PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC, **122**:595–607, 2010 May

Statistics of the Spectral Kurtosis Estimator

GELU M. NITA[1] AND DALE E. GARY[1]

Received 2010 January 7; accepted 2010 February 18; published 2010 April 22

ABSTRACT. Spectral kurtosis (SK) is a statistical approach for detecting and removing radio-frequency interference (RFI) in radio astronomy data. In this article, 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 also are determined for this case.

Online material: color figure, source code

1. INTRODUCTION

Given the expansion of radio astronomy instrumentation to ever-broader bandwidths, and the simultaneous increase in usage of the radio spectrum for wireless communication, radio-frequency interference (RFI) has become a limiting factor in the design of a new generation of radio telescopes. In an effort to find reliable solutions to RFI mitigation, Nita et al. (2007 hereafter, Paper I) proposed the use of a statistical tool, the spectral kurtosis (SK) estimator. Based on theoretical expectations and initial hardware testing, the SK estimator was found to be an efficient tool for automatic excision of certain types of RFI, and due to its conceptual simplicity we suggested that it should become a standard, built-in component of any modern radio spectrograph or FX correlator that is based on field-programmable gate array (FPGA) architecture. Since then, the world's first SK spectrometer, the Korean Solar Radio Burst Locator (KSRBL; Dou et al. 2009), has become operational, and the effectiveness of the SK algorithm for RFI excision has been demonstrated (Gary et al. 2010).

As described in Paper I, an SK spectrometer with N spectral channels accumulates both a set of M instantaneous power spectral density (PSD) estimates, denoted S_1, and the squared spectral power denoted S_2. These sums, which have an implicit dependence on frequency channel $f_k, (k = 0 \ldots N/2)$, are used to compute the averaged power spectrum S_1/M, as well as an SK estimator $\widehat{V_k^2}$, originally defined as

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

which is a cumulant-based estimator of the spectral variability. The "hat" is used to distinguish the estimator from the parent population parameter associated with each frequency channel,

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

where μ_k and σ_k^2 are the frequency-dependent PSD population means and variances, respectively.

In Paper I we studied the statistical properties of the PSD estimates of a normally distributed time-domain signal obtained from its complex discrete Fourier transform (DFT) coefficients, and showed that, in the most general case, at all but the DC ($k = 0$) and Nyquist ($k = N/2$) frequency channels, the real and imaginary parts, A_k and B_k respectively, are correlated zero-mean Gaussian random variables whose variances and correlation coefficients are completely determined by the parent population's PSD mean μ_k and the particular shape of the time-domain windowing function. Under these general conditions, it has been shown that the population spectral variability of the PSD estimates defined by equation (2) is given by

$$V_k^2 = 1 + |W_{2k}|^2, \tag{3}$$

where

$$W_{2k} = \frac{1}{\sum w_n^2} \sum_{n=0}^{N-1} w_n^2 e^{-4\pi i k n/N}, \tag{4}$$

[1] Center For Solar-Terrestrial Research, New Jersey Institute of Technology, Newark, NJ 07102.

ranging from 0 to 1, is the normalized DFT of the squared time-domain window evaluated at the even-indexed discrete frequencies f_{2k}. A separate treatment of the PSD estimates for the DC and Nyquist frequency bins is needed, since those obey different statistics, and we showed that they form a χ^2 distribution with one degree of freedom, identical to the case of time-domain kurtosis (TDK; Ruf et al. 2006). We further showed that the expression given by equation (3) remains valid for these bins, with $|W_0|^2 = |W_N|^2 = 1$ exactly, so that the theory encompasses the case of TDK.

Paper I also showed that, in the most common case of a symmetrical time-domain window, i.e., an even function which has only real DFT coefficients, the DFT coefficients A_k and B_k, used to compute the PSD estimate $\widehat{P_k} = A_k^2 + B_k^2$, become uncorrelated zero-mean Gaussian random variables with variances $\sigma^2_{A_k} = (1 + W_{2k})\frac{\mu_k}{2}$ and $\sigma^2_{B_k} = (1 - W_{2k})\frac{\mu_k}{2}$, and their joint distribution function is a χ^2 distribution with two degrees of freedom. Moreover, we showed that choosing any even windowing function, such as the standard Hanning or Hamming windows, results in W_{2k} values that are practically zero at all but a few frequency bins in the vicinity of the DC and Ny quist channels. In this common case, the variances of the DFT coefficients A_k and B_k become equal, and the probability distribution function of the PSD estimate $\widehat{P_k}$ simplifies to an exponential distribution

$$p(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \tag{5}$$

where x stands for the PSD random variate, and μ should be understood as having a frequency bin dependence.

In the limit of sufficiently large number, M, of accumulations, Paper I provided a first order approximation of the variance of the SK estimator defined by equation (1), which under the conditions leading to the probability distribution function given by equation (5), reduces to the simple expression

$$\sigma^2_{V_k^2} = \begin{cases} \frac{24}{M} + O(\frac{1}{M^2}), & k = 0, \frac{N}{2} \\ \frac{4}{M} + O(\frac{1}{M^2}), & k = 1, \ldots, (\frac{N}{2} - 1) \end{cases}, \tag{6}$$

which was used to define standard, symmetrical RFI detection thresholds of $\pm 3\sigma_{V_k^2}$ around $1 + |W_{2k}|^2$ corresponding to the spectral variability of a normally distributed time-domain signal. If the estimator were itself normally distributed, these thresholds would yield a false alarm rate of 0.135% at both the high and low thresholds.

However, later tests performed with data from the KSRBL instrument in RFI-free observational bands (Gary et al. 2010) have since revealed that the statistical distribution of the estimator is noticeably skewed, even with a fairly large number of accumulated spectra, $M = 6104$. Subsequent Monte Carlo simulations performed for large numbers of SK random deviates, generated for different accumulation numbers ranging from 2 to 20000, showed that, while the variance of the SK estimator asymptotically behaves as predicted by equation (6), and its kurtosis excess approaches a zero value as $1/M$, the skewness of the SK estimator vanishes only as fast as $10/\sqrt{M}$, which is too slow a rate for assuring normal behavior of the SK estimator in the range of interest for practical applications. Moreover, the same simulations suggested that the SK estimator defined by equation (1) has a statistical bias of $1/M$, which in principle may be corrected by a redefinition based on the true statistical nature of the ratio MS_2/S_1^2 that drives its statistical behavior.

Motivated by these practical concerns for RFI detection, we searched the literature and found that the statistical distribution of the ratio MS_2/S_1^2 for a general exponential population has apparently not been fully addressed by any previous work. Given its wide application in many fields (see Nita et al. 2007 and references therein), as well as its central role in our application, we present in § 2 a detailed analysis of the statistical properties of the ratio of the sums S_2 and S_1^2 with the final goal of deriving a reliable analytical expression for computing the false-alarm probabilities associated with any choice of upper and lower RFI detection thresholds. Along the way, we formally prove the key property that the SK estimator is indeed independent of the radio frequency (RF) power, S_1, and in § 3 amend our earlier expression, equation (1), to obtain an unbiased estimator of the PSD spectral variability. In § 4, based on Pearson's analysis of moments (Pearson 1985), we provide analytical and numerical procedures for calculating the SK pdf and associated cumulative function (CF). In § 5 we extend the results to the case of time-domain kurtosis. We summarize the results in § 6.

2. STATISTICS OF EXPONENTIALLY DISTRIBUTED RANDOM VARIABLES

2.1. Linear Correlation Coefficient of the Mean of Squares and the Square of Mean

To derive a first-order approximation of the variance of the SK estimator, we first note that the ratio MS_2/S_1^2 is the same as $m_2'/m_1'^2$, where $m_1' = S_1/M$ and $m_2' = S_2/M$ denote the first and second moments of the PSD estimate about the origin. Therefore, our problem reduces to the problem of finding the statistical properties of this ratio for an *exponential distribution*. Although the general problem of determining the probability distribution function of a ratio of two random variables has a well-established framework addressed by most of the classical textbooks (e.g., Kendall & Stuart 1958, p. 265), the practical applications of this framework usually deal with ratios of independent (uncorrelated) random variables, while the more general case of two correlated random variables has been almost entirely limited to the case of two *normally distributed* random variables (Fieller 1932; Hinkley 1969), for which the joint probability distribution function is exactly known (e.g., Kendall & Stuart 1958, p. 283). To investigate the nature of the joint distribution of the random deviates $m_1'^2$ and m_2', we first show in

Figure 1 a contour plot of the Monte Carlo-generated joint distribution of pairs of random variables representing the squared mean, $m_1'^2 = (S_1/M)^2$ and mean of squares $m_2' = S_2/M$ obtained from sets of M random deviates extracted from an exponential distribution of mean $\mu = 1$. The inset of the figure gives the sample means and standard deviations, $\langle m_1'^2 \rangle = 1.00 \pm 0.03$ and $\langle m_2' \rangle = 2.00 \pm 0.06$, as well as the linear correlation coefficient $r = 0.8946$ defined by

$$r = \frac{\langle (m_1'^2 - \langle m_1'^2 \rangle)(m_2' - \langle m_2' \rangle) \rangle}{\sqrt{\langle (m_1'^2 - \langle m_1'^2 \rangle)^2 \rangle \langle (m_2' - \langle m_2' \rangle)^2 \rangle}}. \tag{7}$$

This illustrates that the squared mean and the mean of the squares are strongly correlated random variables, which is not surprising given that they are constructed from a common set of M exponentially distributed independent random variables.

Without knowing their joint statistical distribution, one may still estimate the linear correlation coefficient

$$\rho(m_1'^2, m_2') = \frac{\mathtt{Cov}(m_1'^2, m_2')}{\sqrt{\mathtt{Var}(m_1'^2)\mathtt{Var}(m_2')}}, \tag{8}$$

from the exact formula for the covariance and variance (covariance with itself) for the sample moments about the origin of any distribution in terms of the corresponding population moments (Kendall & Stuart 1958, p. 229),

$$\mathtt{Cov}(m_q', m_r') = \frac{1}{M}(\mu_{q+r}' - \mu_q'\mu_r'), \tag{9}$$

as well as the first-order approximations (Kendall & Stuart 1958, p. 232) for the variance and covariance of any pair of functions of random variables given by

$$\mathtt{Cov}(f, g) = \sum_{q=1}^{2}\sum_{r=1}^{2} \frac{\partial f}{\partial m_q'}\frac{\partial g}{\partial m_r'}\mathtt{Cov}(m_q', m_r'), \tag{10}$$

where the partial derivatives with respect to the sample moments have to be evaluated in $m_1' = \mu_1'$ and $m_2' = \mu_2'$, respectively. Since in our case we have $f(m_1') = m_1'^2$ and $g(m_2') = m_2'$, these two formulae lead to

$$\mathtt{Var}(m_1'^2) = \mathtt{Cov}(m_1'^2, m_1'^2) = \frac{4}{M}\mu_1'^2(\mu_2' - \mu_1'^2), \tag{11}$$

$$\mathtt{Var}(m_2') = \mathtt{Cov}(m_2', m_2') = \frac{1}{M}(\mu_4' - \mu_2'^2), \tag{12}$$

$$\mathtt{Cov}(m_1'^2, m_2') = \frac{2}{M}\mu_1'(\mu_3' - \mu_1'\mu_2'), \tag{13}$$

which are results that hold for any distribution. For an exponential distribution characterized by $\mu_n' = n!\mu^n$, these general results become

$$\mathtt{Var}(m_1'^2) = \frac{4}{M}\mu^4, \tag{14}$$

$$\mathtt{Var}(m_2') = \frac{20}{M}\mu^4, \tag{15}$$

$$\mathtt{Cov}(m_1'^2, m_2') = \frac{8}{M}\mu^4, \tag{16}$$

which, when entered in equation (8), leads to the exact result

$$\rho(m_1'^2, m_2') = \frac{2}{\sqrt{5}} \simeq 0.8944. \tag{17}$$

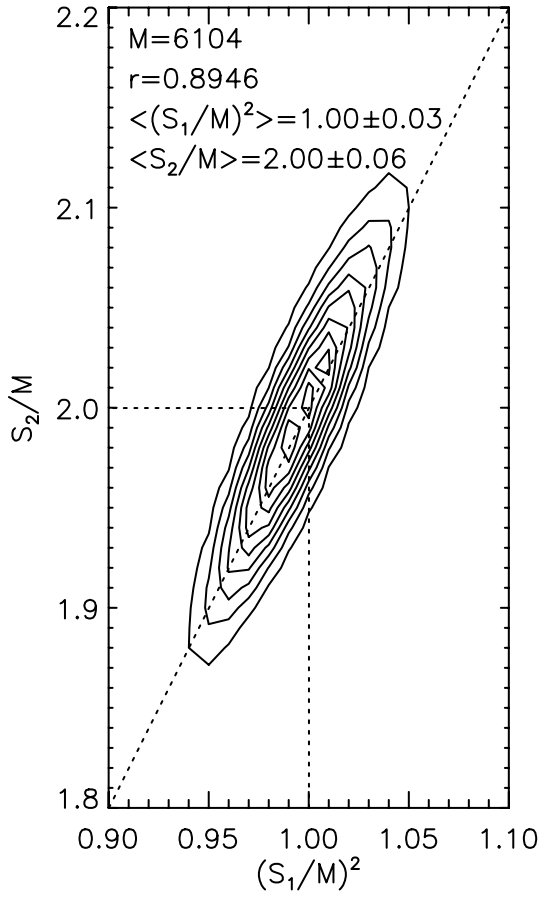

FIG. 1.—Contour plot of the Monte Carlo-generated joint distribution of $N = 327{,}520$ pairs of random variables representing the squared mean $(S_1/M)^2$ and mean of squares S_2/M for sets of $M = 6104$ random deviates extracted from an exponential distribution of mean $\mu = 1$. The contour levels are 10% apart, and the observed means $\langle (S_1/M)^2 \rangle$ and $\langle S_2/M \rangle$, as well as the observed linear correlation coefficient r are indicated on respective means as coordinates.

Thus, we find near-perfect agreement with the observed correlation coefficient $r \simeq 0.8946$ obtained from the numerical simulations presented in Figure 1.

2.2. Central Limit Theorem Approximation

In Figure 1, the quasi-elliptical contours of the numerically simulated joint distribution of the random variables $m_1'^2$ and m_2' look like those characteristic of the joint distribution of a pair of correlated normal variables, for which the pdf and CF have been previously obtained in closed analytical form (Fieller 1932; Hinkley 1969) suitable for accurate estimation of the tail probabilities we eventually are interested in. This approach seems to be justified by the fact that the random variable m_2' exactly satisfies the conditions of the central limit theorem (CLT) (Kendall & Stuart 1958 p. 193), which states that the sample mean of any statistical population characterized by a population mean μ' and variance σ'^2 tends toward a normal distribution as the number of samples M used to compute the mean increases. The same theorem assures that the distribution of the mean m_1' would approach normality in the same manner as m_2', though it remains to be proved whether the two distributions approach normality at the same pace. Moreover, a generalization of the central limit theorem presented by Kendall & Stuart (1958, p. 195) states that the means of any two random variables drawn from populations having defined variances tend toward joint bivariate normality as the number of samples increases. However, once an asymptotic distribution is obtained under such conditions, it remains to be proven whether it is valid for a finite number M of accumulated samples.

We will show later that the means and variances of the parent populations of the random variables $m_1'^2$ and m_2' may be exactly computed for any M. However, to obtain a CLT-based approximation of the $m_2'/m_1'^2$ ratio distribution, we need only the first-order estimates of the variances given by equation (14) and equation (15), which have the required asymptotic behavior $1/M$, and the means μ^2 and $2\mu^2$, respectively, which are at least asymptotically valid as shown by the numerical results presented in Figure 1. The parameters $\theta_1 = \mu^2$, $\theta_2 = 2\mu^2$, $\sigma_1^2 = 4\mu^4/M$, $\sigma_2^2 = 20\mu^4/M$, and $\rho = 2/\sqrt{5}$ may be directly entered in the expression given by Hinkley (1969) for the probability distribution function of the ratio of two correlated random variables, to obtain the CLT approximation for the probability density function of the ratio $v = m_2'/m_1'^2$ as

$$f(v) = \frac{1}{(v^2 - 4v + 5)}\left[\frac{1}{\pi}e^{-M/8} + \frac{1}{2\sqrt{2\pi}}\,\mathrm{Erf}\left(\frac{\sqrt{M/2}}{2\sqrt{v^2 - 4v + 5}}\right)e^{-\frac{M(v-2)^2}{8(v^2-4v+5)}}\right], \quad (18)$$

where $\mathrm{Erf}(z) = (2/\sqrt{\pi})\int_{-\infty}^{z}\mathrm{Exp}(-t^2)dt$ is the well-known standard error function.

Although not explicitly shown here, we find that while this expression does yield nonsymmetric behavior (i.e., nonzero skew), the skew of this asymptotic expression vanishes more rapidly than the value $10/\sqrt{M}$ obtained from simulations, which rules out the use of this approximation for finite M.

2.3. Statistical Moments of the Ratio between the Mean of Squares and the Square of Mean

Although the challenge of finding the true joint distribution of mean of squares and the square of means of a set of M exponentially-distributed independent random variables is itself a problem of theoretical interest, we will show in this section that finding the distribution of their ratio may be attacked from a different perspective, immediately leading to the result we are looking for.

The more mathematically convenient solution we wish to develop is suggested by the property of the population spectral variability (equation [2]), whose value $\sigma^2/\mu^2 \equiv 1$ is by definition uncorrelated with the square of the population mean, μ^2. Consequently, the same property is expected to hold for the ratio $m_2'/m_1'^2$ based on samples of the parent population—a property implicitly assumed by the whole concept of the SK estimator. Figure 2 displays contour levels of the joint distribution of $m_2'/m_1'^2$ and $m_1'^2$ built using the same Monte Carlo data set as in Figure 1. In contrast to the previous figure, the circular shape of these contours, as well as the practically null linear correlation coefficient, suggest that these two parameters are, indeed, uncorrelated.

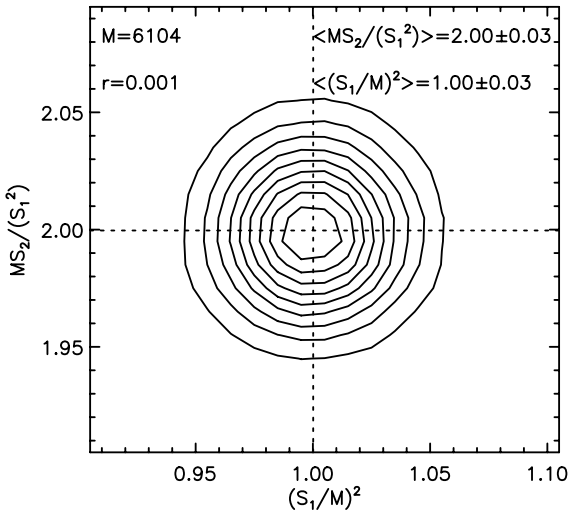

FIG. 2.—Contour plot of the Monte Carlo-generated joint distribution of $N = 327{,}520$ pairs of random variables representing the squared mean, $(S_1/M)^2$ and the ratio of the square of mean and the squared mean MS_2/S_1^2 for sets of $M = 6104$ random deviates extracted from an exponential distribution of mean $\mu = 1$. The contour levels are 10% apart, and the observed means $\langle MS_2/S_1^2\rangle$ and $\langle S_2/M\rangle$, as well as the observed linear correlation coefficient r are indicated.

To analytically prove this essential property, we employ the general formulae given by equations (9)–(13), as well as the identity $\mu'_n = n!\mu^n$, to compute the covariance of the functions $f(m'_2, m'_1) = m'_2/m'^2_1$ and $g(m'_1) = m'^2_1$ as

$$\begin{aligned}\mathrm{Cov}(f,g) &= \frac{\partial f}{\partial m'_2}\frac{\partial g}{\partial m'_1}\mathrm{Cov}(m'_1, m'_2) \\ &\quad + \frac{\partial f}{\partial m'_1}\frac{\partial g}{\partial m'_1}\mathrm{Cov}(m'_1, m'_1) \\ &= \frac{1}{M}\left[\frac{2}{\mu'_1}(\mu'_3 - \mu'_1\mu'_2) - \frac{4\mu'_2}{\mu'^2_1}(\mu'_2 - \mu'^2_1)\right] \\ &= \frac{1}{M}(8\mu^2 - 8\mu^2) = 0. \end{aligned} \tag{19}$$

An alternative way to write the covariance in terms of the statistical expectations of two random variables leads to a useful result in the case of zero covariance, i.e.,

$$\mathrm{Cov}(x,y) = E(xy) - E(x)E(y). \tag{20}$$

Here, we use the expectation values of two random variables x and y defined as

$$E(x) = \int_{-\infty}^{+\infty} x p_x(x)dx \qquad E(y) = \int_{-\infty}^{+\infty} y p_y(y)dy, \tag{21}$$

where

$$p_x(x) = \int_{-\infty}^{+\infty} p(x,y)dy \qquad p_y(y) = \int_{-\infty}^{+\infty} p(x,y)dx \tag{22}$$

are the marginal pdf's of x and y distributed according to the joint distribution $p(x,y)$. Thus, the covariance of m'_2/m'^2_1 and m'^2_1 is

$$\mathrm{Cov}\left(\frac{m'_2}{m'^2_1}, m'^2_1\right) = E(m'_2) - E\left(\frac{m'_2}{m'^2_1}\right)E(m'^2_1) = 0, \tag{23}$$

which immediately leads to the nontrivial result

$$E\left(\frac{m'_2}{m'^2_1}\right) = \frac{E(m'_2)}{E(m'^2_1)}, \tag{24}$$

expressing the fact that the statistical expectation of the ratio between the mean of squares and the squares of the mean of M independent random variables exponentially distributed is given by the ratio of the statistical expectations of quotients computed from their marginal probability density functions.

Moreover, since equation (10) gives for any integer power of n,

$$\mathrm{Cov}(f^n, g^n) = n^2 f^{n-1} g^{n-1}\mathrm{Cov}(f,g) = 0, \tag{25}$$

similar steps to those used to derive equation (24) lead to the more general result

$$E\left[\left(\frac{m'_2}{m'^2_1}\right)^n\right] = \frac{E(m'^n_2)}{E(m'^{2n}_1)}, \tag{26}$$

which states that the same relationship holds for any moment of the distribution.

Since it may be shown that, under certain conditions that we will address later (Kendall & Stuart 1958, p. 111), the complete set of the moments of a probability distribution function uniquely determines that distribution function, the more complex problem of finding the probability density function of the ratio (m'_2/m'^2_1) from the joint distribution function of the quotients may be reduced, at least from the perspective of deriving the analytical expressions of all of its moments, to the much simpler problem of deriving the moments of the quotients from their marginal distribution functions, which we obtain in the following sections.

2.4. The Generalized Gamma Distribution

The expectations $E(m'^{2n}_1)$ may be straightforwardly derived by a simple change of variable performed on the known pdf of the sum of M independent random deviates, $(x = S_1)$, drawn from an exponential population, which is a gamma distribution of integer shape parameter $k = M$ and scale parameter $\lambda = 1/\mu$ originally derived by Erlang (1917):

$$p(x) = \frac{x^{M-1}e^{-\frac{x}{\mu}}}{\mu^M(M-1)!}. \tag{27}$$

However, we provide in this section an alternative derivation of these moments that will help us establish a common framework that we will later use to derive the expectations $E(m'^n_2)$, both of which are needed to solve equation (26).

Stacy (1962) gives a detailed analysis of the properties of the generalized gamma distribution (GGD) defined as

$$f(x,a,d,p) = \frac{p x^{d-1}e^{-(\frac{x}{a})^p}}{a^d\Gamma(d/p)}, \tag{28}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the well-known Euler's Gamma function, which reduces to $(n-1)!$ for integer arguments $z = n$. The GGD defined by equation (28) reduces to the classical gamma distribution, for $p = 1$, and to Erlang's distribution for integer shape parameters $d = M$. Although beyond the scope of this study, it is worth mentioning that, for various combination of parameters, GGD reduces to other classical distributions such as Weibull, Maxwell, and Rayleigh distributions (Lienhard & Meyer 1967). The exponential distribution, which plays the central role in this study, is also a GGD given by $f(x, \mu, 1, 1)$.

One of the main results provided by Stacy (1962) is the moment-generating function of GGD, $\mathbf{M}(t, a, d, p)$, which is given in terms of an infinite analytical series as

$$\mathbf{M}(t, a, d, p) = \frac{1}{\Gamma(d/p)} \sum_{r=0}^{\infty} \frac{(at)^r}{r!} \Gamma\left(\frac{d+r}{p}\right). \quad (29)$$

This provides the population moments about the origin as

$$\mu'_n = \left. \frac{\partial^n \mathbf{M}(t, a, d, p)}{\partial t^n} \right|_{t=0} = \frac{\Gamma(\frac{d+n}{p})}{\Gamma(d/p)} a^n, \quad (30)$$

which reduces to the known result $\mu'_n = n!\mu^n$ for the exponential distribution $f(x, \mu, 1, 1)$.

Since, generally, the moment-generating function of the sum of independent variables is the product of their individual moment-generating functions, equation (29) provides the direct means to compute the moments about the origin of the mean of M independent random variables GGD distributed by using the formula

$$\mu'_n = \left. \frac{\partial^n [\mathbf{M}(t, \frac{a}{M}, d, p)]^M}{\partial t^n} \right|_{t=0}, \quad (31)$$

which results from the change of variable

$$\begin{aligned} f(x, a, d, p)dx &= f\left(\frac{x}{M}, \frac{a}{M}, d, p\right) d\left(\frac{x}{M}\right) \\ &= f\left(y, \frac{a}{M}, d, p\right) dy, \end{aligned} \quad (32)$$

and the identity $\frac{1}{M}\left(\sum_{i=1}^{M} x_i\right) = \sum_{i=1}^{M} y_i$.

Taking into account that only the cross terms resulting in the nth power of t may contribute to the moment μ_n, equation (31) may be reduced to

$$\mu'_n = \frac{1}{[\Gamma(d/p)]^M} \frac{\partial^n}{\partial t^n} \left[\sum_{r=0}^{n} \frac{1}{r!} \Gamma\left(\frac{d+r}{p}\right) \left(\frac{at}{M}\right)^r \right]^M \Bigg|_{t=0}, \quad (33)$$

which, after the convenient scaling $t \to (a/M)t$ of the differential operator, becomes

$$\mu'_n = \frac{(a/M)^n}{[\Gamma(d/p)]^M} \frac{\partial^n}{\partial t^n} \left[\sum_{r=0}^{n} \frac{1}{r!} \Gamma\left(\frac{d+r}{p}\right) t^r \right]^M \Bigg|_{t=0}. \quad (34)$$

However, in the particular case $p = 1$, it may be shown that the pdf of the mean of M independent random variables, $m'_1 = (1/M)\sum_{i=1}^{M} x_i) = S_1/M$, individually distributed according to the GGD function $f(x, a, d, 1)$, is also a GGD given by

$$p(m'_1) = f(m'_1, a/M, Md, 1), \quad (35)$$

which allows writing equation (35) directly in the closed form provided by equation (30).

To prove equation (35), we have to compute the convolution of the individual pdf's of the variables $y_i = x_i/M$, which we do by first computing the Fourier transform of the distribution given by equation (32) (which is its probability generating function; Kendall & Stuart 1958),

$$\Phi(t) = \int_0^{\infty} f\left(y, \frac{a}{M}, d, 1\right) e^{ity} dy = \left(1 - i\frac{at}{M}\right)^{-d}, \quad (36)$$

followed by the inverse transformation of the product of M such probability distribution functions,

$$\begin{aligned} p(m'_1) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(1 - i\frac{at}{M}\right)^{-Md} e^{-im'_1 t} dt \\ &= f(m'_1, a/M, Md, 1). \end{aligned} \quad (37)$$

This result may be further used to derive the distribution of the squared mean of M independent random deviates drawn from a $p = 1$ GGD, by making the change of variable $y = m'^2_1$

$$\begin{aligned} f(m'_1, a/M, Md, 1)d(m'_1) &= f(\sqrt{y}, a/M, Md, 1)d(\sqrt{y}) \\ &= f\left[y, \left(\frac{a}{M}\right)^2, \frac{Md}{2}, \frac{1}{2}\right] dy, \end{aligned} \quad (38)$$

which leads to the distribution

$$p(m'^2_1) = f\left[m'^2_1, \left(\frac{a}{M}\right)^2, \frac{Md}{2}, \frac{1}{2}\right], \quad (39)$$

and, through equation (30), to the expectations corresponding to the denominator of equation (26),

$$E(m'^{2n}_1) = \frac{\Gamma(Md + 2n)}{\Gamma(Md)} \left(\frac{a}{M}\right)^{2n}. \quad (40)$$

To evaluate the expectations corresponding to the numerator of equation (26), we consider a random variable x distributed by $f(x, a, d, 1)$, and perform the change of variable

$$f(x, a, d, 1)dx = f(\sqrt{y}, a, d, 1)d(\sqrt{y}) = f(y, a^2, d/2, 1/2)dy, \quad (41)$$

to obtain the probability density function of the square of a $p = 1$ GGD-distributed random variable, $y = x^2$,

$$p(y) = f(y, a^2, d/2, 1/2), \quad (42)$$

which is a GGD function of noninteger $p = 1/2$, for which the particular results derived from equation (35) no longer apply.

However, since $m_2' = \frac{1}{M}\sum_{i=1}^{M} x_i^2 = \frac{1}{M}\sum_{i=1}^{M} y_i$ is the mean of M random variables distributed according to the GGD function given by equation (42), we may use equation (37) to express the statistical expectations for any power of the random variable m_2' as

$$E(m_2'^n) = \frac{(a/\sqrt{M})^{2n}}{[\Gamma(d)]^M}\frac{\partial^n}{\partial t^n}\left[\sum_{r=0}^{n}\frac{1}{r!}\Gamma(2r+d)t^r\right]^M\Bigg|_{t=0}. \quad (43)$$

2.5. Standard Moments of the Ratio MS_2/S_1^2

Since the exponential distribution is the GGD function $p(x) = f(x,\mu,1,1)$, i.e., $p=1$, $d=1$, and $a=\mu$, equations (40) and (43) give

$$E(m_1'^{2n}) = \frac{(M-1+2n)!}{(M-1)!}\left(\frac{\mu}{M}\right)^{2n}$$

$$E(m_2'^n) = \left(\frac{\mu}{\sqrt{M}}\right)^{2n}\frac{\partial^n}{\partial t^n}\left[\sum_{r=0}^{n}\frac{(2r)!}{r!}t^r\right]^M\Bigg|_{t=0}. \quad (44)$$

Substituting these results into equation (26), and recalling that $m_2'/m_1'^2 = MS_2/S_1^2$, we get

$$E\left[\left(\frac{MS_2}{S_1^2}\right)^n\right] = \frac{M^n(M-1)!}{(M-1+2n)!}\frac{\partial^n}{\partial t^n}\left[\sum_{r=0}^{n}\frac{(2r)!}{r!}t^r\right]^M\Bigg|_{t=0}, \quad (45)$$

which completely determines the statistical properties of the ratio MS_2/S_1^2 in terms of its infinite set of moments about the origin, if (Kendall & Stuart 1958, p. 111) the upper limit of the expression

$$\frac{1}{2n}\left\{E\left[\left(\frac{MS_2}{S_1^2}\right)^{2n}\right]\right\}^{1/2n} \quad (46)$$

is finite.

Since a simple numerical evaluation of equation (46) shows that it is monotonously decreasing while being positively defined, we may conclude that its limit is finite and, therefore, that the probability distribution function of the ratio MS_2/S_1^2 is uniquely determined by its complete set of moments provided by equation (45).

For $n=1$, equation (45) provides the expectation for the ratio MS_2/S_1^2

$$E\left(\frac{MS_2}{S_1^2}\right) = \frac{2M}{M+1}, \quad (47)$$

and, using the general conversion formula (Kendall & Stuart 1958, p. 56) that relates the central moments to the raw moments of any distribution, we get

$$E\left\{\left[\frac{MS_2}{S_1^2} - E\left(\frac{MS_2}{S_1^2}\right)\right]^n\right\} = \sum_{k=0}^{n}\left\{(-1)^k\binom{n}{k}\left(\frac{2M}{M+1}\right)^k E\left[\left(\frac{MS_2}{S_1^2}\right)^{n-k}\right]\right\}. \quad (48)$$

3. AN UNBIASED SK ESTIMATOR AND ITS STATISTICAL MOMENTS

The result expressed by equation (47) may be used to evaluate the expectation of the SK estimator defined by equation (1)

$$E(\widehat{V_k^2}) = E\left[\frac{M}{M-1}\left(\frac{MS_2}{S_1^2}-1\right)\right] = \frac{M}{M-1}\left(\frac{2M}{M+1}-1\right) = \frac{M}{M+1}, \quad (49)$$

which shows that the SK estimator defined in Paper I is, indeed, a biased estimator.

Therefore, we conveniently rescale the estimator originally defined in Paper I by multiplying by $(M+1)/M$ to define a new estimator

$$\widehat{\mathrm{SK}} = \frac{M+1}{M-1}\left(\frac{MS_2}{S_1^2}-1\right), \quad (50)$$

which, for a set of data samples drawn from an exponential distribution, has unit expectation

$$\mu_1' \equiv E(\widehat{\mathrm{SK}}) = 1, \quad (51)$$

the same as the population parameter it is intended to estimate.

To completely determine the statistical properties of the unbiased estimator $\widehat{\mathrm{SK}}$, we derive the formula providing the moments relative to its mean,

$$\mu_n \equiv E\{[\widehat{\mathrm{SK}} - E(\widehat{\mathrm{SK}})]^n\} = \left(\frac{M+1}{M-1}\right)^n E\left\{\left[\frac{MS_2}{S_1^2} - E\left(\frac{MS_2}{S_1^2}\right)\right]^n\right\}, \quad (52)$$

which shows that the central moments of the SK estimator may be written in terms of the corresponding central moments of the MS_2/S_1^2 ratio. Taking into consideration the scaling factor $[(M+1)/(M-1)]^n$ present in equation (52), it is evident that, except for a different mean and variance, which represent the scale parameters of the probability distribution function of the estimator $\widehat{\mathrm{SK}}$, the normalized higher moments $\mu_n/\mu_2^{n/2}$ are identical for the $\widehat{\mathrm{SK}}$ and MS_2/S_1^2 distributions, indicating that their shapes are identical. Note, however, that in both cases, the scale and shape parameters of the distributions are determined by the unique parameter M, which is the only variable

that enters all equations related to the statistical properties of the SK estimator.

Considering now only the standard statistical parameters that may be derived from the first four moments, equations (51) and (52) provide

$$\mu_1' = 1 \qquad \mu_2 = \frac{4M^2}{(M-1)(M+2)(M+3)}$$
$$\beta_1 = \frac{4(M+2)(M+3)(5M-7)^2}{(M-1)(M+4)^2(M+5)^2}$$
$$\beta_2 = \frac{3(M+2)(M+3)(M^3+98M^2-185M+78)}{(M-1)(M+4)(M+5)(M+6)(M+7)} \tag{53}$$

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, $\gamma_2 = \beta_2 - 3$. The first-order approximation in $1/M$ of these results gives expressions

$$\mu_2 \simeq \frac{4}{M} + O\left(\frac{1}{M^2}\right) \qquad \gamma_1 \simeq \frac{10}{\sqrt{M}} + O\left(\frac{1}{M^{3/2}}\right)$$
$$\gamma_2 \simeq \frac{246}{M} + O\left(\frac{1}{M^2}\right), \tag{54}$$

which are in full agreement with the first-order approximation for variance of the SK estimator derived in Paper I, as well as with its asymptotic behavior estimated from numerical simulations. However, the expressions given by equation (53) are exact for any value of M as illustrated in Figure 3, which shows a perfect match between them and the corresponding parameters derived from simulations for $2 \leq M \leq 8196$.

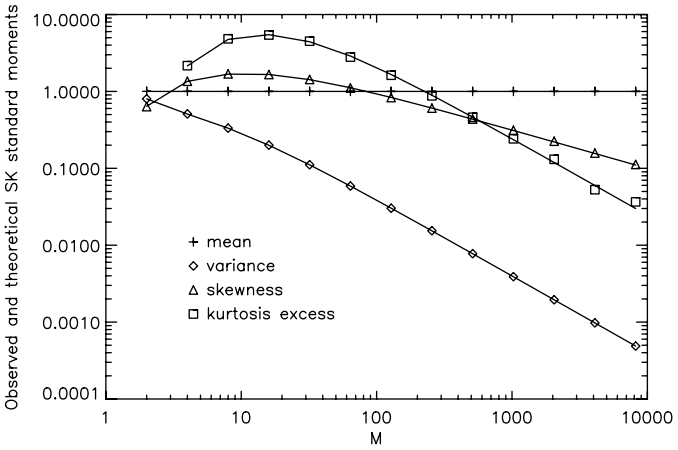

FIG. 3.—Comparison between the standard moments (mean–*crosses*; variance–*diamonds*; skewness–*triangles*; kurtosis excess–*squares*) of the numerically simulated SK distributions with their theoretical expectations (*solid lines*), for different accumulation lengths ($M = 2, 4 \ldots 8196$). Each individual SK distribution corresponding to a particular value of M has been built out of 1,000,000 sums S_1 and S_2. For any value of M, the match of the theoretical expectations and observed random deviates is evident. Note that for $M = 2$, the sample and theoretical kurtosis excess do not appear on the plot due to their negative values (-0.87 and -0.86, respectively).

4. MOMENT-BASED APPROXIMATION OF THE SK PDF AND CF

Deriving the exact expressions of the SK statistical moments is just the first step toward the final goal of determining the cumulative probability function needed to compute the tail probabilities we are interested in. Although, as mentioned before, knowing its exact moments of all orders is theoretically equivalent with knowing the probability distribution function itself, to obtain the latter may involve challenging analytical difficulties without any guarantee of obtaining a closed form solution. Although this approach may be worth investigating in a separate study, here we limit ourselves to approximating the SK distribution to sufficient accuracy that we may derive its tail probabilities for practical applications.

4.1. Pearson's Probability Curves

In his classic work, Pearson (1985) provided a standard approach to the problem of finding accurate analytical approximations to the true distribution functions based on its first four moments derived from observations. Pearson's approach may be straightforwardly applied to the problem of finding an approximation to the $\widehat{\mathrm{SK}}$ distribution, for which we have the advantage of knowing not only its exact first four moments, but also any higher moment that may be subsequently compared with moments of the approximating distribution as a consistency check. We start with Pearson's criterion defined as (Kendall & Stuart 1958, p. 151)

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}, \tag{55}$$

where, in our case, the exact values of the parameters β_1 and β_2 are provided by equation (53). Pearson's criterion, which in the case of the $\widehat{\mathrm{SK}}$ distribution turns out to be a ratio between two polynomials of order 8, is plotted in Figure 4. Three distinct regions, corresponding to $M \in [2, 5]$, $M \in [6, 23]$, and $M \geq 24$, are discriminated according to Pearson's classification as type I, type VI, and type IV, respectively. Since it may be shown that, for $M \to \infty$, k asymptotically approaches from above the limit $k_\infty = 25/64$, these are the only types applicable to this problem. Although the cases corresponding to $M \leq 23$ are of little interest in RFI detection due to the large variances of the estimator, we do not rule out the possibility that this region may be of interest for other applications. For completeness, therefore, we just mention here that Pearson types I and VI pdf's correspond to the standard beta distributions of first and second kinds, respectively, and refer the reader to the original comprehensive study of Pearson (1985), which details how the

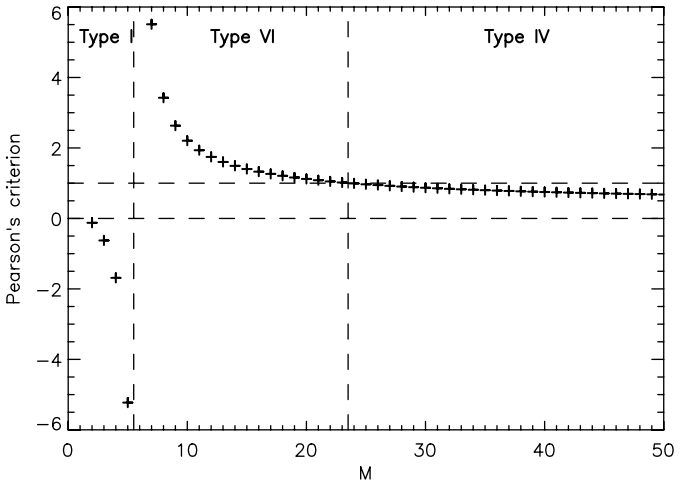

FIG. 4.—Pearson's criterion for $M \in [2, 50]$. The two horizontal lines at $\kappa = 0$ and $\kappa = 1$ are used to discriminate three distinct regions $k < 0$, $k > 1$, and $0 < k < 1$, corresponding to the types I, VI, and IV, respectively, which are separated by the two vertical lines lying between $M = 5$ and $M = 6$ and between $M = 23$ and $M = 24$.

parameters defining these distributions can be related to the observed moments.

4.2. Pearson Type IV Probability Distribution Function

The most general analytical form of the Pearson type IV pdf originally introduced by Pearson (1985), including its nontrivial normalization factor, was given by Nagahara (1999) as

$$p(x) = \frac{1}{a\sqrt{\pi}} \frac{\Gamma(m + i\frac{\nu}{2})\Gamma(m - i\frac{\nu}{2})}{\Gamma(m - \frac{1}{2})\Gamma(m)} \times \left[1 + \left(\frac{x-\lambda}{a}\right)^2\right]^{-m} \mathtt{Exp}\left[-\nu \mathtt{ArcTan}\left(\frac{x-\lambda}{a}\right)\right], \quad (56)$$

where the four parameters m, μ, a, and λ can be expressed in terms of the central moments of the distribution as (Heinrich 2004)

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6} \qquad m = \frac{r+2}{2}$$
$$\nu = -\frac{r(r-2)\sqrt{\beta_1}}{\sqrt{16(r-1) - \beta_1(r-2)^2}}$$
$$a = \frac{1}{4}\sqrt{\mu_2(6(r-1) - \beta_1(r-2)^2)}$$
$$\lambda = \mu - \frac{1}{4}(r-2)\sqrt{\mu_2\beta_1}. \quad (57)$$

It may be shown by simple derivation that the pdf described by equation (56), which is defined on the entire real axis, is unimodal and reaches its maximum at

$$x_{\mathrm{mode}} = \lambda - \frac{a\nu}{2m}. \quad (58)$$

Figure 5 displays the Pearson IV approximations for $M = 32$, 1024, 4096, and 8192. By visual inspection , we can conclude that the Pearson IV approximations accurately reproduce the shapes of the numerically simulated histograms for different orders of magnitude of the accumulation length.

However, to allow a quantitative evaluation of the accuracy of our approximation, we have evaluated the errors of the fifth central moment of the Pearson IV curves, computed according to the recursive formula (Heinrich 2004)

$$\mu_0 \equiv 0; \qquad \mu_1 = 0$$
$$\mu_n = \frac{a(n-1)}{r^2[r-(n-1)]}[-2\nu r\mu_{n-1} + a(r^2+\nu^2)\mu_{n-2}], \quad (59)$$

relative to the exact values provided by equation (52). It was found that the Pearson IV curves, which are based on the exact first four moments of the $\widehat{\mathrm{SK}}$ distribution, reproduce the fifth moment of the true distribution with a relative error that is not larger than 5% for any accumulation length $24 \leq M \leq 1000$ and approaches zero as M increases beyond this interval.

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

$$P(x) = \int_{-\infty}^{x} p(x)dx, \qquad 1 - P(x) = \int_{x}^{\infty} p(x)dx, \quad (60)$$

for which Heinrich (2004) provided the following closed form

$$P(x) = \begin{cases} 1 + P_1(m, \nu, a, \lambda, x), & x < \lambda - a\sqrt{3} \\ P_2(m, \nu, a, \lambda, x), & |x - \lambda| < a\sqrt{3} \\ 1 - P_1(m, -\nu, a, -\lambda, -x), & x > \lambda + a\sqrt{3} \end{cases}, \quad (61)$$

where

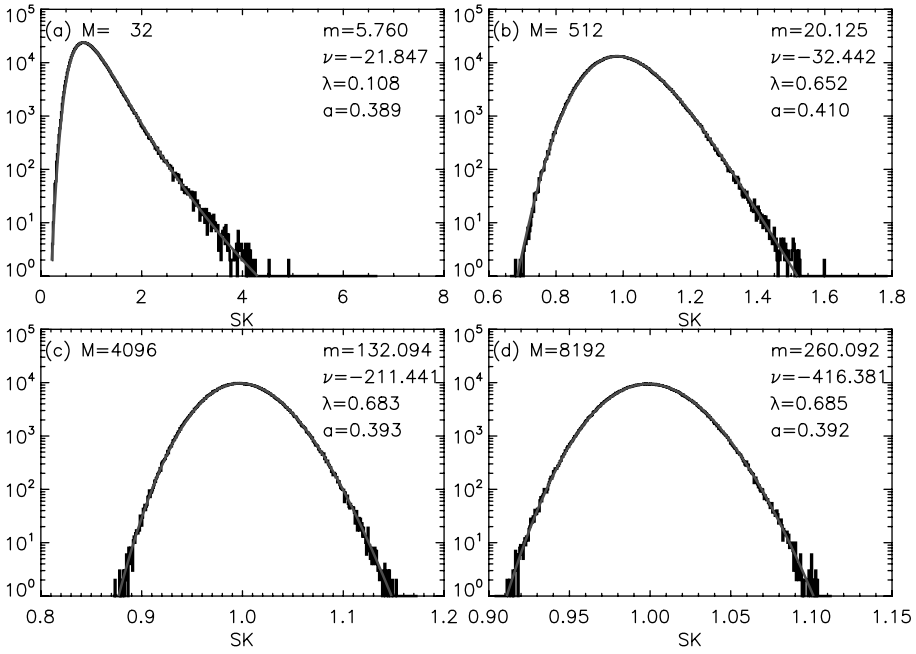

FIG. 5.—Comparison between the SK distributions obtained by numerical simulation for different accumulation lengths M and their corresponding Pearson type IV approximations. The four Pearson type IV parameters computed according to equation (57), m, ν, λ, and a are displayed on each plot. See the electronic edition of the *PASP* for a color version of this figure.

$$P_1(m,\nu,a,\lambda,x) = \frac{a}{2m-1}\left(i - \frac{x-\lambda}{a}\right)p(x)$$
$$\times F\left(1, m + i\frac{\nu}{2}, 2m, \frac{2}{1 - i\frac{x-\lambda}{a}}\right)$$
$$P_2(m,\nu,a,\lambda,x) = \frac{1}{1 - e^{-(\nu+i2m)\pi}} - \frac{ia}{i\nu - 2m + 2}\left[1 + \left(\frac{x-\lambda}{a}\right)^2\right]p(x)$$
$$\times F\left(1, 2-2m, 2-m+i\frac{\nu}{2}, \frac{1+i\frac{x-\lambda}{a}}{2}\right),$$

and

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

is the Gauss hypergeometric series.

The theoretical convergence of equation (61) is assured by the condition $|z| < 1$, (Erdelyi et al. 1953; Abramowitz & Stegun 1965), though its numerical convergence may be a delicate matter for a certain combination of the parameters involved (e.g., Michel & Stoitsov 2008), especially for the parameters m and ν shown in Figure 5, whose absolute values are much larger than unity. However, due to the particularities of the hypergeometric series, which allows many equivalent representations (Erdelyi et al. 1953), the representation of the Pearson type IV CF given in equation (61) is not unique, which leaves open the possibility of finding a more computationally efficient representation tailored for a specific combination of parameters. For example, more recently, Willink (2008), apparently unaware of the previous result provided by Heinrich (2004), found a different representation for the Pearson type IV CF, which is

$$P(m,\nu,a,\lambda,x) = \frac{e^{-[\lambda - i(2-2m)]\Phi}R - 1}{e^{-[\lambda - i(2-2m)]\pi} - 1}, \tag{62}$$

where

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

Although not explicitly addressed by Willink (2008), the representation given by equation (62) is expressed in terms of the ratio of two hypergeometric series that both have to be computed on the complex unity circle, $|z| = 1$, where the hypergeometric series is convergent if, and only if, the condition

$$\Re(\delta - \alpha - \beta) > 0 \tag{63}$$

is strictly satisfied (Erdelyi et al. 1953; Abramowitz & Stegun 1965). Fortunately, this is satisfied for any value of $M > 6$ in our case, since $\delta - \alpha - \beta = 2m - 1$.

Moreover, since the denominator of the ratio R is a hypergeometric series of argument unity, and the condition given by equation (63) is satisfied, it immediately follows (Abramowitz & Stegun 1965 15.1.20) that

$$F(2 - 2m, u, u + 1, 1) = \frac{\Gamma(2m - 1)\Gamma(u + 1)}{\Gamma(2m + u - 1)}, \tag{64}$$

which simplifies the computation. Furthermore, since for the numerator of R the difference $\delta - \beta = u + 1 - 1 = 1$ is a positive integer, the corresponding hypergeometric series is theoretically assured to terminate after a finite number of terms (Abramowitz & Stegun 1965; Erdelyi et al. 1953); disregarding the computational effort involved, this should make it possible, at least in principle, to obtain an exact result.

However, if one wants to avoid the numerical difficulties related to the evaluation of the hypergeometric series, one may choose alternatively to perform a direct numerical integration (equation [60]) of equation (56), which may achieve reasonable accuracy with far less computational effort, especially if tailored integration methods (e.g., Nagahara 1999) are employed.

Figure 6 displays the numerical results for $M = 6104$, computed according to equation (61) (*triangles*), equation (62) (*squares*), and by direct integration of equation (60) (*solid lines*). The hypergeometric series was computed using the *hypergeom* function in Maple 11 (MapleSoft), and the numerical integration was performed using the *int_tabulated* function in Interactive Data Language (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 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 (62) is better than that of equation (61), while the direct numerical integration of CF gives similar results as equation (62). After the peak of the $\widehat{\mathrm{SK}}$ distribution, the numerical accuracy of equation (61) is better than that of equation (62), while the numerical integration of CCF gives similar results as equation (61). Therefore we conclude that the numerical evaluation of equation (62) gives a more accurate estimation of the CF and the numerical evaluation of equation (61) gives a more accurate estimation of the CCF, while the direct numerical integration of equation (60) gives results of comparable accuracy at both sides of the $\widehat{\mathrm{SK}}$ distribution. The lower and higher thresholds displayed by the two vertical solid 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 $1 \pm 6/\sqrt{6104}$ (Fig. 6, *dashed lines*) originally proposed in Paper I. Although this correction seems small in absolute value for the large-M case, e.g., $M = 6104$ illustrated in Figure 6, 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 $\widehat{\mathrm{SK}}$-based RFI rejection algorithm. The correction becomes more important for lower M.

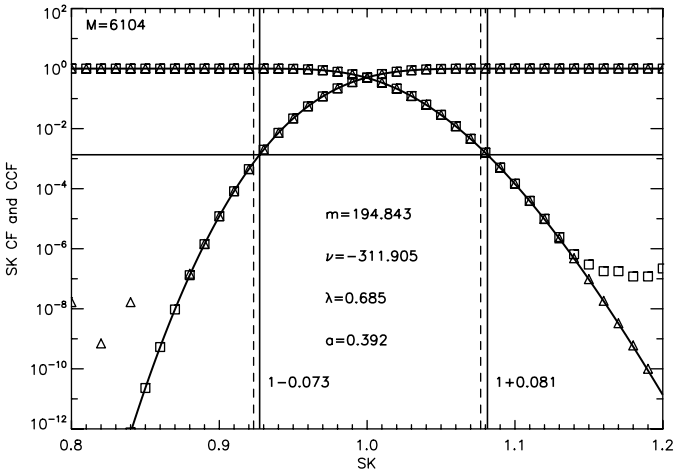

FIG. 6.—Numerical results for $M = 6104$, computed according to equation (61) (*triangles*), equation (62) (*squares*), and by direct integration of equation (60) (*solid lines*). The plot displays both the integral probabilities $P(x)$ (CF-rising curves) and complementary probabilities $1 - P(x)$ (CCF-descending curves) needed to evaluate the RFI thresholds corresponding to a symmetric standard false-alarm probability level of 0.13499% (*horizontal solid line*). The lower and upper thresholds displayed by the two *vertical solid 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 dashed lines*) originally proposed by Nita et al. (2007).

5. THE CONNECTION WITH TIME-DOMAIN KURTOSIS

As shown in Paper I, the DC and Nyquist frequency bins obey statistics identical to the case of a pure time-domain signal. The pdf at these particular frequencies is a χ^2 distribution with one degree of freedom, given by

$$p(x) = \frac{1}{\sqrt{\pi\mu}} x^{-\frac{1}{2}} e^{-\frac{x}{\mu}}, \tag{65}$$

for which the expected value of the spectral variability defined by equation (2) is 2. This is the direct consequence of the purely real nature of the DFT coefficients at these frequency bins. Thus, radio spectrograph designs based on FIR filters produce

data that, while amenable to a kurtosis-based RFI algorithm, require similar considerations to optimize the thresholds for rejection. The use of a time-domain kurtosis (TDK) estimator for the purpose of RFI mitigation has been previously proposed in several studies (Ruf et al. 2006; Johnson & Potter 2009), where the well-known variance of the estimator ($\sim 24/M$; Kendall & Stuart 1958) was employed to derive the detection thresholds needed to discriminate the RFI contamination against a Gaussian background. However, apparently no efforts have been made to investigate the statistical nature of the TDK estimator. We now investigate this as a necessary step toward improving the performance of the time-domain kurtosis estimator.

The derivation of the statistical properties of an SK estimator based on the MS_2/S_1^2 ratio in the case of a FIR filter channelization may be straightforwardly obtained using the same framework employed in the previous section for the DFT-based channelization, starting from the observation that the pdf given by equation (65) is also a GGD,

$$p(x) = f(x, \mu, 1/2, 1). \tag{66}$$

The first three raw moments of this distribution, $\mu_1' = \mu/2$, $\mu_2' = 3\mu^2/4$, and 15 $\mu^3/8$ can be entered in equation (19) to directly prove that the ratio MS_2/S_1^2 and S_1^2 are linearly uncorrelated. This result, however, should have been expected in this case as a direct consequence of the linear independence of any two moments of a Gaussian distribution, which is a fundamental statistical property (Kendall & Stuart 1958) defining such a distribution. Therefore, the moments of the ratio MS_2/S_1^2 may be computed using equation (29), once the moments of the $(S_1/M)^2$ and S_2/M are derived for the particular case of the $f(x, \mu, 1/2, 1)$ GGD.

Since the distribution given by equation (66)) is a GGD with $p = 1$, $d = 1/2$, and $a = \mu$, equations (40) and (43) provide

$$E\left[\left(\frac{S_1}{M}\right)^2\right] = \frac{\Gamma(\frac{M}{2} + 2n)}{\Gamma(\frac{M}{2})}\left(\frac{\mu}{M}\right)^{2n}$$
$$E\left(\frac{S_2}{M}\right) = \frac{(\mu^2/M)^n}{(\sqrt{\pi})^M}\frac{\partial^n}{\partial t^n}\left[\sum_{r=0}^{n}\Gamma\left(\frac{1}{2} + 2r\right)\frac{t^r}{r!}\right]^M\Bigg|_{t=0}, \tag{67}$$

which when entered into equation (26), give

$$E\left[\left(\frac{MS_2}{S_1^2}\right)^n\right] = \frac{M^n\Gamma(\frac{M}{2})}{(\sqrt{\pi})^M\Gamma(\frac{M}{2} + 2n)}\frac{\partial^n}{\partial t^n}$$
$$\times\left[\sum_{r=0}^{n}\Gamma\left(\frac{1}{2} + 2r\right)\frac{t^r}{r!}\right]^M\Bigg|_{t=0}. \tag{68}$$

Therefore,

$$E\left(\frac{MS_2}{S_1^2}\right) = \frac{3M}{M+2}, \tag{69}$$

which asymptotically tends to 3, as expected, since, for a set of time-domain samples obeying a zero-mean normal distribution, the MS_2/S_1^2 ratio is a biased estimator of the distribution kurtosis.

Equation (69) may be used to define an unbiased TDK estimator, $\widehat{K}$, for the DC and Nyquist frequencies of a DFT-based spectrograph, or for any frequency bin of a FIR-filter-based spectrograph as

$$\widehat{K} = \frac{M+2}{M-1}\left(\frac{MS_2}{S_1^2} - 1\right), \tag{70}$$

which has an expectation $E(\widehat{K}) = 2$ for an RFI-free time-domain input.

The general formula providing the central moments of this estimator,

$$\mu_n \equiv E\{[\widehat{K} - E(\widehat{K})]^n\}$$
$$= \left(\frac{M+2}{M-1}\right)^n E\left\{\left[\frac{MS_2}{S_1^2} - E\left(\frac{MS_2}{S_1^2}\right)\right]^n\right\}, \tag{71}$$

may now be used to write down the first four standard moments of its distribution in terms of the accumulation length M as

$$\mu_1' = 2 \qquad \mu_2 = \frac{24M^2}{(M-1)(M+4)(M+6)}$$
$$\beta_1 = \frac{216(M-2)^2(M+4)(M+6)}{(M-1)(M+8)^2(M+10)^2}$$
$$\beta_2 = \frac{3(M+4)(M+6)(M^3+213M^2-474M+368)}{(M-1)(M+8)(M+10)(M+12)(M+14))}, \tag{72}$$

with first order approximations in $1/M$ given by

$$\mu_2 \simeq \frac{24}{M} + O\left(\frac{1}{M^2}\right)$$
$$\gamma_1 \simeq \frac{6\sqrt{6}}{\sqrt{M}} + O\left(\frac{1}{M^{3/2}}\right)$$
$$\gamma_2 \simeq \frac{540}{M} + O\left(\frac{1}{M^2}\right). \tag{73}$$

The Pearson's criterion (equation [55]) shows in this case that the $\widehat{K}$ distribution may be approximated by a Pearson type IV curve for any $M \geq 46$, and therefore the parameters given by equation (72) may be entered in equation (57) to obtain its probability distribution function and compute the appropriate RFI detection thresholds. Figure 7 displays the pdf of the estimator $\widehat{K}$ 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. 6; see Paper I for a more detailed motivation of this choice). Despite its large accumulation length, the estimator $\widehat{K}$ still has a

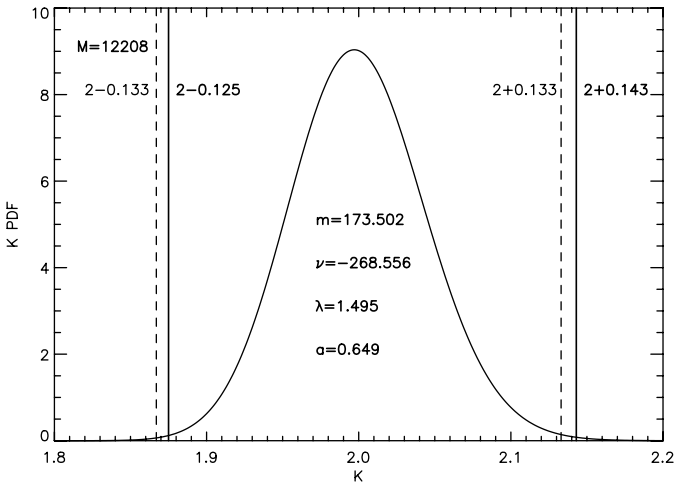

FIG. 7.—K estimator probability distribution function for a FIR filter spectrograph having an accumulation length of $M = 12208$ spectra. The two *solid vertical lines*, having the ordinates $2 - 0.125$ and $2 + 0.143$, represent the RFI detection thresholds corresponding to a symmetric standard false-alarm probability level of 0.13499%, computed using the same method as in Fig. 6. These thresholds have to be compared with the less accurate symmetric thresholds of $2 \pm 3\sqrt{24/12208} = 2 \pm 0.133$ (*vertical dashed lines*).

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.

6. CONCLUSION

In this article, 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. An important result is that we have proved (equation [19]) that the covariance of $m_2'/m_1'^2$ with $m_1'^2$ is zero, which assures the key property of the SK estimator, i.e., that $\widehat{\mathrm{SK}}$ is independent of RF power level (S_1). We also improved the definition of $\widehat{\mathrm{SK}}$ (equation [50]) relative to its original definition (equation [1]) to form an unbiased estimator, and introduced a TDK unbiased estimator (equation [70]) 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 (equations [52] and [71]), and established a common framework that allows accurate estimation of the RFI thresholds based on the first four standard moments of their probabilities distributions (equations [53] and [72]), which, for any accumulation length $M \geq 24$ and $M \geq 46$, respectively, are used to compute the four parameters (equation [57]) that completely determine the Pearson IV approximations (equation [56]) of their true pdf's. Based on these four parameters, which depend only on the accumulation length M, the CF and CCF of the SK or TDK estimators can be computed by using either the closed-form expressions provided by equations (62) and (61), respectively, or by direct numerical estimation of the integrals given by equation (60). 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 et al. (2010) 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 for numerical calculation of the thresholds for any M.

We acknowledge support for this work through NSF grant AST-0908344 and NASA grant NNG06GJ40G to the New Jersey Institute of Technology.

REFERENCES

Abramowitz, M., & Stegun, I. A. 1965, Handbook of Mathematical Functions (New York: Dover)
Dou, Y., Gary, D. E., Liu, Zhiwei, Nita, G. M., Bong, S.-C., Cho, K.-S., Park, Y.-D., & Moon, Y.-J. 2009, PASP, 121, 512
Erdelyi, et al. 1953, Higher Transcendental Functions, Vol. I (New York: McGraw Hill)
Erlang, A. K. 1917, Elektroteknikeren, 13, 5
Fieller, E. C. 1932, Biometrika, 24, 428
Gary, D. E., Liu, Z., & Nita, G. M. 2010, PASP, 122, 595
Johnson, J. T., & Potter, L. C. 2009, IEEE Trans. Geosci. Remote Sensing, 47, 628
Heinrich, J. 2004, Fermilab Collider Detector internal note 6820 (http://www-cdf.fnal.gov/publications/cdf6820_pearson4.pdf)
Hinkley, D.V. 1969, Biometrika, 56, 635
Kendall, M. G., & Stuart, A. 1958, The Advanced Theory of Statistics, Vol. 1 (London: Griffin)
Lienhard, J.ohn H., & Meyer, Paul L. 1967, Q. Appl. Math., 25, 330
Michel, N., & Stoitsov, M. 2008, Comp. Phys. Commun., 178, 535
Nagahara, Y. 1999, Statistics Probability Lett., 43, 251
Nita, G. M., Gary, D. E., Liu, Z., Hurford, G. J., & White, S. M. 2007, PASP, 119, 805
Pearson, K. 1985, Philos. Trans. R. Soc. London A, 186, 343
Ruf, C. S., Gross, S. M., & Misra, S. 2006, IEEE Trans. Geosci. Remote Sensing, 44, 694
Stacy, E. W. 1962, Ann. Math. Statist. Assoc., Vol. 33, 1187
Willink, R. 2008, Austral. NZ J. Stat., 50, 199