Assignment 1

Statistical Learning

Felix Wente - 3337731

November 10, 2022

Part A. Supervised learning

Question 1

(Lasso) Regression is an example of a parametric approach because it assumes a linear functional form for $f(X)$. This strong assumption about its functional form is one of the disadvantages of parametric approaches. The model is not able to capture the underlying relationship between X and y well if the underlying relationship is nonlinear. When looking at the data generating function we can see that there is a nonlinear relationship between the independent variables x_1,x_2,x_3 and the outcome variable y. Thus, a non-parametric model, such as a k-nearest neighbour classifier would fare better with a non-linear relationship as it doesn't have as strong of a bias when assuming the underlying relationship between independent and dependent variable. Given that we only have 3 additional variables in our model which add mere noise, we can assume that the variance of the knn classifier not to be too large and, therefore, will likely have a lowerer error (higher accuracy) than the lasso model as the bias of the lasso model is too large for it to capture the nonlinear relationship between the independent and dependent variable.

Question 2

When including all of the additional noise variables, the relative effectiveness of both models changes. The strength of the non-parametric approach (knn), its flexibility, becomes now its weakness. This is because the knn classifier will have a hard time, due to its flexibility, distinguishing between what is a nonlinear relationship between variables and what is noise. The optimal k chosen by the 10-fold cross validation will likely be a lot higher than in the model with fewer noise variables, to increase the bias of the model. Yet the variance of the model will still be higher than that of the lasso model. While the lasso model will still not be able to fully capture the non linear relationship, it has a much stronger bias and is better at ignoring the noise produced by the irrelevant variables leading to the lasso model having a higher accuracy relative to the knn model.

Question 3

We first create a subselection of the training data to account only for the first 6 variables. Next we normalise these variables to all have a range between 0 and 1. Distance based algorithms such as the KNN are most affected by the range of features. This is because they are using (euclidean) distances between data points to determine their similarity. Next, we run a ten-fold cross validation on the training data for values of k ranging from 1 to 100 to determine the optimal value of k. More precisely this means we split the training data into 10 parts and train the k-nearest neighbour algorithm on 9 parts of the data and test it on the 10th part (validation set). We have 10 repetitions where all 10 individual data parts are the validation set at one point and average the classification accuracy of the classifier on the validation set over all 10 repetitions. This is done for all 100 values of k to determine the best value of k. As can be seen from Figure 1 the optimal k chosen by the 10-fold CV is 21.

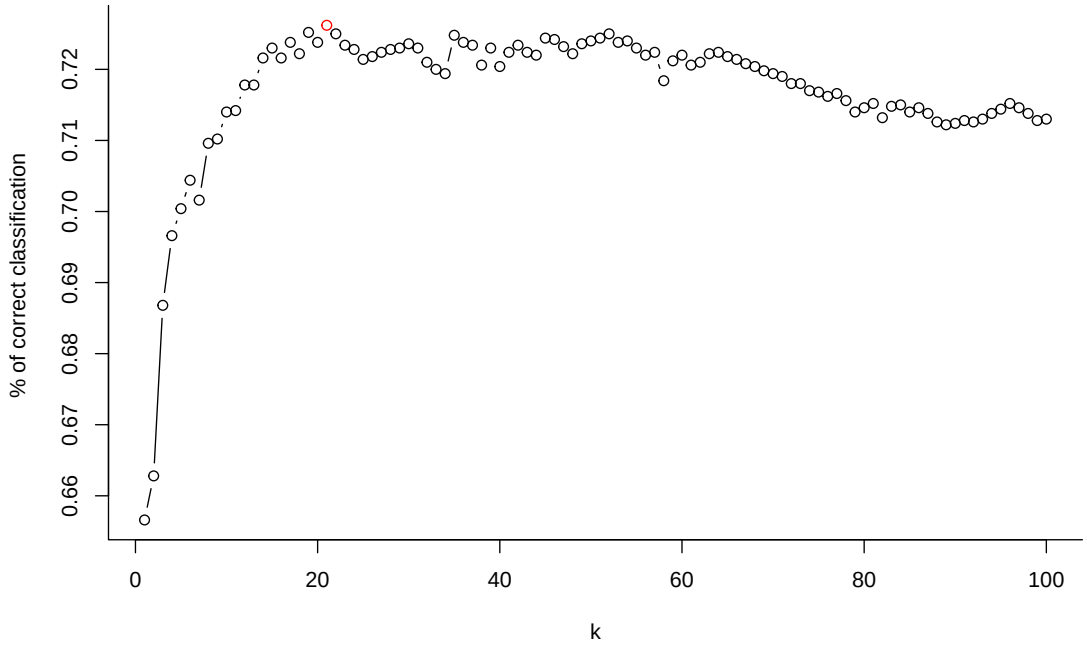

Figure 1: 10-Fold CV accuracies of knn model for given value of k

The optimal k is then used in the final model to predict data from the test input data which are then compared to the test outcome data. The measurement used to compare the performance of the two final models (knn and lasso) on the test data is the percentage of correctly classified outcome variables (1-test misclassification error rate) , i.e., the accuracy of the classifiers.

For the lasso model the procedure was identical. The data was first normalised since the penalties in the lasso model are a function of the size of the coefficients, which depends on the scale of the data. After that a 10-Fold CV is applied to determine the optimal λ. As can be seen from Figure 2 the optimal λ is 0.005057 ($\log(\lambda) = -5.29$). It can be seen that at this level of λ 3 variables would be removed from the model (variable selection).

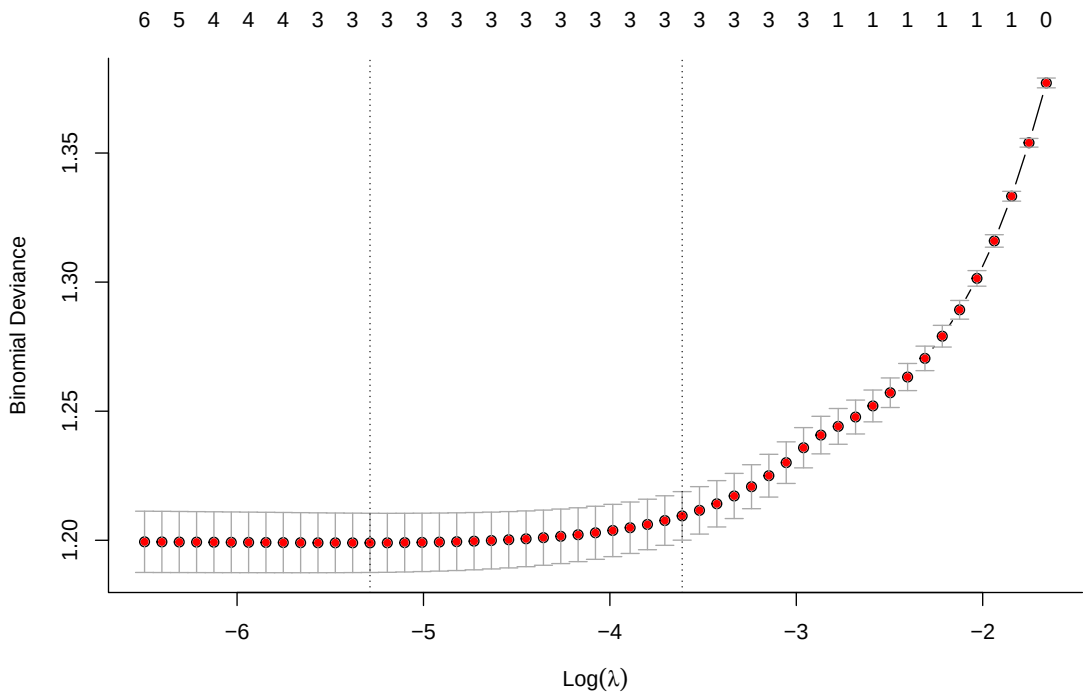

Figure 2: 10-Fold CV Binomial Deviance of Lasso model for given value of λ

The optimal λ is then used in the final model to predict data from the test input data which are then compared to the test outcome data. Again, measurement used to compare the performance of the two final models on the test data is the the accuracy of the classifiers.

The results in Table 1 confirm the expectations of Question 1 that the knn model will do better given its ability to better assume the non linear relationship between the independent variables x_1,x_2,x_3 and the outcome variable y.

Table 1: Test accuracies of models with independent variables x_1-x_6

Model	Accuracy
Knn	0.7218
Lasso	0.6774

Question 4

The procedures for question 3 are identical to question 2, however this time the design matrix includes the variables x_1 - x_{203}. This changed the metaparameters k and λ chosen by the 10-Fold CV as well as the resulting accuracies of the two models.
As expected, the value k choosen by the 10-Fold CV is a lot higher when more noise variables are included, to increase the bias of the model to offset the increased variance. The optimal value k is now at 92 (see also Figure 3).

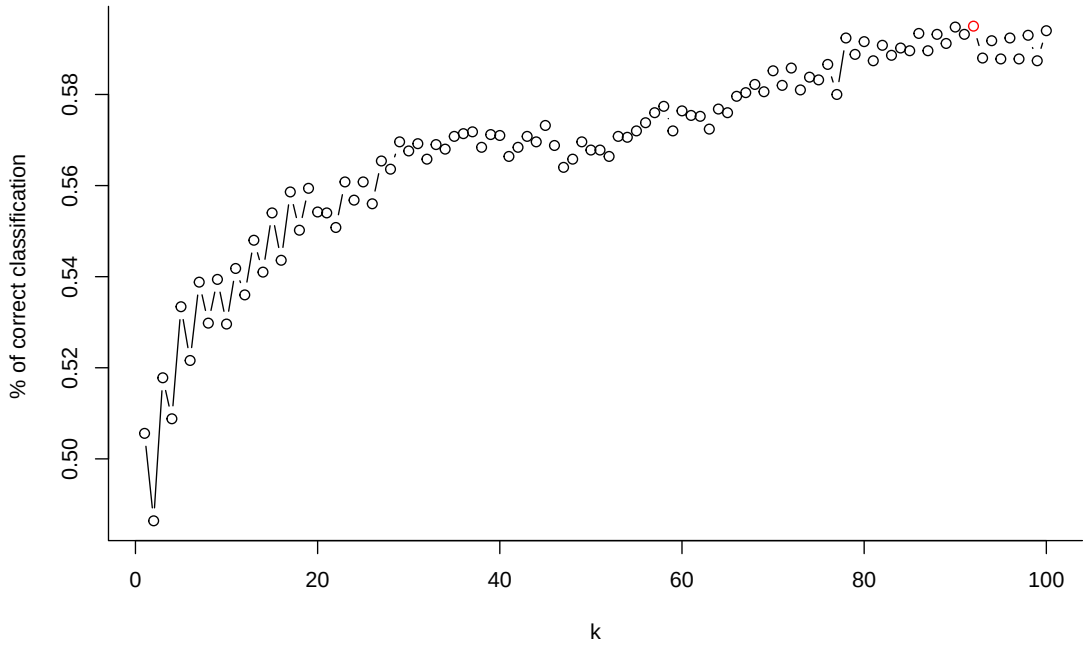

Figure 3: 10-Fold CV accuracies of knn model for given value of k

For the metaparameter λ of the Lasso model the 10-Fold CV has also chosen a higher value of 0.0128 ($\log(\lambda) = -4.357$), as seen in Figure 4. This was to be expected, since the a higher value of λ corresponds to more variables being dropped from the model, which is favourable as there are now more noise variables included in the model. At this λ 15 variables are being dropped from the model.

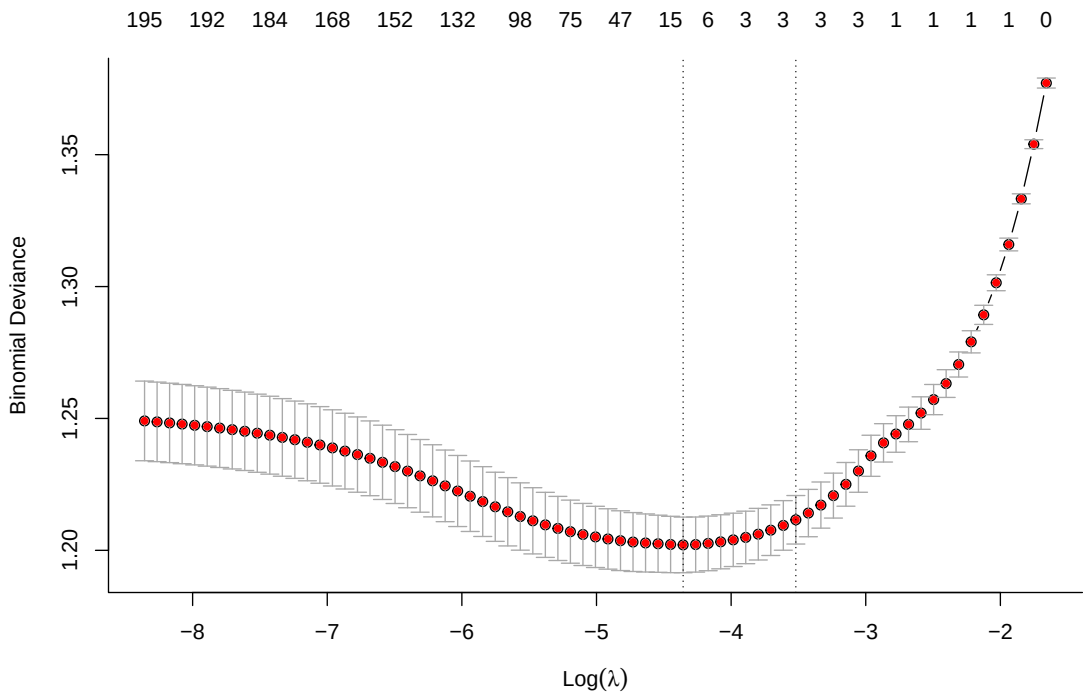

Figure 4: 10-Fold CV Binomial Deviance of Lasso model for given value of λ

As hypothesized in Question 2 the knn model decreased in accuracy as it included the noise in its estimation of the underlying relationship. The higher bias induced by a higher k was not sufficient to counteract the increase in variance of the model caused by over-fitting (subsuming the noise). The accuracy went from 0.7218 (Table 1) to 0.585 (Table 2).

The Lasso model did not change much in accuracy. It still was not able to sufficiently capture the underlying between the independent variables x_1,x_2,x_3 and the outcome variable y. However,as expected, it was successful at separating the noise from the relevant data by dropping (and reducing the coeficcients) of irrelevant variables. The accuracy went from 0.6774 (Table 1) to 0.6796 (Table 2).

Table 2: Test accuracies of models with independent variables x_1-x_{203}

Model	Accuracy
Knn	0.585
Lasso	0.6796

Part B. Unsupervised learning

Question 1

Principal component analysis linearly recombines the existing variables into a new set of uncorrelated variables, while trying to maximise the variance contained in the original data. It groups correlated variables together by giving related variables high weights (loadings) when linearly recombining them while giving lower or negative weights to the ones that are not correlated with these variables. This analysis is appropriate to the interest of the therapists as it will create new variables (dimensions) of similar coping mechanisms on which the patients can be compared and grouped on.
Table 3 shows the proportion of variance of the original data contained within the new created variables.

Table 3: PCA Variance

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.5006	0.9992	0.8391	0.4897	0.4692	0.4280	0.3954	0.3770	0.3206
Proportion of Variance	0.6948	0.1109	0.0782	0.0267	0.0245	0.0204	0.0174	0.0158	0.0114
Cumulative Proportion	0.6948	0.8057	0.8839	0.9106	0.9351	0.9554	0.9728	0.9886	1.0000

The two most common methods when choosing how many of the newly created variables to retain are (1) the Kaiser criterion and (2) observing the elbow of the scree plot. (1) states that variables with an eigenvalue (standard deviation) of equal or larger two 1 should be retained. According to this the first 2 (potentially 3) principal components (PCs) should be retained (Table 3). (2) states that one to the left of the elbow in the scree plot is the right number of PCs to retain. According to this, again the first 2 (potentially 3) PCs should be retained (Figure 5).[1]

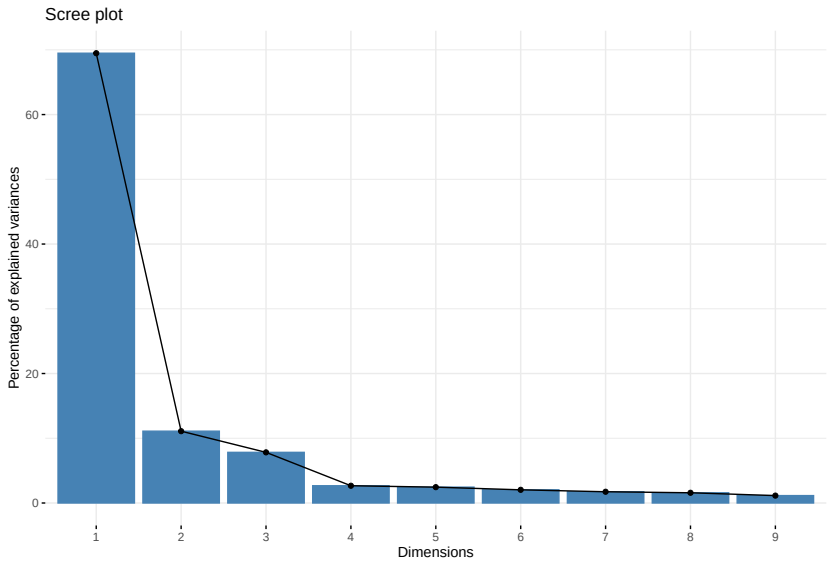

Figure 5: PCA Scree Plot

[1]The analysis of part B has been carried out with both 2 and 3 PCs. As interpretations of biplot and PC loadings are easier in two dimensions only the final clusters in Question 2 will be shown in both 2 and 3 dimensions as not to overload the report with figures.

Table 4: PCA Loadings

	PC1	PC2	PC3
V1	0.37	-0.06	0.12
V2	-0.38	-0.15	-0.04
V3	0.22	-0.29	-0.92
V4	-0.37	-0.19	-0.05
V5	-0.13	0.90	-0.32
V6	-0.36	-0.19	-0.06
V7	-0.36	-0.11	-0.07
V8	0.36	-0.02	-0.02
V9	0.36	-0.05	0.13

In Table 4 we can see the weights the original variables have in the creation of the new variables. Since PC1 includes both high negative and positive weights in the linear recombination of the original variables, the interpretation is easier when looking at the Biplot (Figure 6). We can see that Dimension 1 (PC1) separates the patients by:(1) coping styles (var2) accept, (var4) concentrate on other positive things, (var6) positive reinterpretation, (var7) put in perspective and (2) (var1) self-blame, (var3) rumination, (var8) dramatize, and (var9) blaming others. Dimension 2, on the other hand, seperates the pations into copying by (var5) focusing on planning or not. With those two Dimension we are able to retain 81% of the variation in the original data.

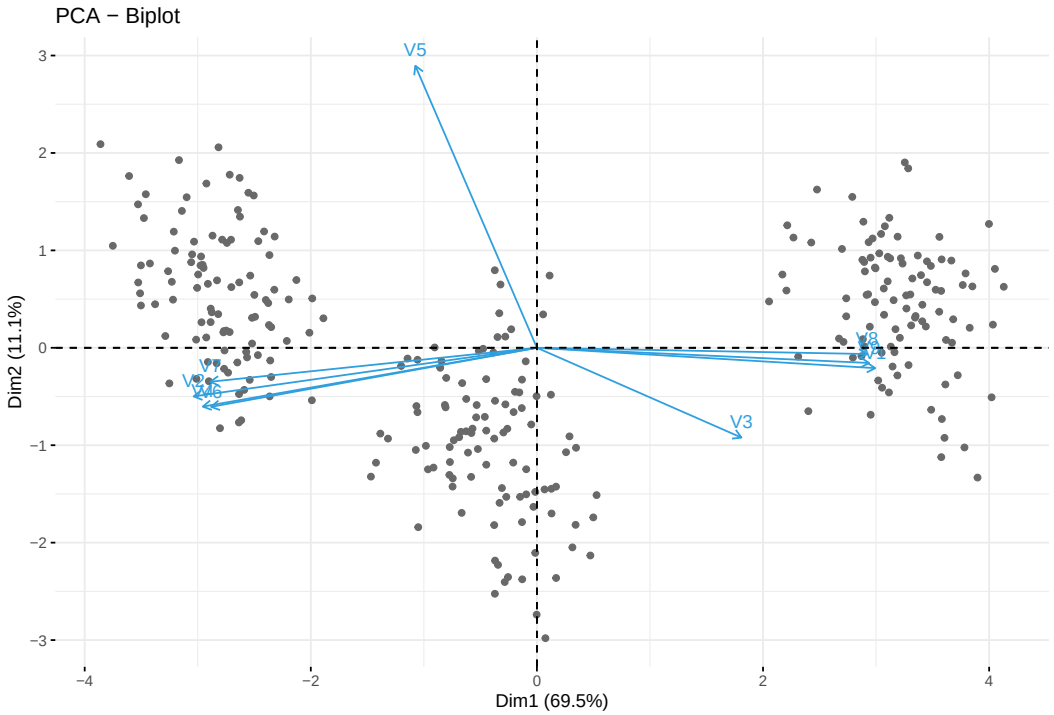

Figure 6: PCA Biplot

Question 2

In order to cluster the patients along the newly created dimensions we can use either hierarchical or k-means clustering, since we have a two dimensional data space in which distances can be computed and both are distance based clustering methods, both methods would be appropriate. For the two methods we have to decide the number of clusters we expect to see in the data since, unlike in supervised learning, we do not have a outcome variable that determines which group participants belong to. By looking at the Biplot from Question 1 (Figure 6) it seems likely to assume that we should choose 3 clusters. To confirm this we look at the gap statistic [1] for a number of cluster up to 10. The intuition behind the gap statistic is that while within cluster dissimilarities (W_k) continuously decreases with increasing number of cluster, when one cuts a natural group, within which the observations are all quite close to each other, W_k reduces less than partitioning the union of two well-separated groups into their proper constituents. The gap statistic contrasts the curve $log(W_k)$ to the curve of uniformly distributed data over a rectangle containing the data. It estimates the optimal number of clusters to be when the gap between the two curves is largest.

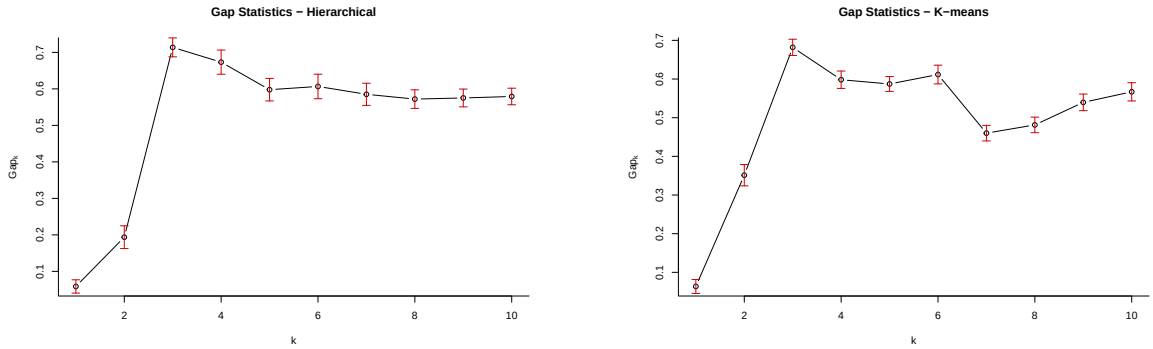

Figure 7: Gap Statistic hierarchical k-means clustering

As seen in Figure 7, the gap statistic for both hierarchical and k-means clustering confirms that the optimal number of clusters is in fact 3. Thus, k-means clustering with $k = 3$ has been performed to assign the patients to one of the three groups. The result of this can be seen in Figure 8. Cluster 1 (top picture Figure 8) has 100 patients and they can be categorized as the group of patients that does not cope by focusing on planing. Cluster 2 has 101 patients and can be categorized as the group of patients that copes predominantly by self-blame, rumination, dramatization, and blaming others. Lastly, cluster 3 has 99 patients and they cope by accepting, concentrating on other positive things, positive reinterpretation and putting things in perspective.

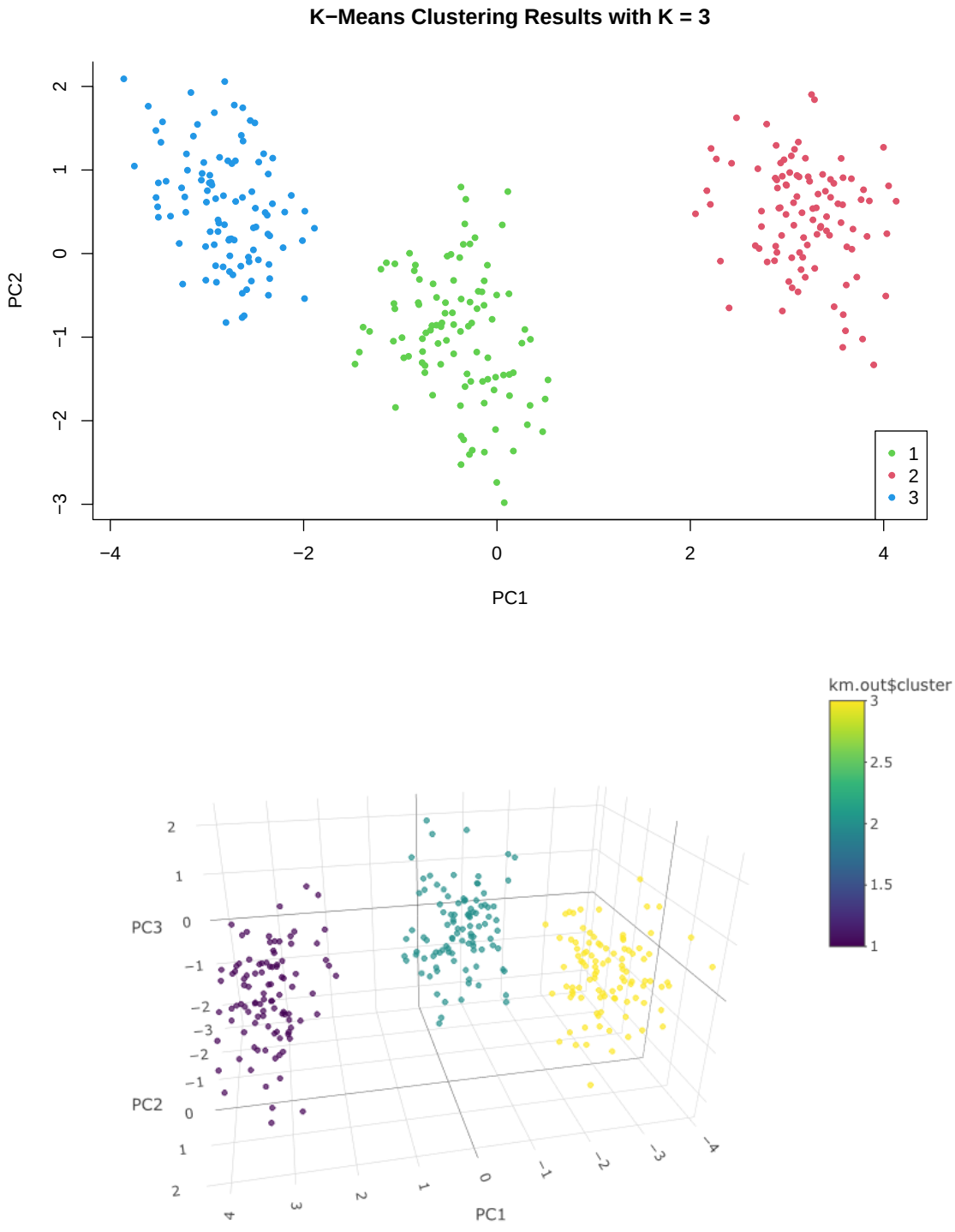

Figure 8: Cluster of patients 2D (top) and 3D (bottom)

References

[1] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.