Technical University of Denmark

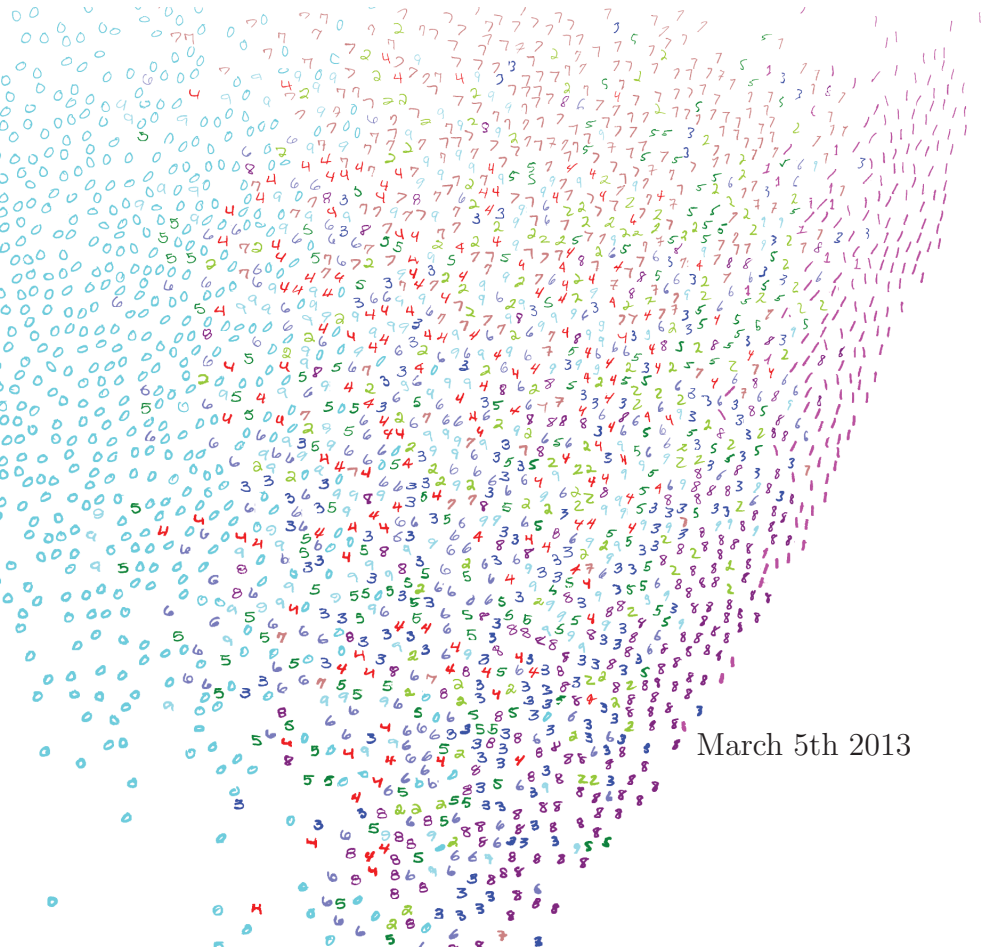
02450 Introduction to Machine Learning and Data Modeling

Assignment 1

Feature Extraction and Visualization

Poul Kjeldager Sørensen
s093294

Martin Kasban Tange
s093280



March 5th 2013

Contents

1 Introduction

This report is made in regard to the course 02450 Introduction to Machine Learning and Data Modeling at DTU (Technical University of Denmark). In this first assignment we are to pick a data set which we can analyse and extract some feature from as well as try to visualize the things we learn about it.

We have chosen a data set containing a lot of hand-written digits which have been digitalized such that each digit can represented by a vector of pixel values. From this we have performed some manipulation and extracted certain features we found to be wide and conclusive enough to be able to do the machine learning we would like to do.

This report goes over which attributes we extract from the raw pixel values, as well as how these can help us tell the different classes in the data set apart. We look at the standard deviation for each class along with the correlation between the classes to get a better picture on how they interact.

Upon doing PCA (Principal Component Analysis) we can get a new look at the data, and determine how many PCs (Principal Components) we need to have a good chance of distinguishing each class from the others. Moreover we can also see how different combinations of PCs are good at telling different classes apart.

All the work carried out in this report are done by Tange, M. K. and Sørensen, P. K. and the code can be found on github[1]. Do note that the code have only been used to generate plots and illustrations for the report and some playing around. There have been done no effort into readability or re-usability of the code.

[1] https://github.com/mktange/IntroMachineLearning

2 Data and Feature Extraction

2.1 Introduction

In this chapter the data are presented and how the features have been extracted are outlined.

2.2 Feature Extraction

In this section it will be explained how the original MNIST data set have been processed to create a feature set other than the raw pixel values. First the original data is outlined.

2.2.1 MNIST Data set

The MNIST Data set consist of 60000 examples of handwritten digits for training and 10000 for testing. It will be shown later how pixels in such images are highly correlated and the real dimension of the data is much lower using PCA analysis. This data set have been widely used in the past and are still used as benchmarks for new algorithms. Some interesting publications have been published the past few years bringing the records down to an error rate of 0.23 percent[1].

The records that been published over the last years are mostly neural networks / convolutional nets and while we at some point in this course will be introduced to these topics, our goal are not to beat this guys who have been working in the field for many years. Instead we want to use this data set because its interesting and something we can relate to.

[1] http://yann.lecun.com/exdb/mnist/

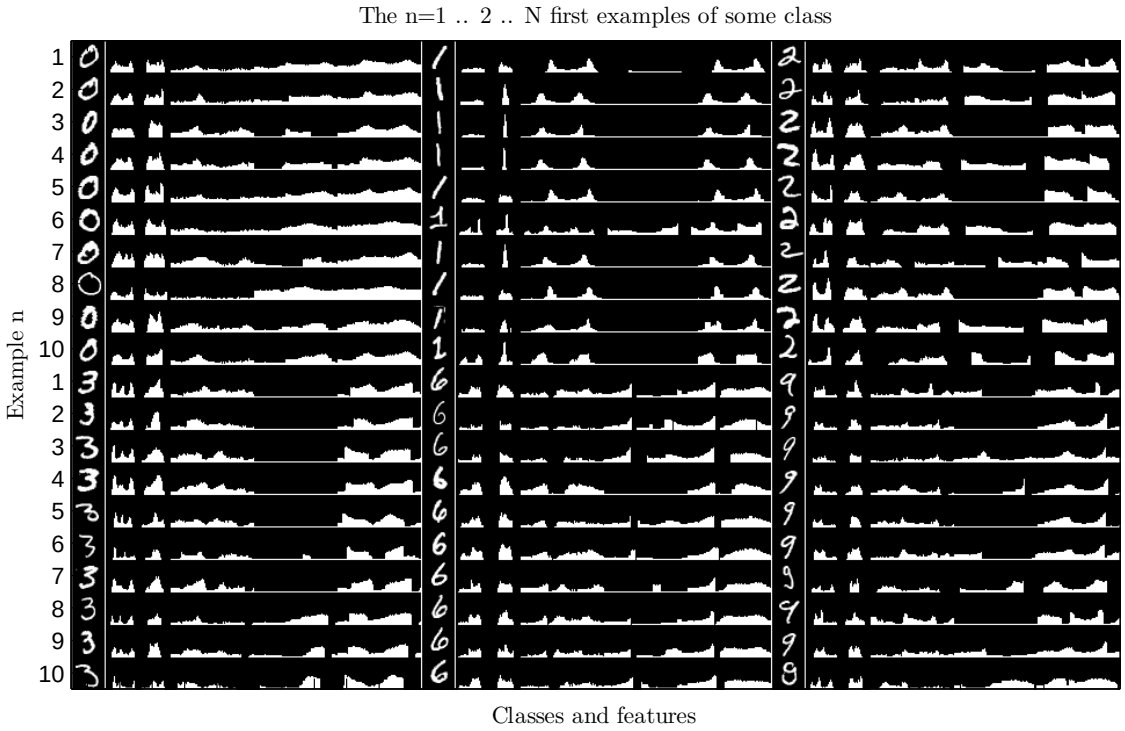


Figure 2.1: 10 Examples of six classes in the data set and the extracted features. Features consist first of vertical[28], horizontal[28] and radial[72] histograms and in-out[72], out-in[72] profiles to a total of 272 features.

In Figure 2.1 ten examples of six picked classes are shown and a lot of variance can be seen for each individual class when looking at the image and some patterns may be detected in the extracted features. The goal is then to get the machine learning algorithm to find the patterns and separate them based on the label information.

2.2.2 Extraction

Feature extraction have been done by computing Structural Characteristics of each 28 by 28 pixel image from MNIST as outlined in [1]. This process extracts a vertical, horizontal and radial histogram concatenated with an in-out and out-in profile.

In Figure 2.1 the features can be seen next to its input image. The radial histogram and profiles are computed as 72 steps of 5 degrees. The profiles are the distance to the first and last on-pixel from the center and out.

This means that all of our attributes are discrete and ratio, since the absence of value in an attributes actually means that there are no pixels in the set of pixels, and the fact that the distance between attribute values are comparable by factors (ratio).

2.3 Machine Learning

One of the tasks for this assignment was to find a data set that could be used through the course for all of the coming machine learning tasks. Following is a short explanation on how we could approach each of these techniques for the future reports.

The main task of this data set is to classify the correct digit label to all of the 10.000 test digits by training a mapping function to go from our input features to a specific class label of zero to

nine.

Regarding regression and associative mining then it could be interesting to find missing attributes. However, since our data set does not have any missing attributes, we could manually remove one or more and try to *re-find* it by these processes.

For anomaly detection it could be interesting to tell if a digit has been abnormally drawn compared to the *normal* or more standard way of drawing it. We could possibly create some different reflections of or changes to some numbers and see if we can detect those from the rest of the data set.

Clustering could be interesting to find within one digit class. There are a lot of different ways to write the same digit, and it could be quite interesting to cluster all the digits of one class to see if there there 1, 2 or K distinctly different ways to write that specific digit. Such analysis could also have been done in this report by doing PCA within a single class, which could give good visualization of the groups for it.

In our data set we have a lot of labelled data, but if that's not the case one could have used clustering for examination of unlabelled data to find information about the distributions of clusters. Then when applying the small amount of labelled data to these clusters one would be able to tell certain characteristics about the test examples close to those distributions.

It was mentioned earlier that publications of the past years are mostly neural networks on our data set and with those networks people have been able to take the raw pixel features and detect the features as the first step in the learning algorithm. This has mainly been referred to as deep learning when searching online and it would also be an interesting topic to dive into.

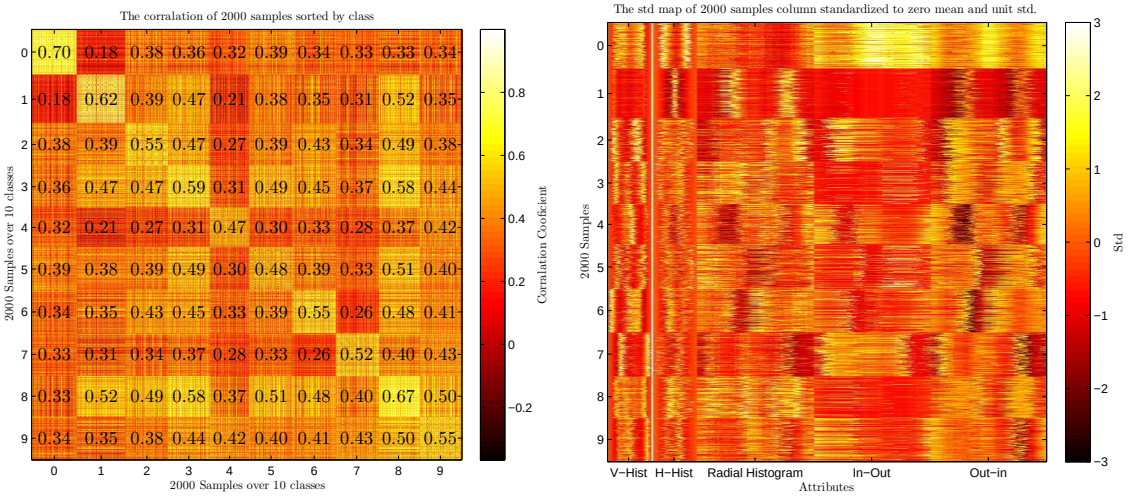
3 Data Analysis

3.1 Introduction

In this chapter a detailed explanation of the attributes will be given. It will be shown how the attributes are somewhat correlated and the real dimensionality of the data will be found using Principal Component Analysis (PCA).

3.2 Basic Statistic Properties

By means of the ACCENT[1] principles some interesting figures have been made in Matlab to illustrate some properties of the data. Some summary statistics have also been omitted as them not contributing any further information for this data set. An example of this is computing the mean and standard deviation(std). Instead the std has been computed from standardized data, for 2000 randomly drawn samples, and plotted in Figure 3.1b.



(a) The correlation of 2000 randomly drawn samples sorted by class. The mean of each class-by-class square have its mean value printed ontop.

(b) The std of all attributes from the same 2000 samples in Figure 3.1a.

Figure 3.1: Conclusion: Classes are clearly correlated and it should be possible to apply classification routines to separate the classes.

From Figure 3.1b some properties can be seen. One example would be the class of zeros in-out profile are significant higher than the average and classes of ones have its horizontal histogram below average near the middle and higher in the middle. Keeping this information in mind and returning to Figure 2.1 this becomes truth when looking at the features. The ten zeros shown earlier have a medium respond for the in-out profile while the five other classes have almost zeros. It also makes sense that ones have high response at the middle because of ones being vertical lines and nothing around them.

[1] Apprehension, clarity, consistency, efficiency, necessity and truthfulness

3.2.1 Correlation

In Figure 3.1a the correlation of samples have been computed and sorted by classes. Furthermore the mean value for each class by class square box have been computed and plotted on top for truthfulness when colours fails. It can been seen that classes has some correlation with the other classes, but important is that each class is mostly correlated with its own class. The fact that some classes do not correlate a lot to itself is likely based on the fact that people tend to write this digit in many different ways (i.e. the digit 4).

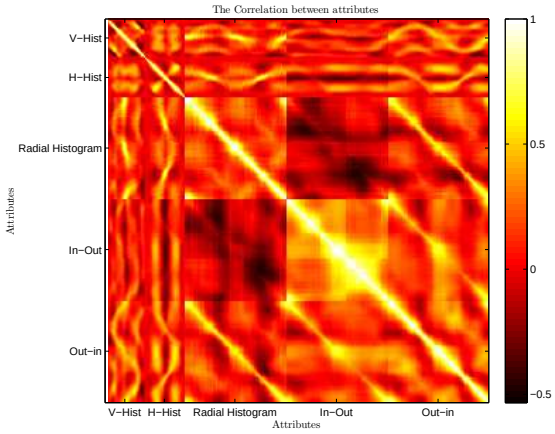


Figure 3.2: The correlation of attributes. Conclusion: Highly correlated and dimension can be reduced.

It can also be very interesting to look at correlation between attributes, as shown in Figure 3.2. It's seen that the attributes are highly correlated in parts of the plot and it also make sense when looking back on the feature generation. The histograms correlate and some kind of symmetry is shown in the figure. This could very well be explained by the fact that a lot of the digits actually also have symmetric parts. Instead of speculating in such relationships, a PCA will be used to find the real dimension of the data.

3.3 Principal Component Analysis

In Principal Component Analysis the goal is to represent N-dimensional data using less then N numbers. This is done by finding M orthogonal directions in which the data have the most variance and ignore the directions in where the data do not variance much. These M principal directions form a lower-dimensional subspace and a N-dimensional data point can be represented by its projections onto these M directions in the lower dimension subspace. The information about where the data is located in the remaining $N - M$ orthogonal directions are therefore lost. Since these do not have much variance, only a little information is lost.

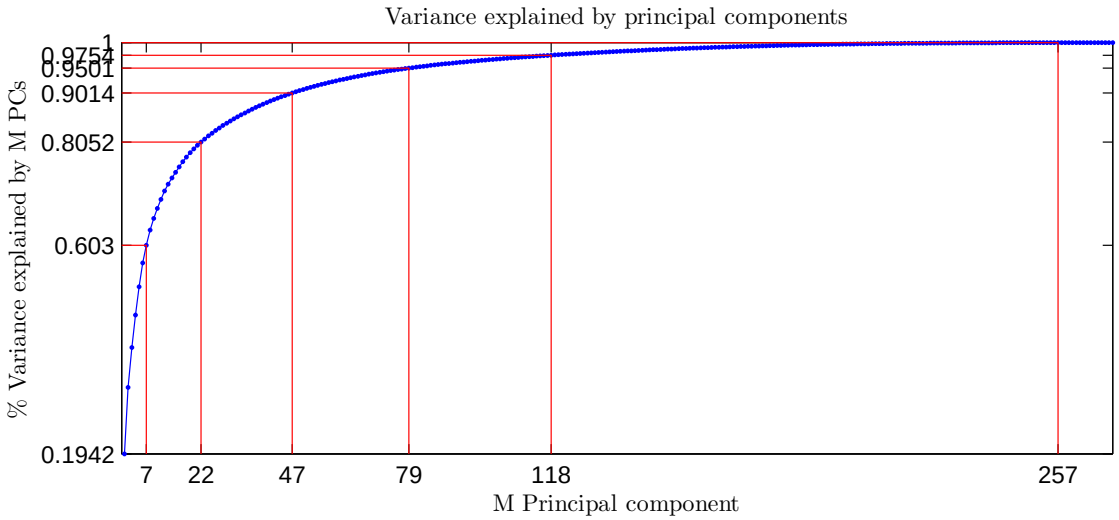


Figure 3.3: The variance explained by M Components.

3.3.1 Percent Variance Explained by M Components

In Figure 3.3 the variance explained by 1 to M principal components are shown where the explained percent for one component are shown as around 19 percent. Then the components needed to explain 60, 80, 90, 96, 97 and 100 percent are marked out on the x axis respective to the percent explained variance on the y axis.

3.3.2 Principal Components Directions

Showing the PCs directions have not been dealt with in this report. We argue that it does not make much sense to show the directions from the original high dimensional data set as each attribute does not have any exact meaning. A better example for such is when we have a data set of attributes with exact meanings like workload and salary. Then a principal component could be directed to watch the two dimensions indicating that people who work hard make more money.

3.3.3 Principal Component Projections

By projecting the data onto the principal components we can by hand separate the classes based on the projections. Our hypothesis is then, that there exist a combination of PCs that will separate most classes. This is to be shown in the next assignments when we learn about classification.

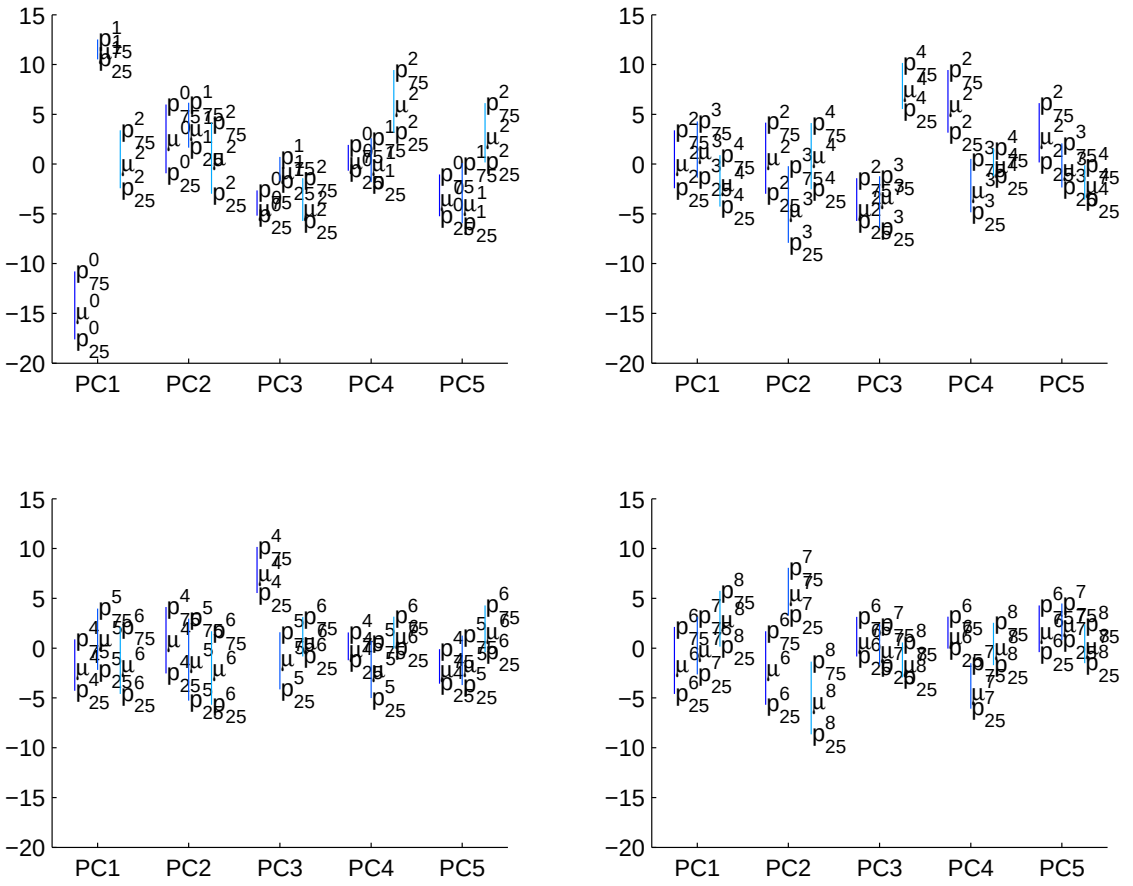


Figure 3.4: Here is nine of the ten classes shown as projections onto the principal components. Data are shown as median μ^{class} and $x = 25, 75$ percentiles as p_x^{class} for the first five components. For each PC three projections distributions are shown in each subplot.

In Figure 3.4 it shows that the first PC are able to distinct class $0, 1$ from each other along with all the other classes. PC three can distinct digit four. PC four can find digits of twos and sevens can be found using PC two. These are some examples that can be verified by human inspection. To classify further digits a linear combination of the projections will be needed and this is the real classification task if PCs was used as features. The problem is the same for our original features, its just not that trivial to separate them by hand as it was done here. Remembering how PCA have created linear combinations of the features such the most variance are in the first PC.

4 Discussion

We have learned from our data set and its features that they are a viable candidate for machine learning purposes and this course through different visualizations of deviation, correlation and PCA on the attributes.

We do not have very meaningful attributes when they stand by themselves, i.e. the amount of pixels in the third row is pretty unimportant negligible. However, together the attributes compliment each other well and makes it very possible to see good correlation within the classes as well as being able to distinguish them and their digits apart.

The data set has also shown to be quite viable for further analysis and for applying various machine learning techniques. We can already see ways of being able to classify most of the numbers with good reliability, and doing further modelling will only help in this regard.

It should also be noted that the data has been projected on to the PCs and plotted with scatter plots, but to comply with the ACCENT principles the projections have been visualized as in Figure 3.4 for easier interpretation. At the end we modified the scatter plot such that it draws the numbers instead of dots, and the more clear result can be seen in Figure 4.1.

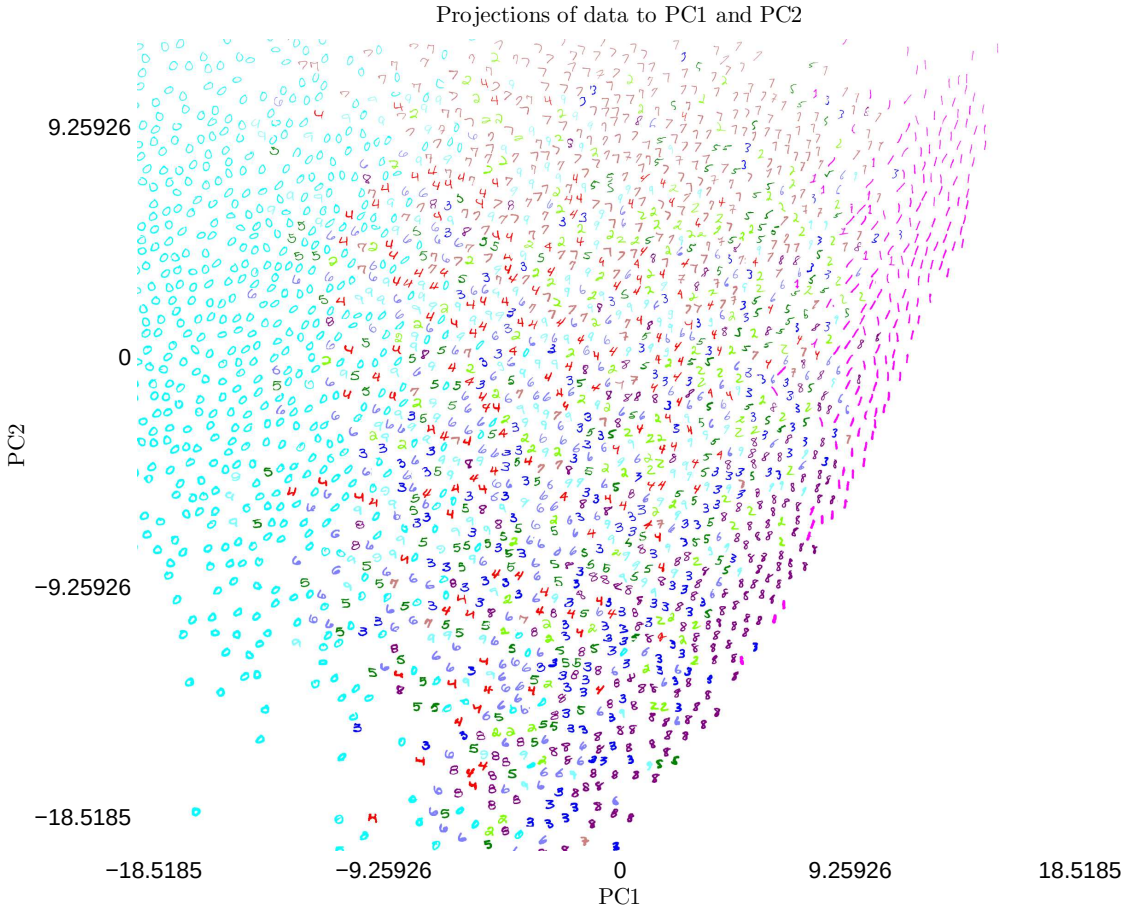


Figure 4.1: Equal amount of samples from each class was sampled, $N = 500$, and then drawn at random if and only if no other digit had been printed at the location in the map.

One should keep in mind that only the median and 25, 75 percentiles are illustrated and there will be some of the data overlapping. This was also why scatter plots hasn't been used, because a decision about how many samples to plot was needed. Too many would lead to a chaotic plot where some samples simply disappear behind others, and too few does not illustrate the spread well enough. Therefore computing summery statistics over the whole data set and showing the properties where preferred is the path we took. It was also experimented with showing mean and confident value given by three times the standard deviation for the distributions instead of the median.

For the PCA the data set was standardized to mean 0 and a single unit standard deviation as this generated more intuitive results to interpretive as humans. A machine learning algorithm should be able to solve it either way if the data had only been normalized to zero mean, so it works out.

4.1 Conclusion

To conclude this report, we will reiterate our answers to the questions raised in the assignment:

- **A description of your data set**
  We have found an interesting data set that is well explored within the machine learning area and there are great baselines to compare with. Furthermore we have given a brief outline about the features we have extracted. We have talked about what interesting tasks that could be applied to it, where classification will be the main task.

- **A detailed explanation about the attributes of the data**
  Together with the feature extraction there has also been given information about the many attributes. In this case, a single attribute does not hold much meaning alone, but combined with the other attributes we can get a wealth of information. This also means that not all summary statistics are interesting for individual attributes. Still we have shown the use of a lot of different statistics in the plots we shown, including median, correlation and therefore also mean and variance. Our data set is high dimensional and a big task for us is to illustrate this in a simple and clear way.

- **Data visualizations based on suitable visualization techniques**
  We have as noted above, used correlation to illustrate some properties of the data set. A PCA have also been carried out which gave some plots that makes it easier to understand the real big variances and clusters of the data set.

- **A discussion explaining what you have learned**
  We have given such discussion and further more experimented with some techniques that are not included as we didn't find the plots living up to the ACCENT principles.

All in all we think we have answered the question raised and think we found good solutions for the topics in this assignment.

Bibliography

[1] E. Kavallieratou, K. Sgarbas, N. Fakotakis, and G. Kokkinakis. Handwritten word recognition based on structural characteristics and lexical support. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 562 – 566 vol.1, aug. 2003.