1. Assignment 3

Principle Components Analysis

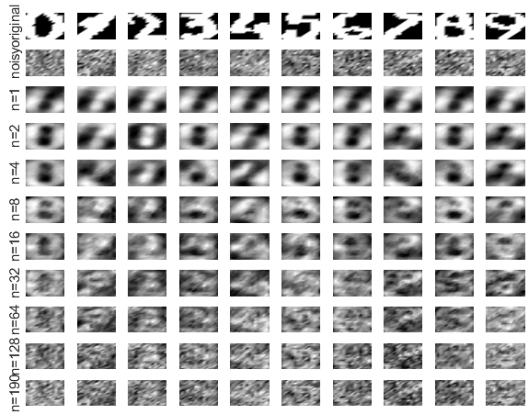
Principle components analysis is a technique which is used to extract useful features from the data. The input data does not contain any labels and the model has to learn from the data. Hence, it is an unsupervised technique to extract features out of the data. The takes of PCA is to perform dimensionality reduction and the being able to construct the original data points as accurately as possible.

This section focuses on denoising the input data by using linear and kernel PCA. Dataset for this section is an artificial dataset which resembles the yin yang pattern. We will inverstigate the impact of number of components on the denoising factor. Figure 1 shows the impact of increasing the number of components. From figure 1 to 6 we start getting better approximation but after that we are considering noise as denoised points. It is clear from the figure with 15 components that instead of getting smooth curve, we get approximation in independent points which are noise.

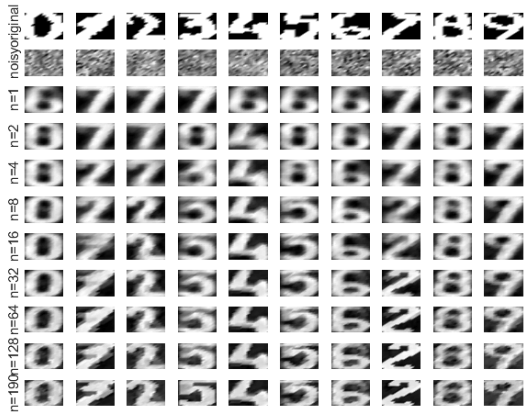
Kernel principle component analysis differs from the linear principle component analysis because it uses kernel induced feature space. The maximum number of principle components in linear pca are equal to the dimension while in kernel PCA, the number of principle components can be equal to the number of training data points.
We can tune the hyper-parameters by inserting all the(to be tuned) parameters in different lists and then try them in combination. As a loss function, one can try to minimize the reconstruction error. Any combination of parameters which gives us least reconstruction error, will be the optimal set of parameters.

Kernel PCA for images denoising

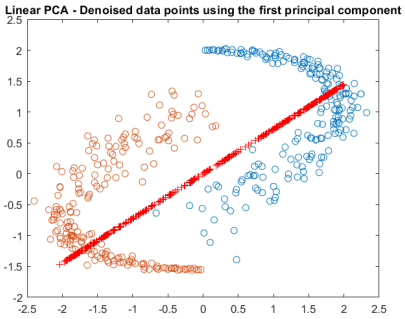
Following figure shows the difference between linear and kernel pca. Linear kernel is not able to approximate the complex shapes. It fails even with large number of principle components. On other hand, kernel PCA performs better as we increase the number of principle components. We have clear distinction between background and foreground. Increasing the number of components results in sharper images.



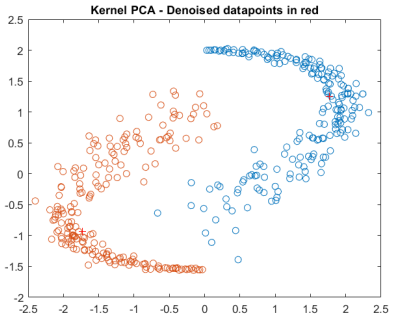
Linear PCA



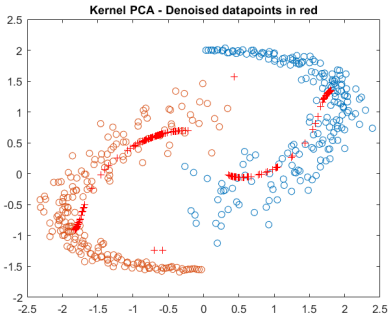
Kernel PCA



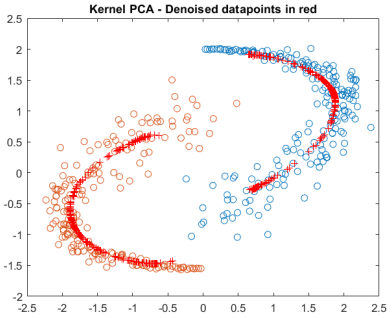
(a) Linear pca



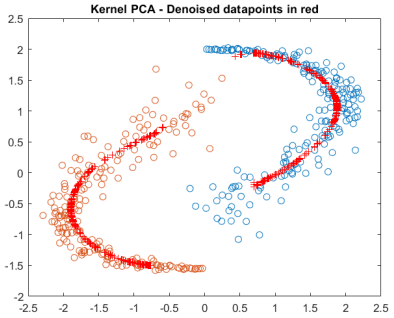
(b) Components = 1



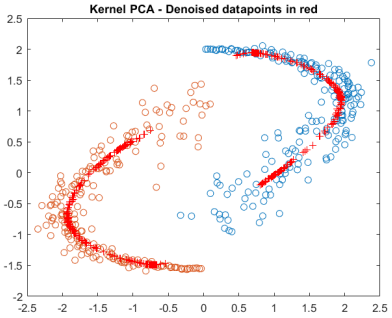
(c) Components = 2



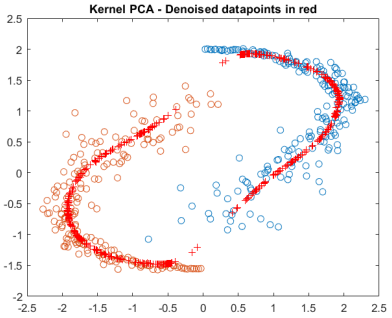
(d) Components = 3



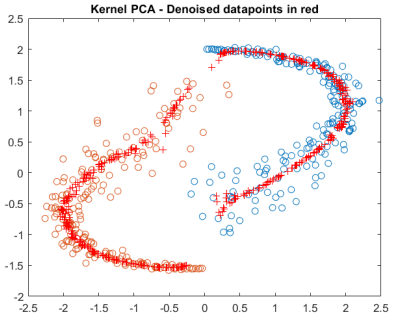
(e) Components = 4



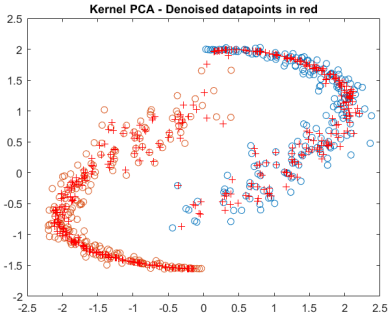
(f) Components = 5



(g) Components = 6



(h) Components = 10



(i) Components = 15

Figura 1: Impact of increasing the components

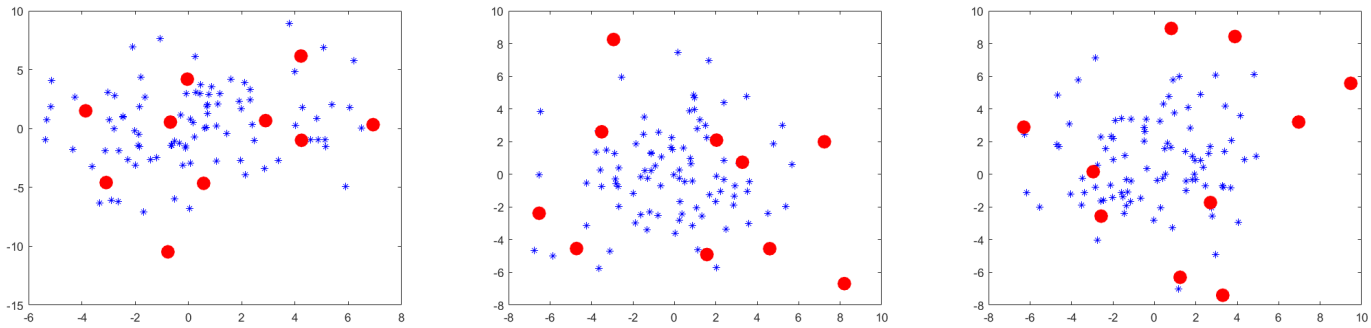
Now we investgate the effects of σ parameter. From our initial experiments, can be seen from figure 4 that inceasing the σ parameter, makes the images blurry while decreasing the sigma value causes the images to become more sharper. However, if we set the sigmafactor = 0,001 then we only get noise as result. No digit is recognizeable. Hence, σ must be chosen carefully.

Fixed size LS-SVM

In this section, we focus on large scale problems. When the dimension of the data is high then it is better to work in dual space but if the data set is huge then it is better to work in primal space. We prefer primal space because when we use a feature mapping(kernel trick) to transform the data in another space then the mapping can become infinite dimensional which results in w vector to of it's own size. In that case, for non-linear problems, we have to estimate the mapping points. Nystrom technique is one way to estimate such mapping and fixed sized ls-svm refers to that fact that we fix number of support vectors in advance.

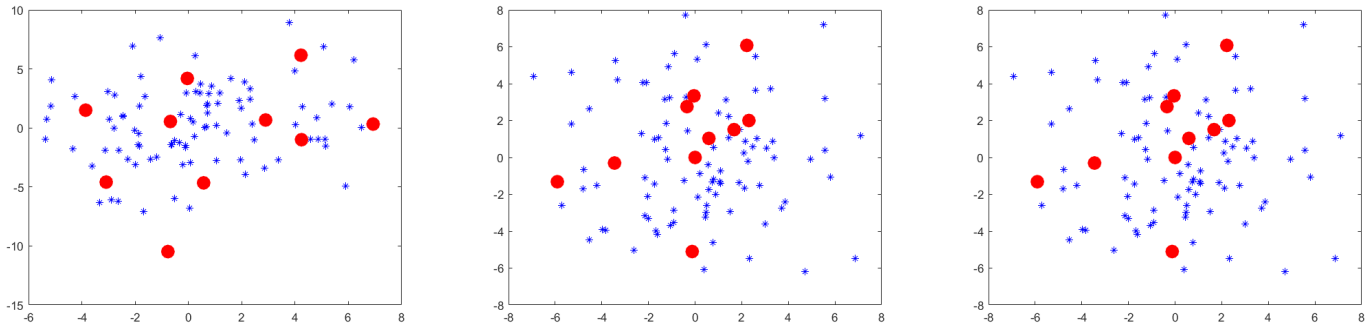
Nystrom choses the datapoints randomly which does not result in an optimal sample. We want to get the sample which gives us more information about the whole data. In this section, we will investigate an entropy based technique to draw the points which will represent the new sample.

As we chose large values of σ, the points lie arround the data points. Chosing smaller of values of σ results in complex shape is formed by to be drawn points. Figure 3 shows that choosing a larger value for σ results in evenly displaced points around the actual points while the points drawn with smaller σ values try to represent high density region.



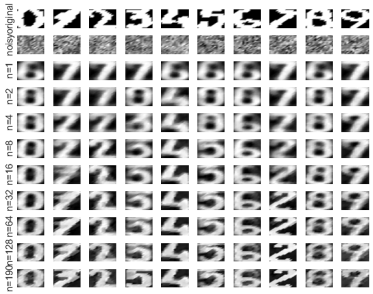
((a)) $\sigma = 1$ ((b)) $\sigma = 10$ ((c)) $\sigma = 20$

Figura 2: Effect of sigma value

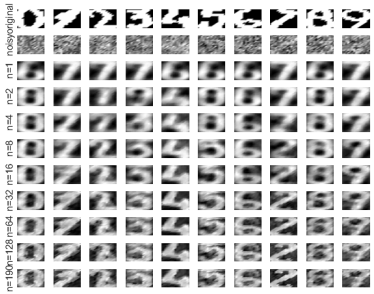


((a)) $\sigma = 1$ ((b)) $\sigma = 0{,}01$ ((c)) $\sigma = 0{,}001$

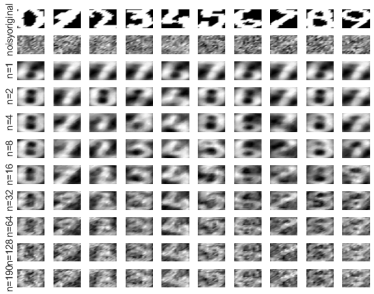
Figura 3: Effect of sigma value



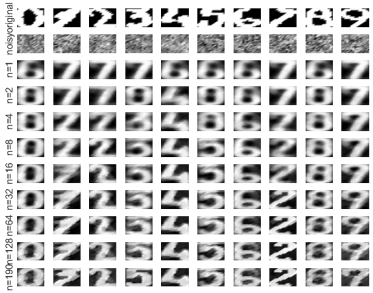
(a) Sigma = = 35



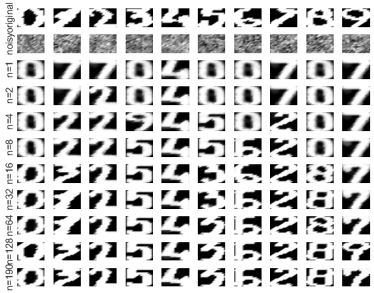
(b) Sigma = 100



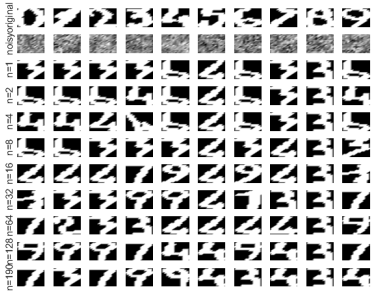
(c) Sigma = 400



(d) Sigma = = 35



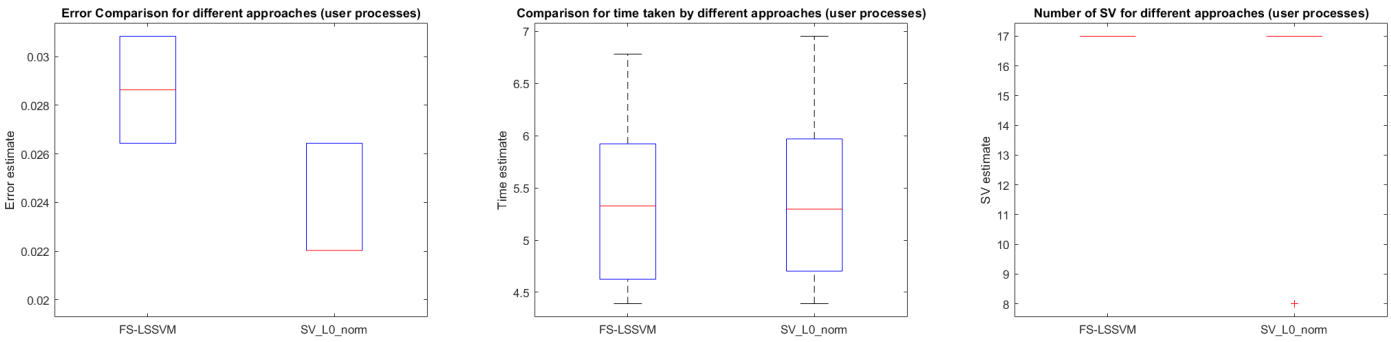
(e) sigmafactor = 0,2



(f) sigmafactor = 0,01

Figura 4: Effects of σ parameter

Next, we use breast_cancer_wisconsin_data for the comparison of fixed size ls-svm and l_0-type approximation. Figure 5 shows the three different quantities. Time and number of support vectors are same.



((a)) Error comparison ((b)) Time comparison ((c)) Number of support vectors

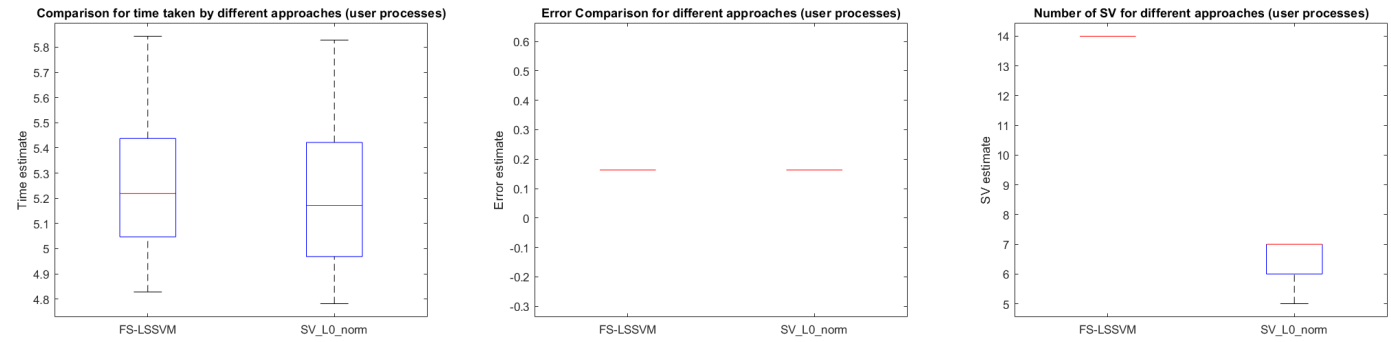
Figura 5: Effect of sigma value

However, there is a big difference between the error estimate. The error estimate lies in the center of box-plot but the error estimate for l_0-type estimate lies in lower bound of the box.

Shuttle dataset

We now use fixed sized svm for the classfication task. We use the shuttle dataset. It is a multivariate dataset with 58000 instance and number of features are 9. In comparison to the dimension of the data, the number of datapoints are huge. So expect that we can obtain better results with fixed size svm. According to the documentation, 80 % of the data belongs to class 1 while the rest of the smallest 5 classes are combined to form outliers. There are in total 7 classes.

Figure 6 shows the comparison between fsls-svm and l_0-type estimation. From the figure 6, one can



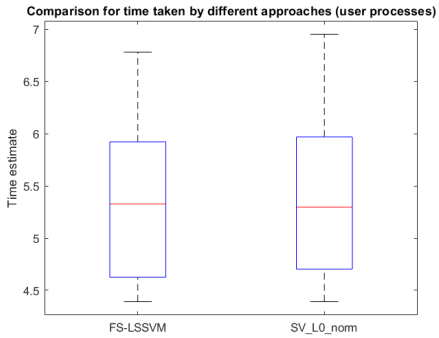
((a)) Error comparison ((b)) Time comparison ((c)) Number of support vectors

Figura 6: Comparison: Performance

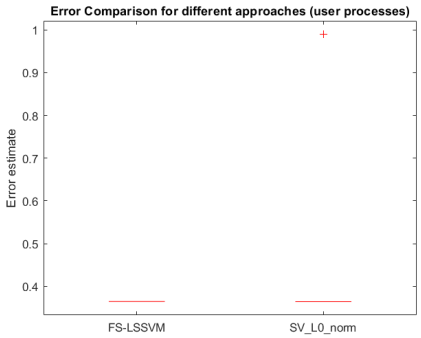
see that fs-lsvm used too many support vectors as compared to l_0-type estimation. However, the time taken and errors are almost equal.

California Housing data

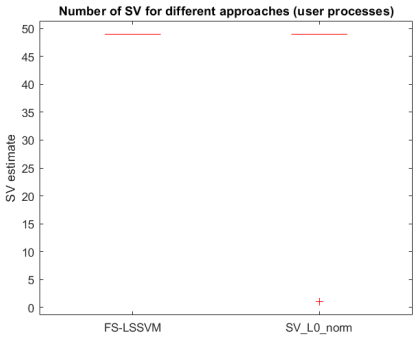
We now consider comparison on California Housing data for the regression purpose. The dataset contains $20,640$ examples and there are 9 variables/features per example. The dimension is again low and number of data points are very high. We now perform experiments again on ls-svm and penalty based estimation. Figure 7 shows the difference between three quantities.



((a)) Error comparison



((b)) Time comparison



((c)) Number of support vectors

Figura 7: Comparison:Performance

From figure 7 one can see that all three quantities are equal.