Project 4

Principal Component Analysis (PCA)

Dana Salem 900191664

Nour Montasser 900191101

---

1. Introduction

1.1 Statement of Problem

The increasing risk of breast cancer severity is more for the women in their entire lifetime (SR, S. C., & Rajaguru, H., 2019). Therefore, the rapid and accurate diagnosis of breast cancer is of great significance for the treatment of cancer (Bian, Kai, et al., 2020). The clinical physicians use microscopic study for the detection and classification of breast cancer conventionally (Documet et al., 2015). However, there are numerous attributes of breast cancer tumor and examining them individually for each patient might slow down the diagnostic process and is computationally expensive, this is known as the “curse of dimensionality”. For that reason, machine learning algorithms that utilize the dimension reduction methods paved an easier way for the analysis and classification of the type of breast cancer. To elaborate, this project will implement the principal component analysis method in order to reduce the number of variables used in the detection of the type of breast cancer tumor, leaving behind the most significant and non-redundant attributes.

1.2 Literature Review

Principal component analysis is one of the key tools that aid in reducing dimensionality to enable more efficient clustering of data points. This tool is also uniquely equipped to handle computationally expensive processes on a given dataset (Abdi & Williams, 2010). In fact, Bian,

Kian et al. conducted a study where PCA was employed on a large breast cancer dataset. The results reduced the variables of the dataset from 30 to 21 attributes which are the most significant features in detecting the tumor and contribute the most to the variation in the data (Bian, Kian et al., 2020). With the dimension reduction prior to their rigorous analysis, the researchers were able to conclude that the PCA improved the time taken to accurately identify breast cancer and provides a theoretical basis for the intelligent diagnosis of breast cancer. While PCA offers valuable insights, it is important to consider the limitations of the technique.To elaborate, one challenge is the interpretability of the transformed principal components (Chao, Yi-Sheng, et al., 2018). In other words, the meaning of each component may not always be straightforward. Furthermore, information loss may occur during dimensionality reduction, potentially discarding important details that could affect the accuracy of breast cancer diagnosis (Chao, Yi-Sheng, et al.,2018). To address these limitations, hybrid approaches combining PCA with other machine learning techniques, such as classification algorithms, can be employed to enhance the diagnostic capabilities of the researchers (Wang, Jiangfeng, et al., 2022).

2. Data Description

2.1 Data Source

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/citation_policy.html

2.2 Summary of Data

The Breast Cancer Wisconsin (Diagnostic) Dataset is a publicly available dataset created by Dr. William H. Wolberg from the University of Wisconsin and was donated to the UCI Machine Learning Repository. It contains measurements of the characteristics of breast cancer tumors

with no missing values. The 569 observations were gathered from a digitized image of a fine needle aspirate (FNA) of breast masses of female patients diagnosed with breast cancer. There are 12 variables in this dataset 2 of which are nominal and categorical and the remaining are quantitative. This dataset will be divided into two classes: malignant patients and benign patients. The description of each variable is given in the below Table 1.0.

2.3 Variables Description

Variable Name	Type	Unit of Measurement	Description
id	Nominal	N/A	This is a unique identifier assigned to each patient.
diagnosis	Categorical	N/A	This is a qualitative variable that identifies the type of the tumor whether it is malignant (M) or benign (B).
radius_mean	Quantitative	Millimeters	This feature measures the mean of distances from the center to points on the perimeter of the tumor. It represents the average size of the tumor.

texture_mean	Quantitative	N/A	This feature measures the mean of gray-scale values in the image of the tumor. It represents the variation in the pixel intensities in the image. It is measured on a scale from 0 to 100.
perimeter_mean	Quantitative	Millimeters	This feature measures the mean perimeter of the tumor, which is the length of its boundary.
area_mean	Quantitative	Millimeters Squared	This feature measures the area of the tumor, which is the total number of pixels inside the boundary. It is measured in square mm.
smoothness_mean	Quantitative	N/A	This feature measures the mean local variation in radius lengths of the tumor. It represents how much the radius of the tumor changes at different points along its boundary. It is measured on a scale from 0 to 1.

compactness_mean	Quantitative	N/A	This feature measures the ratio of the perimeter squared to the area of the tumor, minus 1.0. It represents how tightly the tumor is packed together. It is measured on a scale from 0 to 1.
concavity_mean	Quantitative	N/A	This feature measures the mean severity of concave portions of the contour of the tumor. It represents the amount of concavity in the boundary of the tumor. It is measured on a scale from 0 to 1.
concavepoints_mean	Quantitative	N/A	This feature measures the number of concave portions of the contour of the tumor. It represents the number of inwardly-curved sections in the boundary of the tumor. It is measured on a scale from 0 to 1.

symmetry_mean	Quantitative	N/A	This feature measures how symmetric the tumor is. It represents how similar the left and right halves of the tumor are. It is measured on a scale from 0 to 1.
fractaldimension_mean	Quantitative	N/A	This feature measures the mean "coastline approximation" of the tumor. It represents how much the boundary of the tumor is convoluted. It is measured on a scale from 0 to 1.

Table 1.0 : Description of Variables

3. Data Exploration

Prior to implementing the analysis, it is important to preview the data in order to form some

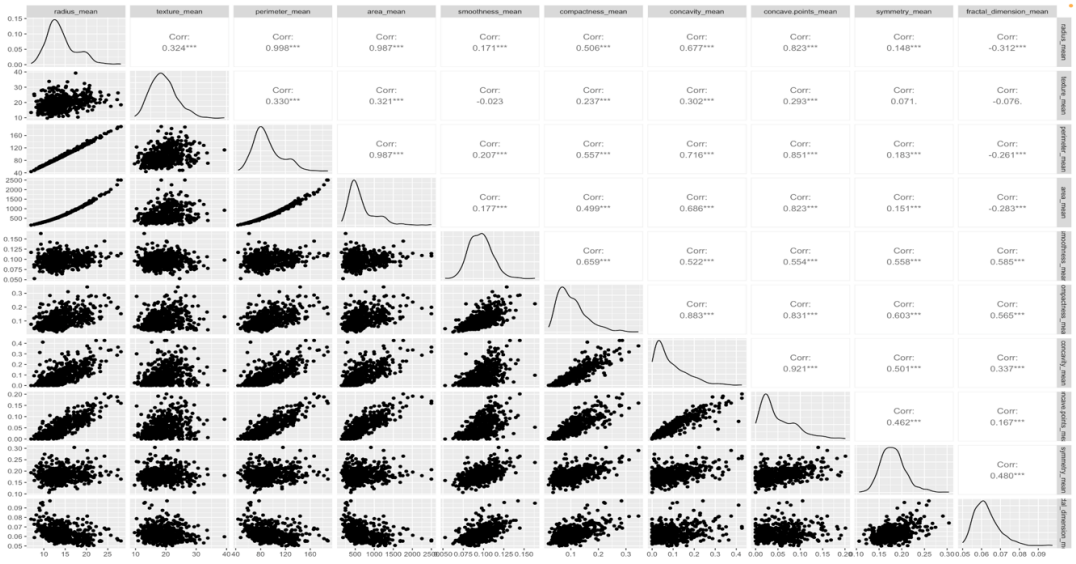

Figure 1.0: Pairwise Scatterplots and Correlations

hypotheses.

For that reason, a scatterplot matrix with kernel density estimates and pairwise correlation coefficients was necessary. From Figure 1.0, the presence of some outlying observations is observed since most of the data points are concentrated in a specific area and few points are further away from the bulk of the data implying that the correlation values might be misleading. Moreover, it can be concluded from the diagonal graphs that most of the variables are relatively skewed to the right indicating the presence of extreme values. Therefore, it is hypothesized performing the classical PCA and the robust PCA will not yield the same results.

More relevant to the analysis at hand is the linear relationships amongst certain variables. Most notably, there appears to be a very strong positive relationship between the radius and area variables, which is expected since the radius is used to calculate the area. Additionally, there is also a strong positive linear relationship between the radius and the perimeter. This is an anticipated result because the radius is used as a part of the calculation of the perimeter .Lastly,

the radius and compactness variables are positively linearly related. To elaborate, compactness is defined as the ratio between surface area and volume where both measures require the use of the radius measure thus dependency is expected. Conclusively, these observations suggest the existence of redundancies within the dataset and pose significant potential for dimensionality reduction.

4. Data Analysis

4.1 Classical Principal Component Analysis

In order to identify the principal components of the dataset at hand, the “princomp” function in R is used. According to the literature, the rule used to identify the number of selected components is provided that the standard deviation of the component is greater than 1, then this component is selected. The results used to identify the number of components is given below in Figure 2.0.

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6      Comp.7      Comp.8      Comp.9      Comp.10
Standard deviation     2.3406384 1.5870456 0.93841099 0.70640600 0.61035989 0.35233755 0.282993481 0.186788096 0.105524692 1.680196e-02
Proportion of Variance 0.5478588 0.2518714 0.08806152 0.04990094 0.03725392 0.01241417 0.008008531 0.003488979 0.001113546 2.823059e-05
Cumulative Proportion  0.5478588 0.7997302 0.88779168 0.93769262 0.97494654 0.98736071 0.995369244 0.998858223 0.999971769 1.000000e+00
```

Figure 2.0 : Output “princomp” Function

It is clear that the standard deviations of the first two components only are greater than 1. For that reason, the selected components are components 1 and 2. Additionally, there is a graphical approach to deciding on which components to select. In this approach, a line plot of the eigenvalues of factors or principal components in an analysis are drawn up in a graph called the scree plot.

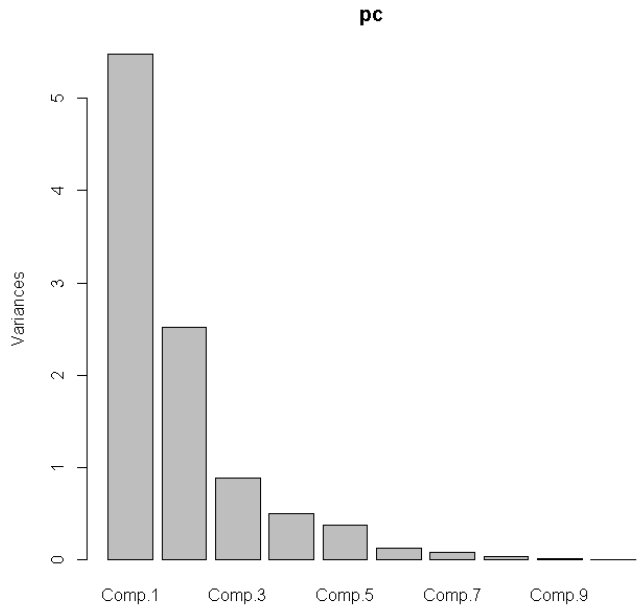

Figure 3.0 : Scree Plot

Similar to the summary of the R function, the results obtained from the scree plot indicate as well that the first two components are the ones that ought to be selected given that they provide the most information about the dataset.

In addition, The cumulative sum of the variances of the principal components, are sorted by largest to smallest variance, are shown in Table 2 to further understand how much variability in the data is explained by each component.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.54785	0.79973	0.88779	0.93769	0.97494	0.98736	0.99536	0.99885	0.99997	1.0000

Table 2.0 : Cumulative Sum of Variances of Principal Components Using Classical PCA

From the table above, it is observed that the first component captures around 55% of the variability in the dataset and the second component explains around 80%. In most literature, it is common to choose the components that explain around 80% of the variability in the data (Jr., T. T., 2022). However, given that a small error in diagnostics would drastically impact the patient, it is better to be more accurate even if it would make it slightly complex and include the third component as well because it explains another 10% of the variability. For that reason, the selected components are the first three components.

4.2 Robust Principal Component Analysis

Since PCA is conducted using the covariance matrix of the data, it is affected by outliers. Hence, the need for robust PCA arises as an essential step in our data analysis process to avoid the impact of outliers on the results. By employing robust estimation techniques and downweighting the influence of outliers, robust PCA allows understanding of the true latent structure of the data and obtains more reliable and meaningful insights. For that reason, a robust outlier detection method, to identify observations that deviate significantly from the overall pattern of the data was implemented. The method is called the BACON algorithm where robust distance measures, such as the robust Mahalanobis distance, are utilized to determine the outlying observations.

The results of executing the mvBACON function in R are reached after 9 iterations where 47 observations out of the 569 observations are outliers. In percentage terms, the data set consists of 8.26% outliers.The displayed graph depicted the robust distances obtained from the Bacon technique on the y-axis, while the corresponding indices of the observations were shown on the x-axis. Additionally, a horizontal line to the graph was added to serve as a threshold for outlier identification.

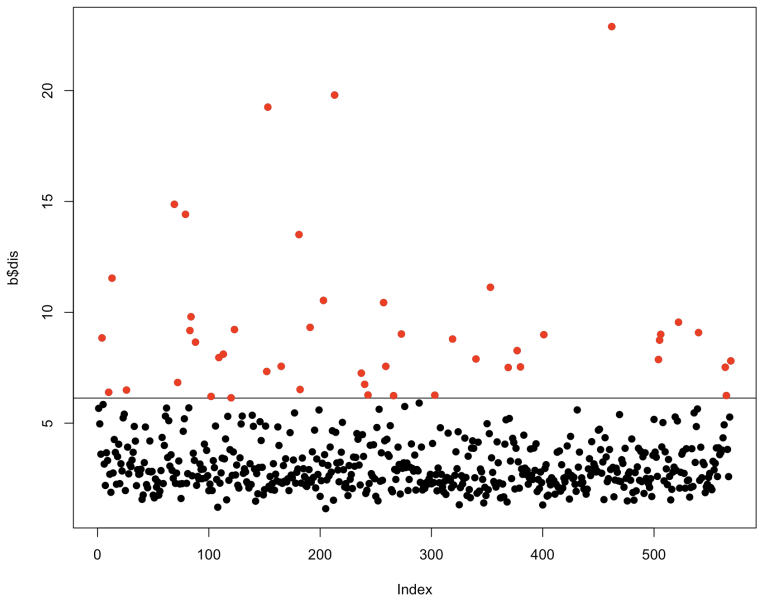

Figure 4.0 : BACON output

Once the outliers were identified, it is important to understand the reason behind the deviations from the rest of the data points before deciding on removing them from the data. However, due to the time constraint, the outliers were automatically removed without analysis. This allowed the exploration of the fundamental relationships and variations in the data without the interference of uncommon observations yielding more accurate and reliable results from the PCA analysis.

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	2.3193844	1.5870941	0.95030056	0.76690441	0.58961459	0.37992535	0.282703255	0.177262675	0.0830899999	1.178551e-02
Proportion of Variance	0.5379544	0.2518868	0.09030712	0.05881424	0.03476454	0.01443433	0.007992113	0.003142206	0.0006903948	1.388983e-05
Cumulative Proportion	0.5379544	0.7898412	0.88014829	0.93896253	0.97372707	0.98816140	0.996153510	0.999295715	0.9999861102	1.000000e+00

Table 3.0: Output of “princomp” Function

Based on the information presented in Table 3.0, we can make the following conclusions. The standard deviations for the first two components are greater than 1, indicating their higher variability and importance in explaining the data. Consequently, we will select these first two components for further analysis. Additionally, the scree plot shown in Figure 5.0 supports this decision, as it suggests that the first two components capture a significant amount of information about the data.

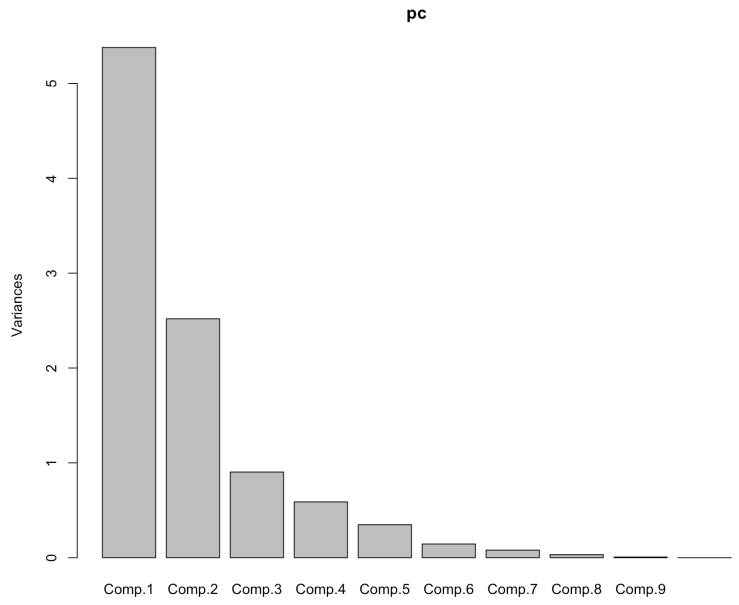

Figure 5.0 : Scree Plot

Furthermore, considering the cumulative sum of variances, we aim to achieve a threshold of 90% explained variance. By examining the cumulative proportion of the variance column, we can observe that selecting the third component would allow us to capture 88% of the data. This choice aligns with our objective of achieving a substantial level of coverage of 90%, as mentioned before, while maximizing the usefulness of the diagnostic data.

Based on the output of the R function, the scree plot, and cumulative proportion of variance, we can conclude that the first few components play a more substantial role in explaining the variance in the data compared to the later components. Therefore, selecting the third component would capture a significant proportion of the variability in the dataset, about 0.88 (88%), while minimizing the loss of information.

4.3 Comparing Results

When comparing the results from the classical PCA and the robust PCA, the initial hypothesis that the output would be different is rejected, indicating that the outliers did not have a significant impact on the results. To elaborate, the standard deviations under each approach slightly varied. However, when the rule of selecting the components that had standard deviations greater than 1 was applied, both methods result in the selection of the first two components only. This implies that the outlying observations had an insignificant impact on the dataset. As for the graphical representation using the scree plot, the graphs for each method did not show any visual differences. And the same conclusion was reached under the classical PCA and the robust PCA which is that the selected components are the first two components only. Lastly, when comparing the cumulative sum of variances under each method, it is noticed that the variability explained by each component under the robust PCA is slightly less than under the classical PCA. This is expected because the classical PCA was performed without removing the outliers and these observations surely played a role in explaining a part of the variability in the data, even if it was minor. But once again, the differences were insignificant and the same results were reached under both methods. And it was decided to include the first three components which contradicted the results reached by the previous test and visual representation because the accuracy with this

medical dataset ought not to be compromised for the simplicity resulting from dimension reduction as explained before. Conclusively, the two approaches reached the same conclusion in practice because the outliers only made up 8% of the dataset; not a significant percentage, so there is no approach better than the other.

5. Conclusion

In conclusion, there were two methods of Principal Component Analysis implemented in this project, the classical PCA and the robust PCA. It was found that both methods reached the same conclusions. To elaborate, according to the test and the graphical representation, it was found that the first two components only ought to be selected. However, when the sum of cumulative variances were observed, it was decided that it is necessary to include the components that provide 90% explanation of the variability in the dataset because with medical data it is better to be more accurate than less complex. For that reason, the third component was also added to the selected components. With that the Principal Component Analysis has indeed reduced the dimensionality of the dataset from 10 attributes to 3 attributes, thus potentially improving the time taken to diagnose the type of breast cancer tumor. However, due to time constraints, we were unable to check the linearity assumption prior to conducting the PCA and it was performed even though it was obvious from the scatterplot that some variables were not linearly related. This implies that some of the correlations computed are meaningless. For that reason, I suggest transforming the data prior to the analysis. In addition, I suggest examining and analyzing the detected outlier prior to removing them completely from the dataset because they might be important data points.

References

Bian, Kai, et al. "RF-PCA: A new solution for rapid identification of breast cancer categorical data based on attribute selection and feature extraction." *Frontiers in genetics* 11 (2020): 566057.

Chao, Yi-Sheng, et al. "Principal component approximation and interpretation in health survey and Biobank data." Frontiers in Digital Humanities 5 (2018): 11.

Chiu, Huan-Jung, Tzuu-Hseng S. Li, and Ping-Huan Kuo. "Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine." IEEE Access 8 (2020): 204309-204324.

Documet P, Bear TM, Flatt JD, et al. The association of social support and education with breast and cervical cancer screening. Health Edu Behav. 2015;42:55–64.

Jr., T. T. (2022). PCA 102: Should you use PCA? how many components to use? how to interpret them?. Medium.

SR, S. C., & Rajaguru, H. (2019). Comparison analysis of linear discriminant analysis and cuckoo-search algorithm in the classification of breast cancer from digital mammograms. Asian Pacific journal of cancer prevention: APJCP, 20(8), 2333.

Wang, Jiangfeng, et al. "A novel combination of PCA and machine learning techniques to select the most important factors for predicting tunnel construction performance." Buildings 12.7 (2022): 919.