

ii) The null hypothesis is rejected when the following criterion meets -

$$m \times p(i)/i \leq \alpha$$

iii) The value of $FDR \leq \alpha$

While collecting the data, we have the freedom to collect it in multiple dimensions. Sometimes, it is the requirement of the case to collect data in higher dimensions (a large number of features / columns). It is pretty useful to have every bit of detail in every dimension of the data. But the problems start in the process of exploratory analysis.

While **exploring the data,** it becomes a tedious task to explore and visualize each and every feature. It is a time-consuming process and also it is hard to correlate the relation and insights in each feature. So, understanding data at an exploratory level becomes a tough deal for the analyst.

To avoid this, one of the foremost requirements while working with high dimension data is to reduce the dimensionality and bring it to a lower dimension so that it becomes understandable and easy to interpret the insights found after exploration and analysis. Imagery is one of the key fields where we especially need to reduce the dimension of the data as image data is usually available in high dimensions. There are two major techniques that we will focus on to reduce the dimensions of the data:

1. **PCA** (Principal Component Analysis)
2. **SNE** (Stochastic Neighbor Embedding)

Let us understand them one by one with their working method.

Principal Component Analysis:

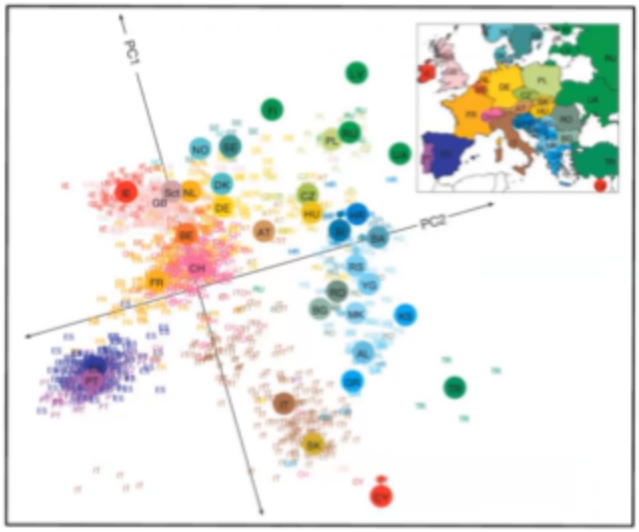
- The principal component analysis is a technique in unsupervised learning that is used to reduce the **dimensionality** of the given data. In general, it is tough to perform statistical analysis on the data with high dimensions because it is extremely time-consuming to do a comparative and descriptive analysis of features available in the data. It is a **linear** dimensionality reduction technique.
- While reducing the dimensions of the data, we nee to make sure we don't lose important information and preserve the **variance** within the data. Ensuring this will make the data contain a similar information as the original data. PCA confirms that most of the **variance** is **preserved** while reducing the dimensions so that the data with reduced dimensions will still give equally trustworthy inferences and insights during any analytical process.
- The principal components found using PCA are **linear combinations** of the original features.
- While performing PCA, the desired possible percentage of information is retained. In general, it is 85-90% of the original information in the data but can vary depending on the problem at hand.

PCA has a diverse range of applications in various industries in the modern era. It is considered the mother of multivariate data analysis because it compresses the size of the data as per the desired

number of dimensions. In the field of imagery, it is used for image compression and much more. Let us understand one of the sample applications of PCA below.

PCA Application

- DNA data is taken over the entire geography of Europe. Letters corresponding to some DNA features are given in the image below. There are a lot of letters that are the same while there are also a huge number of them dissimilar to each other. Using PCA, the existing clusters of similar people can be identified.



- Wherever there is less variation in the data, PCA ignores such dimensions because it is not useful to retain such features that are unable to explain something important about the data.
- PCA is performed on the data and clusters of similar people are created. In the top right corner, the actual map of Europe is shown. Comparing the clusters created by PCA in the above diagram, it can be commented that PCA has performed wonderfully on this data. The sample dimensions PC1 and PC2 are shown in the figure.

The working principle of PCA has been a process to be observed from multiple windows. It has multiple geometric and mathematical interpretations for how it works even though all of them lead to

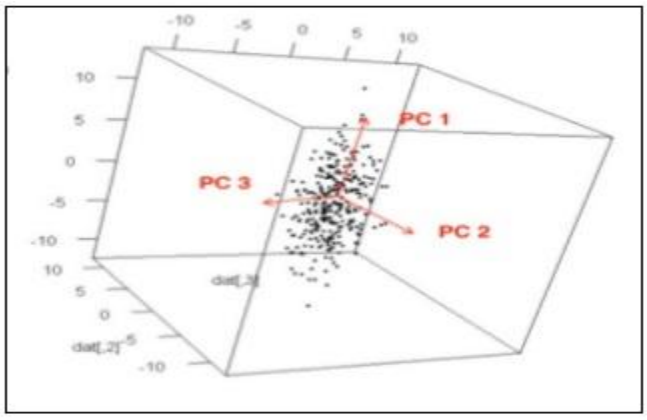
the same conclusion. Studying them will lead to understanding PCA in a much better way. Below are the three working principles of Principal Component Analysis:

1. **Maximum Projection Variance**
2. **Minimize Projection Residuals**
3. **Spectral Decomposition**

Let us understand them one by one in detail -

1. Maximum Projection Variance:

- One of the possible interpretations of the working method of PCA is that it works on the principle of maximum variance projection. It finds directions in the existing high dimensional data along which the variance is maximum. Such directions are called principal components of the data.
- As there exist features that do not contribute much to the variance of the data and due to this they are supposed to not contribute any useful information in the analysis. PCA removes such features by ignoring their direction.



The data used is represented as real numbers in an n-dimensional space, i.e., $X \epsilon R^{nxp}$.

- PC 1: This is the first principal component and it is showing the largest variation along its direction. It can be observed that the spread along the length of PC1 seems to be high.

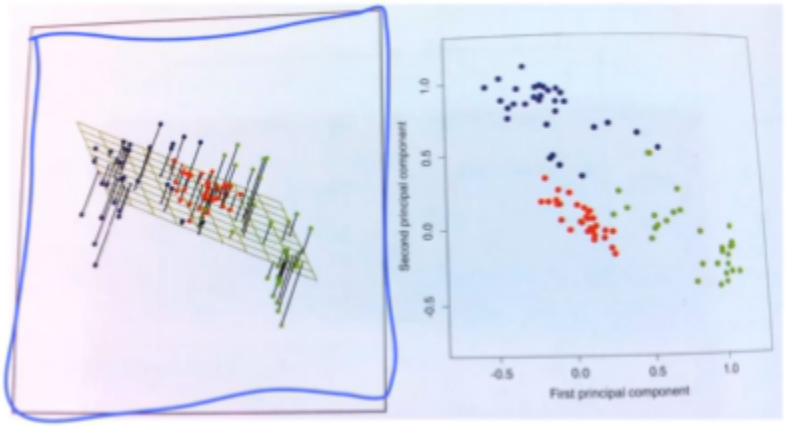
- PC 2: This is the second principal component. It is in a direction perpendicular to PC 1 and it is also showing the largest variance along its direction.

- PC 3: This is the third principal component and it is in another direction perpendicular to both PC 1 and PC 2. Again, it is showing the largest variance along its direction.

2. Minimize Projection Residuals

- The intuition behind this is to reduce the dimensions based on minimizing projection residuals. In the below figure, the original data is in three dimensions. Another plane of 2 dimensions is taken and is set in the first one in such a way that original data finds the minimum residual from this plane. This way one of the dimensions is removed.

Residual: It is the sum of the perpendicular distances of existing points from the 2D plane. Lower residual indicates that removing that dimension won't affect the originality of the data much. In the current scenario, the residual is the sum of all the individual residuals of existing points.



- In the above diagram, the red, blue, and green dots are presenting the output labels of the original data. It seems that there are three existing clusters visible in the figure.

3. Spectral Decomposition

- This is the least intuitive and hard to visually inspect because it is a bit mathematical. From a computational point of view, computers do adopt this method.

- This method starts with fitting an ellipsoid to the data. The axis of the ellipsoid is the eigenvector and the length is the variance. The principal components are found by looking at the covariance matrix and the largest half axis.

- The covariance matrix of a vector X is given as follows:

$$R = (1/n)\, X^T X$$

  Where R is a symmetric matrix.

To find the principal components, the above covariance matrix needs to be split into the product of an orthogonal matrix, a diagonal matrix, and the transpose of the same orthogonal matrix. To do so, we take the help of the Spectral Decomposition theorem that is given below:

Spectral Decomposition Theorem: According to this theorem, every real symmetric matrix R can be decomposed as:

$$R = VAV^T$$

Where R is the covariance matrix, A is the diagonal matrix, and V is the orthogonal matrix.

The matrix V contains the eigenvectors of the matrix R and hence they are principal components of the original data X. The theorem gives all the eigenvectors of the input data. Among them, we pick the largest one and then the next largest, and so on.

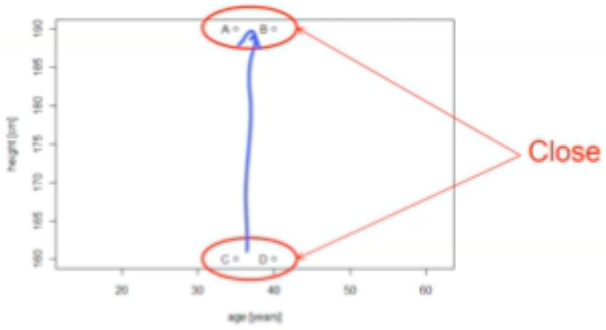
In any n-dimensional data, there is a possibility that features with different units and scale are collected for the sake of analysis. Applying principal components in such cases needs good care of the scale of the data. To get rid of this, we need to understand the effect of scale on the principal components of the data.

Covariance versus correlation - to scale or not to scale:

- Correlation is used when there are features in the data that have different units of measurements. To nullify the effect of scale, normalization is done. This is required because multiplying a certain direction with a huge number will unnecessarily create a high variance in that direction and may make it the direction of the highest variance (Principal component) which it is not.

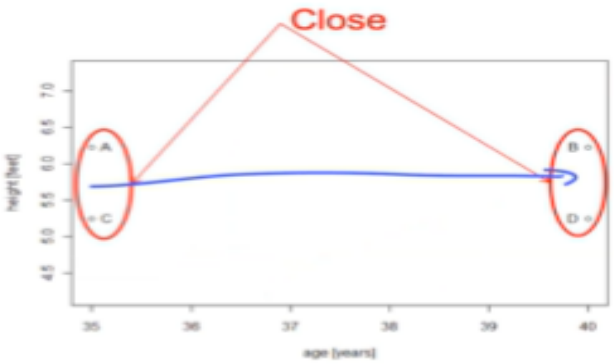
Let us take the help of an example where data of a person's age and height are collected. In the initial stage, the height is given in centimeters, later to inspect the effect of scale we will convert it into feet.

Person	Age(Years)	Height(cm)
A	35	190
B	40	190
C	35	160
D	40	160



Scale is having a huge impact on the principal components. It can even change the direction of the Principal Component. Let us understand it with the above figure. When the height of the person is in cm the principal component is in the vertical direction. Converting it in a different unit (say feet), we can see in the below figure, the principal components are in the horizontal direction. This shows how much it can impact to have a feature in different units.

Person	Age(Years)	Height(feet)
A	-0.87	0.87
B	0.87	0.87
C	-0.87	-0.87
D	0.87	-0.87



This is why, in general, while deploying principal component analysis, we need to apply normalization of features so that the scale effect does not create any unnecessary difference in the outcome.

As we have now seen the importance of using principal component analysis to reduce the dimensionality of the data. Let us now lead to the next method of dimensionality reduction namely **Stochastic Neighbor Embedding**.

Stochastic Neighbor Embedding:

- It is a probabilistic approach to transform objects from high dimension to low dimension. In the given dataset, it takes the individual data points and fits a Gaussian distribution on it. In terms of approach, this method is totally different from the principal component analysis.
- In the lower dimension, it creates another Gaussian in such a way that the two distributions look similar to each other. Two distributions can be considered to be similar if their corresponding dimensions are comparable to each other. In terms of approach this method.
- To describe the similarity between distributions, derivatives of information theory are used. Information theory has many theories that perform similarity computation of distributions. One of these theories is named the **Kullback-Leibler divergence.** It is the method to determine low dimension distribution. Please check the **appendix** to know more about the steps followed in this method.

One of the major problems suffered by most of the embedding methods in reducing the dimensionality of data is the effect of crowding. When the data is reduced from high to low dimensions, then in lower dimensions it becomes crowded. This is because the space available in lower dimensions is less than that in the higher dimension, keeping the count of samples the same in both dimensions. To resolve the problem of crowding instead of using the Gaussian distribution we may choose a flatter distribution that penalizes less when the data comes closer in the lower dimension. A similar variant of SNE is available where we deploy t-distribution instead of Gaussian distribution. It is explained as follows

t-SNE:

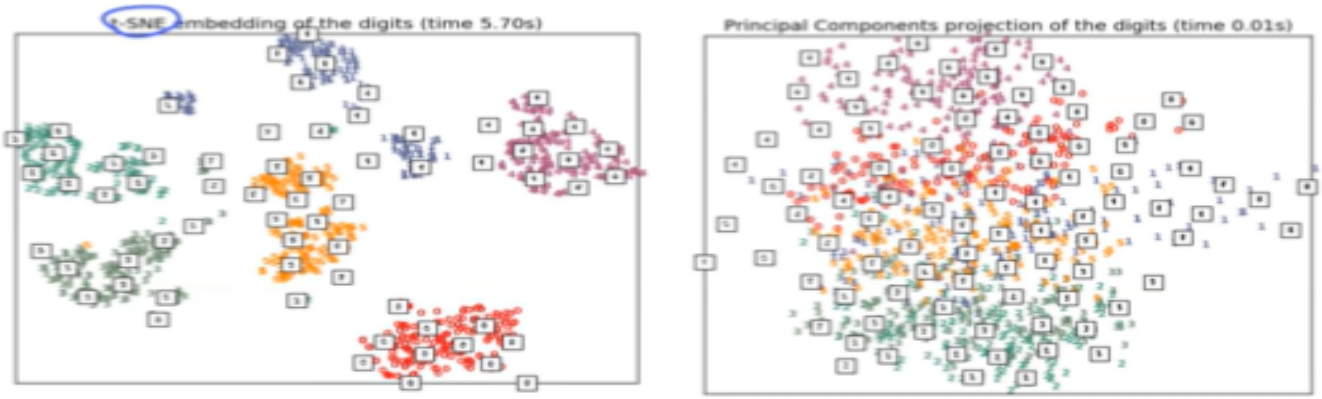
- The t-SNE method differs from normal SNE in terms of the distribution used to replicate the data from higher to lower dimensions. Instead of using Gaussian distribution here, **t-distribution** is deployed on top of every data point. It is a **non-linear** dimensionality reduction technique.

- The major reason behind using the t-distribution is that it resolves the problem of **crowding** better than the Gaussian distribution. A higher dimension, say 100 dimensions, is certainly having much more space than the lower dimension, say 2 dimensions. As t-distribution creates **more space** in the lower dimension, it is preferred over Gaussian distribution. Please check the **appendix** to understand the difference in t-distribution and Gaussian distribution.

Implementation of t-SNE and PCA:

It is observed at some places that t-SNE performs **better** than PCA. Let us see this through an example of **digit recognition**. Here, we have handwritten digits and each digit is put in an 8x8 grid, so there are 64 grey values in each part of the grid. Hence, the input is 64-dimensional data. By reducing the dimensions, it is to be brought into 2-dimensional data.

Both the methods are deployed to perform this task and below is the pictorial representation of the clusters found.



- It can be observed that t-SNE has created **more distinct** clusters than the PCA method.
- One drawback of the t-SNE method is that it takes more time in comparison to PCA. As in the present case, the **time taken** by **PCA** is 0.003 seconds while that by the **t-SNE** method is 2.8 seconds. This is a huge gap in time consumption of the two algorithms, and this difference gets magnified the larger the number of dimensions and the more voluminous your data is.

Remark: The axes in t-SNE have no meaning. They are just embeddings in lower dimensions.

References:

- For a statistics textbook, including controlled experiments and observational studies (chapter 1 and 2) and hypothesis testing (chapter 26-29)
  D.Freedman, R.Pistani, R. Purves. Statistics.2007

- For how to perform testing in R; chapter 4 in P.Dalgaard. Introductory Statistics with R. 2002

- For observational studies and experiments including the HIP study chapter 1 in D.Freedman. Statistical Models: Theory and Practice 2009.

- For selective inference and correcting for multiple hypothesis testing :
  Lecture by Yoav Benjamini, The expert for multiple testing issues:

- http://simons.berkeley.edu/talks/yoav-benjamini-2013-12-11a

- For PCA and other projection methods -

    - B. Everitt & T. Hothorn. An introduction to Applied Multivariate Analysis with R. Springer, 2011.

    - T. Hastie, R.Tibshirani & J. Friedan. The elements of statistical learning: Data mining, inference, and prediction.Springer 2009.

- For t-SNE:

- L. van der Maaten & G. E. Hinton. Visualizing Data using t-SNE JMLR, 2008.

- G. E. Hinton & S. T. Roweis. Stochastic Neighbor Embedding NIPS, 2002.

Appendix

Kullback-Leibler Divergence:

- Given the original dissimilarity matrix D, for each object *i* it computes the probability of picking j as a neighbor as follows -

$$P_{ij} = \frac{e^{-(Dij^2)}}{\sum_{k \neq l} e^{(-Dkl^2)}}$$

Here D_{ij} is the element from the ith row and jth column from the dissimilarity matrix D.

- Now, in the lower dimension space let us compute for each point y_i the probability of picking y_j as its neighbor.

$$Q_{ij} = \frac{e^{-|yi - yj|^2}}{\sum_{k \neq l} e^{-|yk - yl|^2}}$$

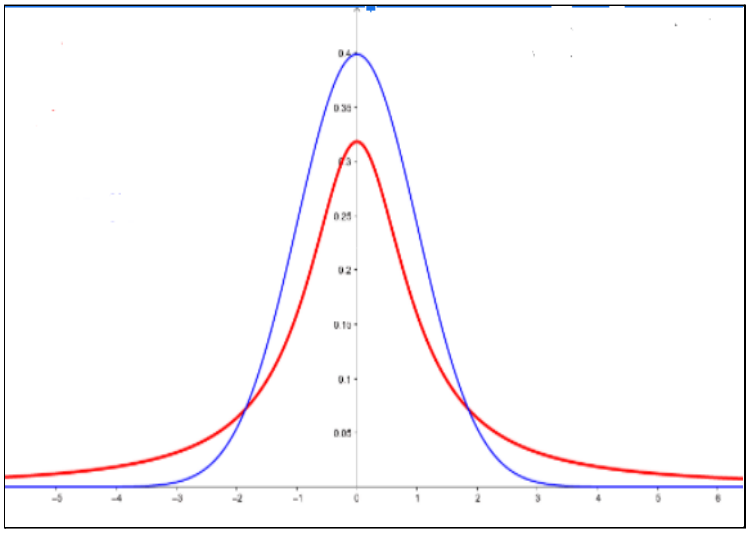
- Minimizing the KL divergence:

$$KL(p||q) = \sum_{k \neq l} Pij \; log(\frac{Pij}{Qij})$$

So, by using the **Kullback-Leibler divergence theorem,** the most suitable and appropriate lower dimension is found that represents the higher dimension in the best possible way.

Difference in t-distribution and Gaussian distribution:

- In the below figure the distribution with blue color is the Gaussian distribution and the one in the red color is the t-distribution.



- Mathematically the working of t-SNE can be defined as follows -

$$Q_{ij} = \frac{(1+|yi-yj|^2)^{-1}}{\sum_{k \neq l} (1+|yi-yj|^2)^{-1}}$$

Here y_i is the ith element in the original high dimension data.

Y_j is the J^{th}term in the lower dimension plane.

Q_{ij} is the probability of picking y_j as the neighbor of y_j.

t-distribution with one degree of freedom is used in the t-SNE method.

It uses Gradient Descent to find optimal parameters.

- In the Gaussian distribution on the tail side, the vertical distance for two equally separated points is more than that in the t-distribution. This causes **more penalization** in the Gaussian distribution. Due to this penalization, there is a chance that close-by points may go into the different clusters which ideally should not happen.

- While reducing the dimension it becomes important to keep points that are **closer** to each other in the high dimension and should remain close in the lower dimension too. They are the ones creating the clusters. Points that are far away, this method does not bother because such points are expected to be in different clusters.

- On the tail side, the t-distribution is **flatter** than the **Gaussian**. Due to this, for points that are closer to each other, the penalization is less in t-distribution in comparison to the Gaussian. When close-by data points come to the lower dimension they are clustered in a better way because the space in t-distribution is stretched more for close points.

- Points that are moderately separated in high dimensions are modeled faithfully by providing a much larger distance in the lower dimension using t-distribution.