SYDE 372 - Lab 2

Nikhil Malhotra – 20368862

Miguel de Arcos – 20400557

# Introduction

The purpose of this lab is to analyze statistical model estimation and classifier aggregation. Specifically 1D and 2D data sets will be analyzed for parametric, nonparametric and sequential estimation. For the 1D a Gaussian, Exponential and Uniform distributions will be investigated utilizing the Maximum Likelihood estimation technique. The Parzen estimation technique will be utilized for the non-parametric case for Gaussian only in both a 1D case and the 2D case. The sequential estimator will only be used for the 2D data sets.

# Model Estimation 1-D Case

The following data sets were utilized for this section:

* variable a - a bunch of Gaussian samples,
* variable b - a bunch of Exponential samples

## Parametric Estimation:

### Gaussian:

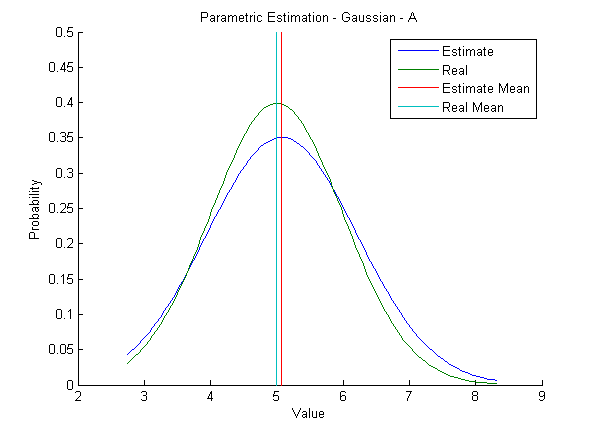


Figure 1: Gaussian ML Estimation on Normal Data

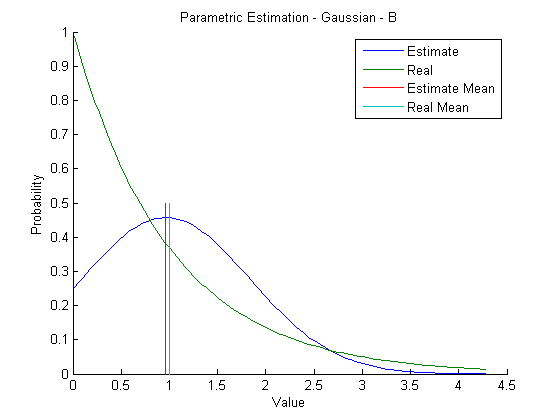


Figure 2: Gaussian ML Estimation on Exponential Data

Observing the graphs above, the closest estimated pdf is for data set a. This is obvious as data set is normally distributed unlike data set b, which is exponentially distributed.

### Exponential:

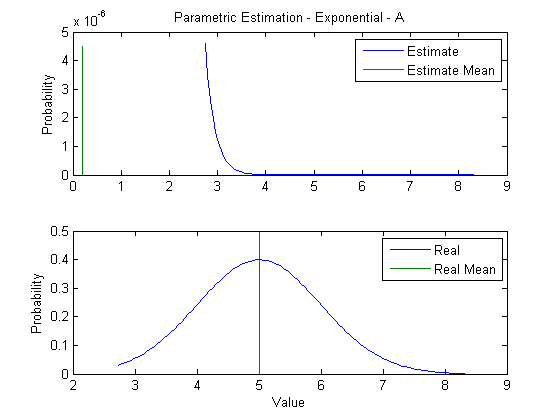


Figure : Exponential ML Estimation on Normal Data

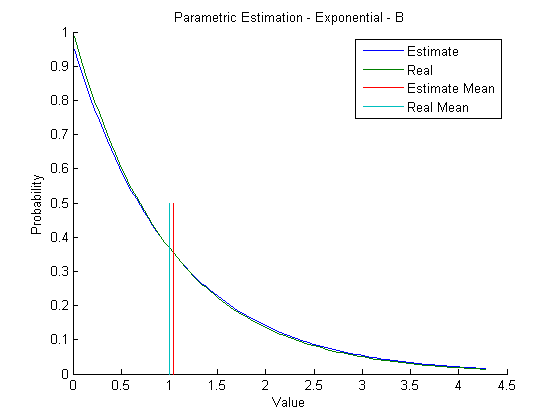


Figure : Exponential ML Estimation on Exponential Data

Observing the graphs above, the closest estimated pdf is for data set b. This is obvious as data set is exponentially distributed unlike data set a, which is normally distributed.

### Uniform:

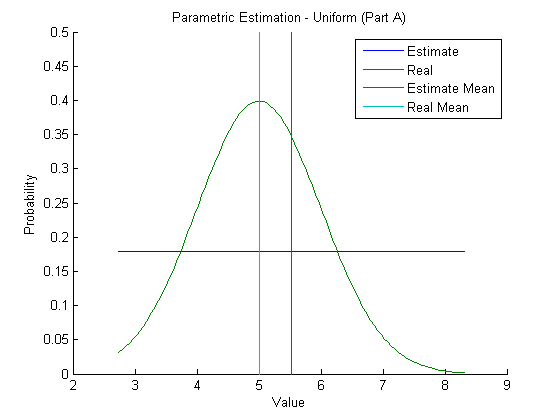


Figure : Uniform ML Estimation on Normal Data

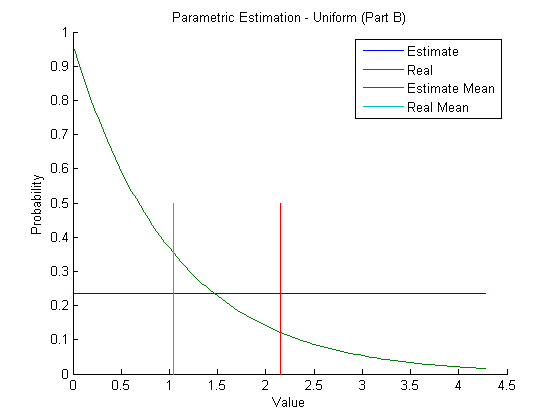


Figure : Uniform ML Estimation on Exponential Data

Observing the graphs above, the closest estimated pdf is for data set a. Although it is closer, it is ill suited for data set a.

## Non-Parametric Estimation:

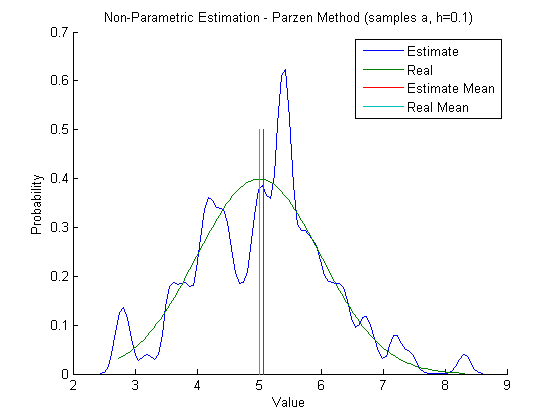


Figure : Parzen Estimation on Normal Data, h =0.1

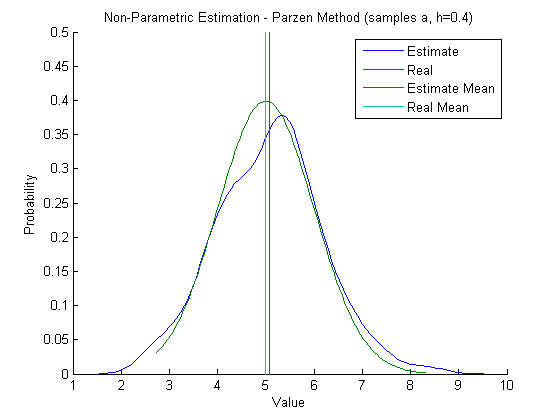


Figure : Parzen Estimation on Normal Data, h =0.4

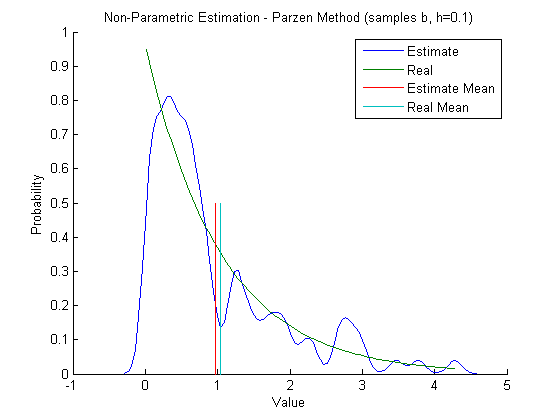


Figure : Parzen Estimation on Exponential Data, h =0.1

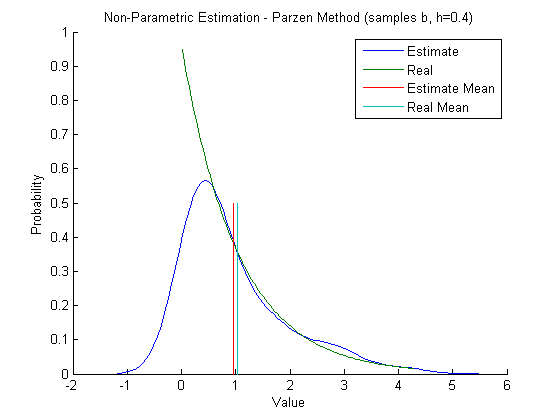


Figure : Parzen Estimation on Exponential Data, h =0.4

Observing the four graphs above it seems Parzen estimation is better suited for data set a, although at a greater value of h it can match to data set b decently at lower probability values.

In general however it is not possible always to use a parametric approach as the estimation assumption must be correct to the data set being analyzed which may not be possible all the time. It is better to use the parametric method when the distribution type of the dataset in question is known. The non-parametric method is preferred for any dataset that does not follow an exponential distribution as it can estimate much better for an entire range of value rather than just the lower probability ones.

# Model Estimation 2-D Case

Three 2 dimensional data sets (al,bl and cl) were utilized for this section.

## Parametric Estimation:

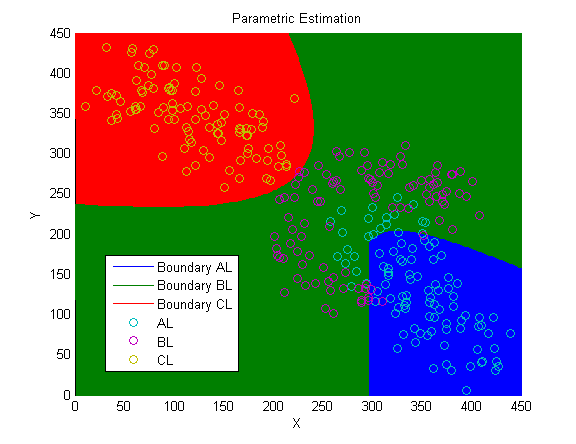


Figure : ML Estimated Classification Boundary

## Non-Parametric Estimation:

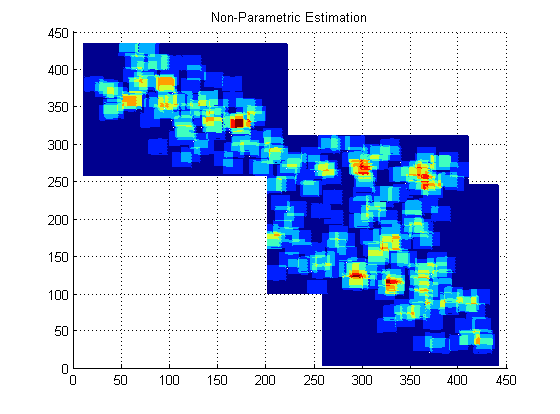
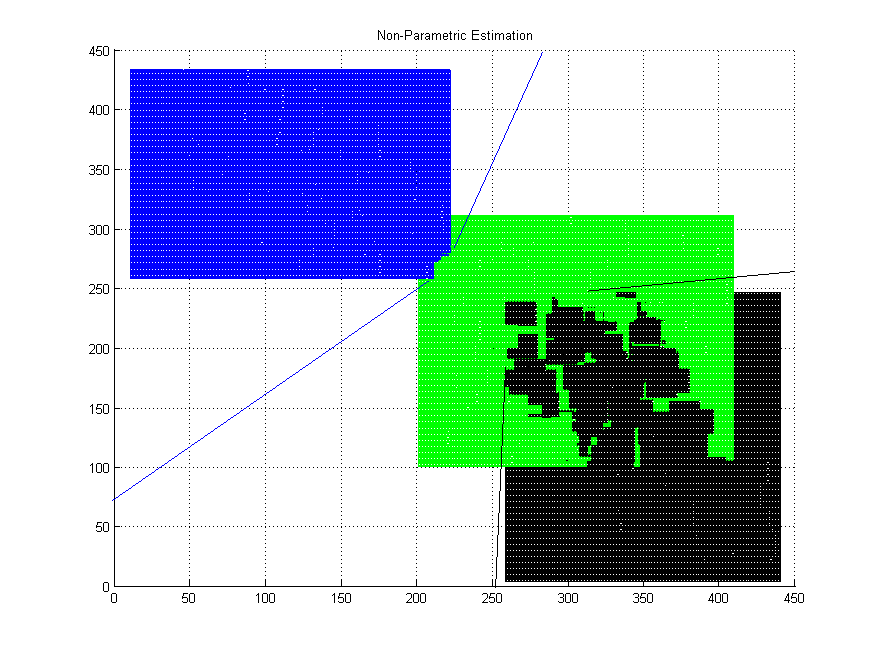


Figure 12: Gaussian Parzen Estimated Classification Boundary

The graph above is an incomplete classification boundary. This question could not be completed, and the boundaries shown are a result of super imposed mesh plots with the larger value visible above the others. It doesn’t show the actual boundary, but in most cases the visible boundaries are where they would end up. The zero values don’t show any boundaries, which is why lines were drawn to demonstrate where we thought the lines would end up.

Observing the parametric and non-parametric graphs above it is obvious that a parametric estimation works well when datasets are easily distinguishable (comparing CL and BL). However when data sets are difficult to distinguish from each other (comparing BL and AL) the parametric estimation generates a boundary of best fit that leaves out a significant chunk of the AL cluster. Where parametric estimation fails non parametric succeeds (based on Figure 12 graphs), as it is observed that the Parzen estimation keeps a higher resolution estimate of micro clusters within the AL and BL clusters in order to generate a more distinguished boundary.

# Sequential Discriminants

The following 3 sequential classifiers were learned:

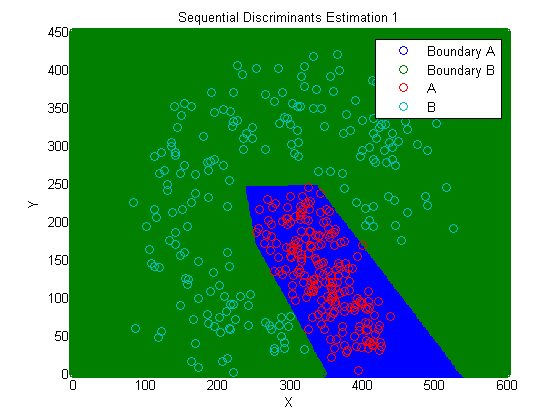


Figure : Sequential Classifier 1

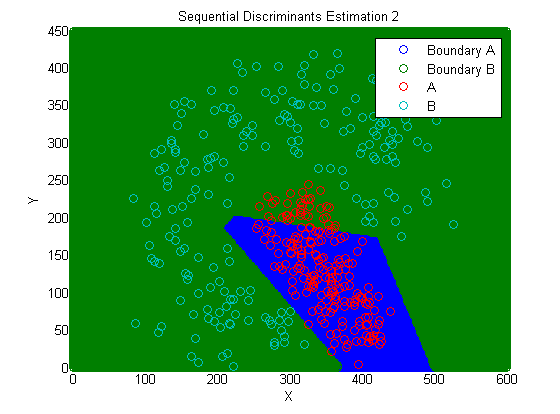


Figure : Sequential Classifier 2

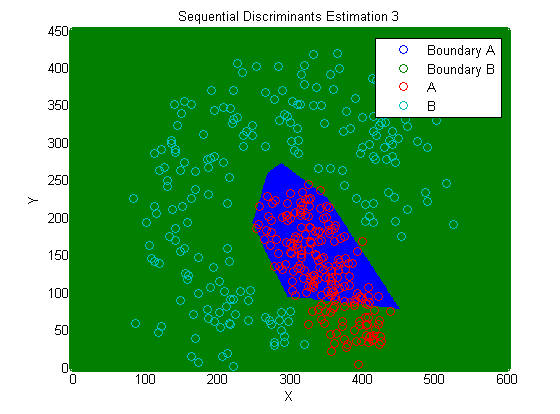


Figure : Sequential Classifier 3

On initial observation across all three classifiers above it is apparent that the probability of error, if training data was to be fed, varies greatly. This can be attributed to the fact that a random point is chosen between both sets and this could result in one data set receiving a shorter list of discriminant functions than the other. A smaller list of discriminants means, according to the algorithm followed, one data set will be completed before the other leaving a random set of samples that have not been utilized to train the classifier. This lack of data results in a partially trained classifier as can be observed in the last two classifiers.

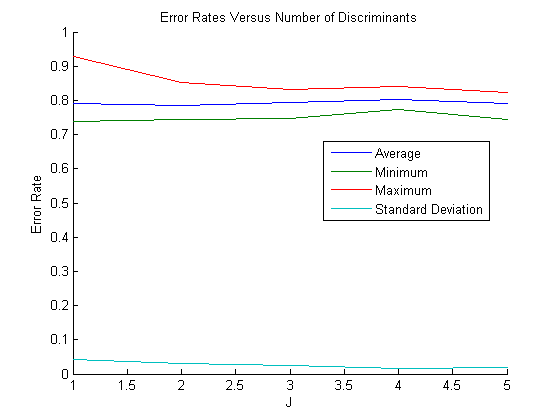


Figure : Error Rates

If the number of points allowed to be tested is limited it would result in a partially trained classifier. Partial training only occurs if the limit is below the number of discriminants required to describe the system perfectly. A fully trained classifier can occur despite a limit if the first few points tested result in a classifier that describes the data set accurately in which case the remaining data points are redundant.

# Summary

After completing this lab it can be concluded that the parametric estimation is an excellent technique for data sets that have known distributions and are fairly easy to distinguish from other data sets. Non- parametric estimation is excellent for data that does not follow regular distributions or where data sets are hard to differentiate. Parametric and non-parametric estimation techniques seem better suited for use for post data collection analysis while sequential estimation seems better suited for use when data is being collected and the system needs to come to conclusions simultaneously. The context of the data being collected however must be known for sequential estimation as it assists greatly in establishing a proper discriminant function, which in turn results in a better trained classifier.