Objectives: Statistical Analysis of a Receiver/Detector

- Receiver collects data using SONAR, RF, IR or LIDAR to determine whether a target is present or absent (Autonomous systems, Machine Vision, Machine learning, Driverless cars, Radar receivers, etc.) in a region being explored. Measurements are corrupted due to noise as well as scattering from intervening media. Two sets of data sets are collected to design an optimum strategy for improved detection, one set with no target present and the second one with target present. Analysis of the input data is carried out by obtaining the area under the ROC curve at the optimum threshold. The statistics of the data are established and signal processing techniques are performed to improve the performance of the receiver.

Data provided:

Two sets of data collected are provided in a single column (100 samples). Each is an excel file with the project #. The first 70 values constitute the 'no-target' data and the remaining 30 values constitute the 'target present' data. The number of samples of 'no target' cases is larger that of the 'target present' ones. These constitute the data under the two hypotheses, H_0 (target absent) and H_1 (target present). Students are required to get the estimated densities of the two sets, see whether the data fits 'Rayleigh', 'gamma', ''Nakagami', or 'Rician' and estimate the parameters. This would require the use of the concepts of random variables, pdf, CDF, parameter estimation and chi square testing.

Students will examine the ROC and obtain the area under the ROC curve and estimate its confidence interval. In this case, simple measure of $\pm$ one standard deviation will be estimated using the method of Hanley & McNeill (J. A. Hanley and B. J. McNeill, "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve", Radiology, Vol. 143, April 1982, pp. 29-36).

$$\sigma(A_z)=\sqrt{\frac{A_z(1-A_z)+(N_a-1)(A_1-A_z^2)+(N_n-1)(A_2-A_z^2)}{N_aN_n}}\ .$$

$$N_n \Rightarrow \text{Number of normal cases} \qquad N_a \Rightarrow \text{Number of abnormal cases}$$

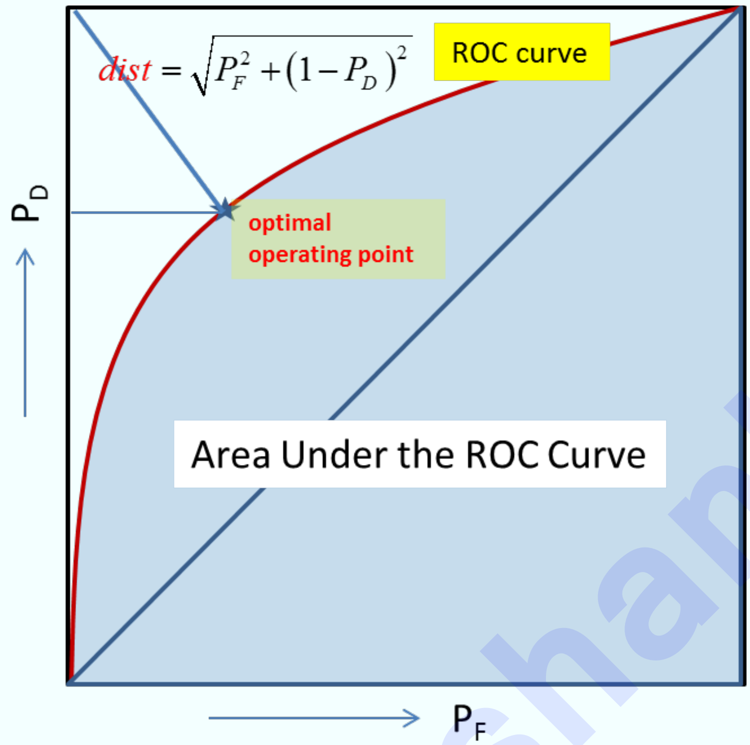
$$\mathrm{A_z} \Rightarrow \text{Area under the ROC curve,} \quad A_1=\frac{\mathrm{A_z}}{2-\mathrm{A_z}},\ A_2=\frac{2\mathrm{A_z^2}}{1+\mathrm{A_z}}$$

In the case of Machine Vision (this project),

$$N_n \Rightarrow \text{Number of samples }\left(\text{target}_{absent}\right) \qquad N_a \Rightarrow \text{Number of samples }\left(\text{target}_{present}\right)$$

ROC is the plot of the probability of false alarm (1-specificity) vs. the probability of detection (sensitivity). ROC is used to obtain the optimal threshold of operation. Optimal threshold is the point where the prob. of False alarm is a **minimum** and the probability of detection is the **maximum**.

This is represented by distance to the top left hand corner where P_D=1 and P_F=0 (best one can get). The optimal operating point occurs when the distance to this corner is the minimum. This gives the point [PF_{min}, PD_{max}]. The threshold value corresponding to this P_F or P_D is the optimal threshold.



Students will determine the optimal threshold for estimating positive predictive value (Bayes' Rule).

$$PPV = \frac{P\left(\text{det\_target}|\text{target}_{present}\right)P\left(\text{target}_{present}\right)}{P\left(\text{det\_target}\right)}$$
$$= \frac{P\left(\text{det\_target}|\text{target}_{present}\right)P\left(\text{target}_{present}\right)}{P\left(\text{det\_target}|\text{target}_{present}\right)P\left(\text{target}_{present}\right)+P\left(\text{det\_target}|\text{target}_{absent}\right)P\left(\text{target}_{absent}\right)}$$

Hypothesis testing: To understand how the receiver will function, there is a need to model the statistics. This requires the estimation of parameters followed by chi square testing. If X is the random variable representing the data, students will test if data matches one of the following densities.

Rayleigh
$$f(x) = \frac{x}{b^2}\exp\left(-\frac{x^2}{2b^2}\right), \quad x \geq 0$$

Nakagami
$$f(x) = \frac{2}{\Gamma(m)}\left(\frac{m}{\Omega}\right)^m x^{2m-1}\exp\left(-\frac{m}{\Omega}x^2\right), \quad x \geq 0, m \geq \frac{1}{2}, \Omega > 0$$

Rician
$$f(x) = \frac{x}{\sigma^2}\exp\left(-\frac{s^2+x^2}{2\sigma^2}\right)I_0\left(\frac{s}{\sigma^2}x\right), \quad x \geq 0, s \geq 0$$

Gamma
$$f(x) = \frac{1}{\Gamma(a)b^a}x^{a-1}\exp\left(-\frac{x}{b}\right), \quad x \geq 0, b \geq 0, a \geq 0$$

Note that $\Gamma(m)$ is the gamma function of m and it is equal to (m-1)! if m is an integer. In Matlab, gamma(m) gives $\Gamma(m)$ For the case of Rician density, $I_0(x)$ the modified Bessel function of the 0^{th} order. In Matlab, besseli(0,x) will give $I_0(x)$. All these densities are available in Matlab.

In Matlab, to get the parameters of a distribution, chi square testing, generation of a new set of samples, use the following command (using the gamma variable as an example)

```
pd=fitdist(dat,'gamma');% parameters of gamma density with data called dat
gam_a=pd.a; gam_b=pd.b; %gamma parameters separated
[h,p,stats]=chi2gof(dat,'CDF',pd);% chi square test for gamma separated
dataN=random('gamma',gam_a,gam_b,N,1);%get a new set gamma samples (N x 1)
-----------------------------------------------------------------------
```

Performance Enhancement

Students will generate two sets of data using random number generators (with the estimated parameters), a set for the 'no target' and a set for the 'target present'. Treating the pair of 'no target' and 'target present' as independent, students can observe the effect of (1) arithmetic mean (2) maximum and (3) geometric mean. In each case, the densities will be estimated along with a repetition of (1) estimation of areas under the ROC curve (2) positive predictive value at the threshold set from the given data (3) a performance index defined as

$$index = \frac{\left| mean(\text{target\_absent}) - mean(\text{target\_present}) \right|}{\sqrt{\text{var}(\text{target\_absent}) + \text{var}(\text{target\_present})}}$$

Project Scoring:	20%	Bayes Rule related calculations.
	20%	Single random variable related calculations
	20%	Joint density (two random variables) related calculations
	20%	Engineering Applications (ROC analysis at different stages)
	20%	Writeup, presentation, discussion, etc.

Project submission will be in three parts (See Next Page)

Part 1 30%

Obtain the ROC, estimate the area under the ROC curve and determine its standard deviation. Obtain the optimal operating point and estimate the Positive Predictive Value at the optimal operating point. Provide the PF and PD ate the optimal operating point. Estimated densities of the two sets, show the threshold obtained from the ROC and shade the areas corresponding to $P_{\text{False Alarm}}$ and P_{miss}. Also, obtain the performance index. For the optimal threshold, obtain the Confusion Matrix. Compare it to the Confusion matrix based on the threshold obtained from the intersection of the two densities.

→ Due Wednesday 5:00 PM, Week # 8

All submissions must contain the Group #, names of the members. A single document titled projectgroup#_part1.pdf must be e-mailed to Dr. Shankar.

Part 2 30%

Conduct chi square tests to determine the best fit of the densities (listed in the write up). In each case, catalog the degrees of freedom, test statistic, statement on accept/reject the specific density, parameters of the densities. State the best fit. Replot the estimated densities and plot the best fits on the same plot. Plot the ROC and superimpose theoretical ROC based on the fit. Compare the values of the areas under the ROC plots.

→ Due Wednesday 5:00 PM, Week # 9

All submissions must contain the Group #, names of the members. A single document titled projectgroup#_part2.pdf must be e-mailed to Dr. Shankar

Part 3 40%

Generate two sets of data for each group ('no Target' and 'Target present') based on the best fits. Now, undertake signal processing using three algorithms, AM, GM and Maximum. IN each case, obtain the ROC plots, area under the ROC curve, Positive predictive value and performance index. For each case (input, MAXM, AM, GM), display the Confusion Matrix. Check whether the results match the values of PPV obtained.

Summarize your conclusions on the whole project (10 sentences or less).

→ Due Wednesday 5:00 PM, Week # 10

All submissions must contain the Group #, names of the members. A single document titled projectgroup#_part3.pdf must be e-mailed to Dr. Shankar

Students will receive Part 1, Part 2, and Part 3 reports with comments and scores.