Variance estimations in the presence of intermittent interference and their applications to incoherent scatter radar signal processing

Qihou Zhou\(^1\), Yanlin Li\(^1\), and Yun Gong\(^2\)

\(^1\)Department of Electrical and Computer Engineering, Miami University, Oxford, Ohio, USA
\(^2\)School of Electronic Information, Faculty of Information Sciences, Wuhan University, Wuhan, China

Correspondence: Qihou Zhou (zhouq@miamioh.edu)

Received: 27 January 2024 – Discussion started: 14 February 2024
Revised: 28 April 2024 – Accepted: 26 May 2024 – Published: 17 July 2024

Abstract. We discuss robust estimations for the variance of normally distributed random variables in the presence of interference. The robust estimators are based on either ranking or the geometric mean. For the interference models used, estimators based on the geometric mean outperform the rank-based ones in both mitigating the effect of interference and reducing the statistical error when there is no interference. One reason for this is that estimators using the geometric mean do not suffer from the “heavy tail” phenomenon like the rank-based estimators do. The ratio of the standard deviation over the mean of the power random variable is sensitive to interference. It can thus be used as a criterion to combine the sample mean with a robust estimator to form a hybrid estimator. We apply the estimators to the Arecibo incoherent scatter radar signals to determine the total power and Doppler velocities in the ionospheric E-region altitudes. Although all the robust estimators selected deal with light contamination well, the hybrid estimator is most effective in all circumstances. It performs well in suppressing heavy contamination and is as efficient as the sample mean in reducing the statistical error. Accurate incoherent scatter radar measurements, especially at nighttime and at E-region altitudes, can improve studies of ionospheric dynamics and composition.

1 Introduction

In radar signal processing and in many other applications, the data samples can often be modeled as a constant superimposed on a normally distributed random variable. The variance of the random process is an important parameter in such applications. In some cases, the variance represents the power of the undesired noise. In other cases, the variance is the desired signal power, such as in our study here on incoherent scatter radar (ISR) signals. Our broad objective is to explore methods that estimate the variance in a normally distributed random variable accurately in the presence of interference. The general problem falls under robust statistics (e.g., Huber and Ronchetti, 2009; Wilcox, 2017). Specifically, we attempt to optimize ISR signal processing using robust estimators.

An ISR, with a large aperture and high transmitting power, measures the electron concentration and other state variables in the ionosphere. Its versatility makes it the most important ground-based instrument for ionospheric studies. Several major ISRs started operation in the 1960s. Readers are referred to Evans (1969) for the principle, capabilities, and comparisons of the early facilities. An ISR typically transmits a binary-phase code to increase the signal-to-noise ratio. The received signal consists of sequences of altitude-dependent in-phase and quadrature voltage samples, which, upon decoding, can be used to obtain a variety of ionosphere parameters such as electron density and electron and ion temperatures (e.g., Zhou et al., 1997; Isham et al., 2000; Hysell et al., 2014). An essential characteristic of the voltage samples is that they are normally distributed, with the variance proportional to the electron density at the corresponding altitude. Because an ISR measures the tiny amount of power scattered off the electrons and ions in space, averaging over 1000 samples is essential to derive ionospheric parameters. In the absence of interference, a simple arithmetic average of the voltage samples squared provides the best estimator for the total power or power spectral density estimates, which
2 Characteristics and comparison of mean power estimators

2.1 Signal and interference models

Let $X$ be an independent identically distributed normal random variable having $N = N_1 N_2$ data samples organized as

$$
X = \left\{ \begin{array}{c}
  x_{11} \cdots x_{1N_1} \\
  \vdots \quad \vdots \\
  x_{N_21} \cdots x_{N_2N_1}
\end{array} \right\}.
$$

For radar and many other digital sampling systems, $X \sim N(0, \sigma^2)$ can be regarded as voltage samples. $Y = \{ x_{11}^2 + \cdots + x_{1N_1}^2, x_{1N_1+1}^2 + \cdots + x_{N_1N_1+1}^2, \ldots, x_{N_21N_2}^2 + \cdots + x_{N_2N_1N_2}^2 \}$ represents the power random variable with $N_2$ elements. Each element in $Y$ is a sample mean of $N_1$ raw power variables, $X^2$.

The expectation of $Y_i$ is $\sigma_i^2$, which is the variance of $X$. We strive to estimate $\sigma_i^2$ most accurately, given samples of $X$.

As there are many types of variances, we will call estimating $\sigma_i^2$ “power estimation” to be specific and to minimize confusion. In the absence of interference, $Y_i$ can be shown to have a gamma probability density distribution (pdf):

$$
f(y; \frac{N_i}{2}, \frac{2\sigma_i^2}{N_i}) = \frac{y^{\frac{N_i}{2}-1}e^{-\frac{y}{2\sigma_i^2}}}{\Gamma\left(\frac{N_i}{2}\right)} \left(\frac{N_i}{2\sigma_i^2}\right)^{\frac{N_i}{2}},
$$

where $\frac{N_i}{2}$ and $\frac{2\sigma_i^2}{N_i}$ are the shape and scale parameters, respectively, and the support of $y$ is $(0, \infty)$ (e.g., Wikipedia, 2024b). The corresponding cumulative distribution function is

$$
F(y; \frac{N_i}{2}, \frac{2\sigma_i^2}{N_i}) = \frac{1}{\Gamma\left(\frac{N_i}{2}\right)} \int_0^y t^{\frac{N_i}{2}-1} e^{-\frac{t}{2\sigma_i^2}} dt,
$$

where $\gamma(x, y) = \int_0^x t^{y-1} e^{-t} dt$ is the lower incomplete gamma function. Distribution function $f(y)$ can also be viewed as a $N_1$-degree chi-squared distribution scaled by $N_1$. The variance of $Y_i$ is $\frac{2\sigma_i^2}{N_i}$. The distribution functions are $N_1 = 1, 2, 8$, which we will study in more detail, are $f(y; \frac{1}{2}, 2\sigma_i^2) = \frac{\sqrt{\pi}}{\sqrt{\pi} \sigma_i^2} y^{\frac{1}{2}} e^{-\frac{y}{2\sigma_i^2}}$, $f(y; 1, \sigma_i^2) = \frac{e^{-\frac{y}{\sigma_i^2}}}{\sigma_i^2}$, and $f(y; 4, \sigma_i^2) = \frac{2^\gamma y e^{-\frac{y}{\sigma_i^2}}}{3\sigma_i^2}$, respectively. At large $N_1$, the pdf is approximately normal, $f(y; \frac{N_i}{2}, \frac{2\sigma_i^2}{N_i}) \sim N(\sigma_i^2, \frac{2\sigma_i^2}{N_i})$. Of particular interest is the case of $N_1 = 2$, which corresponds to the in-phase and quadrature samples in a radar system. The interference is also modeled as a gamma distribution with a shape parameter of $k = 4$ and scale parameter $(a_0 \sigma_0)^2 / k$, which has a mean of $a_0^2 \sigma_0^2$. Since we are mainly concerned with the signal shape parameter being $1/2$ and 1, a larger shape parameter in the interference model makes it easier to differentiate between interference and signal, as the interference is more concentrated around a higher mean value. The interference is equally likely to occur at each data point, with a probability of $p_i = 0.01$, and is always additive to the signal. The total interference power relative to the signal power is thus $p_i a_i^2$. We will mainly consider three cases of interference with $a_i = 2, 6, 18$ to represent low, moderate, and strong interference, respectively.

2.2 Estimators and their characteristics in the absence of interference

The most common estimators are the sample mean, geometric mean, and median. The sample mean of $Y$ is the arithmetic average of $N_2$ samples, i.e., $\bar{Y}_i = \frac{1}{N_2} \sum_{i=1}^{N_2} Y_i$, where $Y_i$ is the sample mean of $X^2$ averaged over $N_i$ samples. With a known shape parameter, the sample mean is the uniformly minimum-variance unbiased estimator (UMVUE) and maximum likelihood estimator (e.g., Siegrist, 2022; Wikipedia, 2024b).
2024a). The geometric mean, $G_N = \left( \prod_{i=1}^{N_2} Y_i \right)^{1/N_2}$, and median, $D_N = \text{med}(Y_1, \ldots Y_{N_2})$, are more resistant to outliers but not effective in reducing the statistical fluctuations. Although the three basic estimators are largely at the opposite ends of efficiency vs. robustness, they can serve as building blocks for other estimators. In the following, we discuss the three basic estimators and compare them with a weighted mean, a hybrid estimator, and two trimmed estimators.

The effectiveness of a power estimator, $Z$, in reducing the statistical fluctuation is measured by the normalized variance

$$R^2(Z) = \frac{N \sigma^2_Z}{2 \mu^2_Z},$$

(3)

where $\sigma^2_Z$ and $\mu_Z$ are the variance and mean of the power estimator while the absolute error is of importance in some cases as well. For the sample mean estimator, $A_N$, its distribution is expressed by Eq. (1) with $N_1$ replaced by $N$. $E(A_N) = \sigma^2_N$ and the variance is $2 \sigma^2_N / N$. The theoretical expectation of $R^2(A_N)$ is thus 1 for the sample mean, which is the lowest that one can obtain. The inverse of $R^2(Z)$ is the efficiency of the estimator. It is of interest to note that since $N$ averages can be expressed as the weighted means of $N_1$ and $N - N_1$ samples, it follows that the convolution of two gamma distributions remains a gamma distribution. This convolution invariance property is also true of most commonly used distributions, including binomial, Poisson, normal, and chi-squared distributions. In general, if the distribution function of the sum or mean remains the same type for different numbers of samples, it is convolution invariant.

The median and its variance do not appear to have a closed form for $N_1$ and $N_2$ in general, although there are closed forms for specific $N_1$ and large $N$. Here we derive the theoretical results for $N_1 = 1, 2, 8, and large N$. For large $N_2$ and an ascending ranking order $K$ relatively close to $N_2/2$, Zhou et al. (1999) show that ranking has an asymptotic normal distribution, with a variance of $\sigma^2_{\text{rank},K} = k/(N_2 f(\mu))$, where $\mu_N$ is the ranking value (e.g., $K = N_2/2$ for median) and $f(\mu)$ is the pdf for the rank random variable, i.e., Eq. (1) for our study here. For the median estimator, the normalized variance is

$$R^2(D_N) = \frac{N_1}{8 f^2(\mu; N_1/2, 2/N_1) \mu^2_2}. $$

(4)

The median can be solved from $F(\mu_2) = 1/2$. For $N_1 = 1$, the median is $2 \text{erf}^{-1}(1/2) \sigma_0^2 = 0.4549 \sigma_0^2$, where erf is the inverse of the error function $\frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. For $N_1 = 2$, the median is $\mu_2 = \sigma_0^2 \ln 2 = 0.6931 \sigma_0^2$. The median for $N_1 = 8$ is $0.9180 \sigma_0^2$, which can be solved from $\gamma(4, 4 \mu_2) = 3$. For large $N_1$, the pdf tends toward normal and the median tends toward $\sigma_0^2$. The $R^2(D_N)$ values for $N_1 = 1, 2, 8, and 100$ and $N = 10000$ are 2.7206, 2.0814, 1.6848, and 1.5760, respectively. In the limiting case of $N_1$ and $N_2$ tending toward infinity, $R^2(D_N) = \pi/2$, indicating that it takes $\pi/2$ times the number of samples for the median operator to achieve the same error as the sample mean. Zhou et al. (1999) also show that taking the 79.7% largest value gives the smallest $R^2$ at 1.5432. (In Zhou et al., 1999, $\pi/2$ in Eqs. 24 and 26 should have been $2/\pi$.)

In Table 1, we list the $R^2$ values and the absolute errors for eight estimators in the null-interference case. The second column is the mean of each estimator without scaling for $\sigma_0 = 1$ (the mean is proportional to $\sigma_0^2$). To compare the different estimators on the same scale, the mean is divided by the respective estimator so that all the estimators in all the cases have a mean of 1 for all subsequent computations in the other columns in Tables 1 and 2. The values not in parentheses listed in the two tables are at least 10000 Monte Carlo simulations with $N = 10000$ for all estimators except $H_N$. The values in parentheses in Table 1 are theoretical predictions that we can derive.

The mean and variance of the geometric mean ($G_N$) can be obtained by first finding the expectation and variance of one element, $Y_1^{1/N_2}$, in the product. The expectation of $Y_1^{1/N_2}$ is

$$E\left( Y_1^{1/N_2} \right) = \int_0^\infty y^{1/N_2} f(y) dy = \frac{\Gamma\left(\frac{N_1}{2} + \frac{1}{N_2}\right)}{\Gamma\left(\frac{N_2}{2}\right)} \left( \frac{2\sigma_0^2}{N_1} \right)^{1/2}. $$

(5)

The second moment of $Y_1^{1/N_2}$ is

$$E\left( Y_1^{2/N_2} \right) = \int_0^\infty y^{2/N_2} f(y) dy = \frac{\Gamma\left(\frac{N_1}{2} + \frac{2}{N_2}\right)}{\Gamma\left(\frac{N_2}{2}\right)} \left( \frac{4\sigma_0^4}{N_1^2} \right)^{1/2}. $$

(6)

Assuming that $Y_i$’s are independent, the expectation, second moment, and variance of the geometric mean are, respectively,

$$E(G_N) = \left( E\left( Y_1^{1/N_2} \right) \right)^{N_2} = \frac{\Gamma\left(\frac{N_1}{2} + \frac{1}{N_2}\right)}{\Gamma\left(\frac{N_2}{2}\right)} 2\sigma_0^2 $$

(7)

$$E(G_N^2) = \left( E\left( Y_1^{2/N_2} \right) \right)^{N_2} = \frac{\Gamma\left(\frac{N_1}{2} + \frac{2}{N_2}\right)}{\Gamma\left(\frac{N_2}{2}\right)} 4\sigma_0^4 $$

(8)

$$\text{Var}(G_N) = \frac{4\sigma_0^4}{N_1^2} \left( \frac{\Gamma\left(\frac{N_1}{2} + \frac{1}{N_2}\right)}{\Gamma\left(\frac{N_2}{2}\right)} - \frac{\Gamma\left(\frac{N_1}{2} + \frac{2}{N_2}\right)}{\Gamma\left(\frac{N_2}{2}\right)} \right). $$

(9)

The normalized variance for the geometric mean, $R^2(G_N)$, is thus

$$R^2(G_N) = \frac{N_1 N_2}{2} \left[ \frac{\Gamma\left(\frac{N_1}{2}\right) \Gamma\left(\frac{N_1}{2} + \frac{1}{N_2}\right)}{\Gamma^2\left(\frac{N_2}{2}\right)} \right]^{N_2} - 1. $$

(10)

This equation is precise for all $N_1$ and $N_2$. $E(G_N)$ and $R^2(G_N)$ values for $N = 10000$ and $N_1 = 1, 2, 8,$ and 100 are listed in Table 1. We are not aware of a precise distribution function for $G_N$ in general. For the asymptotic case of
Table 1. Monte Carlo simulations and theoretical values (in parenthesis) of the mean, $R^2$, and absolute error for eight estimators when there is no interference.

| Method | $N_1$ | Mean (theory) | $R^2$ | $|\text{Error}|$ |
|--------|-------|---------------|-------|-----------|
| $A_N$  | 1     | 1.0000 (1)    | 1.0020 (1) | 0.0113 |
|        | 2     | 1.0000 (1)    | 0.9994 (1) | 0.0112 |
|        | 8     | 1.0000 (1)    | 1.0077 (1) | 0.0113 |
|        | 100   | 1.0000 (1)    | 0.9945 (1) | 0.0113 |
| $D_N$  | 1     | 0.4549; (0.4549) | 2.7149; (2.7206) | 0.0186 |
|        | 2     | 0.6930 (ln2); | 2.0927; (2.0814) | 0.0162 |
|        | 8     | 0.9176 (0.9180) | 1.6980; (1.6848) | 0.0147 |
|        | 100   | 0.9917 (1) | 1.5614; (1.5760) | 0.0150 |
| $G_N$  | 1     | 0.2808; (0.2808) | 2.4841; (2.4672) | 0.0178 |
|        | 2     | 0.5615; (0.5616) | 1.6487; (1.6447) | 0.0144 |
|        | 8     | 0.8780; (0.8780) | 1.1377; (1.1352) | 0.0120 |
|        | 100   | 0.9901; (0.9901) | 1.0028; (1.0100) | 0.0114 |
| $T_{95}$ | 1     | 0.7589; (0.7590) | 1.1480; (1.1423) | 0.0121 |
|        | 2     | 0.8424; (0.8430) | 1.0901; (1.0898) | 0.0117 |
|        | 8     | 0.9317; (0.9320) | 1.0434; (1.0431) | 0.0116 |
|        | 100   | 0.9839; (0.9835) | 1.0198; (1.0178) | 0.0114 |
| $T_{\text{MAD8}}$ | 1     | 0.8742 | 1.6973 | 0.0147 |
|        | 2     | 0.8425 | 1.0769 | 0.0117 |
|        | 8     | 1.0000 | 1.0075 | 0.0113 |
|        | 100   | 1.0000 | 0.9984 | 0.0113 |
| $T_{\text{GEO4}}$ | 1     | 0.9979 | 1.0185 | 0.0114 |
|        | 2     | 0.9884 | 1.0763 | 0.0117 |
|        | 8     | 0.9987 | 1.0210 | 0.0114 |
|        | 100   | 1.0000 | 0.9984 | 0.0113 |
| $W_N$  | 1     | 0.9576 | 1.0419 | 0.0115 |
|        | 2     | 0.9563 | 1.0431 | 0.0115 |
|        | 8     | 0.9888 | 1.0167 | 0.0114 |
|        | 100   | 1.0000 | 0.9995 | 0.0113 |
| $H_N$  | 1     | 0.9576 | 1.0102 | 0.0360 |
|        | 2     | 0.9563 | 1.0178 | 0.0356 |
|        | 8     | 0.9888 | 1.0001 | 0.0357 |
|        | 100   | 1.0000 | 1.0052 | 0.0358 |

large $N_2$, Zhou et al. (1999) show that the geometric mean tends toward the normal distribution, with the variance as

$$\text{Var}(G_N)_{N_2 \to \infty} = \frac{E^2(G_N)\sigma_{ln}^2}{N_2}, \quad (11)$$

where $\sigma_{ln}^2$ is the variance of ln(y). $\sigma_{ln}^2$ is known to equal the trigamma function ($\psi_1(\frac{N_1}{2})$; e.g., Wikipedia, 2024a). Thus,

$$R^2(G_N)_{N_2 \to \infty} = \frac{N_1\sigma_{ln}^2}{2} = \frac{N_1}{2}\psi_1\left(\frac{N_1}{2}\right). \quad (12)$$

For the trigamma function, $\psi_1(\frac{1}{2}) = \frac{\pi^2}{6}$, $\psi_1(1) = \frac{\pi^2}{6}$, and other $\psi_1\left(\frac{N_1}{2}\right)$ values can be found from the recurrence relation $\psi_1(z + 1) = \psi_1(z) - \frac{1}{z^2}$. The asymptotic $R^2(G_N)$ for $N_1 = [1, 2, 8, 100]$ and $N_2 = 10000$ is [2.4674, 1.6449, 1.3529, 1.0101], respectively. They are accurate to the third decimal place compared to the exact values obtained from Eq. (10) for $N_2 = 10000$. For large $N_1$ and $N_2$, $R^2(G_N) \sim 1 + \frac{1}{N_1}$, which gives the number of initial averages, $N_1$, needed to achieve a certain level of efficiency for the geometric mean. The expectation of $G_N$ for large $N_2$ is found to be

$$E(G_N)_{N_2 \to \infty} \sim \sigma_0^2 \left(1 + \frac{2}{N_1N_2}\right) e^{-\frac{1}{N_1} - \frac{1}{N_2} - \frac{1}{N_1N_2}} \sim \sigma_0^2 e^{-\frac{1}{N_1} - \frac{1}{N_2} - \frac{1}{N_1N_2}} \quad (13)$$
In Table 1, we list the theoretical values of the geometric mean and their comparisons with the simulated values for low, moderate, and strong interference. The interference occurrence rate is \( p_\eta = 0.01 \) for all three interference scenarios.

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Method} & \text{\( N_2 \)} & \text{\( a_\eta = 2 \)} & \text{\( a_\eta = 6 \)} & \text{\( a_\eta = 18 \)} & \text{Avg} \\
\hline
& \text{Mean} & \text{\( R^2 \)} & \text{Mean} & \text{\( R^2 \)} & \text{Mean} & \text{\( R^2 \)} & \\
& 100 & 1.0400 & 1.0193 & 1.3600 & 4.8984 & 4.2396 & 36.411 & 14.110 \\
\hline
\text{D}_N & 1 & 1.0237 & 2.7454 & 1.0239 & 2.7388 & 1.0238 & 2.7317 & 2.7486 \\
& 2 & 1.0290 & 2.1082 & 1.0294 & 2.1277 & 1.0295 & 2.1273 & 2.1211 \\
& 8 & 1.0394 & 1.7302 & 1.0555 & 1.9460 & 1.0554 & 1.9503 & 1.8755 \\
& 100 & 1.0400 & 1.6002 & 1.2779 & 8.9737 & 3.4042 & 111.64 & 40.738 \\
\hline
\text{G}_N & 1 & 1.0278 & 2.4724 & 1.0488 & 2.5614 & 1.0717 & 1.6730 & 2.2356 \\
& 2 & 1.0316 & 1.6650 & 1.0701 & 1.8519 & 1.1167 & 2.2247 & 1.9139 \\
& 8 & 1.0375 & 1.1544 & 1.1467 & 2.0017 & 1.3376 & 5.1570 & 2.7710 \\
& 100 & 1.0399 & 1.0310 & 1.3147 & 4.3308 & 2.9238 & 42.091 & 15.818 \\
\hline
\text{T}_95 & 1 & 1.0358 & 1.1711 & 1.0430 & 1.2202 & 1.0430 & 1.2239 & 1.2051 \\
& 2 & 1.0380 & 1.1079 & 1.0561 & 1.2533 & 1.0562 & 1.2498 & 1.2037 \\
& 8 & 1.0394 & 1.0594 & 1.1450 & 3.5711 & 1.7359 & 94.667 & 33.099 \\
& 100 & 1.0400 & 1.0394 & 1.3243 & 4.8256 & 3.7994 & 41.530 & 15.798 \\
\hline
\text{T}_{MAD8} & 1 & 1.0284 & 1.7843 & 1.0080 & 1.6774 & 1.0080 & 1.6714 & 1.7110 \\
& 2 & 1.0398 & 1.0910 & 1.0056 & 1.1300 & 1.0020 & 1.0801 & 1.1004 \\
& 8 & 1.0399 & 1.0143 & 1.1085 & 3.1179 & 1.0003 & 1.0893 & 1.7405 \\
& 100 & 1.0400 & 1.0193 & 1.3601 & 4.9145 & 4.2330 & 39.909 & 15.281 \\
\hline
\text{T}_{GEO4} & 1 & 1.0380 & 1.0343 & 1.0087 & 1.1344 & 0.9993 & 1.0159 & 1.0615 \\
& 2 & 1.0280 & 1.0885 & 0.9996 & 1.1170 & 0.9981 & 1.0328 & 1.0794 \\
& 8 & 1.0400 & 1.0287 & 1.0758 & 2.4126 & 1.0032 & 1.1384 & 1.5266 \\
& 100 & 1.0400 & 1.0310 & 1.3600 & 4.9090 & 3.9775 & 54.270 & 20.070 \\
\hline
\text{W}_N & 1 & 1.0390 & 1.0625 & 1.0074 & 1.1054 & 1.0001 & 1.0429 & 1.0703 \\
& 2 & 1.0400 & 1.0558 & 1.0115 & 1.1304 & 1.0098 & 1.0415 & 1.0759 \\
& 8 & 1.0400 & 1.0246 & 1.1078 & 3.1480 & 1.0001 & 1.0996 & 1.7574 \\
& 100 & 1.0391 & 1.0223 & 1.3501 & 4.9301 & 4.0907 & 41.158 & 15.703 \\
\hline
\text{H}_N & 1 & 1.0392 & 1.0236 & 1.0162 & 1.1090 & 1.0092 & 1.0462 & 1.0596 \\
& 2 & 1.0377 & 1.0447 & 1.0124 & 1.1290 & 1.0112 & 1.0447 & 1.0728 \\
& 8 & 1.0407 & 1.0199 & 1.1247 & 3.3169 & 1.0101 & 1.1043 & 1.8206 \\
& 100 & 1.0409 & 1.0231 & 1.3648 & 4.9272 & 4.1395 & 41.780 & 15.910 \\
\hline
\end{array}
\]
normalized variance for the trimmed mean is thus

\[ R^2(T_\beta) = \frac{\sigma^2_G + (1 - \beta)(y_\beta - \mu)^2}{2\beta\mu^2}. \quad (15) \]

In the following examples, \( N = 10000 \), \( \beta = 0.95 \), and \( \sigma_0 = 1 \). If \( N_1 = 2 \), then \( y_\beta = 3.843 \), \( \mu_\beta = 0.7590 \), \( \sigma^2_\beta = 0.7747 \), and \( R^2(T_{05}) = 1.1423 \). If \( N_1 = 2 \), we have \( y_\beta = 2.995 \), \( \mu_\beta = 0.8422 \), \( \sigma^2_\beta = 0.5027 \), and \( R^2(T_{05}) = 1.0898 \). When \( N_1 = 8 \), then \( y_\beta = 1.9384 \), \( \mu_\beta = 0.9320 \), \( \sigma^2_\beta = 0.1645 \), and \( R^2(T_{05}) = 1.0431 \). For \( N_1 = 100 \), we have \( y_\beta = 1.2435 \), \( \mu_\beta = 0.9835 \), \( \sigma^2_\beta = 0.0153 \), and \( R^2(T_{05}) = 1.0178 \). As seen in Table 1, the \( R_2 \) values agree with the simulation very well. It is of interest to note that \( R_2 \) is not 1/0.95 = 1.05 as intuition might suggest. It varies from 1.142 at \( N_1 = 1 \) to 1.018 at \( N_1 = 100 \). When \( N_1 = 1 \), the tail is long and has more variations, leading to a large \( R^2 \) value – a tail-wagging-the-dog situation. More averaging makes the tail more stable and \( R^2 \) smaller. This phenomenon, and the effects of a “fat tail” or heavy tail, are extensively discussed by Resnick (2007) and Taleb (2022).

To estimate a parameter robustly, we can attempt to identify outliers and exclude them from the average. Most of the outlier classifying methods involve estimating a nominal deviation and using it in a threshold to detect outliers. The median absolute deviation (MAD), defined as MAD = \( \text{med}(|Y_1 - \text{med}(Y))| \), is most frequently used to detect outliers (Huber and Ronchetti, 2009). Since only a small fraction of the ISR data is contaminated most of the time, we will classify a data point having 8 MADs above the median as an outlier. The sample mean of all non-outlier points is the outlier classifying methods involve estimating a nominal deviation and using it in a threshold to detect outliers. The sample mean performs best when there is no interference, although its efficiency is no less than 95.6%. If the constant 40 is changed to 60, the worst \( R^2 \) becomes 1.031, but the weighted mean is less effective in mitigating the effect of outliers. The mean and standard deviation of \( T_{GE04} \) are chosen because of their general accuracy and computing efficiency.

Knowing whether interference exists can help mitigate its effect. For a gamma distribution, we cannot associate the existence of outliers with interference with certainty, as there are outliers even when there is no interference. Since the expectation of \( R^2 \) for the sample mean is known in the null-interference case, a deviation from the expectation indicates that the underlying process may contain interference. As the sample mean performs better when there is no interference, an expedient strategy to reduce the variance is to combine the sample mean when no interference is detected with another estimator that is effective in mitigating the interference. We have used \( N = 10000 \) for the asymptotic case for all the estimators discussed above. In combining different estimators, a smaller \( N \) value is preferred so that the combined estimator will not be dominated by the interference-mitigating estimator in the presence of interference. We can also define a mixed \( R^2 \) that uses the mean of the \( T_{GE04} \) estimator and the normally defined variance. Such a mixed \( R^2 \) is more sensitive to outliers, but its variance is larger. Simulations show it does not cause a material difference from \( R^2(A_N) \) using the sample mean and standard deviation. Because of its simplicity, we choose \( R^2(A_N) \) as the criterion to determine if the data samples follow the desired process. The decision rule for this hybrid estimator, \( H_N \), is that if \( R(A_N) \) is less than 2 standard deviations above the mean, it uses the sample mean, otherwise the weighted mean is used. The performance of such a combined or hybrid estimator compares well to the other estimators. In Table 1, \( N = 1000 \) for the hybrid estimator, \( H_N \).

As seen in Table 1, all the order-based estimators (\( D_N \), \( T_{05} \), and \( T_{MAD8} \)) perform better as \( N \) increases. The tail-wagging-the-dog phenomenon discussed for \( T_{05} \) above is also applicable to \( D_N \) and \( T_{MAD8} \), as they also truncate the largest values. Although \( T_{GE04} \) is also a trimmed mean, the tail does not control \( R^2 \) in the same manner as in the order-based estimators because the length of the tail depends on the largest values. Large sample values increase the geometric deviation, which diminishes the chance of a large sam-

Q. Zhou et al.: Robust radar power estimations
Because of their interference-mitigating ability, efficiency in reducing statistical fluctuation, and computational efficiency. When \( p_\eta \) is less than 0.005, \( W_N \) (by extension, the combination of \( W_N \) and \( A_N \)) outperforms \( T_{\text{GEO4}} \) for all interference levels. In cases of prevalent contamination (e.g., \( p_\eta > 10\% \)) one can combine order-based estimators (such as the median or trimmed mean) with the sample mean.

### 3 Application to incoherent scatter radar signal processing

In this section, we apply four estimators to incoherent scatter total power and Doppler velocity processing and compare their performance. The example incoherent scatter radar data were taken at the Arecibo Observatory, Puerto Rico, on 11–12 September 2014. The total power is used to derive the electron density. The Doppler velocity is the same as the neutral wind velocity below about 115 km, but it also depends on the electric field and ion-neutral collision frequency above this altitude. Readers are referred to Zhou et al. (1997) and Isham et al. (2000) for further description of the Arecibo ISR, especially concerning E-region signal processing.

#### 3.1 Total power processing

The most common way to obtain the total power and hence electron density in the ionosphere using an ISR is to transmit a 13-baud Barker code with a total pulse length duration less than 52 \( \mu \)s. Barker code is chosen because of its minimized sidelobe. The lack of longer Barker codes is not a severe limitation due to the finite correlation time of the ionosphere. The 13-baud Barker data we use here have a baud length of 2 \( \mu \)s, making the range resolution 300 m. In-phase and quadrature voltage samples from each pulse are stored for post-processing. An inter-pulse period of 10 ms was used so that range aliasing is negligible. As the antenna was pointing vertically, range and altitude are interchangeable here. Although the sampling range in the data was from 60 to 766 km, we mostly focus on the altitude range from 90 to 150 km, where interference is most severe. The raw voltage samples were decoded using a matched filter.

Figure 1 shows the averaged power returns as a function of time and altitude using the sample, trimmed, \( T_{\text{GEO4}} \), and hybrid means. Because the radar samples are in in-phase and quadrature pairs and larger \( N_1 \) contaminates more data samples, \( N_1 \) is chosen to be 2. The last panel shows the normalized standard deviation \( R(A_N) \) for the sample mean, whose expectation is 1 when there is no interference. For each data point, we first average 250 pulses using the method indicated in the title and then average four such groups arithmetically for a total of 1000 pulses. Using a smaller number of pulses makes the memory requirement less stringent and the trimmed mean more efficient. The ionosphere signal is largely characterized by smooth temporal and spatial varia-

Figure 1. Range–time–intensity (RTI) plots of incoherent scatter total power returns on 11–12 September 2014. The first four panels, starting from the top, are the power return of the sample mean, the trimmed mean at the 95% level, the trimmed mean based on the geometric deviation, and a hybrid method, respectively. The last panel is the normalized standard deviation.

The top panel in Fig. 1 shows the result of arithmetically averaging 1000 pulses (i.e., the sample mean). All types of interference show up prominently, as the method does not filter out any contamination. The trimmed mean (second panel) cleans up the first part of the heavy contamination in box A but is not effective against the second part, most likely because more than 5% of the pulses were contaminated. $T_{GEO4}$ and the hybrid method largely filters out the contamination in box A and reveal the underlying sporadic layer despite the heavy contamination. Although $T_{GEO4}$ appears to handle all the contamination as well as the hybrid method does, it is slightly inferior to the latter in reducing statistical error, as seen in the later part of this section. The only residue contamination not filtered out is around 22:30 LT. None of the methods is effective in removing it completely, and all three robust estimators appear to perform the same. As the total power of the interference is relatively low, interference may permeate most of the pulses, making it very difficult to remove it from each pulse. For this type of interference, one way is to find the mean at non-ionosphere heights and subtract it from the entire profile. Noise samples are available at Arecibo. Background noise is not subtracted here to focus on the effect of robust estimators in this study. The trimmed
A. Zhou et al.: Robust radar power estimations

Figure 2. (a) The square of the relative error in the sample mean method normalized to that of the hybrid method. (b) The normalized variance. The yellow color in (a) indicates that the sample mean has a larger error than the hybrid method.

Figure 3. Mean relative errors (in base-2 logarithms) of the sample mean, trimmed mean, and \( T_{\text{GEO4}} \) normalized to that of the hybrid method (red and blue lines, respectively). The black line is 6 times the logarithm (base 2) of the mean \( R \). The time duration averaged for all the lines in the figure is from 07:00 to 13:00 LT on 12 September 2014.

mean, \( T_{\text{GEO4}} \), and hybrid methods are all effective at removing meteor interference, which typically does not last more than 50 ms at Arecibo, i.e., 5 pulses (Zhou and Kelly, 1997).

Other than the most obvious interference highlighted in boxes A and B, no other contamination appears to be obvious. The \( R \) value in the region indicated by box C has elevated values, indicating likely contamination. Yet there appears to be little difference between the sample mean result in the top panel of Fig. 1 and the results from the robust estimators. One effect of the interference is that it increases the statistical error, which is more difficult to see from the RTI plot. To estimate the statistical error, we use the difference in the power minus the average power of the surrounding 15 points in height and 5 points in time as a proxy for the error. The square ratio of the sample mean error to the error in the hybrid method is displayed in Fig. 2a. The corresponding \( R(A_N) \) is displayed in Fig. 2b. Larger statistical error from the sample mean in the region indicated by box C in Fig. 1 is quite evident. Although \( R(A_N) \) is not linearly related to the error, elevated \( R(A_N) \) is a robust indicator of contamination. This is also evidenced from 01:00 to 03:00 LT in Fig. 2, where sporadic elevations of \( R(A_N) \) and statistical errors are seen to be correlated.

An estimator needs to be efficient when there is no interference. Figure 3 shows the ratio of the sample mean and \( T_{95} \) errors to the hybrid error as well as the corresponding standard deviation \( R(A_N) \) averaged between 07:00 to 13:00 LT, during which period contamination is minimal above 120 km (as seen in Fig. 2). The error in the hybrid estimator is virtually the same as that in the arithmetic average. The error in the \( T_{95} \) estimator is 1.036 times the error in the hybrid estimator, which is in good agreement with the simulated value of \( \sqrt{1.09/1.018} = 1.035 \). Similarly, the error in \( T_{\text{GEO4}} \) is slightly smaller than that in \( T_{95} \), which is also in good agreement with the simulation results shown in Table 1. The mean \( R(A_N) \) correlates with the elevated error in the region of 90–120 km. We also note that the mean \( R(A_N) \) above 120 km is 0.997, which is slightly below the expected value of 1. Although the deviation is small, it is statistically significant. This may be caused by the bias in the receiving channels or the finite dynamic range of the analog-to-digital converters.

3.2 Power spectrum processing and Doppler velocity comparisons

The power spectral density (PSD) of an ISR is obtained by transmitting a coded long pulse (CLP), 440 \( \mu \)s in our case. The baud length is 2 \( \mu \)s, making the bit number of the pulse 220. The inter-pulse period is 10 ms as in the Barker data. The bit sequence is random for each transmitted pulse. The PSD is obtained by the Fourier transform of the data multiplied by the complex conjugate of the code. The characteristics of the CLP are discussed by Sulzer (1986). The averaging of the PSD at each frequency component is identical to that of the total power in the above section, which can be viewed as the center frequency component.

Figure 4 shows the Doppler velocity derived from the four estimators using the phase of the auto-correlation function. The vertical ion drift in the altitude range of 90–150 km is typically less than 50 m s\(^{-1}\) above Arecibo. Below 120 km, the plasma drift is the same as the neutral wind because of the complete coupling between ions and neutral molecules. During the daytime, there are sufficient signals above 95 km to obtain continuous spatial and temporal velocities. During the nighttime, it is only possible to obtain velocities within thin ionization layers. While ion velocity with fine height and time resolutions is of great geophysical interest (e.g., Zhou
et al., 1997; Hysell et al., 2014), our focus here is to study the relative accuracy of the velocities obtained from different estimators.

Comparisons of the velocity results largely follow those of the total power. The sample mean fails in boxes A and B. Additionally, during the sunrise hours when the ionospheric signal is low and the meteoric interference is strong, the sample mean can only yield valid velocities occasionally while the robust estimators can obtain the velocities continuously in altitude and time. As in the total power estimation, the trimmed mean does not yield valid results in the second part of box A from 21:30 to 22:30 LT, while the hybrid and $T_{\text{GEO4}}$ methods appear not to be affected by the interference very much.

To compare the statistical fluctuations, we use the altitude difference in the velocity divided by the square root of 2 as a proxy for velocity error. Figure 5 shows the altitude variation in the velocity error during 08:00–10:00 LT as well as 14:30–16:30 LT on 12 September. All the robust estimators have essentially the same error at each altitude, while the sample mean has a much larger error around 100 km. The error in the sample mean converges to those of the robust estimators above 145 km. The diminishing error difference in the sample mean with increasing altitude is due to the long pulse length (440 µs) used. A characteristic of the CLP pulse is that the interference at one altitude is uniformly spread across the entire bandwidth randomly at other altitudes. A meteor echo at 100 km increases the spectral power fluctuations with diminishing strength up to 166 km. Meteoric influx peaks at 06:00 LT and varies strongly with the local time. The daily variation in meteoric flux is quantitatively analyzed by Zhou et al. (1995) and Li and Zhou (2019). It can also be qualitatively seen in Fig. 2b. The larger error in the sample mean during 08:00–10:00 LT is a reflection of the strong meteoric flux. Although the afternoon period suffers from meteoric interference and radio contamination, as seen in Fig. 2, both of them are weak. Statistical averaging of 6000 pulses is able to even out the spectral power fluctuation to such a degree that all the estimators produce the same velocity. For spectral processing, the most important factor is the total amount of noise power, while the percentage of pulses contaminated is often more important in total power processing.

Overall, we see that the $T_{\text{GEO4}}$ and hybrid estimators accurately and consistently improve velocity and total power measurements over the sample mean, which are important for studying the E-region dynamics and composition. The availability of nighttime velocities will help reduce the large
error in the measurement of atmospheric tides in the E region (Zhou et al., 1997; Gong et al., 2013). Accurate measurement of the power spectrum and total power will facilitate all E-region studies, especially those concerning the climatology and dynamics of sporadic E and intermediate layers (Zhou et al., 2005; Hysell et al., 2009; Raizada et al., 2018; Gong et al., 2021). Of particular importance are the vertical wind and ion composition of the E region, which have not been studied much due to a lack of quality data.

4 Summary and conclusion

We have discussed several robust estimators to compute the variance of a normally distributed random variable, $X$, to deal with interference. This variance is the same as the mean of the power variable, $X^2$. The effectiveness of an estimator is described by the normalized standard deviation, $R$. We derive the theoretical $R$ values for the median, geometric mean, and trimmed mean of gamma distributions, which result from averaging the power random variables. We discuss and compare another four estimators through simulations for various interference scenarios. Robust estimators found in the literature are typically rank-based (e.g., the median, trimmed mean, and median absolute deviation). We have used the geometric mean and geometric deviation as two basic parameters in assessing the likelihood of a data point being contaminated. The methods based on the geometric mean have two advantages over the rank-based ones: they are less susceptible to the large uncertainties in the tail part of the distributions and they are computationally more efficient. For the interference model used, the $T_{\text{GEO4}}$ estimator, which is based on the geometric mean, is particularly effective as a stand-alone estimator when there is no initial average. Another effective estimator based on the geometric mean is the weighted mean. The $R$ value of the sample mean can be used to assess whether the process conforms to the expected distribution. This knowledge allows us to combine the sample mean with other robust estimators to mitigate contamination and achieve statistical accuracy.

We apply three robust estimators to incoherent scatter power and velocity processing, along with the traditional sample mean estimator. We show that the performance of estimators with real data agrees well with simulations. In the total power processing, the trimmed mean performs mostly well except when the contamination is very heavy. The $T_{\text{GEO4}}$ estimator performs almost as well as the hybrid method in mitigating interference. The hybrid method performs the best at mitigating interference as well as at reducing statistical errors. For Doppler velocity processing, the same conclusion can be drawn in cases of frequent interference. When the interference is weak, all the robust estimators appear to perform well. For the Arecibo ISR data, the sample mean has larger statistical errors even for data that may not appear to contain obvious interference. This highlights the need for robust estimation to process or reprocess decades of E-region data taken at Arecibo. The hybrid estimator is most advantageous under all circumstances. This conclusion is likely applicable to other incoherent scatter radars as well. While the interference characteristics differ at each radar site, the study provides a foundation to optimize robust estimation, which is an essential step in many data processing applications.

Data availability. The raw data are available upon request from the Texas Advanced Computing Center (https://tacc.utexas.edu/research/tacc-research/arecibo-observatory/, last access: August 2023). The processed and the raw data can be obtained from the corresponding author.

Author contributions. QZ conceptualized the framework, formulated the theory, and drafted and finalized the paper. YL did the data analysis and contributed to the data visualization. YG contributed to the conceptualization and verification of the theory and edited the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

https://doi.org/10.5194/amt-17-4197-2024
Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgement. We thank the former staff members of the Arecibo Observatory for collecting the data.

Financial support. This research has been supported by the US National Science Foundation, Directorate for Geosciences (grant no. 2152109).

Review statement. This paper was edited by Meng Gao and reviewed by two anonymous referees.

References

