Real CFAR

Estimation of Prescribed False Alarm Rate Thresholds and ROC Curves From Local Data Using Tolerance Regions

Edward C. Real
Advanced Technology Division
Sanders, A Lockheed Martin Company
Nashua, NH

Prof. Donald W. Tufts
Electrical Engineering Department
University Of Rhode Island
Kingston, RI
Abstract

We present a method for estimating threshold values for signal detection and classification systems in which a prescribed value of false alarm probability is needed. The threshold values are determined directly from real observed test statistic data without knowledge of the probability distribution of the data. This surprising result arises from our use of tolerance regions from nonparametric statistics. We use this same approach to compute ROC curves from real data. We demonstrate our method with data taken from a two color IR focal plane array.
Definition of a Tolerance Region

• Let x denote an N-dimensional random vector, or more generally, a random sample. Let R denote any fixed region of a given population. Define the coverage of R as the proportion of the population which lies within R. The coverage of R is expressed mathematically as:

\[ C(R) = P(x \in R) \]

• A tolerance region is a random region having a specified probability (say 1 - \( \alpha \)) that its coverage is at least a specified value (say c).

• The tolerance region is random because the end points of the intervals which specify the region are functions of the observed data.
Why Tolerance Regions Work

• Consider a sample of n independent real numbers \( \{X_k\}, k = 1, 2, \ldots, n \) drawn from a particular distribution with cumulative distribution function \( F(x) \).

• We order these numbers from smallest to largest \( X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)} \) where \( X_{(r)} \) is the \( r \)-th order statistic.

• \( F(x) \) is the probability that any random variable \( X \) is less than or equal to \( x \).

• The probability that \( k \) out of \( n \) independent random variables are less than or equal to \( x \) is

\[
\binom{n}{k} F(x)^k (1 - F(x))^{n-k}
\]
Why Tolerance Regions Work

• For a given sample of n random variables, the random variable $X_{(r)}$ can be less than or equal to x if:

$$X_{(r)} \leq x \leq X_{(r+1)} \leq \cdots \leq X_{(n)}$$

but also if:

$$X_{(r)} \leq X_{(r+1)} \leq x \leq X_{(r+2)} \leq \cdots \leq X_{(n)}$$

or if:

$$X_{(r)} \leq X_{(r+1)} \leq X_{(r+2)} \leq \cdots \leq X_{(n)} \leq x$$
Why Tolerance Regions Work

• Taking all of these possibilities into account, the probability that the r-th order statistic $X_{(r)}$ is less than or equal to $x$ is:

$$P(X_{(r)} \leq x) = \sum_{k=r}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

• Similarly, the probability that the interval $(X_{(r)}, X_{(v)})$ contains the quantile $x$ is:

$$P(X_{(r)} < x < X_{(v)}) = \sum_{k=r}^{v-1} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$
Why Tolerance Regions Work

• To illustrate the independence of the above equations from the underlying distributions, let $x$ be the 95-th percentile point (quantile) of some distribution defined by $F(x)$. That is, let $F(x) = 0.95$. The probability that $x$ is between $X_{(r)}$ and $X_{(v)}$ is:

$$P(X_{(r)} < x < X_{(v)}) = \sum_{k=r}^{v-1} \binom{n}{k} 0.95^k (1 - 0.95)^{n-k}$$

• Independent of the distribution!
Estimating The Quantile

• For the case of $n = 100$ samples we see that the interval bounded by $X_{(95)}$ and $X_{(96)}$ is the most probable for the quantile $x = 0.95$ with a probability of 0.1830.

• An estimate of the quantile can be made by simply averaging the bounding order statistics $X_{(95)}$ and $X_{(96)}$.

• The interval between $X_{(95)}$ and $X_{(96)}$ is the maximum likelihood estimate for the location of the quantile.

• The lower boundary of the maximum likelihood interval $I$ can be computed directly as:

$$I = [(n + 1)F(x)]$$
For the case of $n = 500$ samples we see that the interval between $X_{(475)}$ and $X_{(476)}$ is the most probable for the quantile $x = 0.95$ with a probability of 0.0819.
Tolerance Region Adaptive CFAR

Uses Un-normalized Data

Sample Support = 460 samples
Designed PFA = 0.01
Estimated PFA = 0.0057
Magnitude Error = 0.0043
Percent Error = 43%
Split Window Normalized CFAR

Uses Normalized Data

Sample Support = 460 samples
Designed PFA = 0.01
Estimated PFA = 0.0173
Magnitude Error = 0.0073
Percent Error = 73%

- This technique consists of a sliding split window region for estimating the local mean and standard deviation.
- This estimate of the mean is then subtracted from the target sample. The result is normalized by the standard deviation estimate.
- This normalized result is then compared to a pre-determined threshold based on assumptions about the statistics of the normalized sample (e.g. it is stationary and distributed as N[0,1])
• The difference in performance lies in the departure from the Gaussian assumption.
• Even after normalization the data is still non-Gaussian.
• The elongated tail causes the errors in the PFA estimate.
• The skewness of the distribution is a function of the sample support used in estimating the variance (n = 460 in this case).
Tolerance Region Adaptive CFAR

Sample Support = 4604 samples
  Designed PFA = 0.001
  Estimated PFA = 0.0003
  Magnitude Error = 0.0007
  Percent Error = 70%
Split Window Normalized CFAR

Normalized Data

Sample Support = 4604 samples
  Designed PFA = 0.001
  Estimated PFA = 0.0062
  Magnitude Error = 0.0052
  Percent Error = 520%

- The increased variability of the the normalized data is due to the larger number of samples included in the support region.
- Since the original data is non-stationary, increasing the support region blends the changing statistical regions.
- The resulting mean and standard deviation estimates are therefore not sufficiently local to produce a stationary output, increasing the error.
This results in performance degradation because the Gaussian assumption is clearly invalid.

Thus the stationarity of the data must be kept in mind when increasing the sample support in order to achieve better estimates of the mean and standard deviation.

However, smaller sample supports do not necessarily lead to optimal performance....
• Reducing the support region (to \( n = 460 \) samples) brings the mean and standard deviation estimates more in line with the interval over which the data can be assumed to be stationary, but it also increases their variance.

• Moreover, although the PFA estimate is improved (0.0046 or 360\% error) it is still not as good as the tolerance interval CFAR estimate (0.0003 or 70\%).
• The tolerance region based adaptive CFAR algorithm outlined above can also be used to estimate probability of detection for a given constant false alarm rate.
• Thus it is useful in estimating ROC curves such as those shown above.
• The algorithm can be adapted for multi-dimensional problems such as multi-color image sequences.
FPGA based, noise statistics independent, CFAR algorithm

- Requires no “arithmetic” operations
- Accurate false alarm estimation
- No assumptions regarding the distribution of the noise or clutter.
- Cascadable for efficient FPGA implementation

“RAAC” FPGA Based, Transient RF Algorithms
- “Chirped” signal detection
- Signal Parameter Measurement
- FIR filters
Conclusions

- We have demonstrated the utility of tolerance regions in:
  - Estimating adaptive CFAR thresholds
  - Estimating ROC curves

- These tolerance region based techniques are useful because:
  - They are independent of the underlying distribution
  - They are computationally efficient since they require no arithmetic operations (adds, subtracts, multiplies or divides) only compares
  - They are well suited to high speed hardware implementations (e.g. FPGA’s or ASIC’s).
  - Whenever one wishes to design or evaluate systems based on real data, these techniques can be applied.