Team Name – Machine Learners

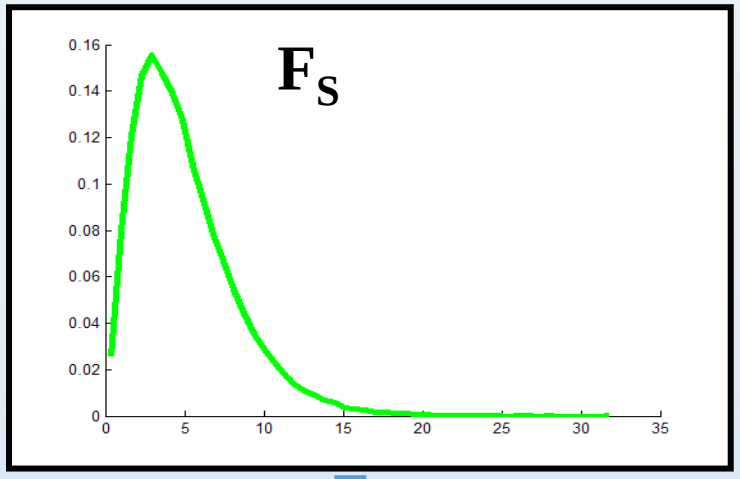
Statistical Change Detection for Multi-Dimensional Data

Course Instructor: Prof P Balamurugan
Course Code: IE-506
Machine Learning Principles and Techniques

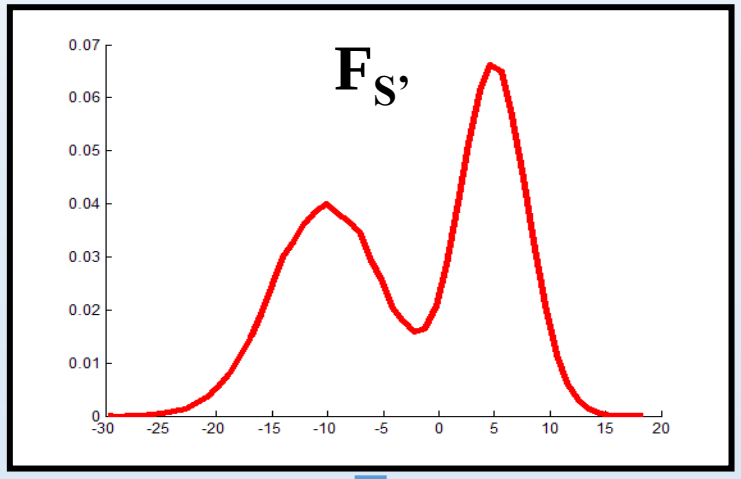
Name	Roll Number
Mayur Dhanawade	23M1512
Shivam Negi	23M1508

Contents

Problem Explanation.



Source: [3]



Source: [3]

Unknown distributions

Data Set (S)

Baseline data

Data Set (S')

Recently observed data

Multi-dimensional space

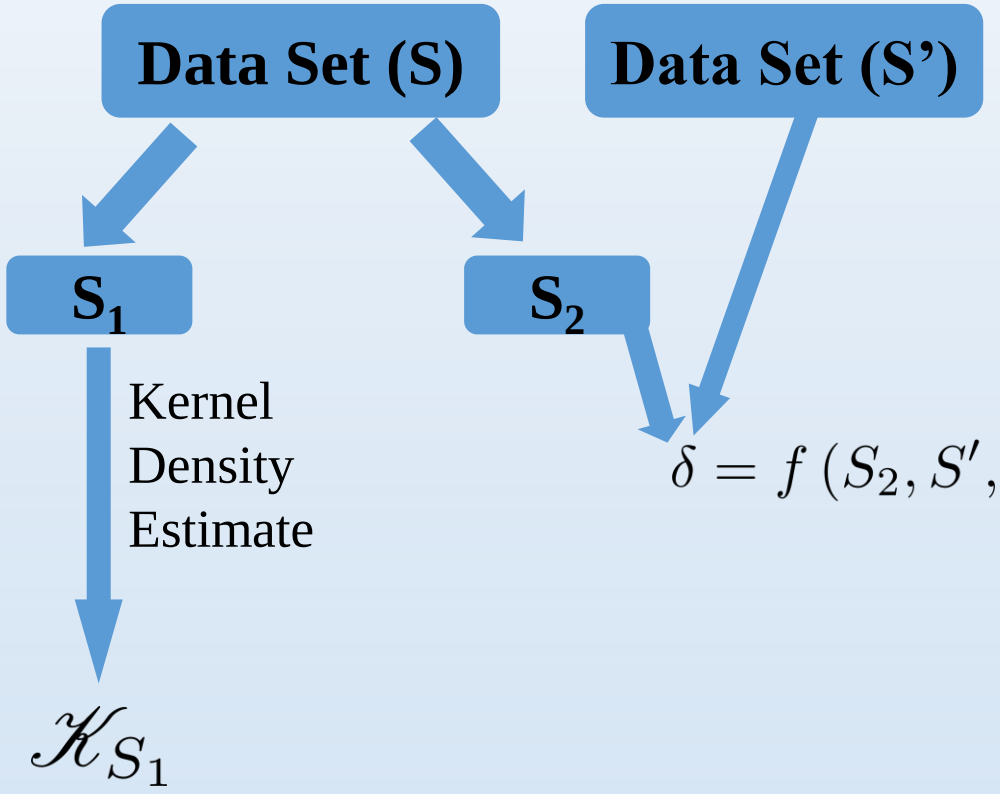
Introduction to Prior Work

- For uni-dimensional data, many existed tests, such as K-S test, chi-square test and many more.

- Only two tests to detect a generic distributional change in multi-dimensional space.

  - Cross-match test [2]:
    - analyzing distances between observations.
    - minimize the total distance within pairs.
    - cross-match statistic.
    - computationally expensive.

Density test high-level overview



Step 1: Gaussian kernel density estimate of S_1.

Step 2: Define and calculate test statistic δ.

$$\delta = f\left(S_2, S', \mathscr{K}_{S_1}\right) = LLH\left(\mathscr{K}_{S_1}, S'\right) - \frac{|S'|}{|S_2|} \times LLH\left(\mathscr{K}_{S_1}, S_2\right)$$

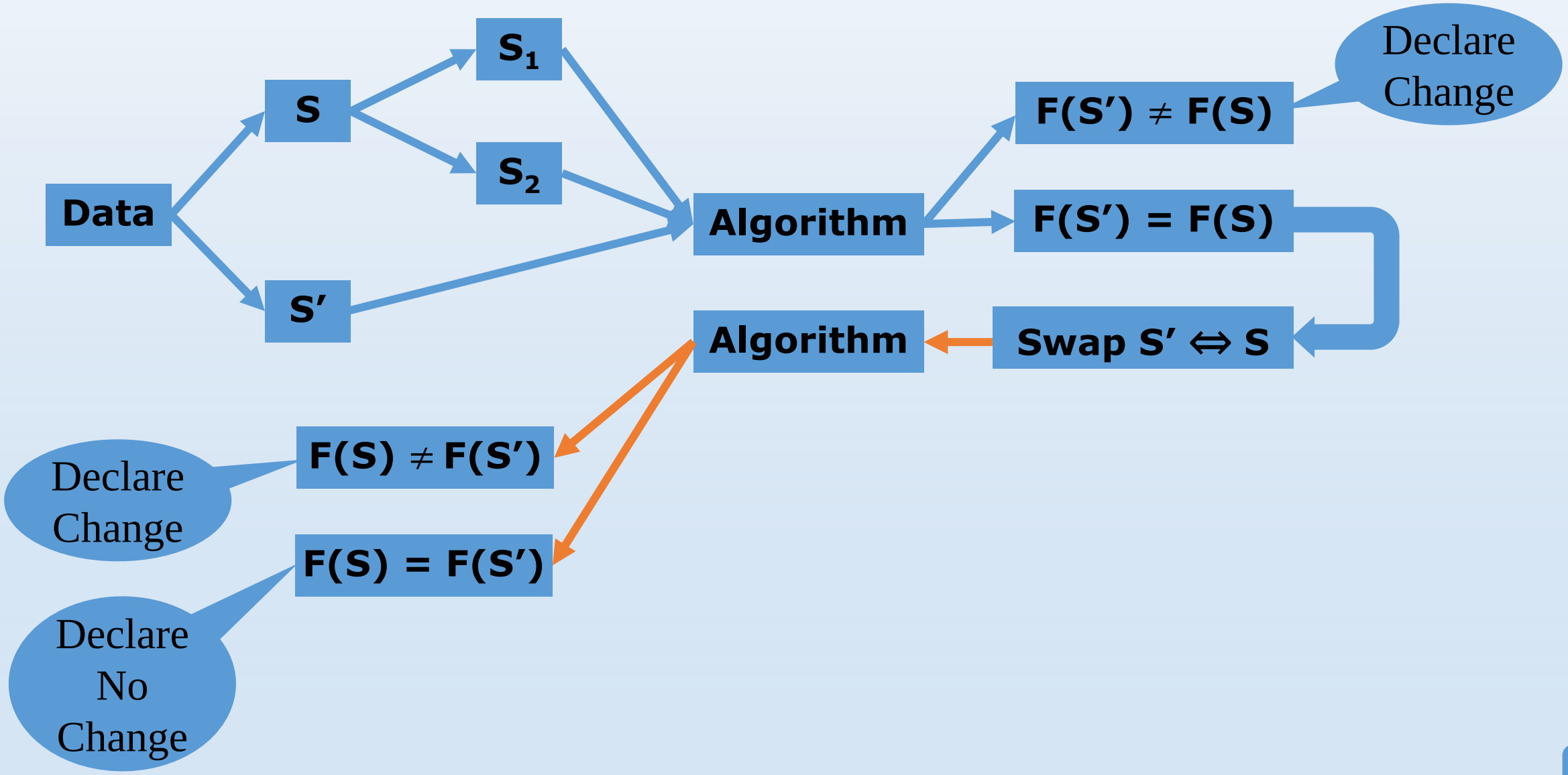
Step 3: Calculate the critical value, compare δ with critical value.

- If δ < critical value, we will declare a change.
- Otherwise, it means no change.

Work done Before the mid-term

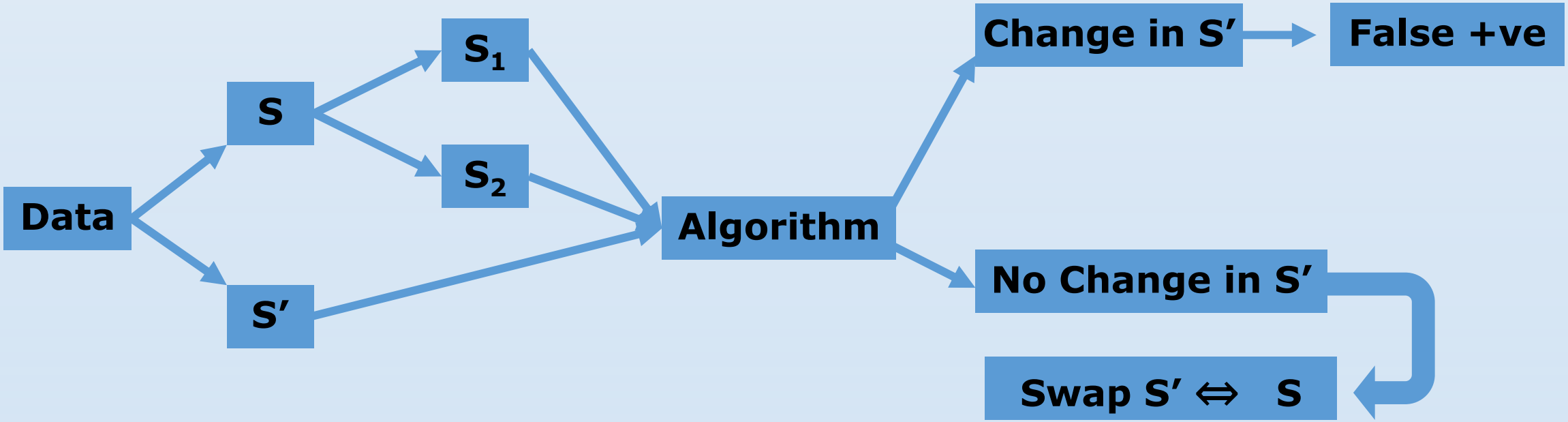
- Full Implementation of Code from scratch.
- Worked on EL Nino(5D) dataset.
- Done Two way test.
- Done Experiment 1 (Calculating false positive rate).
- Done Experiment 2 (Calculating false negative rate).
- All the result were in the acceptable range.

Two way density test implementation



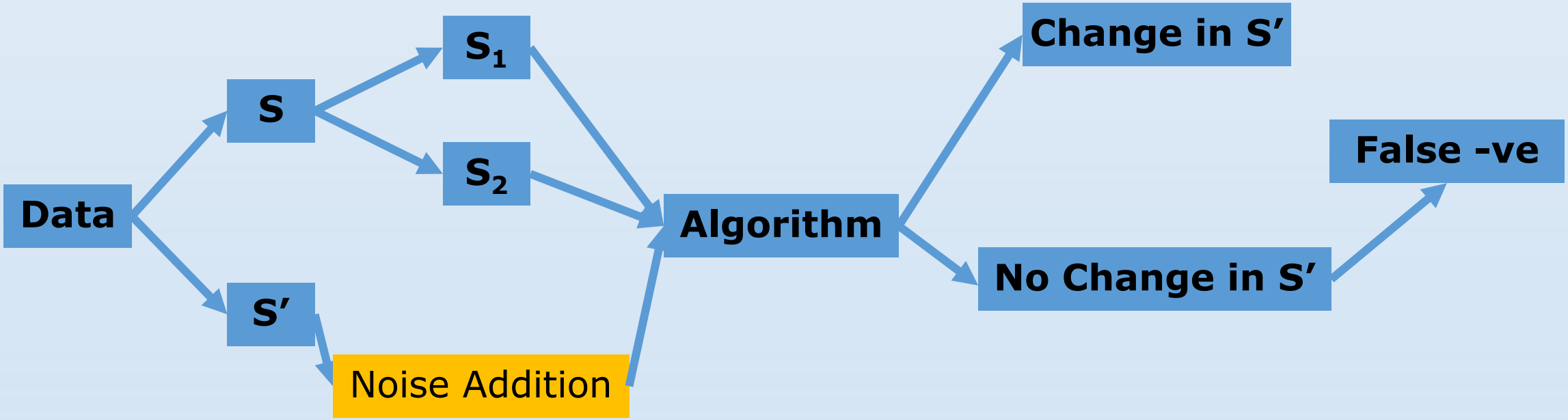
Experiment 1

- False Positive Rate Estimation
  - Test for samples from the same distribution.
  - Run for 20 instances.
  - Test distributional change in each instance using two way test.
  - False positive rate = (Number of times it detects change when actually there is no change) / 20.



Experiment 2

- Add some noise to original data. Then check for change detection.



Major comments given in mid-term project review

Instructor's Comments:
Statistical Change Detection for Multi-Dimensional Data problem motivation -- ok
prior work: some tests are mentioned and they are not explained
method in paper: density test:-- split and test procedure explained to compare densities between two splits
log likelihood function: explained carefully Algorithm was explained in detail Good effort in slides
Test statistic: -- explained nicely test statistic and its distribution: clearly explained allowable type 1 error: double density check -- very clearly explained (overall very good description) code written from scratch

Experiments to be showcased for end-term review:

- high dimensional dataset should be tried
- show scalable solutions which can work on high dimensions
- try mnist dataset and examine your algorithm for different class based data

Report Comments: It is advisable to refrain from using first-person pronouns such as "we" or "our" when discussing the author's work in the report. Instead, consider using phrases like "the author's proposed...". Avoid using screenshots in the report. The naming convention is not followed.

Actions taken to incorporate those comments

1) Explained prior work again.

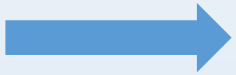
2) Modify report :
- Removed screen shot from report and write algorithms in latex in report.
- Read the report again and improved.

3) Experiments to be showcased for end-term review:
- Tried on High dimensional dataset.
- Tried scalable solutions which can work on high dimensions.
- Tried image dataset (mnist) and examined our algorithm for different images.

Work done After the mid-term

Worked on higher dimensional data.

Body Fat dataset
250 rows, 15 Columns



New Data
20000 rows, 15 Columns

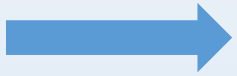
1. Randomly Select Data point (A) from original data.
2. Find A's five nearest neighbours A1, A2, A3, A4, A5.
3. Then A, A1, A2, A3, A4, A5 are averaged to produce a new data point.

Repeat this process 20000 times to bump up data size.

- Done Experiments on new data.

Worked on higher dimensional data.

Boston dataset
506 rows, 20 Columns



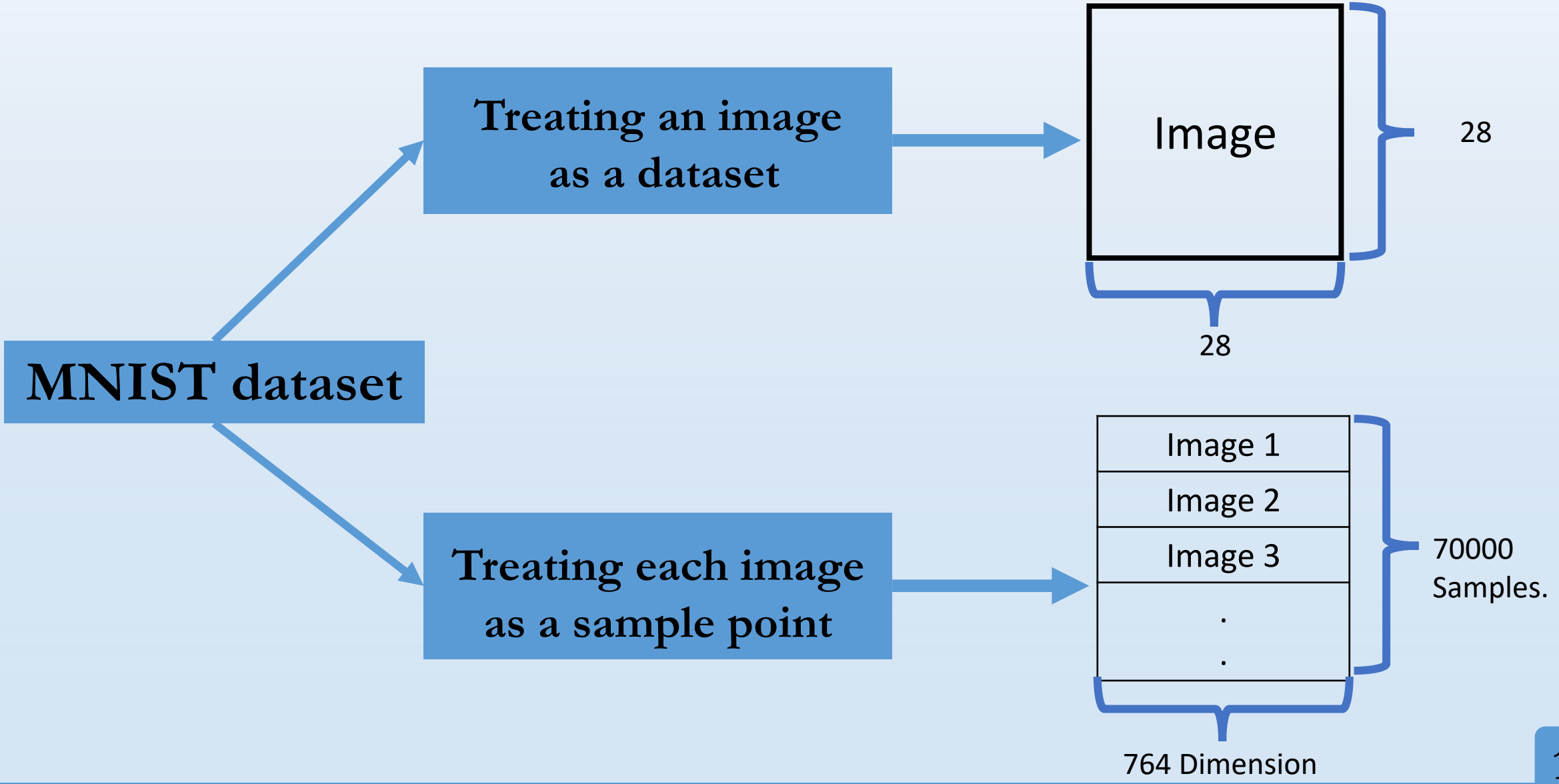
New Data
20000 rows, 16 Columns

1. Randomly Select Data point (A) from original data.
2. Find A's five nearest neighbours A1, A2, A3, A4, A5.
3. Then A, A1, A2, A3, A4, A5 are averaged to produce a new data point.

Repeat this process 20000 times to bump up data size.

- Done Experiments on new data.

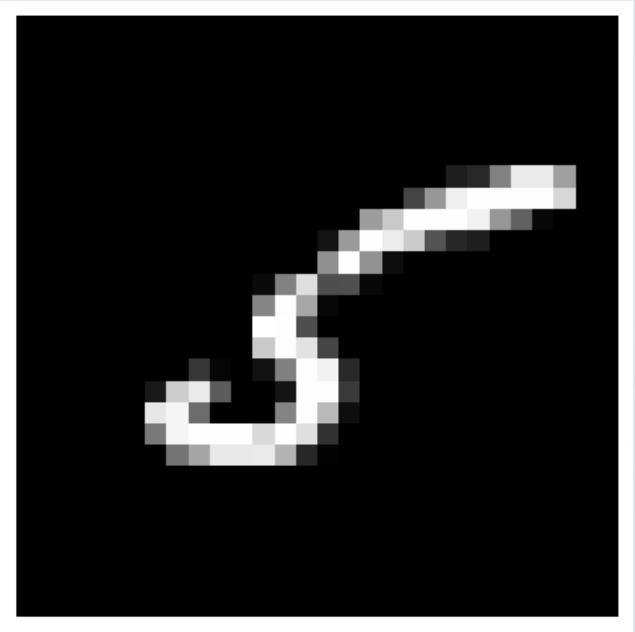
Worked on Image dataset (MNIST).



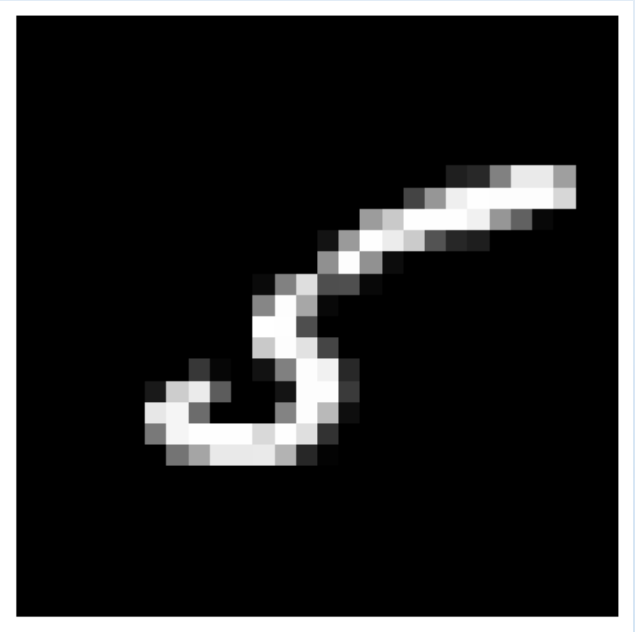
Treating an image as a dataset

- Change detection between two same images.

S1

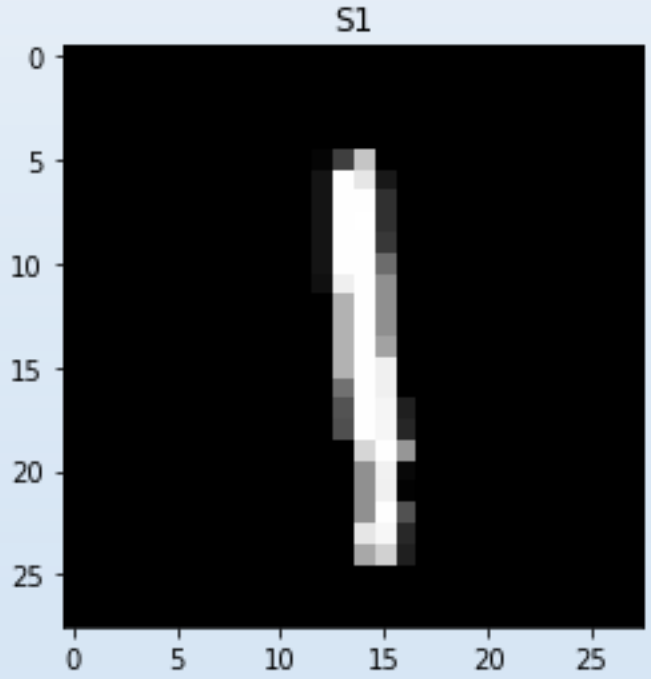


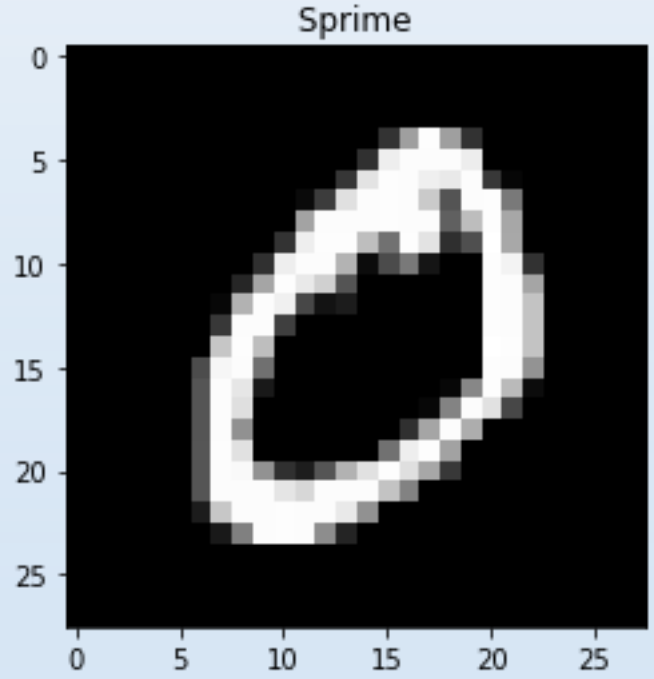
Sprime



Treating an image as a dataset

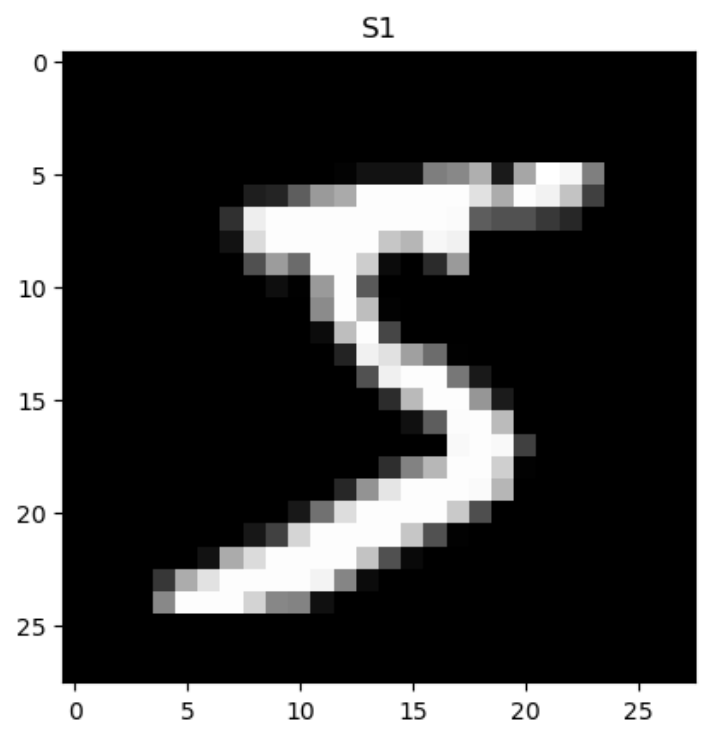
- Change detection between two different images.



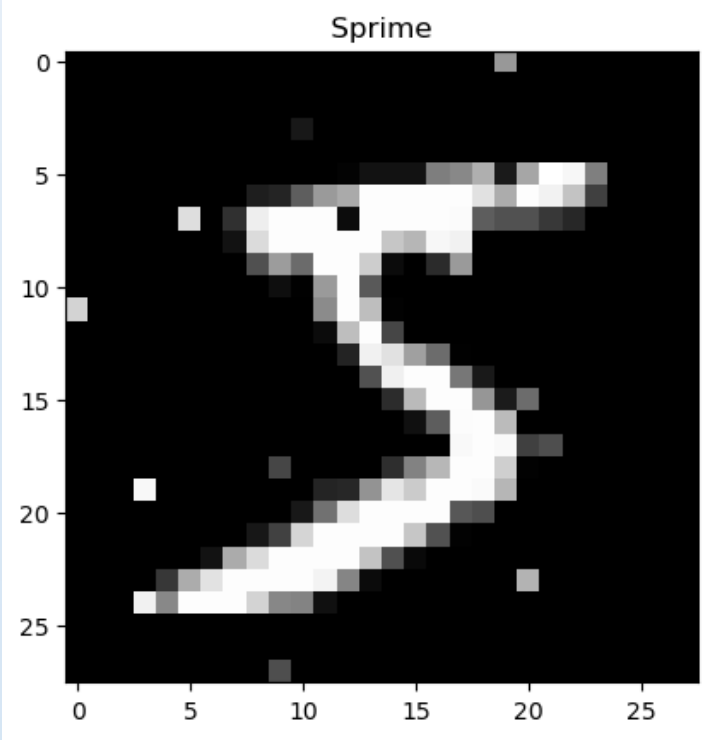


Treating an image as a dataset

- Change detection between original image and noisy image.



Original Image



Noise added Image

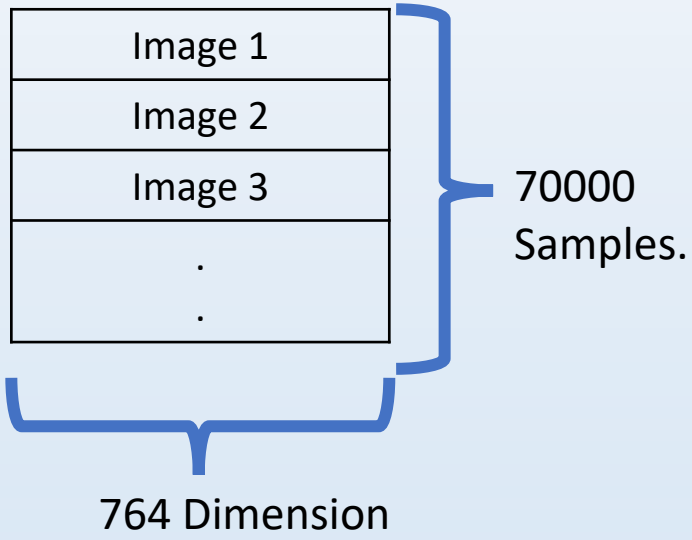
Treating each image as a sample point

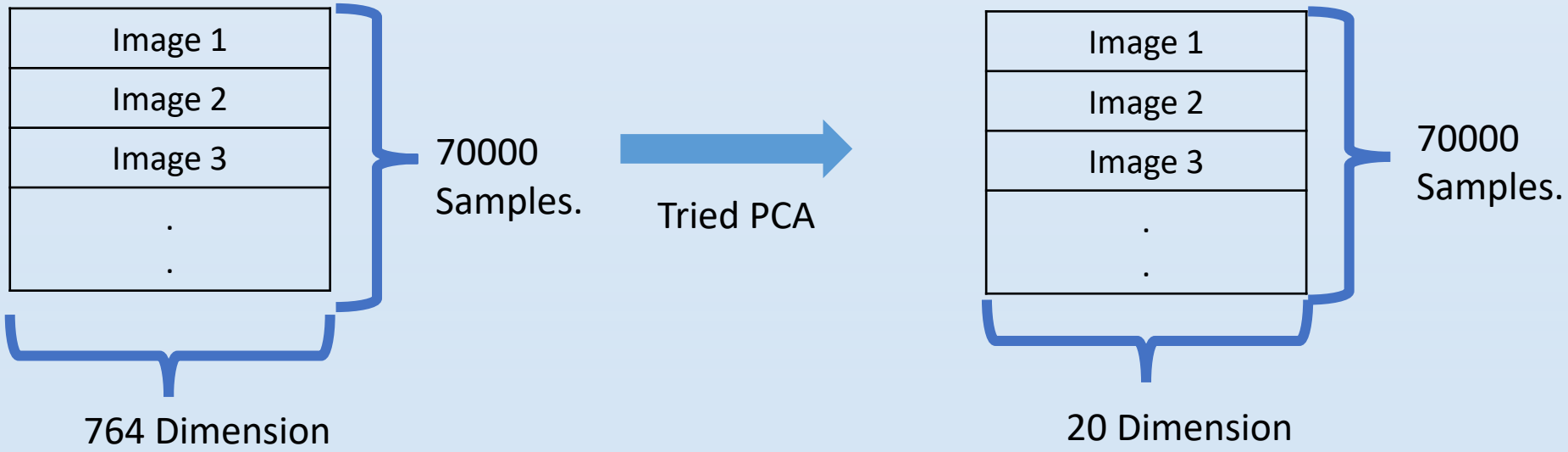
Challenges faced:

- Algorithm becomes extremely slow because of higher dimensionality.
- Authors have worked on at max 26 dimensional dataset for this paper.

Things we tried:

- Converted each image into binary image and then formed the dataset.
- Then removed those columns which had all zeros.
- Tried PCA to reduce dimensions (But we are unable to get expected results using this method).





Experiment Results

Results with Bodyfat dataset (15 dimension):

- Algorithm is successfully declaring that there is no change when S and S' have same distribution.
- Declaring change when S and S' have different distribution.
- False positives we got is 0 in 10 instances. (Experiment 1)
- False negative rates (Experiment 2):
  - For full dimensional noise we got 0 in 10 instances.
  - For one dimensional noise we got 1 in 10 instances.
  - For scaling single dimensional noise we got 0 in 10 instances.

Results with image dataset (mnist 28 × 28 pixel images):

- Algorithm is successfully declaring that there is no change when two images are same.
- Successfully detecting the change when we compare one type of image with another and also when we add some noise to the same image (noise as low as 2% in some cases, 3% in almost all the cases)

Results with Boston dataset (16 dimension):

- Algorithm is successfully declaring that there is no change when S and S' have same distribution.
- Declaring change when S and S' have different distribution.

Conclusions

- Conducted experiments to evaluate the performance of the density test. As the results of experiments show, false positive rate and false negative rate within the acceptable range as per paper.
- Acknowledged that the density test is not the most efficient method in terms of running time, especially for higher dimensional dataset (larger than 30 dimensions).

Possible future directions

- Finding the optimal way to handle the covariance matrix.
- Finding a way to reduce the time complexity of these algorithm.
- Proper analysis of whether PCA can be used can be done.

Work done by the team.

Shivam Negi (23M1508):

- Data preprocessing.
- Algorithm 2.
- Experiment 2.1
- Experiment 2.2
- Worked on Higher dimensional dataset.

Mayur Dhanawade (23M1512):

- Algorithm 1
- Algorithm 3.
- Two way test.
- Experiment 1.
- Worker on Image dataset (mnist).

References

[1] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. 2007. Statistical change detection for multi-dimensional data. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07). Association for Computing Machinery, New York, NY, USA, 667–676. https://doi.org/10.1145/1281192.1281264, Accessed 5th Mar. 2024

[2] Rosenbaum, Paul R. "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency." Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 67, no. 4, 2005, pp. 515–30. JSTOR, http://www.jstor.org/stable/3647642, Accessed 30th Mar. 2024.

[3] Image Material: https://www.merlot.org/merlot/viewMaterial, Accessed on: 5th Feb, 2024.

[4] Alicia Horsch towardsdatascience.com, Accessed on: 5th Feb 2024.

[5] Dataset (EL-Nino) link, https://www.kaggle.com/datasets/uciml/el-nino-dataset, Accessed on: 5th Feb 2024.

[6] Dataset (mnist) link, https://www.kaggle.com/datasets/hojjatk/mnist-dataset, Accessed on: 3th May 2024.

[7] Central Limit Theorem: web.stanford.edu, Accessed 2th Feb 2024.

Thank You