4
Clustering Analysis of Constellation Signature

4.1 INTRODUCTION

Likelihood-based methods require a channel model and need to evaluate function values corresponding to each considered modulation in the modulation pool. Feature-based approaches need to optimize thresholds and hierarchy of features for comparison in a decision tree. Most of the approaches are based on modulated signal components, but the modulation type can be best identified using the constellation diagram. As the constellation points form clusters in the I-Q plane, modulation order can be obtained by estimating the correct number of clusters. As AWGN channel is considered in which density is greater around ideal constellation points, density-based clustering called Ordering Points to investigate the clustering structure (OPTICS) is used for estimating the order of modulation. The proposed technique can recognize M-QAM, M-ASK, and M-PSK modulation schemes. Linear regression and k-means clustering are used for domain estimation, i.e., ASK, PSK, or QAM. The algorithm is implemented in LabVIEW and tested on real-time signals. The system developed does not require any information of symbol rate, carrier frequency, and any training phase to set thresholds and detects the type of modulation blindly in real-time. Modulated RF signals are generated by NI PXIe-5673 (RF transmitter). NI PXI 5600 is used to downconvert the RF signal, and NI PXI-5142 (100 MS/s OSP digitizer) to sample the downconverted signal.

4.2 CLUSTERING ALGORITHMS

4.2.1 Ordering points to investigate the clustering structure (OPTICS)

Different partitioning-based clustering algorithms require the number of cluster centers to be estimated like k-means, k-medoids, etc. Hierarchical clustering also needs to set a threshold to stop dividing the data into clusters in case of divisive hierarchical clustering. Hence OPTICS clustering is used. It estimates the order of modulation based on the number of higher density regions. It requires data points and a parameter represented by *minpoints* as inputs and return three values corresponding to each data point, i.e., *Core Distance*, *Reachability Distance (RD)* and the *Order* in which the data points are processed. To clearly understand OPTICS, the following definitions are needed:

Core Distance(CD): Core distance of any i^{th} point is the minimum distance of the radius from the data point i such that, k (*minpoints*) points come under the circle drawn from data point i.

Reachability Distance(RD): Reachability Distance of any j^{th} object is the maximum distance between j^{th} object and the nearest object and the core distance of the nearest object. Figure 4.1 shows *CD* and *RD* for *minpoints* equal to 3. OPTICS is a density-based clustering algorithm and can be considered as an extension of the Density-based spatial clustering of applications with noise (DBSCAN) algorithm. Clusters with different densities that are not well separated from each other will not be identified as a separate cluster with DBSCAN. This problem is solved in the OPTICS algorithm explained in [Daszykowski *et al.*, 2002].

Steps to calculate CD, RD, and Order of processing:

Step1 Randomly, the first object is selected, called the current object. This object is marked as

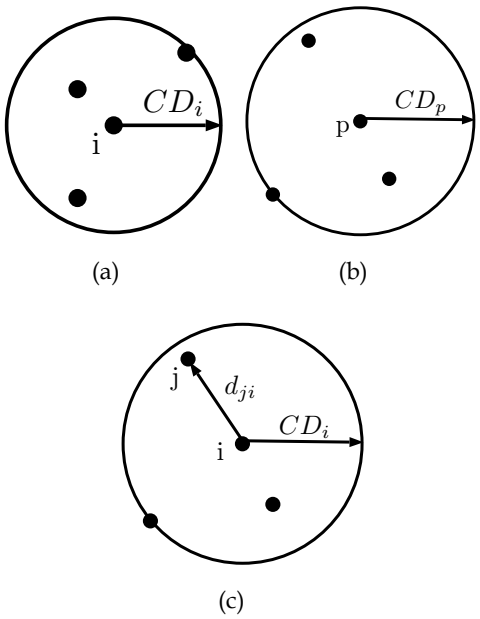


Figure 4.1 : For minpoints = 3,(a) core distance of the i_{th} object (CD_i) (b) Core distance of p_{th} object (c) Reachability distance of j_{th} object (RD_j), RD_j = max(d_{ji}, CD_i)

processed, and its Reachability distance is taken as a large value.

Step2 RD for all unprocessed objects is calculated with respect to the current object.

Step3 In this step, the data object whose RD is minimum is marked processed and is considered as the current object. If present RD of any object that is not processed is less than the previous RD, it is updated with the new lower RD. Order of all the processed objects is saved.

Step4 : Go to step 2 till all the objects are processed.

4.2.2 k-means Clustering

k-means is the partition-based clustering algorithm. To calculate centroids using k-means, the following steps are applied:

Step1 *k* initial centroids i.e., seed points are randomly considered from the data points. Here *k* is given by order of modulation.

Step2 Assign each data point to the nearest centroid. Means of the assigned points in all *k* clusters are calculated, and centroids are updated for each cluster with the calculated mean.

Step3 Repeat Step2 until no more reassignment is possible.

4.3 ESTIMATION OF MODULATION ORDER

Once we get RD and Order of processing for all the data points, peaks from the graph between RD and Order of processing are calculated. The input to the OPTICS is the constellation points shown in Figure 4.2(a). There are 1000 data points of 8 QAM at 10 dB SNR in the AWGN channel. Figure 4.2(b) shows the plot between RD and Order of processing with minpoints 50. The

optimum value of the number of clusters is obtained by the number of peaks in the plot between RD and the order of processing of objects shown in Figure 4.2(c). The number of peaks extracted is 7 for modulation order 8 because the RD of the initial point selected is made large hence the actual number of clusters will be 1 greater than the estimated number of peaks.

Sometimes in peak detection, the lower variations are also estimated as a peak. To avoid the detection of false peaks, a threshold value (taken as 60% of the mean of two highest values) is set. Peaks greater than the threshold value are taken into consideration. As the order of modulation cannot be any arbitrary number, it can be only in the power of 2; the number of clusters (number of peaks + 1) nearest to 2^N is considered as an actual order of modulation. All the simulated graphs for modulation schemes BPSK, QPSK, 8PSK, 8QAM, and 4ASK are obtained with 1000 symbol points and *minpoints* (input parameter for OPTICS clustering) are taken as 50. For higher-order modulation schemes (16QAM, 32QAM, 16PSK, and 8ASK), results are simulated with 2000 data points and *minpoints* as 30.

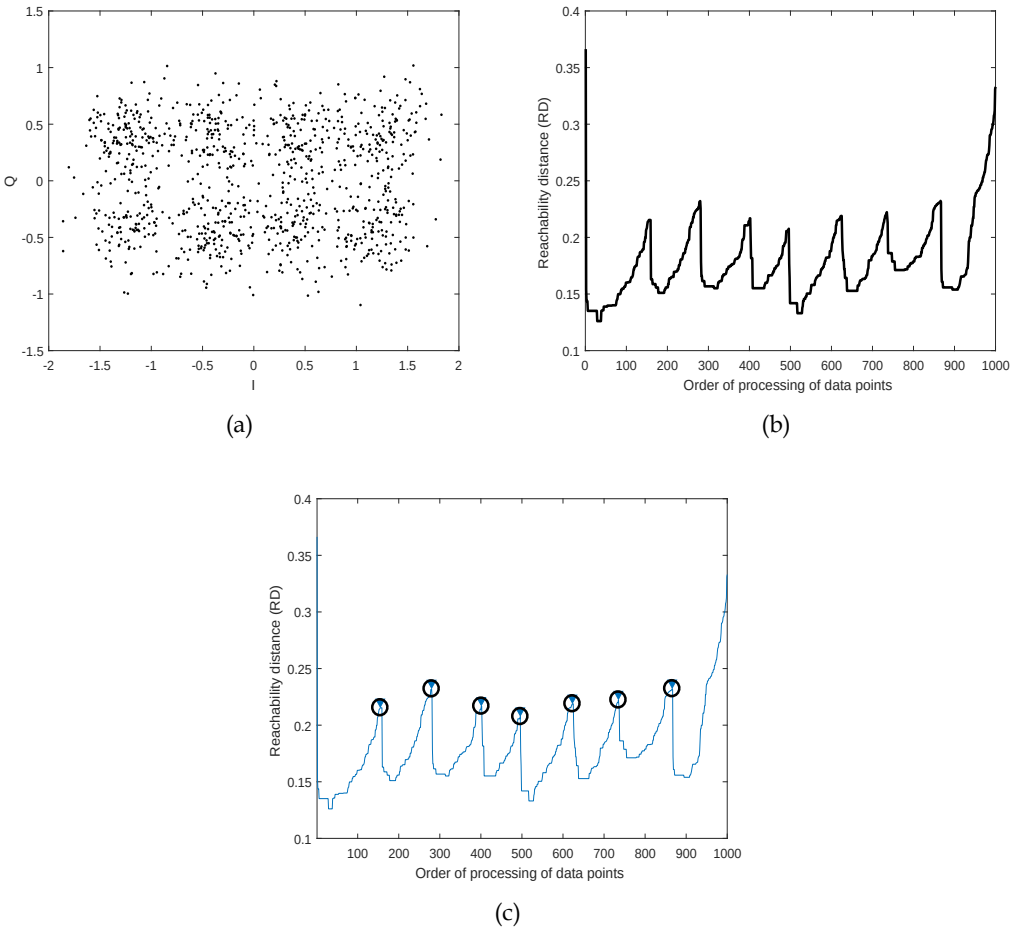


Figure 4.2 : (a)8 QAM at 10 dB SNR with 1000 constellation points. (b) Reachability Distance versus Order of processing of constellation points for (a). (c) Peak detection of RD versus order of processing obtained in (b).

4.4 ESTIMATION OF MODULATION DOMAIN

Identifying the modulation type of the signal from the pool of PSK, QAM, and ASK is done using linear regression and k-means clustering. Figure 4.3 shows the block diagram for the

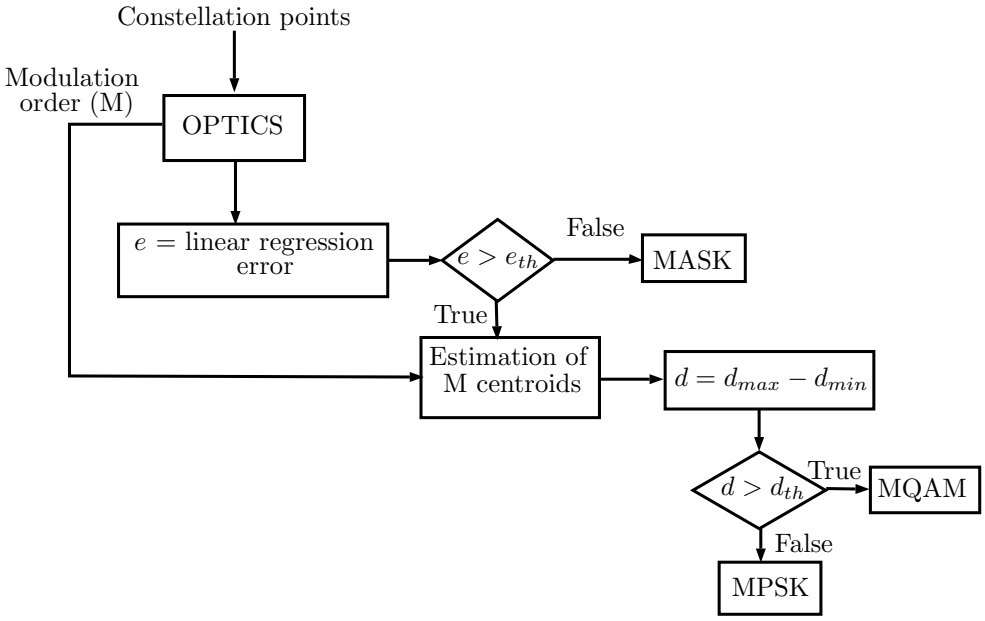


Figure 4.3 : Proposed Algorithm for Modulation Domain Classification.

modulation domain classification, and algorithm 1 explains the process for the same. To decide between ASK or PSK and QAM, a straight line with slope b_2 and intercept b_1 is regressed on the given data points (in the present case, data extracted is constellation points say (x_k, y_k), k=1,2,3..N). Thus the regression line is given by

$$Y = b_1 + b_2 X \tag{4.1}$$

where, b_1 and b_2 are constants. We try to minimize the squared error given by

$$e(b_1, b_2) = \sum_{k=1}^{N} \varepsilon_k^2 = \sum_{k=1}^{N} (y_k - (b_1 + b_2 x_k))^2 \tag{4.2}$$

Minimizing 'e' with respect to b_1 and b_2 leads to the following values of slope and intercept given in Equations (4.3) and (4.4).

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.3}$$

$$b_1 = \bar{y} - b_2 \bar{x} \tag{4.4}$$

Where $\bar{x}$ and $\bar{y}$ are the means of (x_k, y_k), for k=1,2,3..N respectively. b_1 and b_2 are obtained directly from the constellation points extracted, and square error is calculated and divided by the number of constellation points. Figures 4.4(a) and (b) show a line regressed on the 8ASK constellation and 8QAM constellation respectively. From the figures, it is clear that error in the case of ASK is less compared to QAM and the same follows for PSK as shown in Figure 4.4(c). On this basis, ASK is differentiated from PSK and QAM.

If the value of 'e' falls under PSK and QAM class, k centroids are estimated using the k-means clustering algorithm where k is given by the modulation order estimated in section 4.3.

Once we get the final k centroids, their absolute values are calculated. The difference in maximum absolute value (d_{max}) and minimum absolute value (d_{min}) is calculated. Figure 4.5 shows d_{max} and d_{min} values for PSK and QAM constellation. This difference for PSK will be significantly less as compared to the QAM constellation. An optimum threshold differentiates PSK and QAM.

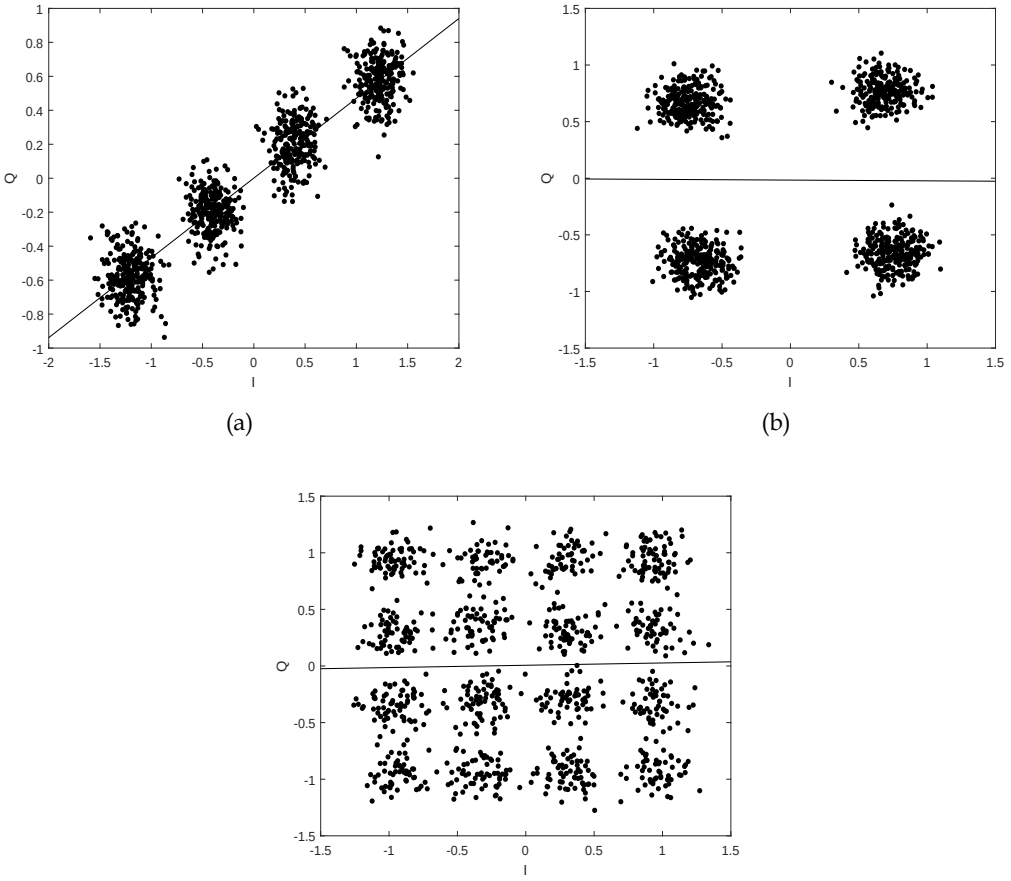


Figure 4.4 : Linear regression on different constellations

Algorithm 1: Modulation Domain Classification

Calculate error 'e' through linear regression and normalize it with respect to length of the data points.
if e $\leq e_{th}$ **then**
  Modulation is ASK
else
  Calculate M number of centroids using k-means clustering algorithm where M is modulation order calculated using OPTICS.
  $Centroids = C_i \quad i = 1, 2, \cdots M.$
  Calculate $d = max(|C_i|) - min(|C_i|) \quad i = 1, 2, \cdots M.$
  if d $\geq d_{th}$ **then**
    Modulation domain is QAM
  else
    Modulation is PSK
  end if
end if

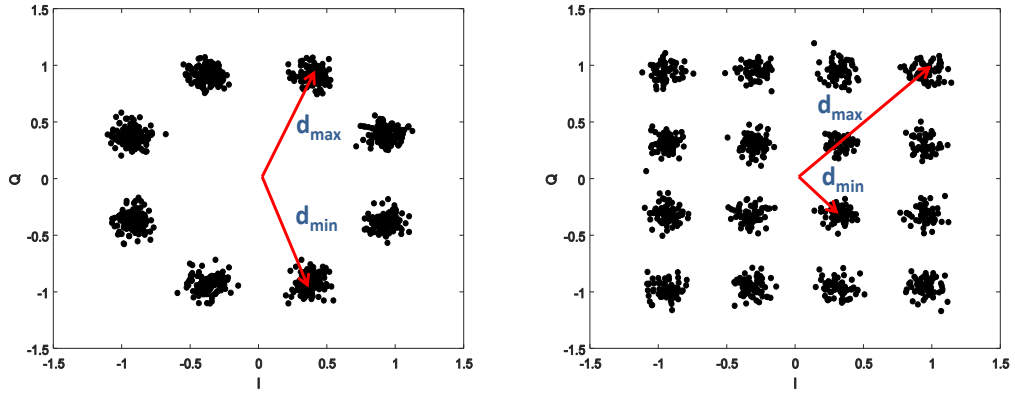


Figure 4.5 : Classification of PSK and QAM using centroid values

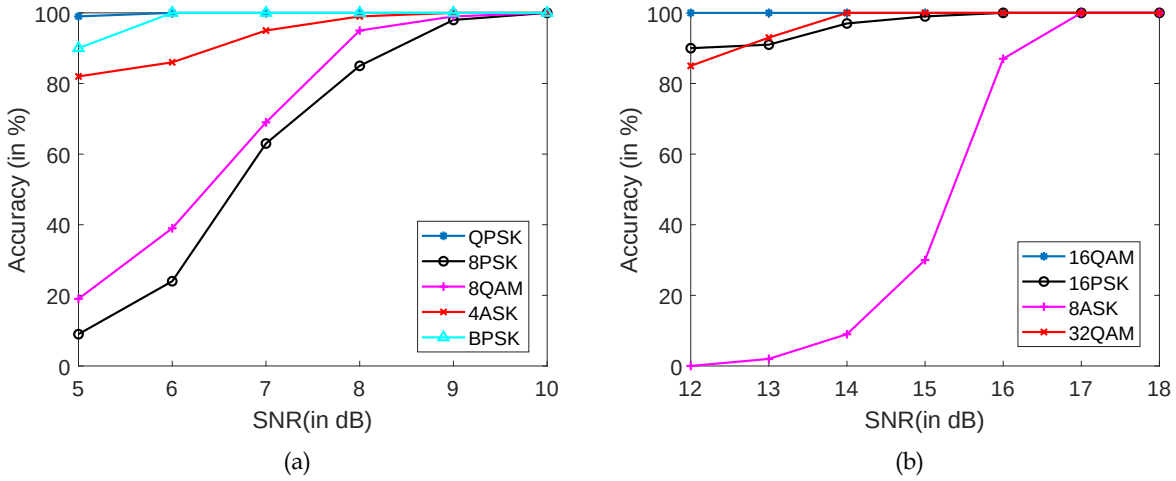


Figure 4.6 : Accuracy versus SNR for different modulation schemes

4.5 RESULTS

In this section, the performance of the proposed method for AMC has been analyzed in the AWGN channel and is compared with the existing methods in the literature. Almost all the methods in the literature require to train the classifier for setting the thresholds and weights for training models but in our proposed approach, no weights and thresholds have to be set and it works without any information of the SNR.

4.5.1 Proposed Classifier

All the simulations are performed in MATLAB. In Figure 4.6(a), a classification problem among BPSK, QPSK, 8PSK, 8QAM, and 4ASK is considered under the AWGN channel from 5dB to 10dB. It is simulated for 1000 symbol points with the value of minpoints (input parameter of the OPTICS clustering algorithm) as 50. It is observed that BPSK and QPSK are classified with 100% accuracy above 6 dB SNR. 8PSK, 8QAM, and 4ASK are classified reliably above 9dB SNR. Similarly, Figure 4.6(b) shows the accuracy of 4 modulation schemes i.e., 8ASK, 16PSK, 16QAM, and 32QAM.

Table 4.1 : Confusion matrix at 10dB SNR

	BPSK	QPSK	8PSK	8QAM	4ASK
BPSK	100	0	0	0	0
QPSK	0	100	0	0	0
8PSK	0	0	100	0	0
8QAM	0	0	0	100	0
4ASK	0	0	0	0	100

These schemes' are found out with 2000 symbols and minpoints value 30, as 32 order modulation scheme is considered with all the symbols being equiprobable. It is observed that 16QAM and 32QAM are accurately classified above 14dB SNR, while 16PSK is classified accurately above 16dB. ASK uses only one dimension for symbols; hence its accuracy is less compared to other schemes and reaches 100% above 17 dB SNR. Table 4.1 shows the confusion matrix of BPSK, QPSK, 8PSK, 8QAM, and 4ASK. It shows that at 10 dB SNR, all the modulation schemes under consideration are recognized correctly.

4.5.2 Comparison to related work

Table 4.2 compares the classification probability of Naive Bayes Classifier [Wong and Nandi, 2001], Support Vector Machine [Wong *et al.*, 2008] and, the proposed method. Naive Bayes classifier and SVM are given for 4 classes i.e., BPSK, QPSK, 16QAM, and 64QAM. For fair comparison, classification results for the same configuration are calculated. For 1024 and 2046 number of symbols, the value of minpoints is considered as 9 and 17, respectively. Table 4.2 shows that Naive Bayes and SVM classifies with the accuracy of 96.66% and 97.89%, respectively while the proposed method classifies with an accuracy of 99.5% at 14 dB SNR. On increasing the number of symbols to 2046 at 14 dB, the Naive Bayes classifier classifies with 99.43%, while the proposed method achieved the classification accuracy of 100%. This shows that the new method has outperformed the Naive Bayes classifier and SVM classifier. It is also shown in[Aslam *et al.*, 2012] that cumulants-based classification for these four classes achieve 98.2% accuracy at 20 dB (compared to 100% using the proposed method), 87.1% at 15 dB (compared with 100% in our approach). These results clearly show the improvement in the classification accuracy.

Table 4.3 compares the classification accuracy of KNN, GP-KNN, GP-Tree, from [Aslam *et al.*, 2012] and our method. The proposed classifier is simulated for the same configuration used in [Aslam *et al.*, 2012]. When the number of symbols is 1024, KNN, GPKNN, and GP-Tree achieve 97%, 99%, and 99.31% accuracy, respectively, while the proposed method achieves 99.6% accuracy at 15 dB SNR. From table 4.3, it is clear that our method achieves better accuracy compared to KNN, GP-KNN, and GP-Tree. Other than the accuracy, to provide a final training model, GP-KNN and SVM require iterative approaches that consume significant processing time compared to our approach which works without any training. [Mirarab and Sobhani, 2007] achieved 95% accuracy at 15 dB SNR with 2000 symbols while we achieved 100% accuracy. Sometimes using cumulants and higher-order statistics, separation of the modulation scheme is not possible as there is no optimum threshold defined like for 16QAM and 64QAM explained in [Abdelmutalab *et al.*, 2016]. In the proposed algorithm, the modulation order calculated through OPTICS returns the number of clusters nearest to 2^N. Therefore, it can classify any higher-order modulation scheme like 64QAM, 128QAM, 256QAM, etc. For example, if the transmitted modulation is 256QAM, and the estimated number of peaks is between 193 and 383, it will give the order as 256. So the margin of error is very

Table 4.2 : Comparison of Naive classifier and SVM with proposed classifier at different SNR

N	SNR	**Probability of correct classification**		
		Naive	SVM	Proposed
1024	14	96.66	97.89	99.5
	16	97.29	98.65	100
	18	97.76	98.93	100
	20	98.05	98.89	100
2046	14	99.43	99.04	100
	16	99.42	99.26	100
	18	99.4	99.36	100
	20	99.66	99.4	100

Table 4.3 : Comparison of KNN, GP-KNN, GP-tree and proposed classifier

N	SNR	**Probability of correct classification**			
		KNN	GP-KNN	GP-tree	Proposed
1024	15	97	99	99.31	99.6
	20	98	100	99.55	100
2048	15	100	100	-	100
	20	100	100	-	100

less as far as the constellation extracted is good. In a practical implementation of the algorithm, if the value of Oversampling factor is non-integer, the signal is resampled with the integer multiple of symbol rate and then downsampled.

…