Research Report
No. 72

CFAR detection with non-Gaussian and dependent data

Tim Hesterberg

Draft, Last revised
July 10, 1998

Acknowledgments: This work was partially supported by Navy Phase I SBIR Award No. N68936-97-C-0150.

MathSoft, Inc.
1700 Westlake Ave. N, Suite 500
Seattle, WA 98109–9891, USA
Tel: (206) 283–8802
FAX: (206) 283–6310

E-mail: timh@statsci.com

CFAR detection with non-Gaussian and dependent data

Tim C. Hesterberg

Contents

CFAR detection with non-Gaussian and dependent data

Tim C. Hesterberg

Abstract

We discuss methods for setting a threshold for the output of a filter, with particular reference to non-Gaussian and dependent data. For the Gaussian case, we discuss the use of thresholds based on t distributions with reduced degrees of freedom. For the non-Gaussian case, we show that the use of t distributions is inappropriate, and suggest alternatives based on parametric families of distributions, with location, scale, and shape parameters. For fixed shape parameters but unknown location and scale parameters, the thresholds can be determined using a Monte Carlo technique, using variance reduction techniques to improve the computational efficiency by a factor of 1800. Analytical approximations are also discussed. For unknown shape parameters, approaches discussed include a bias-corrected plug-in approach, a shrinkage method, and a Bayesian method.

Key Words: Local CFAR detector, False alarm probability, Matched filter, SNR.

1 Introduction

In this report, we discuss constant false alarm rate (CFAR) target detection. The basic problem is to use training data in order to determine a threshold, such that if a new observation exceeds the threshold it will be considered as a possible target. The goal is to set a threshold that obtains a predetermined small false alarm probability.

We assume that a "training set" $\vec{v} = (v_1, \ldots, v_n)$ of n observations is available, each with the same mean μ, variance σ^2, and cumulative distribution function F. These should be random, containing no signal. In addition there is an observation to be tested, y. Our goal is to set a threshold C that matches a pre-specified false alarm rate α, i.e.

$$P(y > C) = \alpha \tag{1}$$

if y contains no signal and is from the same distribution as the v's. The ideal threshold (if y is independent of $\vec{v}$) would be $C = F^{-1}(1 - \alpha)$. Unfortunately in practice F is unknown, so C must be determined from $v_1, \ldots, v_n$.

The v's and y may be the output of a nonlinear or linear filter, such as a matched filter or wavelet coefficients. Some of the methods described here would also apply if the v's are pixel values.

One simple situation is the Gaussian white noise situation, where F is a Gaussian distribution with unknown mean and variance, and y and all the v's are independent. Then

the choice

$$C = \bar{v} + s(1 + 1/n)^{1/2} t_{\alpha,n-1} \tag{2}$$

satisfies (1) exactly, where $\bar{v} = n^{-1} \sum_{i=1}^{n} v_i$, $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (v_i - \bar{v})^2$, and $t_{\alpha,n-1}$ is the critical value for a Student's t-distribution with $n-1$ degrees of freedom. Then $P(y > C) = P(\frac{y-\bar{v}}{s\sqrt{1+1/n}} > t_{\alpha,n-1}) = \alpha$.

In this discussion we consider possible methods when

- there is dependence among the v's
- there is dependence between y and the v's
- the distribution is non-Gaussian, with known shape, and
- the distribution is non-Gaussian, with unknown shape.

in Sections 2, 3, 4, and 6, respectively. The intermediate Sections 5 and 7 involve both dependence and non-Gaussian shapes. We conclude with a discussion on choosing suitable parametric families, and on estimating parameters for those families, in Section 8.

(Singer and Sasaki (1995); Singer and Sasaki (1996)) discussed using critical values based on t distributions with modified degrees of freedom for the case of dependence among the v's. We extend their ideas in that case and to the case when there is dependence between y and the v's. We indicate why it is inappropriate to use the same method for the non-Gaussian case.

We give a critical discussion of those articles and others from the CFAR literate in a separate report (Hesterberg (1998)).

The v's and y may depend on multispectral images, as in (Scheffè et al. (1994)).

2 Gaussian Distributions, dependence within the training sample

In this section we assume that the v's may be dependent, with a multivariate Gaussian distribution, but that y is still independent of the v's. This is the case if y is from a different image than the v's (unless there is dependence between images). In some cases it is also true if y is in the same image as the v's, if y is sufficiently distance from any of the v's and the dependence structure has certain forms.

Dependence among the v's has three effects:

- the distribution of $\bar{v}$ is different; it still has mean μ, but a different variance,
- the distribution of s^2 is different; it still still has mean approximately equal to σ^2, but has a different variance, and

- $\bar{v}$ and s^2 are dependent.

If thresholds are set without taking the dependence into account, the actual false alarm rate will differ from the desired value, by $O(n^{-1})$.

Some of the moments for $\bar{v}$ and s^2 are:

$$\begin{aligned}
\mathrm{Var}(\bar{v}) &= \sigma^2 n^{-2} \sum_{i,j} \rho_{ij} && (3)\\
\mathrm{E}(s^2) &= (\sigma^2 - \mathrm{Var}(\bar{v}))\frac{n}{n-1} && (4)\\
\mathrm{Var}(s^2) &\doteq \frac{2\sigma^4}{n-1}\left(n^{-1}\sum_{i,j}\rho_{ij}^2\right) && (5)\\
\mathrm{Cov}(\bar{v}, s^2) &= 0\\
\mathrm{Cov}(\bar{v}^2, s^2) &\geq 0
\end{aligned}$$

Two of these deserve further comment. To obtain (5), we note that $s^2 \doteq n^{-1}\sum(v_i-\mu)^2$, and in particular $\mathrm{Var}(s^2) \doteq (n/(n-1))\mathrm{Var}(n^{-1}\sum(v_i-\mu)^2)$; this holds exactly in the independent case. Because the v's are multivariate Gaussian, $\mathrm{Corr}((v_i-\mu)^2,(v_j-\mu)^2) = \rho_{ij}^2$. Putting these together yields

$$\mathrm{Var}(s^2)(n-1)/n \doteq \mathrm{Var}(n^{-1}\sum(v_i-\mu)^2) = 2\sigma^4 n^{-2}\sum_{i,j}\rho_{ij}^2$$

which simplifies to (5).

The positive covariance between $\bar{v}^2$ and s^2 also deserves comment. A geometric argument (not given here) indicates that the correlation is positive in general, in particular whenever the direction $(1, 1, \ldots 1)$ does not coincide with any of the principle components of the covariance matrix. The geometric argument suggests that the conditional expected value of s^2 given $\bar{v}$ would be of the form $\mathrm{E}(s^2|\bar{v}) = a + b(\bar{v}-\mu)^2$ for some constants a and b. Calculating these constants and the covariance requires tedious algebra, and we have not done so. Our intuition is that the dependence between $\bar{v}$ and s^2 is relatively unimportant in practice if n is reasonably large, because the distribution of $\bar{v}$ is relatively unimportant for setting thresholds; we discuss this point in greater detail in Section 4.2. We ignore the dependence between $\bar{v}$ and s^2 in the sequel.

We approximate

$$\frac{y-\bar{v}}{cs}$$

as having a t-distribution with some reduced degrees of freedom, where

$$c = (1 + n^{-2}\sum_{i,j}\rho_{ij})^{1/2}\left(\frac{(n-1)}{n}(1 - n^{-2}\sum_{i,j}\rho_{ij})\right)^{1/2}$$

is a constant that incorporates the variance of $y - \bar{v}$ and corrects for the bias of s^2.

The degrees of freedom should be determined by how accurately s^2 estimates σ^2, in particular by the relationship

$$\frac{2}{df} = \frac{\mathrm{Var}(s^2)}{\mathrm{E}(s^2)^2}. \tag{6}$$

This is the relationship used for determining degrees of freedom for many statistical situations when t distributions are used as approximations. It coincides with the classical answer for degrees of freedom when the classical answer applies. Note that the degrees of freedom are larger when s^2 is more accurate (when it has a smaller variance). Simplifying (5) and (6) yields

$$df \doteq (n-1)\left(n^{-1}\sum_{i,j}\rho_{ij}^2\right)^{-1}. \tag{7}$$

This reduces to the usual degrees of freedom $n-1$ when the v's are independent.

(Singer and Sasaki (1996)) use t-distributions with reduced degrees of freedom, for the special case with $\mathrm{E}(v) = 0$. They use a formulation equivalent to (6), but estimate $\mathrm{Var}(s^2)$ using values of s^2 obtained from multiple images, rather than using (7) and a single image.

2.1 Estimating correlations

To estimate the correlations ρ_{ij} we may make use of the fact that the v's typically have a two-dimensional structure such that (temporarily switching to double subscripts) $\mathrm{Corr}(v_{s,t}, v_{s+a,t+b})$ depends only on a and b, so we may estimate the correlations using autocorrelation methods. Furthermore, the correlations tend to be near zero unless a and b are small, so that in practice a relatively small number of correlations need to be estimated—the others may be treated as zero.

The estimated autocovariance is

$$\widehat{\mathrm{Cov}}_{a,b} = n^{-1}\sum(v_{s,t} - \bar{v})(v_{s+a,t+b} - \bar{v})$$

where the sum is taken across all pairs of observations with lag (a, b). This follows the usual practice in computing autocorrelations and autocovariances, of using a divisor n, although this gives estimates which are biased toward zero, particularly for larger lags. For target detection this is appropriate because covariances for larger lags should tend to be small, and using the biased estimate reduces the variability in the estimates and produces more robust results from calculations that are based on the covariances.

The following S-PLUS function that calculates autocovariances for all "lags" (pairs of (a, b)), making use of the two-dimensional fast Fourier transform for speed:

```
acf2 <- function(x){
  # Estimate bivariate autocovariance function.  Missing values allowed.
```

```
  d <- dim(x)
  if(length(d) != 2)
    stop("x is not a matrix")
  pad.x <- array(0, 2*d-1)
  pad.x[1:d[1], 1:d[2]] <- x - mean(x, na.rm=T)
  pad.na <- array(0, 2*d-1)
  pad.na[1:d[1], 1:d[2]] <- !is.na(x)
  pad.x[which.na(pad.x)] <- 0
  fft.pad.x <- fft(pad.x)
  conv.x <- Re(fft( fft.pad.x * Conj(fft.pad.x), T))/prod(2*d-1) # Im ~ 0
  x.acf.fft <- conv.x[1:d[1], c(seq(d[2]+1,2*d[2]-1), 1:d[2]) ]
  dimnames(x.acf.fft) <- list( 0:(d[1]-1), -(d[2]-1):(d[2]-1) )
  fft.na <- fft(pad.na)
  conv.na <- Re(fft( fft.na * Conj(fft.na), T))/prod(2*d-1) # Im ~ 0
  N <- conv.na[1:d[1], c(seq(d[2]+1,2*d[2]-1), 1:d[2]) ]
  x.acf.fft / (length(x) + N -
       outer( d[1] - 0:(d[1]-1), d[2] - abs(-(d[2]-1):(d[2]-1))))
}
```

The function allows missing values, which is important if 'y' is from the same image as $\vec{v}$. With missing values the denominator is adjusted by subtracting the number of missing values from n.

An alternate approach to estimating correlations requires working at the pixel level instead of the filter output level. For example, if the pixel values are $\vec{x}$ and the filter is linear, $\vec{v} = M\vec{x}$, then $\mathbf{\Sigma}_v = M\mathbf{\Sigma}_X M^T$, where $\mathbf{\Sigma}_X$ is the known or estimated covariance matrix for the pixel values.

2.2 Improvements on $\bar{v}$ and s^2

In this section we obtain more accurate estimates of μ and σ. The potential gain in accuracy may be substantial if the observations in $\vec{v}$ have moderate to high multiple correlation. Otherwise the gains would be modest, and the extra complication of these estimates argues against their use.

When the variables are dependent, it is in general possible to improve on $\bar{v}$ as an estimate of μ, and on s^2 as an estimate of σ^2. The optimal estimate of μ is

$$\hat{\mu} = \sum_{i=1}^{n} w_i v_i$$

where

$$\vec{w} = c\mathbf{\Sigma}_v^{-1}(1, 1, \ldots, 1)$$

and where c normalizes the weights to sum to 1. (The correlation matrix may be substituted for $\mathbf{\Sigma}_v$). With positive autocorrelations the weights are higher for observations near the edges and particularly the corners.

A fast approximation is obtained by setting

$$w_i = c(\sum_j \rho_{ij})^{-1}.$$

This is essentially a weighted average, with smaller weights on those observations which are most redundant (highest correlations with other observations). The approximation should be accurate when the correlations are not too large. The approximation should be more robust than the optimal estimate, less sensitive to minor perturbations in the estimated covariances. We have observed that the weights used by the optimal estimate can be negative, which is not true of the fast approximation.

Example 1 In this example, $n = 64$ and the v's are from an 8 by 8 array, with an autoregressive correlation structure,

$$\text{Corr}(v_{s,t}, v_{s+a,t+b}) = .5^{|a|+|b|}.$$

Then the optimal weights $w_{s,t}$ are:

```
0.040 0.020 0.020 0.020 0.020 0.020 0.020 0.040
0.020 0.010 0.010 0.010 0.010 0.010 0.010 0.020
0.020 0.010 0.010 0.010 0.010 0.010 0.010 0.020
0.020 0.010 0.010 0.010 0.010 0.010 0.010 0.020
0.020 0.010 0.010 0.010 0.010 0.010 0.010 0.020
0.020 0.010 0.010 0.010 0.010 0.010 0.010 0.020
0.020 0.010 0.010 0.010 0.010 0.010 0.010 0.020
0.040 0.020 0.020 0.020 0.020 0.020 0.020 0.040
```

while the approximate weights $w_{s,t}$ are:

```
0.024 0.019 0.017 0.017 0.017 0.017 0.019 0.024
0.019 0.015 0.014 0.014 0.014 0.014 0.015 0.019
0.017 0.014 0.013 0.012 0.012 0.013 0.014 0.017
0.017 0.014 0.012 0.012 0.012 0.012 0.014 0.017
0.017 0.014 0.012 0.012 0.012 0.012 0.014 0.017
0.017 0.014 0.013 0.012 0.012 0.013 0.014 0.017
0.019 0.015 0.014 0.014 0.014 0.014 0.015 0.019
0.024 0.019 0.017 0.017 0.017 0.017 0.019 0.024
```

The gains from using weighted estimates in place of μ and σ^2 are modest, unless the autocorrelations are high. For Example 1 the optimal weighted estimate of μ has variance

about 9% smaller than does $\bar{v}$, or 6% smaller using the approximate weights. The gains would be smaller for larger matrices or smaller correlations.

A further shortcut is available for the approximate weights, with a fast implementation even for large matrices—to set autocorrelations for longer lags to zero. For some small integer k, say $k = 3$, let $M_{a,b} = \mathrm{Corr}(v_{s,t}, v_{s+a,t+b})$ if $|a| \leq 3$ and $|b| \leq 3$, and 0 otherwise. Then the weights (before normalizing to sum to 1) are

$$w_{s,t} = \sum_{a=-k}^{k} \sum_{b=-k}^{k} M_{a,b} I\left((s+a, t+b) \text{ is in the image}\right)$$

where I is the indicator function. The weights for interior locations are identical, only those locations within k of an edge receive different weight.

The analogous weighted estimates of σ^2 substitute ρ_{ij}^2 for ρ_{ij} (the "optimal" estimate may not be quite optimal here, due to complications arising from estimating μ simultaneously with σ^2).

The more accurate estimates would result in thresholds of the form

$$C = \hat{\mu} + \hat{\sigma}(1 + \widehat{\mathrm{Var}}(\hat{\mu}))^{1/2} t_\alpha$$

where the degrees of freedom are determined using (6).

3 Gaussian Distributions, dependence between the training sample and the test observation

In this section we assume that there may be dependence between y and the v's. For instance, they may be from different parts of the same image, a sequence of correlated images, or from multiple frequency bands for the same image as in (Scheffè et al. (1994)). We assume that the joint distribution is multivariate Gaussian, with common mean μ and variance σ^2. We also assume that a signal would affect only the value of y, not any of the v's. This may be an important limitation; relaxing it is possible but is beyond the scope of this discussion.

The most important change in this case is that thresholds should be based on the estimated values for $\mathrm{E}(y|\vec{v})$ and $\mathrm{Var}(y|\vec{v})$ rather than on the estimated values for μ and σ^2. This makes possible much tighter thresholds, and a much greater chance for detecting a target. The degrees of freedom should also change, but this is less important.

The ideal threshold (if μ and all covariances were known) would be

$$C = \mathrm{E}(y|\vec{v}) + z_\alpha \sqrt{\mathrm{Var}(y|\vec{v})} \tag{8}$$

Let $\mathbf{\Sigma}_v$ denote the covariance matrix of the v's and $\mathbf{\Sigma}_{v,y}$ the column vector of covariances between the v's and y, with transpose $\mathbf{\Sigma}_{y,v}$ Then the conditional moments for y given $\vec{v}$ are

$$\mathrm{E}(y|\vec{v}) = \mu + \mathbf{\Sigma}_{y,v} \mathbf{\Sigma}_v^{-1} (\vec{v} - \vec{\mu}) \tag{9}$$

and

$$\mathrm{Var}(y|\vec{v}) = \sigma^2 - \mathbf{\Sigma}_{y,v}\mathbf{\Sigma}_v^{-1}\mathbf{\Sigma}_{v,y}, \tag{10}$$

where $\vec{\mu}$ consists of μ repeated n times.

In practice μ, σ^2, $\mathbf{\Sigma}_v$ and $\mathbf{\Sigma}_{v,y}$ are unknown and must be replaced by estimates. The threshold will be

$$C = m + t_\alpha \sqrt{\widehat{\mathrm{Var}}(y - m|\vec{v})}$$

where $m = \hat{\mathrm{E}}(y|\vec{v})$. Here m may be calculated using (9), replacing μ with either $\bar{v}$ or the a weighted average $\hat{\mu}$ discussed in Section 2.2. Similarly, the estimation of $\widehat{\mathrm{Var}}(y - m|\vec{v})$ may be based on s or $\hat{\sigma}$.

Computing the conditional means and variances requires inverting $\mathbf{\Sigma}_v$ or $\hat{\mathbf{\Sigma}}_v$, which imposes a computational burden if n is large. This can be mitigated with little loss of accuracy (as long as multiple correlations are not large) by conditioning on only a subset of the v's—those with the highest correlation with y, or spatially nearest to the observation being tested.

Preliminary numerical results comparing the conditional to unconditional thresholds are promising. These results are for the same situation as Example 1, an 8 by 8 matrix with autoregressive correlation structure. Let $y = v_{4,4}$ (from the middle of the matrix), and exclude $v_{4,4}$ from $\vec{v}$ for calculating the conditional distribution of y. To reduce the computational burden, we condition only on the eight observations adjacent to the one being tested.

Table 1 contains estimated detection probabilities, using unconditional and conditional thresholds. Detection probabilities are substantially higher using conditional thresholding, e.g. 90% instead of 23% for a signal-to-nose ratio of 3 with known moments. The first two lines are when moments are known, the last two lines are when autocorrelations and moments are estimated, using $\bar{v}$ and s^2 rather than weighted versions. For both known and estimated moments, detection probabilities are substantially higher using conditional distributions.

4 Non-Gaussian Distributions, known shape

In this section we assume that the v's and y are all independent, from a distribution F with unknown mean and standard deviation but known shape, which need not be Gaussian. While it is unlikely in practice that the shape would be known, this is useful as a starting point in investigating non-Gaussian distributions. Furthermore, a threshold obtained using a shape that is close to the true shape would be more accurate than one based on a Gaussian shape.

Let M be location statistic based on $v_1, \ldots, v_n$, e.g. the sample mean, median, or other robust statistic. A location statistic satisfies $M(a + b\vec{v}) = a + bM(\vec{v})$ for any constants a and b, and is typically an estimate of the center of F. Similarly let S be any scale

Table 1: Probability of Detection, Unconditional and Conditional Thresholds

		SNR					
Moments	Method	0	1	2	3	4	5
Known	Unconditional	0.0001	0.0033	0.0428	0.2361	0.6106	0.8999
	Conditional	0.0001	0.0201	0.3499	0.8999	0.9984	1.0000
Estimated	Unconditional	0.0000	0.0006	0.016	0.14	0.45	0.76
	Conditional	0.0000	0.0031	0.095	0.55	0.938	0.9983

Probability of detection for a signal in the [4,4] location in Example 1, for thresholds calculated without and with conditioning on the values of adjacent locations. The first two rows are for known means and covariances, and are calculated exactly. The second two use means and covariances estimated from remaining observations, and are estimated from a simulation. The number of digits for each entry reflects the standard error for that entry—standard errors are the same order of magnitude as the last digit shown.

statistic, e.g. sample standard deviation or interquartile range. A scale statistic satisfies $S(a + b\vec{v}) = bS(\vec{v})$, and typically estimates the spread of F.

An exact threshold (one for which the false alarm probability is exactly α) is given by

$$C = M + ST_\alpha$$

where T_α is the $1 - \alpha$ quantile of the distribution of

$$T = (y - M)/S. \tag{11}$$

Note that

$$P(y > c) = P(y > M + ST_\alpha) = P((y - M)/S > T_\alpha) = \alpha$$

and that the distribution of T does not depend on the unknown location and scale.

In some cases the value of T_α may be calculated exactly; for example, for the Gaussian family with $M = \bar{v}$ and $S = s$, $T_\alpha = t_{\alpha,n-1}\sqrt{1 + 1/n}$. In other cases T_α may be approximated analytically, or estimated using simulation.

In Table 2 are simulation results for three shapes of distributions—Gaussian, t with 8 degrees of freedom, and exponential. In all cases it is possible to determine thresholds such that the false alarm rate is the the desired value, 0.0001 in this case. The thresholds depend both on the shape of F, and the location and scale estimates, and are not directly comparable (a large threshold does not indicate a poor method). The probability of detection increases as the signal-to-noise ratio increases, although rather slowly for the long-tailed distributions (t and exponential).

The power of this procedure, i.e. the probability of detection when there is a signal, depend to some extent on how accurate M and especially S are. In particular, if F has

Table 2: Probability of Detection, Known Distribution Shapes

			SNR					
Shape	Method	Threshold	0	1	2	3	4	5
Gaussian	$\bar{v}, s$	4.0	0.0001	0.0027	0.032	0.178	0.503	0.826
	robust	4.5	0.0001	0.0026	0.032	0.177	0.501	0.824
t(df=8)	$\bar{v}, s$	5.9	0.0001	0.0004	0.0018	0.0102	0.056	0.227
	robust	6.7	0.0001	0.0004	0.0018	0.0102	0.056	0.228
exp.	$\bar{v}, s$	9.5	0.0001	0.0003	0.0007	0.0020	0.0055	0.0148
	robust	10.9	0.0001	0.0003	0.0007	0.0020	0.0055	0.0148

In each case (Gaussian, t, and exponential), the shape of F is known, but the location and scale are estimated from a sample of size 64. SNR is the signal-to-noise ratio (signal divided by σ). Location and scale are estimated either using mean and standard deviation, or more robust weighted versions of these. Number of digits shown are based on the standard errors of the entries.

long tails then these should be robust estimates. For example, the coefficient of variation (standard deviation divided by average) of the scale estimates is

	Gaussian	t	exponential
s	0.093	0.114	0.170
robust	0.094	0.108	0.160

The "robust" estimate is a weighted standard deviation, with smaller weights given to more extreme observations to make the estimate more robust. It is slightly less accurate for the Gaussian distribution than is the usual unweighted sample standard deviation s, but is more accurate for the other two distributions. However, there appears to be (surprisingly) little difference in Table 2 between using the simple estimates of location and scale ($\bar{v}$ and s) and more robust versions. We conjecture that this is because the distribution of T (11) is determined largely by the distribution of y than by M and S. We return to this issue in Section 4.2.

In Table 3 we summarize what happens when thresholds are calculated using the incorrect shape. If the analyst sets thresholds based on a Gaussian shape but the true shape has longer tails, then the probability of detection is high, and in particular is much higher than the desired false alarm rate when there is no signal, e.g. 12 times too high if the true shape is t and 93 times too high if the true shape is exponential. When the actual shape is Gaussian but the thresholds are based on either of the other distributions, the probability of detection is very low. We emphasize the importance of this result—if thresholds are set assuming a certain distributional shape, and the actual distributional shape is different, the actual false alarm rate could be very different from the desired rate.

Table 3: Probability of Detection, Known Distribution Shapes

Actual Shape	Threshold based on	SNR 0	1	2	3	4	5
Gaussian	Gaussian	0.0001	0.0027	0.0318	0.1779	0.5034	0.8262
	t(df=8)	0.0000	0.0000	0.0003	0.0056	0.0481	0.2151
	exp.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004
t(df=8)	Gaussian	0.0012	0.0062	0.0362	0.1768	0.5142	0.8371
	t(df=8)	0.0001	0.0004	0.0018	0.0102	0.0563	0.2265
	exp.	0.0000	0.0000	0.0000	0.0001	0.0003	0.0017
exp.	Gaussian	0.0093	0.0253	0.0688	0.1870	0.4707	0.8174
	t(df=8)	0.0018	0.0049	0.0133	0.0362	0.0983	0.2540
	exp.	0.0001	0.0003	0.0007	0.0020	0.0055	0.0148

Location and scale are estimated using mean and standard deviation of samples of size 64. Thresholds are calculated assuming that the true shape is one of three shapes (Gaussian, t, and exponential), and those thresholds are used for all three shapes. SNR is the ratio of the size of the signal divided by the standard deviation of the actual distribution. Standard errors for the table entries are comparable to those in Table 2.

4.1 Simulation Approximations

It is straightforward to use simulation to estimate T_α. The computer time required to obtain answers of a given accuracy can be reduced by the use of Monte Carlo variance reduction methods. We used "conditioning" with great effect in our simulations, and describe that procedure here.

Let

$$\begin{aligned} h(c) &= P((y-M)/S > c) \\ &= P(y > M + cS) \\ &= \mathrm{E}_{M,S}\left(P(y > M + cS | M, S)\right) \\ &= \mathrm{E}_{M,S}\left(1 - F(M + cS)\right)) \end{aligned} \tag{12}$$

Then the estimate of $h(c)$ based on conditioning and N Monte Carlo replications is is

$$\hat{h}(c) = 1 - N^{-1} \sum_{i=1}^{N} F(M_i + cS_i).$$

Note that the random distribution of y does not enter here. In effect, because M and S are relatively difficult to compute, requiring generating a whole sample of $\vec{v}$ of possibly correlated observations, we make maximum use of each pair (M, S) by matching it with an infinite number of values of y, analytically.

To estimate T_α, we solve the nonlinear equation

$$\hat{h}(c) = \alpha$$

numerically, using Newton-Raphson equation solving, keeping the N pairs (M_i, S_i) fixed.

This procedure is very effective, reducing the number of replications by a factor of about 1800. We used $N = 1000$ Monte Carlo replications. The standard error of $h(c)$ at $c = \hat{T}_\alpha$ in our simulation is $7.4E^{-6}$, $2.9E^{-6}$, and $6.2E^{-6}$, for the Gaussian distribution, t, and exponential distributions, respectively (where E^{-1} signifies 10^{-6}). Without using conditioning, the standard errors would have been $3.2E^{-4}$ in each case, and about 1.8 million replications would be necessary to reduce the standard error to $7.4E^{-6}$.

It may be practical to use the simulation method on-line, using conditioning and possibly also importance sampling. If standard errors larger than those obtained here are acceptable, then the number of replications may be reduced; e.g. $N = 10$ replications would be adequate to obtain standard errors of about $2.9E^{-5}$ for the t_8 distribution.

Note that because the distribution of T and the value of $h(c)$ do not depend on the unknown location and scale, we may use any convenient values of the location and scale parameters when doing the simulations.

4.1.1 Avoiding explicit threshold determination

Given a particular value y and corresponding M and S, it is not necessary to explicitly find a threshold T_α in order determine whether $y > M + ST_\alpha$. Instead simulation or another method can be used to estimate $h((y - M)/S)$. If this probability is smaller than α, then y exceeds the threshold, otherwise not. No Newton-Raphson equation solving is necessary in this case.

In addition, inequalities may be available to determine whether $h((y - M)/S) < \alpha$ without evaluating h. For example, thresholds based on a parametric family of t-distributions are always larger than those based on the Gaussian family. If y fails to exceed a Gaussian threshold if would also fail to exceed a t threshold, so h need not be evaluated at all.

4.2 Analytic Approximations

In this Section we obtain an analytical approximation for T_α. We begin, however, by motivating this approximation by discussing what are the most important factors in determining the distribution of T. This discussion also touches on a related issue, why it is inappropriate to use t distributions for non-Gaussian situations, and hence pointless to estimate degrees of freedom for those situations.

Our goal is to determine T_α, the $1 - \alpha$ quantile of $T = (y - M)/S$. The following proposition guides our analysis:

Proposition *the distribution of T is determined largely by the distribution of the single value y,*

because the variability in y is much greater than the variability in either M or S. M and S are less variable because they are obtained from n observations instead of just one. For example, Table 4 shows the variability in the three terms, for the Gaussian, $t(df = 8)$ and exponential distributions. The final two columns in the table will be discussed below, in Section 4.4.

Table 4: Variance of terms in T

	Threshold	Variances				
Distribution	T_α	y	$\bar{v}$	s	$T_\alpha s$	$\bar{v} - T_\alpha s$
Gaussian	4.02	1	0.015	0.008	0.136	0.149
t(df=8)	5.95	1.33	0.020	0.017	0.599	0.608
Exponential	9.55	1	0.015	0.028	2.533	2.254

The variance of y is exact. Other quantities are estimated from a simulation.

The proposition relates to an important question: can we use quantiles of t distributions as estimates of T_α, possibly with reduced degrees of freedom? This was our practice in earlier sections, where the underlying distribution was Gaussian; then y and the numerator of $T = (y - \bar{v})/s$ are Gaussian, and variability in the denominator s just smears things out slightly, causing the overall shape to be that of a t distribution.

But when F is not Gaussian, the distribution of y is not Gaussian, no matter how large n is. Therefore, *we will not use t distributions for non-Gaussian problems.* This also makes it pointless to estimate degrees of freedom for non-Gaussian problems. As a side comment for those familiar with statistics, the reason why common one-sample and two-sample t statistics may be used in non-Gaussian problems is that the numerator is an average, not a single observation. For example, a common one-sample procedure is to reject $H_0 : \mu = 0$ if $\bar{x}/(s/\sqrt{(}n)) > t_\alpha$; the numerator $\bar{x}$ has approximately a Gaussian distribution if n is large.

We can, however, obtain other approximations for T_α. As in the case of simulation approximation, we evaluate $h(c)$ approximately for any c, then solve $h(c) = \alpha$ numerically to estimate the threshold. Note that $h(c)$ is independent of the location and scale of the distribution, so without loss of generality we assume that these are known.

We describe here two methods of approximating T_α, based on the proposition. First suppose that the variability in M and S is negligible; then

$$h(c) = P(y > M + cS) \doteq P(y > \mathrm{E}(M) + c\mathrm{E}(S)) = 1 - F(\mathrm{E}(M) + c\mathrm{E}(S));$$

setting this equal to α yields a first approximation for the threshold

$$c^{(1)} = \frac{y_\alpha - \mathrm{E}(M)}{\mathrm{E}(S)}.$$

where $y_\alpha = F^{-1}(\alpha)$.

Next, let $Q = M + cS$. We use the same conditioning idea as with the simulation approximation,

$$\begin{aligned} h(c) &= P(y > Q) \\ &= \mathrm{E}_Q\left(1 - F(Q)\right)) \\ &\doteq \mathrm{E}_Q\left(1 - F(\mu_Q) - f(\mu_Q)(Q-\mu_Q) - (1/2)f'(\mu_Q)(Q-\mu_Q)^2)\right) \\ &= 1 - F(\mu_Q) + 0 - (1/2)f'(\mu_Q)\mathrm{Var}(Q) \end{aligned} \tag{13}$$

where the second equality uses the independence of y and Q (y is independent of the v's) and the approximation uses a Taylor series expansion of $1 - F(q)$ about the point $\mu_Q = \mathrm{E}(Q)$. Note that when $c = c^{(1)}$ that $\mu_Q = y_\alpha$.

The next approximation is obtained by performing a single Newton-Raphson step toward solving $h(c) = \alpha$, of the form $c^{(2)} \doteq c^{(1)} - \frac{h(c^{(1)})-\alpha}{h'(c^{(1)})}$. We have $h(c^{(1)}) \doteq \alpha - (1/2)f'(y_\alpha)\mathrm{Var}(Q)$. We also need the derivative of h; differentiating (13) with respect to c, but dropping a term involving $\mathrm{Var}(Q)$, yields $h'(c^{(1)}) \doteq -f(\mu_Q)\mathrm{E}(S) = -f(y_\alpha)\mathrm{E}(S)$, giving the second approximation

$$c^{(2)} = c^{(1)} - \frac{f'(y_\alpha)\mathrm{Var}(Q)}{2f(y_\alpha)\mathrm{E}(S)} \tag{14}$$

In this approximation, $\mathrm{Var}(Q)$ is evaluated using $Q = M + c^{(1)}S$. The approximations in our three examples are:

Distribution	Simulation	$c^{(1)}$	$c^{(2)}$
Gaussian	4.02	3.71	3.97
t(df=8)	5.95	5.64	5.97
Exponential	9.55	8.36	9.49

For this table the moments of M and S needed to estimate $\mathrm{Var}(Q)$ were obtained by simulation; in practice they would be approximated analytically.

The accuracy of (14) requires that $\mathrm{Var}(Q)$ be small. More accurate approximations are possible by keeping more terms in the Taylor series expansions, and using addition steps of the Newton-Raphson root finding procedure.

4.3 Edgeworth and Saddlepoint Approximations

Other analytical approximations are possible. Note that $h(c) = P(y - M - cS > 0)$; the quantity $H = y - M - cS$ is a linear combination, so analytical methods designed for such linear combinations may be useful in evaluating $h(c)$. Then, as before, we solve numerically to find the value c for which $h(c) = \alpha$.

One well-known method is Edgeworth approximations. These require only moments of the distribution of H. However, Edgeworth approximations are notoriously inaccurate in

the tails of distributions, as when trying to estimate small probabilities. They are also inaccurate when applied to distributions which are not approximately Gaussian, which may be the case here. We do not recommend Edgeworth approximations here.

Saddlepoint approximations are more accurate. They do require that the cumulant generating function $K(t) = \log(\mathrm{E}(e^{tH}))$ be known (this is the log of the moment generating function). Unfortunately, they cannot be used in problems where the moment generating function does not exist, such as for t-distributions. Furthermore, when variables are dependent (M and S, and possibly y) it may be very difficult to obtain the cumulant generating function. Finally, saddlepoint approximations may not be accurate enough; they work well for the distribution sample means, even for quite small samples, but the distribution of H is determined largely by the distribution of a single observation.

4.4 Relative importance of y and S

We discuss in this section the need to use caution in relying on the proposition.

The proposition indicates that the distribution of $T = (y - M)/S$ is determined largely by the distribution of y, if n is large. That is true for the most part; M and S have relatively little effect on most of the distribution of T. However, the behavior of S may play a much bigger role in obtaining the right tail, the extremely large values of T that determine the threshold T_α. In particular, small values of S make the ratio large.

Recall that $h(c) = P(y - M - cS > 0)$, and that the threshold is that value T_α for which $h(T_\alpha) = \alpha$. This means roughly that the distribution of $y - M - T_\alpha S$, and especially its $1 - \alpha$ quantile, are important for determining thresholds. When α is small, then T_α is large, e.g. $4.02, 5.95, 9.55$ when $\alpha = 1E^{-4}$ for the Gaussian, t_8, and exponential distributions, respectively. Then the variance of $T_\alpha S$ can be large relative to the variance of y, see for example Table 4, where $\mathrm{Var}(T_\alpha s)$ is 2.5 times as large as $\mathrm{Var}(y)$ for the exponential distribution.

This may indicate that it is inappropriate to rely on the proposition in determining thresholds, because the variability in S contributes substantially to the distribution of $y - M - T_\alpha S$. On the other hand, the analytical approximations $c^{(2)}$, which are based on the proposition, seem to perform quite well. It may be that an analysis based on variances overstates the importance of $T_\alpha S$ to the important tail, the right tail of $y - M - T_\alpha S$. The reason is that distribution of S is often skewed, so it may have a large variance, but still not extend very far in the left tail, and hence not supply many of the small values of S that induce large values of both $y - M - T_\alpha S$ and $(y - M)/S$. This should be investigated further.

5 Non-Gaussian Distributions, dependence

Many of the methods discussed in Section 4 extend naturally to the case where the v's are dependent, if the dependence structure is known. For any location and scale statistics M and S, the desired false alarm rate is achieved if detection occurs when $y > M + T_\alpha S$, where T_α is the $1-\alpha$ quantile of $T = (y-M)/S$. The simulation method for estimating this quantile can be used easily, providing a way to generate the random dependent samples $\vec{v}$ is available. The conditioning method for doing the simulations quickly applies easily. The analytic approximations in Section 4.2 apply as well, though estimating the moments of M and S may be more difficult.

We applied the methods of Section 4 to the case where the $n = 64$ training observations are from an 8 by 8 matrix, as in Example 1. We also used monotone transformations to obtain exponential and t_8 samples with dependence, of the form $v_{ij} = F^{-1}(\Phi(z_{ij}))$, where Φ is the standard Gaussian distribution function and z_{ij} is a Gaussian bivariate autoregressive process. The approximate thresholds are:

Distribution	Simulation	$c^{(1)}$	$c^{(2)}$
Gaussian	4.59	3.72	4.31
t(df=8)	6.50	5.89	6.54
Exponential	12.97	8.76	11.97

These thresholds are slightly larger than those for independent data, as expected. For this table the moments of M and S needed to estimate $\mathrm{Var}(Q)$ were obtained by simulation; in practice they would be approximated analytically.

The power in these examples for various signal-to-noise ratios is

		SNR					
Shape	Threshold	0	1	2	3	4	5
Gaussian	4.6	0.0001	0.0021	0.022	0.123	0.378	0.703
t(df=8)	6.5	0.0001	0.0004	0.0020	0.0112	0.056	0.200
exp.	13.0	0.0001	0.0003	0.0007	0.0020	0.006	0.013

This table parallels Table 2, except for dependent data, and using $\overline{v}$ and s only.

The case where y is dependent on the v's is beyond the scope of this discussion. This would involve conditional distributions for y given the v's, and may be intractable depending on the dependence structure.

6 Non-Gaussian Distributions, unknown shape

In this section we assume that the v's and y are all independent, from a distribution with unknown mean, standard deviation, and shape.

It is difficult or impossible to get satisfactory results for this case, if there is no information on the shape of the distributions, unless there is an extremely large amount of data available. The desired thresholds exceed the $1 - 1E^{-4}$ quantile of the distribution of individual observations; to accurately estimate such quantiles in a completely nonparametric fashion would require that n be of the order of multiple tens of thousands. We presume in our discussion that less data is available. In return, we require that some information on the shape of the distributions is available.

In this section we assume that the actual shape falls within some parametric family, with unknown location, scale, and shape (e.g. skewness and kurtosis) parameters. While it is unlikely in practice that the true distribution would fall *exactly* within this family, it should be possible to pick a family that approximates the true distribution. In particular, as an analyst gains experience with different sensor setups, he or she can learn what distributional shapes tend to arise from those sensors.

In this discussion we use a shifted gamma family as an example, with density $f(x) = c(x - x_0)^{a-1} \exp(-(x - x_0)/b)$ for $x > x_0$, where c is a normalizing constant and x_0, a, and b are parameters. The skewness is $2a^{-1/2}$. The family can be extended to negative skewness by replacing x with $(-x)$ about zero.

Let θ denote the shape parameter, which may be vector-valued (e.g. skewness and kurtosis). Let $\hat{\theta}$ be an estimate of the shape parameter, and M and S be location and scale statistics.

We now describe three method for setting quantiles:

- plug-in,
- bias-corrected plug-in, and
- Bayesian and shrinkage methods.

6.1 Plug-in approach

The "plug-in" approach substitutes a shape parameter estimated from a training sample for the unknown shape parameter, then proceeds as in Section 4. Let $T_\alpha(\theta)$ be the $1 - \alpha$ quantile of T if θ were known. One example is shown in the left panel of Figure 6.1. The "plug-in" approach substitutes $\hat{\theta}$ for the unknown θ, yielding the threshold

$$M + T_\alpha(\hat{\theta})S. \tag{15}$$

$T_\alpha(\hat{\theta})$ could be computed on-line after $\hat{\theta}$ is estimated, by simulation or analytic methods. Or the quantile $T_\alpha(\theta_j)$ could be pre-computed and stored for selected values θ_j, $j = 1, \ldots$, after which $T_\alpha(\hat{\theta})$ can be quickly obtained on-line by interpolation. We use the latter method.

There are two potential problems with the plug-in approach: variance, and bias. Shape parameters are often difficult to estimate, particularly because shape parameter estimates

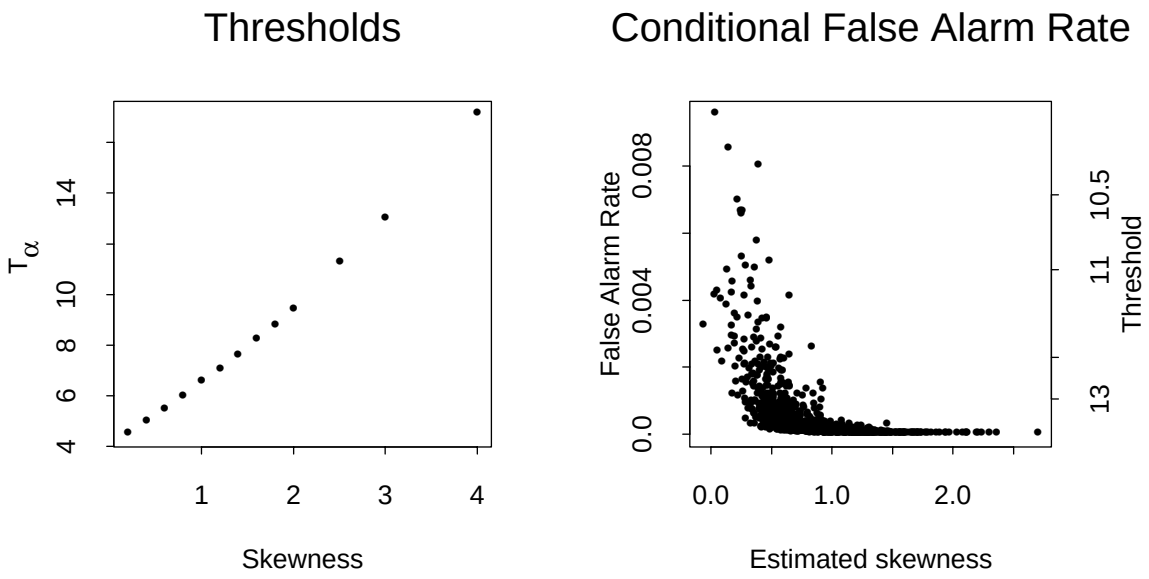

Figure 1: In the left panel, for each value of skewness, T_α is estimated from 1000 samples of $n = 64$ each, with $M = \bar{v}$ and $S = s$, from the shifted gamma family of distributions. In the right panel, each point corresponds to a training sample of 64 observations. The conditional false alarm rate is the probability that a new observation exceeds the threshold estimated from the sample, using the plug-in approach.

often depend on higher moments such as sample skewness and kurtosis, and these are notoriously difficult to estimate from data. For instance, for the shifted gamma family, the mean and standard deviation of the skewness from samples of size 64 (estimated using simulation with 1000 samples), and of the corresponding shape parameters, are given in Table 5. The estimated skewness tends to be smaller than the true skewness, the estimated shape parameters tend to be larger than the true values, and the standard deviations are substantial, sometimes greater than the averages.

Table 5: Randomness of shape parameters $\hat{\theta}$

θ = Skewness	0.2	0.6	1	1.4	2	3	4
mean($\hat{\theta}$)	0.17	0.52	0.85	1.23	1.67	2.32	2.99
std($\hat{\theta}$)	0.29	0.33	0.37	0.50	0.57	0.72	0.91
$\theta = a = 4\text{Skewness}^{-2}$	100	11.11	4.00	2.04	1.00	0.44	0.25
mean($\hat{\theta}$)	5.2E4	7.2E3	26.41	4.20	1.97	0.95	0.57
std($\hat{\theta}$)	9.8E5	1.6E5	238.70	4.28	1.48	0.53	0.32

The large variance of the estimated shape parameters causes substantial variance in the thresholds actually used. The variance and bias both contribute to incorrect false alarm rates. For instance, when the true skewness is 1.0, $T_\alpha = 6.5$, the ideal threshold would be $y_\alpha = 15.9$. If the skewness were known, the thresholds $C = \bar{v} + T_\alpha(\theta)s$ have mean 16.9, variance 2.8, and false alarm rate $\alpha = 1E^{-4}$. With skewness unknown, the plug-in thresholds $C = \bar{v} + T_\alpha(\hat{\theta})s$ have mean 16.2 and variance 10.2, and false alarm rate $4.5E^{-4}$, four and a half times the desired rate. The high false alarm rate is due primarily to those training samples in which the skewness was underestimated, which tend to have smaller thresholds and higher conditional false alarm rates; as shown in the right panel of Figure 6.1.

6.2 Bias-corrected plug-in approach

One remedy for the bias in false alarm rates is to adjust the thresholds. In particular, instead of using the plug-in thresholds $C = M + ST_\alpha(\hat{\theta})$, we suggest thresholds

$$C = M + ST^*_\alpha(\hat{\theta}) \tag{16}$$

where $T^*_\alpha(\hat{\theta})$ is an adjusted version of $T_\alpha(\hat{\theta})$, chosen to give the approximately the desired false alarm rate over a range of values of θ.

For example, one adjustment has $T^*_\alpha(\hat{\theta}) = q(T_\alpha(\hat{\theta}))$ for some smooth monotone increasing function q, which satisfies $q(t) \geq t$. The choice $q(t) = 1.2 + t$ gives the desired estimated false alarm rate in the previous example, when the true skewness is 1, and $q(t) = -3.5 +$

$2.58t - .13t^2$ does so for skewness = 0.4, 1, and 1.6. The coefficients of the polynomial q were found by a numerical root-finding procedure in S-PLUS.

6.3 Bayesian and Shrinkage methods

The high variability in shape estimates $\hat{\theta}$, and corresponding variability in thresholds, can be remedied by making the estimates less variable. Ideally, this parameter could be estimated using more data, a larger n. Unfortunately, the size of the training sample is fixed. But other information may be available, in particular experience with shape parameters from past images from the same sensor. We describe here a Bayesian procedure which combines prior information and information from the training sample.

Let $\pi(\theta)$ be a prior density for the shape parameter θ. This should encompass past experience with the shape parameters. Basically this says that even before we see the training sample, the estimated probability that the shape parameter falls in the range (a, b) is $\int_a^b \pi(t)dt$. Let $f(\hat{\theta}|\theta)$ be the density for $\hat{\theta}$ given θ; the mean and variance of these distributions for various values of θ are given in Table 5 for the shifted gamma family. Then by Bayes theorem, the posterior distribution that combines the prior information and the information from the sample is

$$\pi(\theta|\hat{\theta}) = \frac{\pi(\theta) f(\hat{\theta}|\theta)}{\int_{-\infty}^{\infty} \pi(t) f(\hat{\theta}|t) dt}. \tag{17}$$

Then, the probability of a false alarm is

$$\begin{aligned} P(y > M + Sc|\hat{\theta}) &= \mathrm{E}_{\theta|\hat{\theta}}(P(y > M + Sc|\theta)) \\ &= \int_{-\infty}^{\infty} \pi(\theta|\hat{\theta}) P(y > M + Sc|\theta)) d\theta. \end{aligned}$$

The value of $c = c(\hat{\theta})$ for which the false alarm probability equals α can be found by a numerical root-finding procedure. Finally, the threshold is

$$M + Sc(\hat{\theta}).$$

The thresholds $c(\hat{\theta})$ can be computed off-line for various values of $\hat{\theta}$, then interpolation can be used on-line. The computations should not be very expensive. The inner conditional probability $P(y > M + Sc|\theta)$ can be estimated by conditional simulation, and the integral evaluated numerically. A small number of replications would suffice for each conditional simulation, if different random numbers are used for each θ and the numerical integration procedure is of the form $\int g(x)dx = \sum w_i g(x_i)$ with $\sum w_i = 1$ and each w_i positive and roughly equal in value, because then simulation errors tend to average out; the variance of the result is $\sum w_i^2 \mathrm{Var}(\hat{g}(x_i))$.

Note that the prior distribution $\pi(\theta)$ is a distribution for θ, not for $\hat{\theta}$. The idea is that associated with each image is a true θ, which is estimated with some variability and bias

by $\hat{\theta}$. The values of θ are in fact unobservable, but would have (much) smaller variance than does $\hat{\theta}$. For example, if both the bias and variance of $\hat{\theta}$ given θ are constant, then $\mathrm{Var}_\pi(\theta) = \mathrm{Var}(\hat{\theta}) - \mathrm{Var}(\hat{\theta}|\theta)$; here $\mathrm{Var}(\hat{\theta})$ could be estimated from historical data and $\mathrm{Var}(\hat{\theta}|\theta)$ analytically or by simulation.

π could be chosen subjectively, or chosen based on historical data. Suppose that historical observations $\hat{\theta}_j$ are available, $j = 1, \ldots, J$. Here are three ways of setting π:

- A quick-and-dirty prior, useful if the conditional variance $\mathrm{Var}(\hat{\theta}|\theta)$ is constant, is Gaussian with mean $\overline{\theta} = J^{-1} \sum \hat{\theta}_j$ and variance $(J-1)^{-1} \sum (\hat{\theta}_j - \overline{\theta})^2 - \mathrm{Var}(\hat{\theta}|\theta)$.

- If the conditional variance $\mathrm{Var}(\hat{\theta}|\theta)$ is not constant, then a variance-stabilizing transformation may be used to make the conditional variance constant. Let η, be related to θ by a monotone nonlinear transformation chosen so that $\mathrm{Var}(\hat{\eta}|\eta)$ is constant. An approximate variance-stabilizing transformation is given by the indefinite integral (evaluated numerically) of $\frac{d\eta}{d\theta} = \mathrm{Var}(\hat{\theta}|\theta)^{-1/2}$. Then the quick-and-dirty Gaussian prior can be assigned to η.

  There is another argument for this nonlinear transformation. It is an empirical rule that variance-stabilizing transformations usually improve normality. Typically that would mean that the distribution of $\hat{\eta}$ given η would be more Gaussian. It could also mean that a Gaussian prior is more reasonable for η than for θ. It is easy to construct examples for which this is true.

- The third approach to setting π is iterative. Let π^0 be a first guess for the prior distribution, then let

$$\pi^{k+1}(\theta) = J^{-1} \sum_{j=1}^{J} \pi^k(\theta|\hat{\theta}_j).$$

  This should converge fairly quickly (though we have not yet tried this).

Here is an artificial example for which analytical results are available; this is valuable for the insight it provides. Suppose the $\pi(\theta)$ is a Gaussian distribution with mean μ_θ and variance σ_θ^2, and that $f(\hat{\theta}|\theta)$ is a Gaussian distribution with mean $\theta + b$ (bias b) and variance $\sigma_{\hat{\theta}}^2$. Then the posterior distribution $\pi(\theta|\hat{\theta})$ is Gaussian with mean $\frac{\sigma_{\hat{\theta}}^2 \mu_\theta + \sigma_\theta^2(\hat{\theta}-b)}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}$ and variance $\frac{\sigma_{\hat{\theta}}^2 \sigma_\theta^2}{\sigma_{\hat{\theta}}^2 + \sigma_\theta^2}$. The mean is of the form $\lambda\mu_\theta + (1-\lambda)(\hat{\theta} - b)$, and corresponds to subtracting the bias b from $\hat{\theta}$, then shrinking a fraction λ of the distance toward μ_θ. The variance of the result is smaller than both the variance of the prior distribution and the variance of the parameter estimate, and so reflects the increased accuracy that results from combining information.

This motivates a simpler alternative to the Bayesian formulation—to shrink the parameter estimate $\hat{\theta}$ a fraction λ of the distance toward a pre-specified point θ_0, reducing variability

by a fraction $(1-\lambda)^2$. The threshold would be

$$M + T_\alpha(\lambda\theta_0 + (1-\lambda)\hat{\theta}).$$

In effect, shrinkage estimates provide a compromise between assuming that θ is known and unknown; the former leads to thresholds that are less variable but biased if θ is not equal to the hypothesized value, while the latter gives results that are highly variable and therefore also biased. θ_0 could be the average value of $\hat{\theta}$ from historical data, and λ chosen to optimize the results when applied to the historical data.

7 Non-Gaussian Distributions, dependence and unknown shape

Many of the ideas discussed in Section 6 extend naturally to the case where the v's are dependent, if the dependence structure given θ is known. For instance, the bias-corrected plug-in method would use any location and scale statistics M and S, and a detection would occur if $y > M + T^*_\alpha(\hat{\theta})S$. The process of obtaining T^* would parallel the process used for independent data, beginning with the calculation of $T_\alpha(\theta)$ (e.g. using conditional simulation), then adjusting.

If the dependence structure is unknown, then two approaches are possible. One is to assume that the dependence structure falls within a parametric family. For example, the data could obtained by a monotone transformation of the form $v_{ij} = F^{-1}(\Phi(z_{ij}))$, where z_{ij} is a Gaussian bivariate autoregressive process with unknown autocorrelation parameter(s); then the correlation parameter(s) can be treated like shape parameters, with similar methods applied, including bias-correction for plug-in estimates, and Bayesian methods. Another approach uses a version of bootstrapping with nonparametric dependence structure but individual values generated consistently with a parametric family.

The case where y is dependent on the v's is beyond the scope of this discussion. This would involve conditional distributions for y given the v's, and may be intractable depending on the dependence structure.

8 Parametric Families and Maximum Likelihood

Much of the preceding discussion involves parametric families. To obtain good results, it is important that the parametric family in use approximate the real distribution, particularly in the extreme tail of the distribution. The use of flexible families, with shape parameters estimated from the training sample and from historical data, makes it easier to achieve this.

Note that the family chosen will be "wrong," in that the actual distribution will *not* be a member of the family. In the words of George Box, "All models are wrong, but some are

useful." Here the family is useful if it approximates the shapes of distributions observed in practice, particularly in the desired tail.

Because the parametric family is "wrong", it is important to use parameter estimation methods which are not sensitive to features of the parametric family. In practice, this means to avoid maximum likelihood (ML) estimation methods, which tend to be very sensitive to model violations and odd distributional features implied by a particular family.

For instance, the shifted gamma family has the feature that $P(v_j < x_0) = 0$. ML estimates of the parameters would never have $\hat{x}_0 > \min(v_j)$, the smallest observation in the training sample. Indeed when the ML estimate of the skewness is less than 2, the ML estimate is exactly that observation; the ML parameter estimate is completely determined by $\min(v_j)$, even though that observation provides very little information about the right tail of the distribution.

In practice the true distribution may have no lower limit, but we may still wish to use the shifted gamma family, because it fits the actual distribution well, particularly in the right tail. Parameter estimation methods other than ML may produce estimates which are "impossible" according to the model, but which fit the data better for our purposes.

Method of moments estimation is robust to features in the model, but can be highly variable with long-tailed distributions. This is particularly a problem for shape estimates determined by sample skewness and kurtosis. Other robust procedures should be considered.

9 Conclusion

The general method used in this discussion is to let thresholds be of the form $M+ST_\alpha$, where M and S are location and scale estimates, and T_α is the $1-\alpha$ quantile of the distribution of $T = (y - M)/S$. This method gives exactly the desired false alarm rate for a family with known shape, whether or not Gaussian, and can be extended to families with unknown shape by various methods; of these, the bias-corrected and Bayesian methods promise to give nearly the correct false alarm rates.

In contrast, some other approaches give actual false alarm rates much higher than desired. For example, the approach of estimating parameters of a distribution, substituting those parameters for the unknown parameters, and using the $1-\alpha$ quantile of the result as a threshold gives incorrect coverage. In the case of Gaussian distributions, the resulting quantile would be $\bar{v}+z_\alpha s$, where z_α is the $1-\alpha$ quantile of a standard Gaussian distribution. Contrast that to (2). When $n = 64$ and $\alpha = 1E^{-4}$, the actual false alarm rate with this substitution approach is 23 times the desired value.

The parametric approach used here is sensitive to a poor choice of parametric family, that does not match the true distribution. The choice of family can be refined with experience. Nonparametric approaches do not seem feasible, if thresholds are to be determined from relatively small amounts of data, for small false alarm rates.

Approaches outlined here should be computationally feasible, with moderate requirements for off-line computation and small requirements for on-line computation.

Methods outlined here extend naturally to the case of dependent observations in the training sample, as long as the dependence structure given the shape parameters is known.

For dependence between the test observation and the training sample, we discussed one promising method for multivariate Gaussian situations, based on conditional distributions. The conditional distribution idea should also be used for non-Gaussian situations. That is a topic for future investigation. Depending on the nature of the dependence structure, other methods described here may or may not extend easily to this situation.

References

Hesterberg, T. C. (1998). Comments on the CFAR literature. Research department, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109. Unfinished.

Scheffè, M., Blane, M. M., and Cooper, D. B. (1994). Multicovariance Matched Filter for Target Detection and Background Recognition. In *Signal and Data Processing of Small Targets*, volume 2235 of *SPIE Proceedings*, pages 203–218.

Singer, P. F. and Sasaki, D. M. (1995). The Heavy Tailed Distribution of a Common CFAR Detector. In *Signal and Data Processing of Small Targets*, volume 2561 of *SPIE Proceedings*, pages 124–140.

Singer, P. F. and Sasaki, D. M. (1996). Estimating the Degrees of Freedom for a Common CFAR Detector. In *Signal and Data Processing of Small Targets*, volume 2759 of *SPIE Proceedings*, pages 90–97.