PROCEEDINGS OF SCIENCE

Statistics of the SK estimator

Gelu M. Nita[*†]
New Jersey Institute of Technology
E-mail: gnita@njit.edu

Dale E. Gary
New Jersey Institute of Technology
E-mail: dgary@njit.edu

Spectral Kurtosis (SK) is a statistical approach for detecting and removing radio frequency interference (RFI) in radio astronomy data. In this study, the statistical properties of the SK estimator are investigated and all moments of its probability density function are analytically determined. These moments provide a means to determine the tail probabilities of the estimator that are essential to defining the thresholds for RFI discrimination. It is shown that, for a number of accumulated spectra $M \geq 24$, the first SK standard moments satisfy the conditions required by a Pearson Type IV probability density function (PDF), which is shown to accurately reproduce the observed distributions. The cumulative function (CF) of the Pearson Type IV is then found, in both analytical and numerical forms, suitable for accurate estimation of the tail probabilities of the SK estimator. This same framework is also shown to be applicable to the related Time Domain Kurtosis (TDK) estimator, whose PDF corresponds to Pearson Type IV when the number of time-domain samples is $M \geq 46$. The PDF and CF are determined for this case also.

RFI mitigation workshop - RFI2010,
March 29-31, 2010
Groningen, the Netherlands

[*]Speaker.
[†]This work was supported by NSF grant AST-0908344 to NJIT

1. Introduction

The Spectral Kurtosis estimator ($\widehat{SK}$) was originally proposed by Nita et al. [7] as a statistical tool for real-time radio frequency interference (RFI) detection and excision in a Fast Fourier Transform (FFT) radio spectrograph. The first spectrograph designed for $\widehat{SK}$, the Korean Solar Radio Burst Locator [KSRBL; 1], demonstrated the effectiveness of the SK algorithm, but also revealed the need for a more accurate calculation of the theoretical RFI detection thresholds than initially proposed. Consequently, Nita & Gary [8] derived the exact analytical expressions for the statistical moments of $\widehat{SK}$ and, based on its first four standard moments, assigned to it a Pearson Type IV probability curve [9], which was shown to be in very good agreement with the Monte Carlo simulated $\widehat{SK}$ probability distribution function (PDF), as well as with the distribution derived from direct experimental observations made with the KSRBL instrument [2].

As extensively described in the previous papers, what makes an SK spectrograph with N spectral channels distinct from a traditional one is the fact that it accumulates not only a set of M instantaneous power spectral density (PSD) estimates, denoted S_1, but also the squared spectral power denoted S_2. These sums, which have an implicit dependence on frequency channel f_k, are used to compute the averaged power spectrum $\langle P\rangle = S_1/M$, as well as the quantity

$$\widehat{SK} = \frac{M+1}{M-1}\left(\frac{MS_2}{S_1^2} - 1\right), \tag{1.1}$$

which is a cumulant-based estimator of the spectral variability corresponding to the signal parent population,

$$V_k^2 = \frac{\sigma_k^2}{\mu_k^2}, \tag{1.2}$$

where μ_k and σ_k^2 are the frequency-dependent PSD population means and variances, respectively. For a normally distributed time domain signal, i.e. an RFI-free signal, Nita & Gary [8] showed that the estimator given by equation (1.1) is unbiased, i.e. $E(\widehat{SK}) = V_k^2 = 1$.

This presentation compiles the main results of Nita & Gary [8] leading to the definition of $\widehat{SK}$ given by equation (1.1), and provides a practical guide for computing the RFI thresholds with a predefined false alarm probability (PFA) level.

2. The Unbiased Spectral Kurtosis and Time Domain Kurtosis Estimators

The unbiased estimator $\widehat{SK}$ may be derived based on the properties of the generalized gamma distribution [GGD, 10] defined as

$$f(x,a,d,p) = \frac{px^{d-1}e^{-(\frac{x}{a})^p}}{a^d\Gamma(d/p)}, \tag{2.1}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the well known Euler's Gamma function. The expected sample moments about origin corresponding to a GGD function are given by

$$E(x^n) = \frac{\Gamma(\frac{d+n}{p})}{\Gamma(d/p)}a^n, \tag{2.2}$$

which reduces to the known result $E(x^n) = n!\mu^n$ in the particular case of an exponential distribution, $f(x,\mu,1,1)$.

As shown by Nita & Gary [8], if one considers a set of M independent random variables that are are individually distributed according with a $p = 1$ GGD function, $f(x,a,d,1)$, (which is nothing else than a standard gamma distribution), the probability distribution of the mean $\langle x\rangle = \Sigma_{i=1}^{M} x_i$ is the GGD function

$$f(\langle x\rangle, a/M, Md, 1), \tag{2.3}$$

and the probability distribution of the squared mean is the GGD function

$$p(\langle x\rangle^2) = f\Big[\langle x\rangle^2, \Big(\frac{a}{M}\Big)^2, \frac{Md}{2}, \frac{1}{2}\Big], \tag{2.4}$$

which, making use of equation (2.2), provides the expectations of an arbitrary power of the squared mean,

$$E(\langle x\rangle^{2n}) = \frac{\Gamma(Md+2n)}{\Gamma(Md)}\Big(\frac{a}{M}\Big)^{2n}. \tag{2.5}$$

Although a closed form for the PDF of the mean of squares $\langle x^2\rangle = \Sigma_{i=1}^{M} x_i^2$ has not been found, Nita & Gary [8] proved that the statistical moments of the mean of squares can be exactly computed according with the formula

$$E(\langle x^2\rangle^n) = \frac{(a/\sqrt{M})^{2n}}{[\Gamma(d)]^M}\frac{\partial^n}{\partial t^n}\Big[\sum_{r=0}^{n}\frac{1}{r!}\Gamma(2r+d)t^r\Big]^M\Big|_{t=0}. \tag{2.6}$$

For the particular cases of the exponential and χ^2 distributions, $f(x,\mu,1,1)$ and $f(x,\mu,1/2,1)$ respectively, Nita & Gary [8] proved that $\langle x\rangle^2/\langle x^2\rangle$ and $\langle x^2\rangle$ are uncorrelated random variables, i.e. that they have null covariance, which immediately led to the exact moments of the ratio between the mean of squares and square of mean, which can be expressed as

$$E\Big[\Big(\frac{\langle x^2\rangle}{\langle x\rangle^2}\Big)^n\Big] = \frac{E(\langle x^2\rangle^n)}{E(\langle x\rangle^{2n})}. \tag{2.7}$$

However, as we will show here, the null covariance of $\langle x\rangle^2/\langle x^2\rangle$ and $\langle x^2\rangle$ is an intrinsic property of any $p = 1$ GGD function $f(x,a,d,1)$, which makes equation (2.7) hold for any value of d. To prove this key property, we make use of the expression

$$cov\Big(\frac{\langle x\rangle^2}{\langle x^2\rangle}, \langle x^2\rangle\Big) = \frac{2}{M}\Big[\frac{E(x^3)}{E(x)} - 2\frac{E(x^2)^2}{E(x)^2} + E(x^2)\Big] \tag{2.8}$$

that is valid for any particular PDF [Eq.19, 8], in which we enter the explicit moments provided by equation (2.2) to obtain

$$cov\Big(\frac{\langle x\rangle^2}{\langle x^2\rangle}, \langle x^2\rangle\Big) = 0. \tag{2.9}$$

Therefore, provided that the observable x is distributed according to a GGD function $f(x,a,d,1)$, and taking in consideration equations (2.5) and (2.6), the statistical moments of the ratio between

the mean of squares and square of mean given by equation (2.7) may be written in the compact form

$$E\Big[\Big(\frac{\langle x^2\rangle}{\langle x\rangle^2}\Big)^n\Big]=\frac{M^n\Gamma(Md)}{\Gamma(d)^M\Gamma(Md+2n)}\frac{\partial^n}{\partial t^n}\Big[\sum_{r=0}^{n}\frac{1}{r!}\Gamma(2r+d)t^r\Big]^M\Big|_{t=0}, \tag{2.10}$$

which is independent of the scaling parameter a. Nita & Gary [8] used the particular form of equation (2.10) corresponding to $d=1$ to derive the SK estimator given by equation (1.1), and the particular form corresponding to $d=1/2$ to derive a time domain kurtosis (TDK) estimator

$$\widehat{TDK}=\frac{M+2}{M-1}\Big(\frac{MS_2}{S_1^2}-1\Big), \tag{2.11}$$

both of them being unbiased estimators of the spectral variability corresponding to the underlying probability distribution $f(x,a,d,1)$, which according to equation (1.2) is

$$V^2=\frac{E(x^2)-E(x)^2}{E(x)^2}=\frac{1}{d}, \tag{2.12}$$

an expression that is 1 for the $d=1$ (SK) case, and 2 for the $d=1/2$ (TDK) case.

3. The Pearson Type IV Approximation of the Spectral and Time Domain Kurtosis Estimators

Using equation (2.10), the first standard moments of the $\widehat{SK}$ estimator given by equation (1.1) may be written as

$$\begin{aligned}\mu_1' &= 1;\ \mu_2=\frac{4M^2}{(M-1)(M+2)(M+3)} \\ \beta_1 &= \frac{4(M+2)(M+3)(5M-7)^2}{(M-1)(M+4)^2(M+5)^2};\ \beta_2=\frac{3(M+2)(M+3)(M^3+98M^2-185M+78)}{(M-1)(M+4)(M+5)(M+6)(M+7)},\end{aligned} \tag{3.1}$$

where $\beta_1=\mu_3^2/\mu_2^3$ and $\beta_2=\mu_4/\mu_2^2$ are directly related to the more commonly used skewness, $\gamma_1=\sqrt{\beta_1}$ and kurtosis excess, $\gamma_2=\beta_2-3$. Similarly, the first standard moments of the $\widehat{TDK}$ estimator are

$$\begin{aligned}\mu_1' &= 2;\ \mu_2=\frac{24M^2}{(M-1)(M+4)(M+6)} \\ \beta_1 &= \frac{216(M-2)^2(M+4)(M+6)}{(M-1)(M+8)^2(M+10)^2};\ \beta_2=\frac{3(M+4)(M+6)(M^3+213M^2-474M+368)}{(M-1)(M+8)(M+10)(M+12)(M+14))}.\end{aligned} \tag{3.2}$$

The appropriate PDF approximation for the $\widehat{SK}$ and $\widehat{TDK}$ estimators may be found by investigating the Pearson's criterion defined as [5, p. 151]:

$$\kappa=\frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}, \tag{3.3}$$

where the exact values of the parameters β_1 and β_2 are provided by equations (3.1) and (3.2), respectively. Pearson's criterion indicates that for $M\geq 24$ the $\widehat{SK}$ estimator may be approximated by a Pearson Type IV probability curve. The same PDF approximation may be used for the $\widehat{TDK}$ estimator for $M\geq 46$.

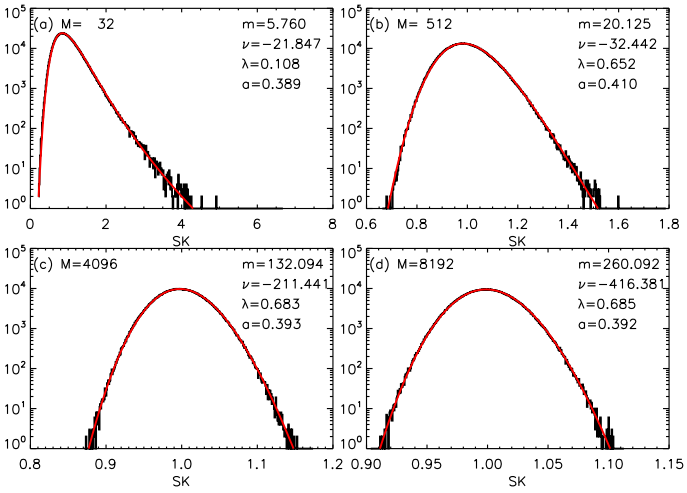

Figure 1: Comparison between the SK distributions (black lines) obtained by numerical simulation for different accumulation lengths M and their corresponding Pearson Type IV approximations (red lines). The four Pearson Type IV parameters m, ν, λ, and a are displayed on each plot.

3.1 Pearson Type IV Probability Distribution Function

The most general analytical form of the Pearson Type IV PDF originally introduced by Pearson [9], including its non-trivial normalization factor, was given by Nagahara [6] as

$$p(x)=\frac{1}{a\sqrt{\pi}}\frac{\Gamma(m+i\frac{\nu}{2})\Gamma(m-i\frac{\nu}{2})}{\Gamma(m-\frac{1}{2})\Gamma(m)}\left[1+\left(\frac{x-\lambda}{a}\right)^2\right]^{-m}\mathtt{Exp}\left[-\nu\mathtt{ArcTan}(\frac{x-\lambda}{a})\right], \tag{3.4}$$

where the four parameters m, μ, a, and λ may be expressed in terms of the central moments of the distribution as [4]:

$$\begin{aligned} r &= \frac{6(\beta_2-\beta_1-1)}{2\beta_2-3\beta_1-6};\ m=\frac{r+2}{2};\ \nu=-\frac{r(r-2)\sqrt{\beta_1}}{\sqrt{16(r-1)-\beta_1(r-2)^2}} \\ a &= \frac{1}{4}\sqrt{\mu_2(6(r-1)-\beta_1(r-2)^2)};\ \lambda=\mu-\frac{1}{4}(r-2)\sqrt{\mu_2\beta_1}. \end{aligned} \tag{3.5}$$

Figure 1 compares the Pearson IV approximations of the $\widehat{SK}$ PDF with the Monte Carlo simulated distributions for $M=32$, 1024, 4096, and 8192. By visual inspection, we may conclude that the Pearson IV approximations accurately reproduce the shapes of the numerically simulated histograms for different orders of magnitude of the accumulation length.

To compute the tail probabilities of the Pearson Type IV PDF, one has to compute the cumulative function $P(x)$, (CF), and the complementary cumulative function, $1-P(x)$, (CCF),

$$P(x)=\int_{-\infty}^{x}p(x)dx;\ 1-P(x)=\int_{x}^{\infty}p(x)dx, \tag{3.6}$$

for which Heinrich [4] provided the following closed form:

$$P(x)=\begin{cases} 1+P_1(m,\nu,a,\lambda,x), & x<\lambda-a\sqrt{3} \\ P_2(m,\nu,a,\lambda,x) & |x-\lambda|<a\sqrt{3} \\ 1-P_1(m,-\nu,a,-\lambda,-x), & x>\lambda+a\sqrt{3}, \end{cases} \tag{3.7}$$

where

$$P_1(m,\nu,a,\lambda,x)=\frac{a}{2m-1}\left(i-\frac{x-\lambda}{a}\right)F\left(1,m+i\frac{\nu}{2},2m,\frac{2}{1-i\frac{x-\lambda}{a}}\right)p(x)$$

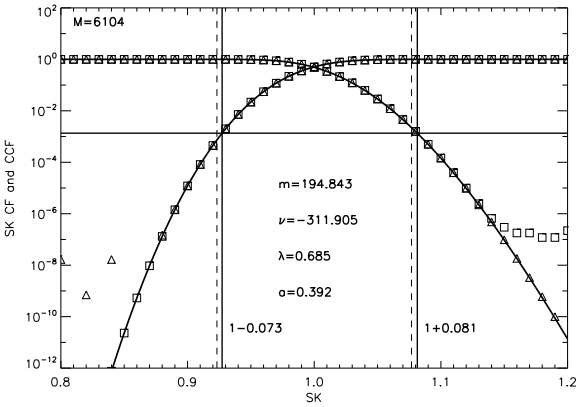

Figure 2: $\widehat{SK}$ threshold computation for $M = 6104$. The lower and upper thresholds displayed by the two vertical lines have been estimated as the intersection points of the horizontal and numerical integration lines. Their values of $1 - 0.073 = 1 - 5.6799/\sqrt{6104}$ and $1 + 0.081 = 1 + 6.3596/\sqrt{6104}$, respectively, have to be compared with the symmetric thresholds of $1 \pm 6/\sqrt{6104}$ (vertical dotted lines).

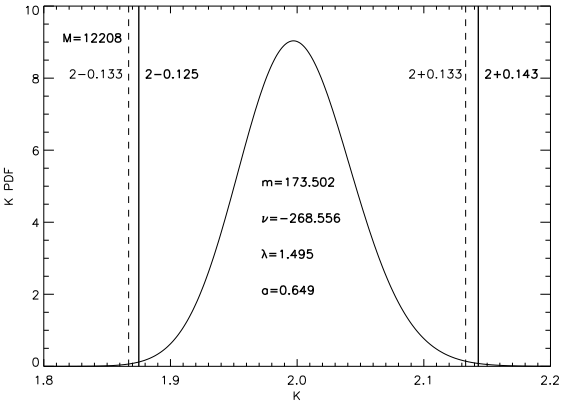

Figure 3: $\widehat{TDK}$ threshold computation for $M = 12208$. The two solid vertical lines, having the ordinates $2 - 0.125$ and $2 + 0.143$, represent the RFI detection thresholds corresponding to a symmetric standard false alarm probability level of 0.13499%. These thresholds have to be compared with with the less accurate symmetric thresholds of $2 \pm 3\sqrt{24/12208} = 2 \pm 0.133$ (vertical dotted lines).

$$P_2(m,\nu,a,\lambda,x) = \frac{1}{1-e^{-(\nu+i2m)\pi}} - \frac{ia}{i\nu-2m+2}\left[1+\left(\frac{x-\lambda}{a}\right)^2\right] \times F\left(1,2-2m,2-m+i\frac{\nu}{2},\frac{1+i\frac{x-\lambda}{a}}{2}\right)p(x),$$

and

$$F(\alpha,\beta,\delta,z) = 1+\frac{\alpha\beta}{1!\delta}z+\frac{\alpha(\alpha+1)\beta(\beta+1)}{2!\delta(\delta+1)}z^2+\ldots=\sum_{k=0}^{\infty}\frac{\alpha_{(k)}\beta_{(k)}}{k!\delta_{(k)}}z^k \tag{3.8}$$

is the Gauss hypergeometric series.

Alternatively, instead of equation (3.7), one may use the formula provided by Willink [11]

$$P(m,\nu,a,\lambda,x) = \frac{e^{-[\lambda-i(2-2m)]\Phi}R-1}{e^{-[\lambda-i(2-2m)]\pi}-1}, \tag{3.9}$$

where

$$\Phi = \frac{\pi}{2}+\arctan\left(\frac{x-\lambda}{a}\right);\ u = -m-\tfrac{i}{2}\nu;\ R = \tfrac{F(2-2m,u,u+1,e^{i\Phi})}{F(2-2m,u,u+1,1)}.$$

However, if one wants to avoid the numerical difficulties related to the evaluation of the hypergeometric series, one may choose to perform a direct numerical integration (equation [3.6]) of equation (3.4), which may achieve reasonable accuracy with far less computational effort, especially if tailored integration methods, [e.g. 6], are employed.

Figure 2 displays the numerical results for $M = 6104$, computed according to equation (3.7), (triangular symbols), equation (3.9), (square symbols), and by direct integration of equations (3.6),

(solid line). The hypergeometric series where computed using the *hypergeom* function in Maple 11 (MapleSoft), and the numerical integration was performed using the *int_tabulated* function in IDL 6.4 (ITT). The plots display both CF (rising) and CCF (descending) needed to evaluate the RFI thresholds equivalent to normal distribution's $\pm 3\sigma$ level (probability 0.13499%, horizontal solid line). It may be concluded that, in the region of interest, all three methods provide similar numerical results. However, it was found that, for SK values well before the distribution peak, the numerical accuracy of equation (3.9) is better than that of equation (3.7), while the direct numerical integration of CF gives similar results as equation (3.9). After the peak of the $\widehat{SK}$ distribution, the numerical accuracy of equation (3.7) is better than that of equation (3.9), while the numerical integration of CCF gives similar results as equation (3.7). Therefore we conclude that the numerical evaluation of equation (3.9) gives a more accurate estimation of the CF and the numerical evaluation of equation (3.7) gives a more accurate estimation of the CCF, while the direct numerical integration of equations (3.6) gives results of comparable accuracy at both sides of the $\widehat{SK}$ distribution. The lower and higher thresholds displayed by the two vertical lines have been estimated as the intersection points of the horizontal and numerical integration lines. Their values of $1-0.073 = 1-5.6799/\sqrt{6104}$ and $1+0.081 = 1+6.3596/\sqrt{6104}$, respectively, are compared with the symmetric thresholds of $1 \pm 6/\sqrt{6104}$ (vertical dotted lines) originally proposed by Nita & Gary [8]. Although this correction seems small in absolute value for the large-M case, e.g. $M = 6104$ illustrated in Figure 2, we calculate that, compared with the symmetric thresholds, the new thresholds account for 67% less rejection of valid data as false RFI occurrences at the upper bound of the distribution, and provide better rejection of true RFI signals of low signal to noise ratio at the lower bound. In combination, the result is an overall better performance of the $\widehat{SK}$–based RFI rejection algorithm. The correction becomes more important for lower M. Figure 3 displays the PDF of the $\widehat{TDK}$ estimator corresponding to an accumulation length of $M = 12208$, chosen to match the same frequency and time resolution of a DFT-based spectrograph with $M = 6104$ (the example used in Fig. 2; see Nita et al. [7] for a more detailed motivation of this choice). Despite its large accumulation length, the estimator $\widehat{K}$ still has a noticeable skewness, which needs to be properly considered in order to obtain the false alarm probability levels equivalent to $\pm 3\sigma$ for a normal distribution. Compared with the symmetric thresholds of $2 \pm 3\sqrt{24/12208}$, the new thresholds would reject 71% less valid data at the higher end of the distribution, while the shifted lower threshold would improve the sensitivity of RFI detection at the lower end of the distribution.

4. Conclusion

We have investigated the statistical properties of the SK estimator and determined analytical expressions for its PDF and CF with the goal of improving the selection of thresholds for RFI discrimination. We also improved the definition of $\widehat{SK}$ (equation [1.1]) relative to its original definition [7] to form an unbiased estimator, and introduced a TDK unbiased estimator (equation [2.11]) to be used for RFI detection at the DC and Nyquist frequency bins of a DFT-based spectrograph, or at any frequency bin of a FIR–based spectrograph. We have derived closed form analytical expressions for the complete set of the central moments of the SK and TDK estimators, and established a common framework that allows accurate estimation of the RFI thresholds based on the first four standard moments of their probability distributions (equations [3.1] and [3.2]), which, for any

accumulation length $M \geq 24$ and $M \geq 46$, respectively, are used to compute the four parameters (equation [3.5]) that completely determine the Pearson IV approximations (equation [3.4]) of their true PDFs. Based on these four parameters, which depend only on the accumulation length M, the CF and CCF of the SK or TDK estimators may be computed by using either the closed forms expressions provided by equations (3.9) and (3.7), respectively, or by direct numerical estimation of the integrals given by equation (3.6). Compared to the symmetrical thresholds originally suggested in Paper I, the procedure described in this study properly takes into account the intrinsic skewness of the probability density functions of the SK and TDK estimators, which provides better overall RFI detection performance for either small or large accumulation lengths. These theoretically established results are shown in Gary, Liu & Nita [2, 3] to be exactly obeyed by data taken in the KSRBL spectrometer hardware implementation of the algorithm, where the improvement in RFI excision by use of these modified thresholds is confirmed. The modified thresholds become ever more important when a smaller number M of accumulations is used. A simple procedure has been written in IDL (Interactive Data Language) for numerical calculation of the thresholds for any M, and is available upon request from the authors.

References

[1] Y. Dou, D. E. Gary, Z. Liu, G. M. Nita, S. -C. Bong, K. -S. Cho, Y. -D. Park, & Y. -J. Moon, *The Korean Solar Radio Burst Locator (KSRBL)*, PASP, 121, 512

[2] D. E. Gary, Z. Liu, & G. M. Nita, *A Wideband Spectrometer with RFI Detection*, PASP, 122, 560

[3] D. E. Gary, Z. Liu, & G. M. Nita, *Hardware Implementation of an SK Spectrometer*, in this proceedings, `PoS(RFI2010)020`

[4] J. Heinrich, *A Guide to the Pearson Type IV Distribution*, Collider Detector at Fermilab internal note 6820, http://www-cdf.fnal.gov/publications/cdf6820_pearson4.pdf

[5] M. G. Kendall & A. Stuart, *The Advanced Theory of Statistics Vol.1*, Griffin, London 1958

[6] Y. Nagahara, *The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters*, Statistics & Probability Letters, 43, 251

[7] G. M. Nita, D. E. Gary, Z. Liu, G. J. Hurford, & S. M. White, *Radio Frequency Interference Excision Using Spectral Domain Statistics*, PASP, 119, 805

[8] G. M. Nita & D. E. Gary, *Statistics of the Spectral Kurtosis Estimator*, PASP, 122, 595

[9] K. Pearson, *Contributions to the Mathematical Theory of Evolution.-II. Skew Variation in Homogeneous Material*, Philos. Trans. R. Soc. London A, 186, 343

[10] E.W. Stacy, *A Generalization of the Gamma Distribution*, Ann. Math. Statist. Assoc., 33, 1187

[11] R. Willink, *A Closed-Form Expression For The Pearson Type IV Distribution Function*, Austral. NZ J. Stat., 50(2), 199