SYDDANSK UNIVERSITET

Data Preprocessing

Bjarki Sigurdsson
Abdulrahman Abdulrahim
Miguel de la Colina

March 21, 2017

ABSTRACT

THE PURPOSE OF THE EXERCISE WAS TO EXAMINE THE EFFECTS OF PREPROCESSING ON kNN CLASSIFICATION. WE SPECIFICALLY TESTED PRINCIPAL COMPONENT ANALYSIS (PCA) AND MIN-MAX NORMALIZATION. OTHER PREPROCESSING METHODS USED WERE THOSE PROVIDED IN THE PREVIOUS EXERCISE BY THE `LOADIMAGE.R` SCRIPT, NAMELY IMAGE SMOOTHING AND DPI DOWNSAMPLING.

1 PCA

For this part of the exercise we used PCA to reduce the dimensionality of the data and observed the effect on kNN classification of using the first principal components resembling 80, 90, 95 and 99% of the total variance. We tested for classification accuracy and computation times for varying k.

1.1 SINGLE-PERSON DATA

Information on the first 10 principal components, in order, is shown in table 1. These correspond to the data from member 1 in group 5. This is how we determine the breakpoints for 80, 90, 95 and 99% of the total variance. They are 13, 23, 36 and 77, respectively. This is also illustrated in figure 1.

Table 1: The first 10 principal components.

PC#	Standard Deviation	Proportion of Variance	Cumulative PV
1	.7646	.1876	.1876
2	.6878	.1518	.3994
3	.6365	.1300	.4695
4	.4804	.0741	.5435
5	.4290	.0591	.6026
6	.3723	.0445	.6471
7	.3162	.0321	.6791
8	.2864	.0263	.7055
9	.2689	.0232	.7287
10	.2662	.0228	.7514

Figure 2 shows the results from setting k to 50, varying the number of principal components according to the breakpoints mentioned and measuring the average execution time of the kNN classification step. We see that there is a linear relationship between the two.

Figure 3 shows the classification accuracy for varying numbers of principal components and varying k. We see that the behaviour is similar for varying k and that the effect of using more principal components is negligible above 90% of total variance.

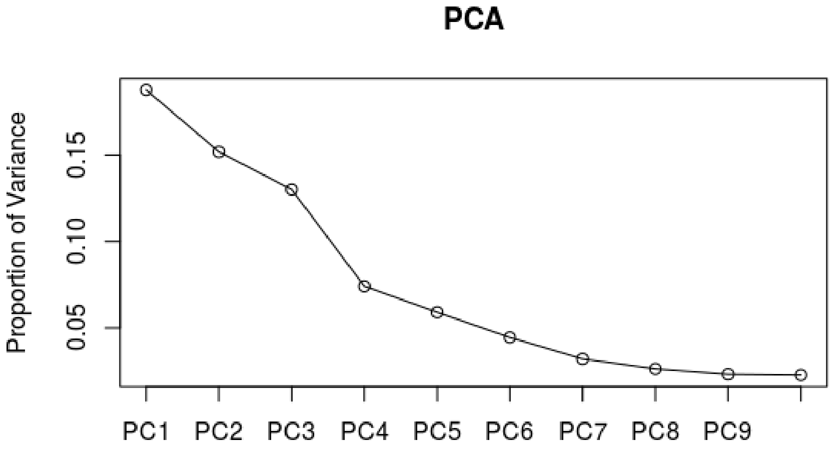


Figure 1: Scree-plot of the first 10 principal components.

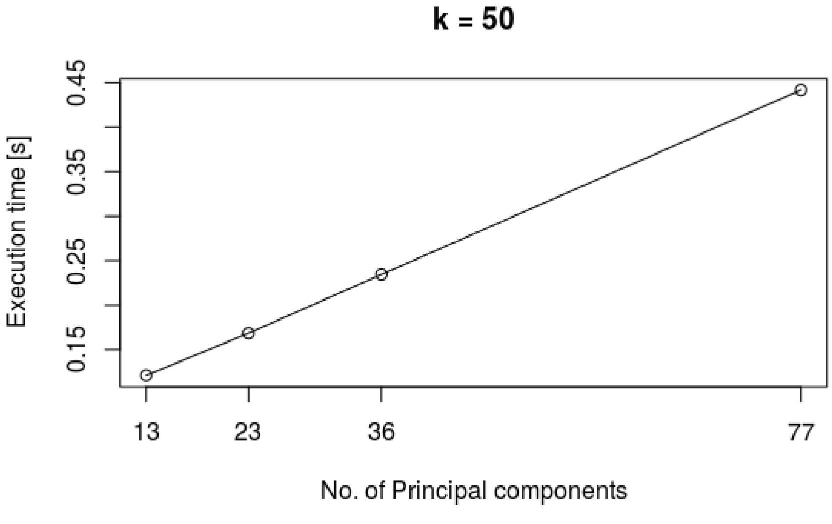


Figure 2: Timing results with a varying number of principal components.

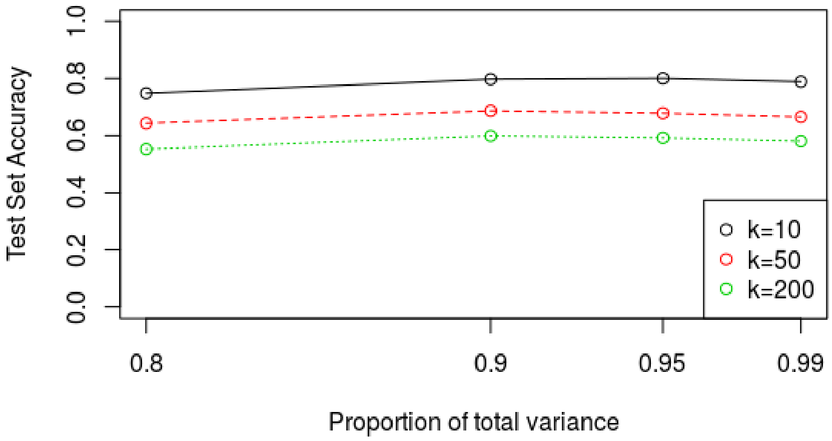


Figure 3: Classification accuracy results for varying k and Proportion of Variance.

1.2 Multi-person data

Information on the first 10 principal components, in order, is shown in table 2. In this case the breakpoints are 20, 31, 45 and 89 for 80, 90, 95 and 99%, respectively. This is illustrated in figure 4.

Table 2: The first 10 principal components.

PC#	Standard Deviation	Proportion of Variance	Cumulative PV
1	.8013	.1336	.1336
2	.7457	.1157	.2493
3	.5915	.0728	.3220
4	.5542	.0639	.3859
5	.5369	.0600	.4459
6	.4652	.0450	.4909
7	.4434	.0409	.5318
8	.4335	.0391	.5709
9	.3806	.0301	.6010
10	.3757	.0294	.6304

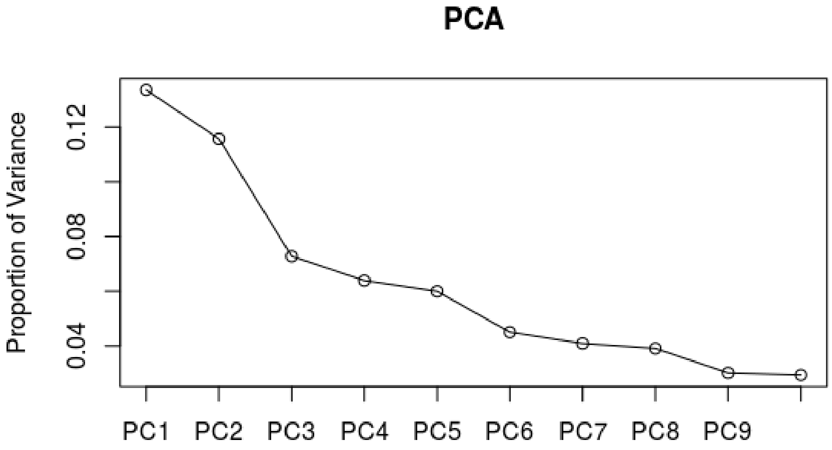


Figure 4: Scree-plot of the first 10 principal components.

Figure 5 shows the results from setting k to 50, varying the number of principal components according to the breakpoints mentioned and measuring the average execution time of the kNN classification step. Again we see that there is a linear relationship between the two. This time around the timings were less accurate due to time constraints and the size of the dataset.

Figure 6 shows the classification accuracy for varying numbers of principal components

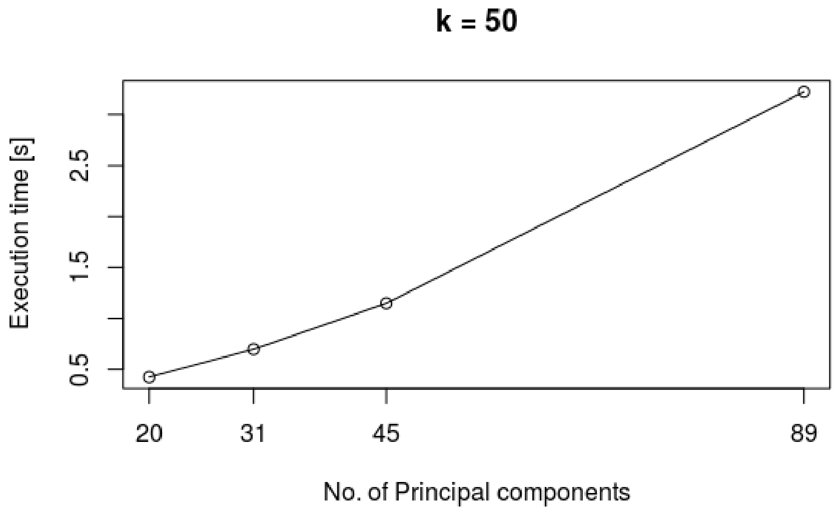


Figure 5: Timing results with a varying number of principal components.

and varying k. This is similar to what we saw for the single-person dataset.

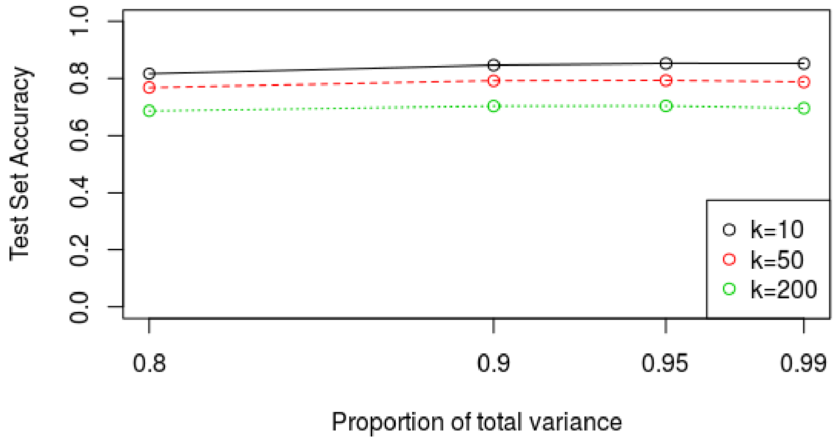


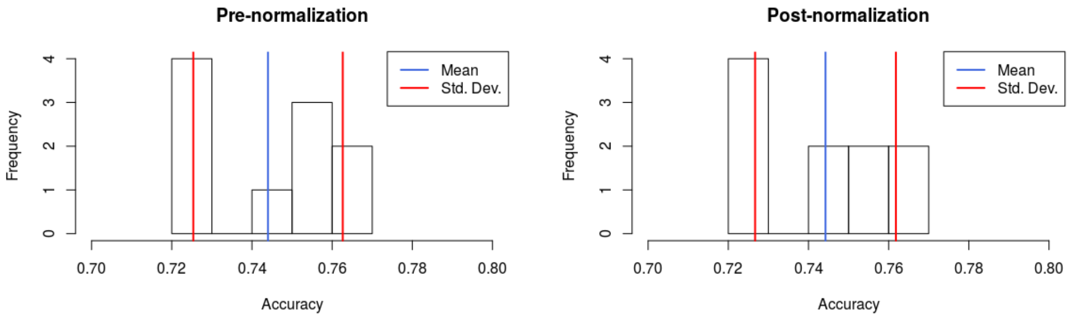
Figure 6: Classification accuracy results for varying k and Proportion of Variance.

2 Normalization

In this section we analyze the effects of applying Min-Max Normalization before the PCA step (pre-normalization) or after (post-normalization). To this end, we perform a cross-validation of the kNN classifier and compare the results.

2.1 Single-person data

Figure 7 shows the results of cross-validation on pre-normalized and post-normalized data. We see that the results are near-identical. This is logical as PCA depends on the eigenvalue decomposition of the covariance matrix and is therefore not significantly altered by normalization via robust methods.



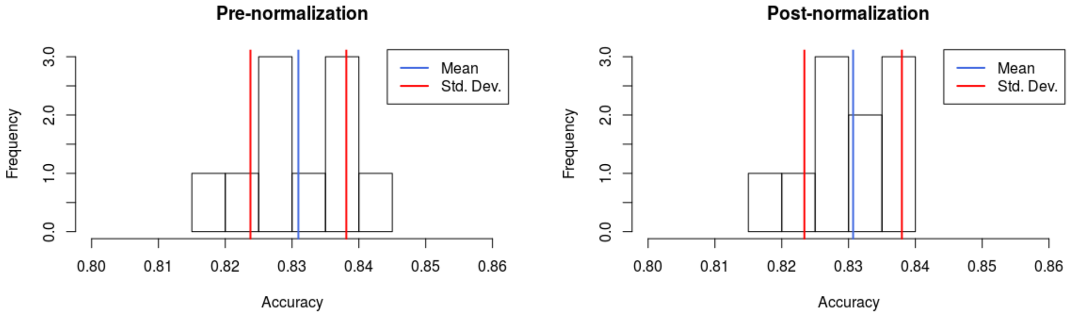
(a) Pre-normalization

(b) Post-normalization

Figure 7: Histogram of cross-validation results with k=50 using 36 principal components (95% of total variance).

2.2 MULTI-PERSON DATA

Figure 7 shows the results of cross-validation on pre-normalized and post-normalized data. Again the results are near-identical.



(a) Pre-normalization

(b) Post-normalization

Figure 8: Histogram of cross-validation results with k=50 using 45 principal components (95% of total variance).