Principal Component Analysis Of The Wi-Fi Signal Strength of Smart phones in Indoor Locations

CONCORDIA UNIVERSITY

Quality Systems Engineering

INSE 6220: Advanced Statistical Approaches to Quality

Kartik Ghagre

Student ID: 40160545

Abstract- Principal Component analysis (PCA) is a dimensionally reduction technique that enables to identify correlations and patterns in data set without loss of any important information. This technique is applied to analyze the collected dataset in indoor space by observing signal strengths of Wi-Fi signals visible on a seven Smartphone's. The data was collected to perform the experimentation of Wi-Fi signal strength in one of the indoor locations.

Keywords: PCA, Wireless networks strength and classifications

I. INTRODUCTION

Over the years, much research is based on the gathering and analysis of measurement data. Data-sets are, at least, intermediate results in many research projects. For some time data-sets weren't even published and even if they were published it was mostly as a (not re-usable) by-product of the publication. But an interesting phenomenon might be observed here: data-sets (often in combination with models and parameters) are becoming more important themselves and can sometimes be seen as the primary intellectual output of the research. Publishing and preserving data-sets should therefore seriously be considered. But the dataset complexity determines how difficult a given dataset to classify. Since complexity is a nontrivial issue, it is typically defined by a number of measures. To resolve this issue, multiple dimensionality reduction techniques have been proposed such as feature extraction and selection. So for clearly understanding the data without the complexity and losing the important information the principal component analysis has been used (PCA).

Principal component analysis (PCA) helps to learn about the data sets. It reduces the dimensionality of the data set while retaining as much as possible the variation in the data set. It makes possible the singular value decomposition to compute the principal component analysis.
In this report, we evaluate PCA algorithms on the data set and these algorithms classify the reduced data into categories and helps in better interpretation and understanding of the results of Wi-Fi signal strength.

II. PRINCIPAL COMPONENT ANALYSIS

Principal component is the dimensionality reduction technique for probability and statistics and it is still very commonly used in data science and machine learning applications. When have a big data that might have some statistical distribution and want to uncover the low dimensional pattern to build

models off it. PCA is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. The principal components are eigenvectors of the data's covariance matrix.
Let X be n observation vectors on a random vector x= (X1, X2…Xp) with p quality characteristics is a p-dimensional column vector, which represents the vector of quality characteristics means. The transformation is defined by (The centered data matrix)

$$Z = A.X$$

The p×p covariance matrix S of the centred data matrix is computed as

$$S = \frac{1}{n-1} X'X$$

The columns of Z are called principal components and they represent the maximum possible variance from X.
Thus the j^{th} principal component is

$$Z_j = a_1 jX_1 + a_2 jX_2 + \ldots + a_p jX_p$$

And the Eigen values are arranged in descending order as

$$\lambda_1 > \lambda_2 > \ldots \lambda_p.$$

Then compute the transformed data matrix Z=X.A of size n×p. This matrix contains the original data in a new coordinate system defined by the PCs. Rows of Z represents observation and columns of Z represent PC scores. Eigen values are the variances of the columns of the PCs. Thus the first PC scores contribute the maximum possible variability in the data, and each succeeding component score for as much of the remaining variability as possible.

III. CLASSIFICATION ALGORITHMS

After the reduction of the complex data set into simpler data sets, the next step is to classify the data into categories. In statistical data, the data set classified into different classifications to determine the signal strength. The whole data set classified into categories and these classification methods or algorithms are known as classifiers. The wireless signal strength classified as a weak or strong would define by the 7 variables (Indoor locations). There are several classification algorithms in the literature. The decision on which algorithm to choose would depend on the properties of data.
In this project, we will use two classification algorithms: Logistic regression and Naïve Bayes. Results will be compared.

In this project, one of the classifications will be used to categorise the data namely, Logistics regression. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.
In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Also logical regression is used when the data size is quite large and only the most valuable or contributing variables are included in the data set. Some of the data sets have more than two categories, thus in that case multinomial logical regression or simple logical regression can be used depending on the scale of measurement. Thus, logical regression method predicts that a certain observation x belongs to a certain class v. Now consider a set of attributes where each observation Xi is represented by (x1, x2 , , ,xp) and a binary target function f where f (x) = $v \in$ V, the probability can be expressed as

$$\mathrm{p}\,(X_i \in V_1) = \frac{1}{1+e^{-\beta^T x}}$$

Therefore, the logistic regression classifier defines the attributes to the right class.

Bayes Classifier, like other bayesian learning algorithm, calculates the probability of hypotheses given data and outputs the optimal hypothesis. The Naive Bayes classifier assumes that feature vectors are independent. Unlike many other learning

algorithms, Naive Bayes classifier do not search explicitly the optimal hypotheses in the Hypothesis space, but instead it estimates the most likely hypotheses by counting frequencies of different attributes, Target combinations in the train set.

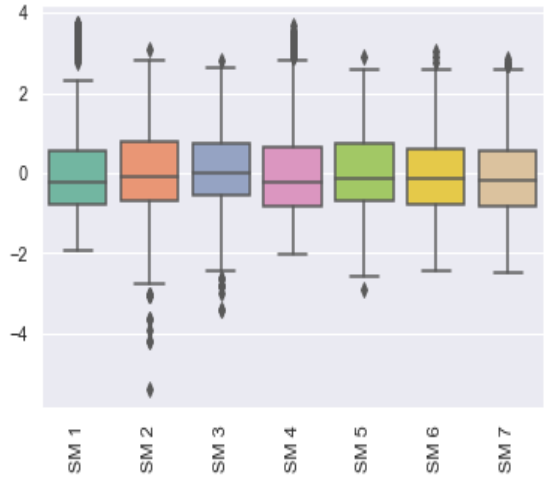
IV. EXPERIMENTAL RESULTS

A. Date set description:

The development of machine learning languages and dimension using methodologies have made it easy to interpret and understand the result of complex data and helped in various such departments. It helps to study of computer algorithms that improve automatically through experience. Wi-Fi is the wireless network protocol, commonly used for local area. It uses multiple parts of the IEEE 802 protocol family and is designed to interwork seamlessly with its wired sibling Ethernet. Compatible devices can network through wireless access points to each other as well as to wired devices and the Internet. The different versions of Wi-Fi are specified by various IEEE 802.11 protocol standards, with the different radio technologies determining radio bands, and the maximum ranges, and speeds that may be achieved. Wi-Fi most commonly uses the 2.4 gigahertz (120 mm) UHF and 5 gigahertz (60 mm) SHF ISM radio bands; these bands are subdivided into multiple channels.

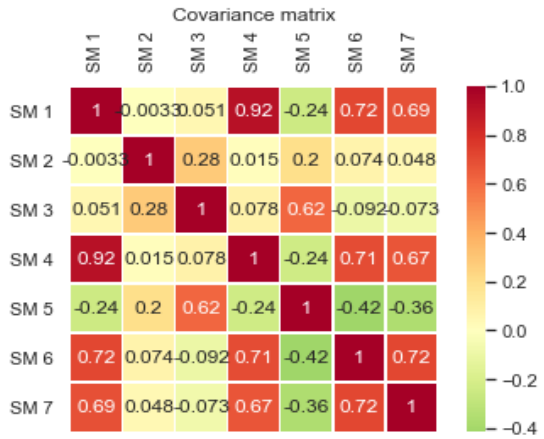
Wi-Fi signal strength is tricky. The most accurate way to express it is with milliwatts (MW), but you end up with tons of decimal places due to Wi-Fi's super-low transmit power, making it difficult to read. Ultimately, the easiest and most consistent way to express signal strength is with dBm, which stands for decibels relative to a milliwatt.

We classified the data sets into various locations at indoor place and by observing signal strengths of seven Wi-Fi signals visible on a Smartphone. There is 2000 wireless signal strength with 7 different smart phones in various indoor locations. The data is then standardized and the distribution of different attributes is represented in a box plot diagram as in fig 1. Also, the figure shows that the data is normally distributed and we can proceed with the data set for further analysis. Also, we can see some outliers for some of the attributes but can still be used and doesn't cause any errors in the analysis. We used data for analysis, outliers doesn't cause any effect.

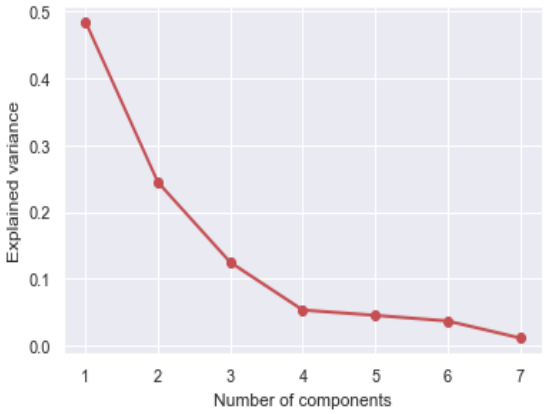


Box plot Fig. 1

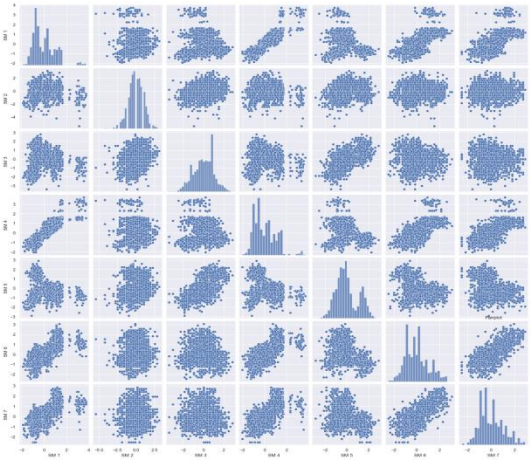
The covariance matrix of the data is shown in Fig. 2. From this covariance matrix plot we can see that the percentage of smart phone 1(SM1) is positive correlated with all variables except SM2 and SM5. It is negative correlated with SM2 and SM5. The correlation between SM4 with all devices is positive except one device with SM5 is going negative. The highest negative correlation can be seen between SM5 and SM6. Similarly, the highest positive correlation exists between SM4 and SM1.These relations can also be seen and confirmed by Pair plot diagram as in fig 3.



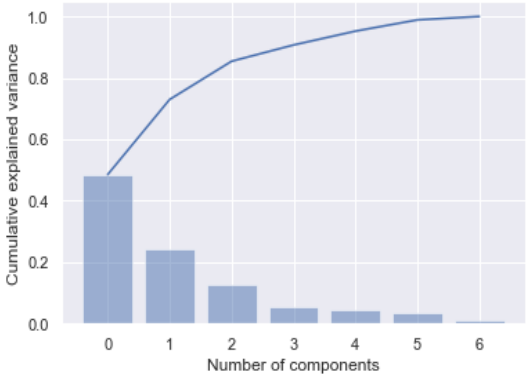
Covariance matrix Fig. 2



Scree Plot Fig. 4



Pair Plot Fig. 3



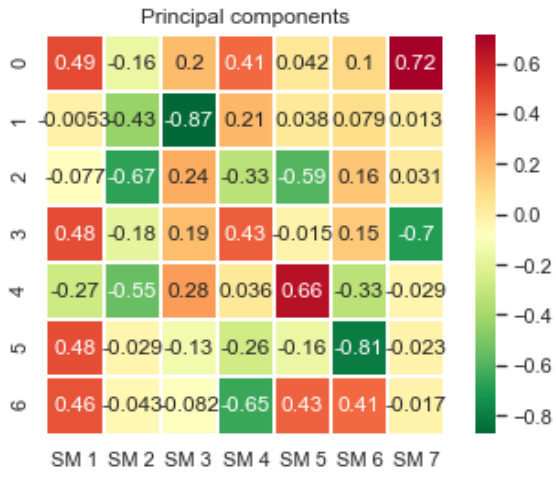
Pareto Plot Fig. 5

The scattering in the pair plot diagram clearly highlights the relations between the four feature vectors. It can be seen that the highest positive correlation between smart phone 1 and smart phone 4 has a linear shape. Also, the highest negative correlation can also be seen between smart phone 6 and smart phone 5.

The above plots explained the variance by the three PCs are L1 = 48.5%, L2= 24.4 %, and L3= 12.4%. Notice, that PC1, PC2 and PC3 combined account for 85.3 % of the variance in the data. Based on the explained variance by PC1, PC2 and PC3 and also from the scree and Pareto plots, it can be deducted that the lowest- dimensional space to represent the signal strength data. The Eigen vector matrix is shown in fig 6.

B. PCA FOR DIMENSIONALITY REDUCTION

We apply PCA on above prescribed data set. The PCA is used to reduce the dimension of the data. The above data set has 7 vectors and that can be reduces to feature vector less than 7. This reduced number of vectors will be computed by using Scree Plot diagram or Pareto diagram. Fig 4 shows the analysis in the form of Scree plot diagram and fig.5 shows the analysis in the form of Pareto diagram.



Eigen Vector Matrix Fig. 6

The eigenvectors Matrix is given by:

$$\begin{pmatrix} 0.49 & -0.16 & 0.2 & 0.41 & 0.04 & 0.1 & 0.72 \\ -0.005 & -0.43 & -0.87 & 0.21 & 0.03 & 0.08 & 0.01 \\ -0.07 & -0.67 & 0.24 & -0.33 & -0.59 & 0.16 & 0.03 \\ 0.48 & -0.18 & 0.19 & 0.43 & -0.01 & 0.15 & -0.7 \\ -0.27 & -0.55 & 0.28 & 0.03 & 0.66 & -0.33 & -0.02 \\ 0.48 & -0.03 & -0.13 & -0.26 & -0.16 & -0.81 & -0.02 \\ 0.46 & -0.04 & -0.08 & -0.65 & 0.43 & 0.41 & -0.01 \end{pmatrix}$$

Therefore, the first two components are given by:
First PC;
$Z_1 = 0.49X_1 - 0.005X_2 - 0.07X_3 + 0.48X_4 - 0.27X_5 + 0.48X_6 + 0.46X_7$.

We can see that the first PC is essentially the contrast between SM1, SM4, SM6, SM7 and SM2, SM3, SM5. This is evidenced by the positive coefficient for SM1, SM4, SM6, SM7 and the negative coefficients for SM2, SM3, and SM5 are to be considered.

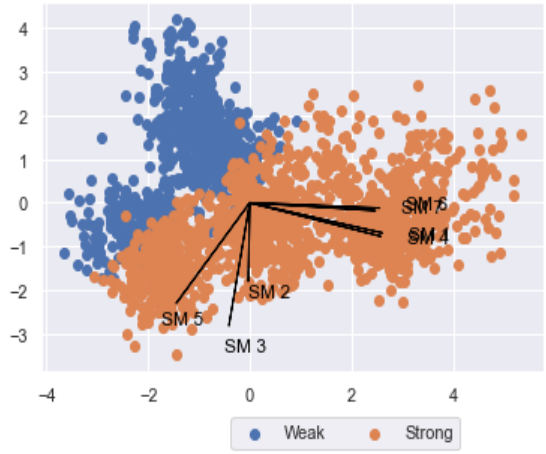
Second PC:
$Z_2 = -0.06X_1 - 0.43X_2 - 0.67X_3 - 0.18X_4 - 0.55X_5 - 0.029X_6 - 0.043X_7$.

In 2nd PC, it appears to be a contrast between all variables negatively. This is evidenced by the negative coefficient for all smart phones from 1 to 7. Further analysis is shows by Scatter pot diagram fig. 7 which shows the plot between two principal components.

Scatter Plot coefficients Of PC1 and PC2. Fig 7

The above scatter plot helps understand which variables have a similar involvement within PCs. As it can be seen in fig. 7, the SM1, SM4, SM6, SM7 is located on the right of the plot, while the other variables are located on the left of the plot. It also confirms that SM1 and SM4 have almost same involvement whereas, SM6 and SM7 have same.

A Biplot allows information on both samples and variables of a data matrix to be displayed graphically. The axes in the biplot represent the principal components (columns of A), and the observed variables (rows of A) are represented as vectors. Each observation (row of Z) is represented as a point in the biplot. The colour of point signifies to the particular class these points belong to.



BiPlot diagram Fig. 8

The Biplot clearly shows that the first principal components have 3 negative coefficients SM2, SM3 and SM5 and 4 positive coefficients SM1, SM4, SM6 and SM7. That corresponds to 3 vectors directed into right half of the plot, and 4 vectors directed into the left half of the plot, respectively. Also, it can be seen that there is all negative coefficients for the second principal coefficients which corresponds to horizontal axis. This analysis confirms the description of the two equations Z1 and Z2. The rays represent the nature of the feature vector by which they affect the principal component. It can be seen that rays in the same direction are related to each other positively and rays that are in the

different direction are negatively correlated with the other rays. From the Biplot, SM6 and SM7 vectors are more on positive side, whereas SM4 and SM1 have negative closer to above 2 positive vectors. All signal strength from the data has a point in the biplot and the location of these points shows the score of each observation for the two principal components. The strength is also assigned a specific colour which divides them into two categories Weak and strong. The Wi-Fi signal strength has ranges from -10 dBM to -98 dBM. The higher the signal strength, strong the signal connectivity whereas, the lower the signal strength Weak the signal connectivity. The blue colour signifies the weak signal strength and the orange colour signifies the strong signal strength. From the biplot, majorly the weak signal plots occupy the left side of the biplot and can be easily distinguished from the strong ones. Also, it can be noticed that the SM6 and SM7 two are closely placed in the biplot and are hard to differentiate because of their common attributes. But on closer look it can be seen the majority of points lying towards negative side close to vertical axis.

V. HOTELLING'S T^2 CONTROL CHART

Hotelling's T^2 control chart is a widely used tool for monitoring simultaneously several related quality characteristics of a process. A $\bar{X}$ chart is basically a chart that is used for determining that if a measurement with only one variable has gone out of statistical control or not. But when there is more than one variate in any data set, T2 control chart is used. Therefore, we can say that T2 control charts are used to detect shifts in a mean of more than one interrelated variable. The T2 statistic is given by:

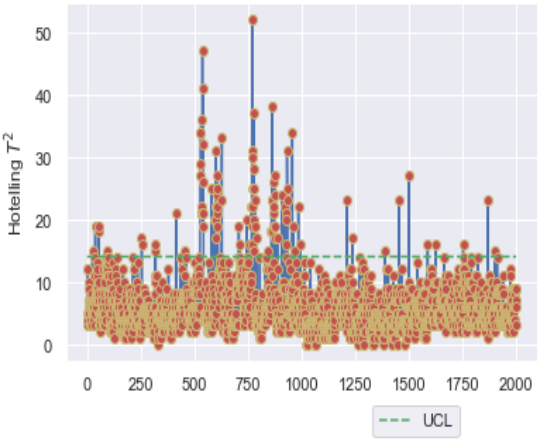
$$T^2 = n\,(\bar{x} - \hat{\mu})'\,\hat{\Sigma}^{-1}(\bar{x} - \bar{\bar{x}})$$

Also, for phase I the upper control limit and for phase II the upper control limit is calculated as:

$$UCL = \frac{p\,(m-1)(n-1)}{mn-m-p+1}\,F_{\alpha,\,p,\,mn-m-p+1}$$

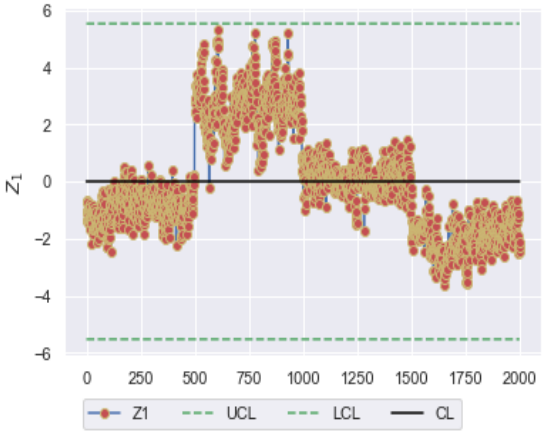
$$UCL = \frac{p\,(m+1)(n-1)}{mn-m-p+1}\,F_{\alpha,\,p,\,mn-m-p+1}$$

The hotelling T^2 chart displays all the signal strength samples.



Hotelling T^2 control chart Fig. 9

The Hotelling's T2 control chart displays the data sets which are out of the control limit. From the control charts, mostly points are in the limits but few of the points are in the out of control region.



T^2 control chart Fig. 10

The control chart for first component shows that every signal is under control limits.

VI. CLASSIFICATION

In this classification, we applied two classification algorithms on original data. Logistic regression and Gaussian Naive Bayes. The goal is to asset the impact of PCA on two different classifiers and to compare the performance of the models. In total 6 experiments have been conducted:

- Logistic Regression
  - ✓ Original Dataset
  - ✓ All principle Components
  - ✓ First two components
- Naive Bayes
  - ✓ Original Dataset
  - ✓ All principle Components
  - ✓ First two components

Fold Cross Validation Accuracy						
Fold	Logistic Regression			Naive Bayes		
	Original	All Components	First Two Components	Original	All Components	First Two Components
1	0.9675	0.9675	0.9775	0.9775	0.9500	0.9700
2	0.9925	0.9925	0.9475	0.9600	0.9525	0.9400
3	0.9925	0.9925	0.9475	0.9925	0.9700	0.9500
4	0.9575	0.9575	0.9600	0.8225	0.8900	0.9075
5	0.9700	0.9700	0.9150	0.5800	0.6750	0.6775
Average	0.9760	0.9760	0.9495	0.8665	0.8875	0.8890

Fold Cross validation Fit Time						
Fold	Original	All Component	First Two Components	Original	All Componen	First Two Component
1	0.0170	0.0170	0.0110	0.0060	0.0050	0.0050
2	0.0250	0.0170	0.0100	0.0050	0.0050	0.0060
3	0.0170	0.0200	0.0090	0.0060	0.0050	0.0030
4	0.0190	0.0170	0.0140	0.0050	0.0030	0.0030
5	0.0210	0.0170	0.0110	0.0050	0.0030	0.0030
Average	0.0198	0.0176	0.0110	0.0054	0.0042	0.0040

Classification Fig. 11

The performance and accuracy for the classifier can be determined by the fit time data and its accuracy. The average fit time data or the average accuracy level can be calculated manually. Logistic Regression performed on the full dataset has the highest accuracy. In accuracy table, the first two components have performed best in all components.

VII. CONCLUSION

Principal component analysis is an analyzing algorithm used for data with various variables or data sets with more than one feature vectors and thus for the analysis of Wi-Fi signal strength classification this analysis was used as the data has various feature vectors. The wireless signal strength data set with 2000 strength was first introduced to PCA algorithm to reduce the complexity of the data and after reducing the vectors the results showed that only three components were required to carry out 85.3% features of data. After this process, the analysis was carried out by the help of Hotelling's T2 chart to find out points which are out of the control limits. After the PCA analysis the data set was analyzed with classifiers which classified the data into two categories of the signal strength which are Weak and strong connectivity. After the classification the performance of the classifier was also carried out to check whether the classifier is reliable or not by checking the accuracy level of the classifier. Thus, the data set was successfully analyzed by PCA and was classified using Logical Regression and Gaussian Naïve Bayes.

VIII. REFERENCES

1. A. Ben Hamza, "Advanced Statistical Approaches to Quality", Unpublished.
2. https://en.wikipedia.org/wiki/Wi-Fi
3. https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization
4. https://en.wikipedia.org/wiki/Principal_component_analysis
5. https://en.wikipedia.org/wiki/Logistic_regression
6. https://en.wikipedia.org/wiki/Data_set
7. https://en.wikipedia.org/wiki/Statistical_classification
8. https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=176