Analysis of the effect of a TheraTech drug on the treatment of cancer patients.

Matthew Parker
Department of Chemistry, University of Bristol.

(Dated: October 30, 2024)

This report analyses the effectiveness of a new TheraTech drug on the improvement of symptoms in a sample of 2000 cancer patients. Principal Component Analysis (PCA) and Gaussian Mixture Models were used to reduce the number of dimensions of the provided data to allow detection and visualisation of patterns within the data, and to classify the improvement of patients. PCA showed that the first two components explained 82% of the variance, with markers 1 and 2 being the most important in the determination of clusters. Overall, it was calculated that 71.55% of patients improved after treatment. However, this result may be unreliable, as it relies on the assumption that all patients grouped in the same cluster as a single patient who improved significantly also experienced improvement.

INTRODUCTION

In the UK, over 167,000 deaths each year are caused by cancer, with approximately half of all patients diagnosed with cancer dying within ten years. [1] Treatment of cancer is evidently very important, as cancer is one of the largest causes of death in the UK. [2] Cancer markers are substances that can be detected in tissue, blood, urine, and other bodily fluids, and are key for treatment of cancer patients as they are used to monitor cancer progression and response to treatment. Since these markers can be influenced by external factors unrelated to cancer, multiple markers are often used to ensure that observed changes are truly linked to the disease rather than these external influences. However, some markers show more distinctive results when the treatment is working and are better indicators of the effectiveness of a drug. [3]

In this report, data provided by TheraTech was analysed to assess the effectiveness of a drug on the treatment of cancer patients. Measurements for six different markers were taken both before and after treatment started for 2000 patients, with the pre- and post-treatment data provided in two separate dataframes. The exact details of the measurements were not provided, and instead were given generic names - marker 0 through to marker 5 – to maintain anonymity. All the provided data was pre-processed and normalised against other factors, to ensure any changes observed were purely a result of the drug, rather than being from any other factors.

Principal Component Analysis (PCA) was an important analysis technique used throughout this project. This is a technique used to reduce the number of dimensions of a dataset typically down to 2 or 3 dimensions, while maintaining as much of the variation of the dataset as possible. By reducing dimensionality, PCA makes it easier to visualise complex data, revealing underlying structures or patterns that may relate to the treatment outcomes. In this project, the dataset was reduced from 6 dimensions to 2 dimensions.

Gaussian Mixture Modelling (GMM) is a clustering technique that models the data as a mixture of several Gaussian distributions, which was used in tandem with PCA to distinguish distinct clusters in the data. By applying GMM, the analysis can identify patterns in how the patients reacted to the treatment, potentially distinguishing patient groups based on how they reacted to the drug.

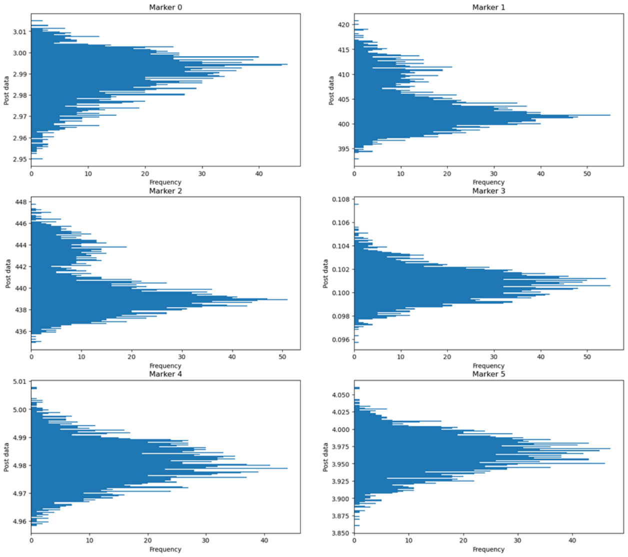

FIG. 1. Distributions of post-treatment data as histograms for each marker.

ANALYSIS AND DISCUSSION

Initially, the pre- and post-treatment data for each marker were plotted as individual histograms to see if any obious patterns emerged. Figure 1 shows the histograms of the post-treatment data for each marker.

The pre-treatment data for each marker generally followed normal distributions. If the drug had no effect on the patients, the post data for each marker would be expected to also follow a normal distribution. However, for markers 1 and 2, the data has a clear bimodal distribution with two distinct peaks. In the plot of pre- against post-treatment in figure 2, this reveals itself as two clusters. The histogram for marker 0 also looks to have a skewed distribution as it tails off towards lower values of the post-treatment data. Markers 4, 5, and 6 show no obvious patterns.

These initial patterns hint at the effect the drug has, as it is

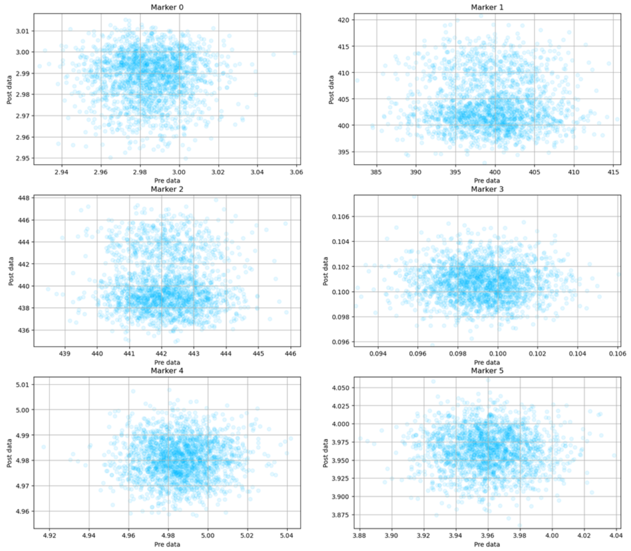

FIG. 2. Pre- vs post-treatment data for each marker

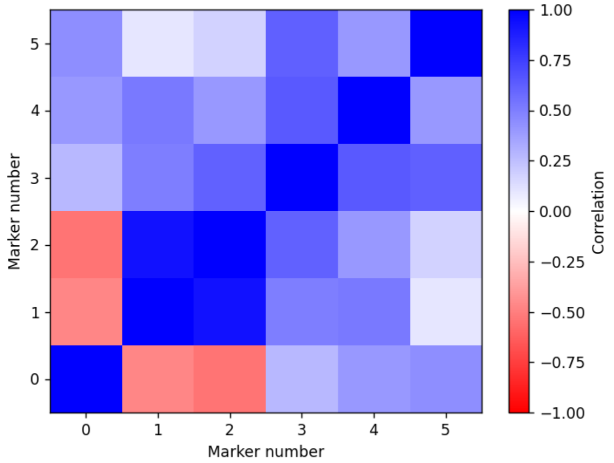

FIG. 3. Colour map of correlation values between markers. Blue represents positive correlation, red represents negative correlation, and white represents no correlation. The squares on the positive diagonal all have a value of exactly 1 as the markers are perfectly correlated with themselves.

possible the data for patients who improved and unimproved split into separate clusters for markers 1 and 2, and caused the skewed distribution in marker 0.

Figure 3 shows a grid of the correlation values between each combination of markers which can be used to see if any of the markers react to the drug in the same way. The diagonal values are all exactly 1, as these are the same data points plotted against each other. This figure shows a strong correlation between markers 1 and 2, as well as an anticorrelation between markers 1 and 2 with marker 0. Furthermore, marker 5 has very low correlation with markers 1 and 2.

PCA was then done on the dataset for each marker to reduce the dimensions of the datasets from six (one for each marker) to two dimensions.

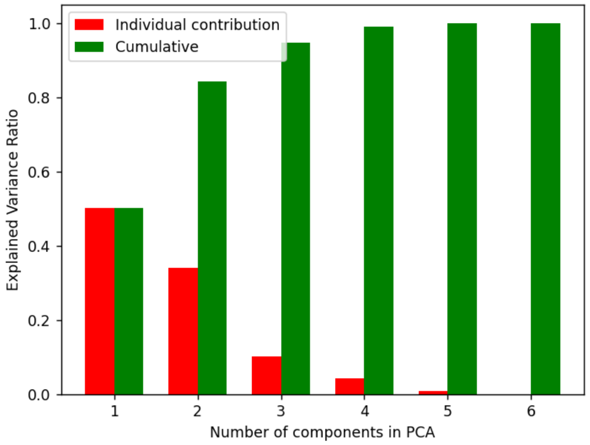

FIG. 4. Bar chart showing the explained variance ratio from each of the components in the principal component analysis. Red bars show the individual contribution from each component, whilst the green bars show the cumulative contribution up to and including each component.

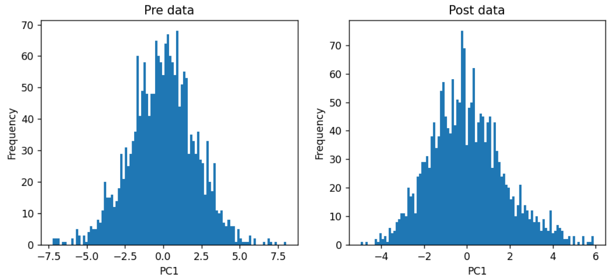

FIG. 5. Comparison of histograms showing the first principal component of both the pre and post data. 100 bins were used in each figure.

Figure 4 shows the variance explained by each principal component, as well as the cumulative value up to and including each value. With one principal component used, only 50% of the variance of the dataset is accounted for, which is relatively insignificant.

This loss of variance using just one principal component can be seen in the histogram of the first principal component (PC1) for the post dataset, shown in figure 5. It has a similar distribution to the pre-treatment data, and does not provide any meaningful insights into the dataset.

However, when two principal components are used, the explained variance reaches 82%, which is significantly greater. Figure 6 shows the first principal component (PC1) plotted against the second principal component (PC2) for both the pre- and post-treatment datasets. The post-treatment data produces two clear clusters which suggests the data for the patients has split into two separate groups - likely related to whether the patients improved or didn't after taking the treatment. Performing PCA on the pre-treatment data confirmed that the clusters were due to the treatment, as no clusters

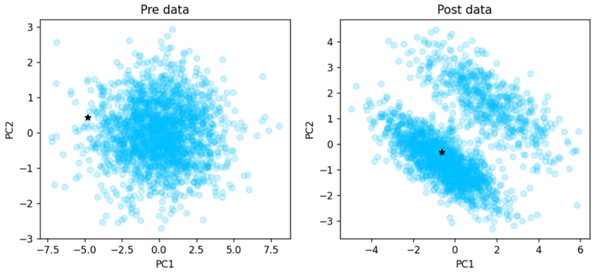

FIG. 6. Figures showing the first two principal components plotted against each other for the pre data and the post data. The post data clearly shows two distinct clusters, whereas the pre data has no distinguishable pattern. The data points for ‘patient zero’ are marked with a black dot.

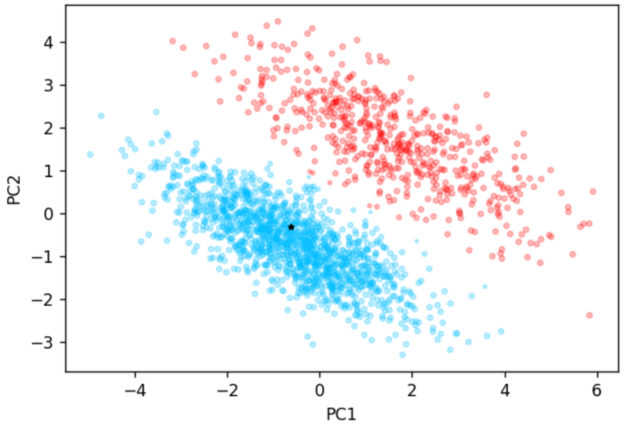

FIG. 7. Figure showing the two clusters from plotting the first two principal components against each other, after the Gaussian Mixture Model has been applied. The size of the dots corresponds to the probability of each data point being in its designated cluster, with smaller dots representing lower certainty for which cluster that data point will fall into. The black dot represents the data point of ’patient zero’.

formed. This shows there were no distinct groups before the treatment and suggests that the clusters are purely due to the patients’ response to the TheraTech drug.

Figure 7 shows the clusters formed using a Gaussian Mixture Model, which was then used to quantify the number of data points in each of these two clusters. This calculation revealed that there were 1431 patients in one cluster, and 569 patients in the other cluster. These values are the same depending on whether the clusters are calculated using two or three principal components.

The patient who was in the first row of the dataset (’patient zero’) was revealed to have improved significantly after taking the TheraTech drug and is represented in figure 6 by a black dot. This information about patient zero was used to determine which cluster represented patients who improved and patients who didn’t. This may be a large assumption but it is the only available reference for our analysis. With this assumption in mind, an improvement rate of 71.55% was determined, as patient zero was contained in the larger cluster.

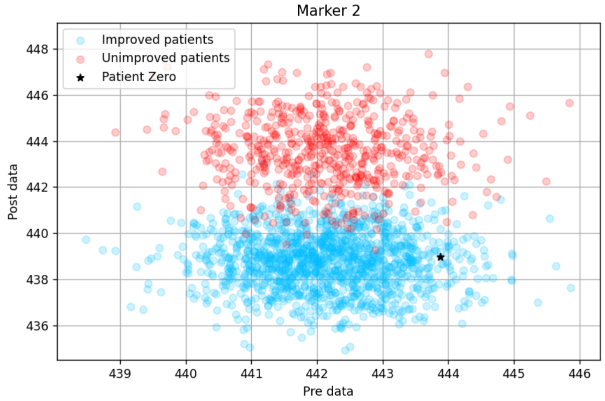

FIG. 8. Example plot of pre data against post data for marker 2 showing improved patients in blue and unimproved patients in red. The two clusters formed are roughly split between improved and unimproved patients, as predicted. The datapoint for ’patient zero’ is represented with a black dot, showing it is inside the improved patients cluster as expected.

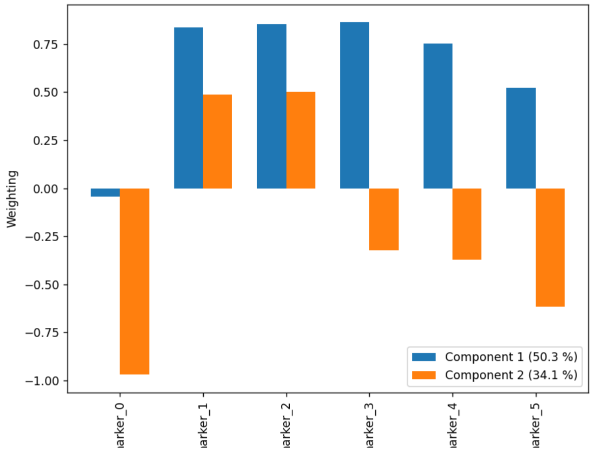

FIG. 9. Bar chart showing the loading of each marker for each of the first two components.

With this information, the knowledge of which patients improved and which didn’t was used to show how the data varied in the initial plots. Figure 8 shows the pre- against post-treatment plot for marker 2, which clearly shows the improved and unimproved patients in separate clusters.

Figure 9 shows the loadings of the markers on each of the first two principal components. This shows which markers had the most influence in determining the principal components. Bars with larger magnitudes indicate stronger relationships whereas low magnitude loadings are less relevant in understanding that specific principal component. [4]

This information can be clearly plotted on a biplot, as shown in figure 10. The relationship between the biplot and the loadings bar chart can be seen with the arrow for marker 0 which has very little dependency on PC1, but varies significantly with PC2. This can also be seen clearly in the bar chart

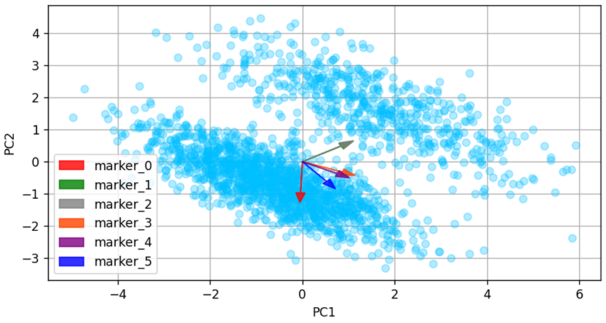

FIG. 10. Biplot of the first principal component plotted against the second principal component. The arrows represent the loading vectors for each marker, showing the direction and magnitude of its influence on each of the first two principal components.

for marker 0.

The biplot also shows that markers 1 and 2 were most significant in determining the clusters, as these arrows have a large component pointing between the clusters. The markers 3, 4, and 5 point parallel to the bodies of the clusters, which shows that their main contributions were to the width of the clusters, rather than the separation between them. This is evidently less useful, as the separation of the clusters was what was used to determine which patients improved and which didn't. This shows that markers 1 and 2 were the most important in determining the clusters. Marker 0 also contributed significantly, but not quite as much as the other two markers.

CONCLUSIONS

Overall, this analysis has provided insights into the effectiveness of TheraTech's drug for treating cancer patients by examining data across six anonymised markers, measured both before and after treatment. Initial exploration showed that while the pre-treatment data was normally distributed across all six markers, post-treatment data developed different distributions. In particular, markers 1 and 2 formed a bimodal distribution, and marker 0 formed a skewed distribution. This suggested the emergence of two distinct patient groups following treatment, which appeared to correlate with different responses to the drug.

Principal Component Analysis was used to reduced the complexity of the data, and revealed that 82% of the dataset's variance could be explained by the first two principal components. When these components were plotted against each other, they separated into two distinct clusters, with a Gaussian Mixture Model used to calculate a quantitative value for the number of patients in each cluster. 'Patient zero', who showed a significant improvement after treatment, was used to determine which cluster represented patients who improved and which patients did not. This was used to find the value for the percentage of patients who improved as 71.55%, showing that a majority of patients responded positively to the treatment.

Further analysis of the loadings in PCA indicated that markers 0, 1, and 2 were the most influential in defining the separation between clusters, with markers 3, 4, and 5 having minimal impact. These findings imply that specific markers were more accurate indicators of the response of patients to the treatment.

In conclusion, the findings indicate that TheraTech's drug provides an improvement in the majority of patients. However, to get a more accurate understanding of the impact of the drug on the patients, more data should be taken. In particular, it would be useful if the improvement of each patient was provided, rather than just the information that 'patient zero' improved significantly. Relying on a single patient as the basis for all the results introduces significant risks, as this patient could be a statistical outlier, potentially placed within the wrong cluster.

REFERENCES

[1] *Cancer survival statistics*, `https://www.cancerresearchuk.org/health-professional/cancer-statistics/survival`, (accessed 25/10/2024).

[2] *Death registration summary statistics, England and Wales: 2022*, `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathregistrationsummarystatisticsenglandandwales/2022`, (accessed 25/10/2024).

[3] I. Tothill, *Seminars in Cell Developmental Biology*, 2009, **20**, 55-62.

[4] S, Wold, K. Esbensen, P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37-52.