Spacing and shape of peaks in nonparametric spectrum estimates

Charlotte L. Haley and David J. Thomson
Queen's University, Jeffery Hall, University Avenue, Kingston, Ontario, Canada

Abstract
In the spectral analysis of time series, the goal is often to determine the frequency and power contributions of periodic components in noise. Single and multiple taper spectrum estimates are approximately chi-squared distributed, so assigning significance to high peaks is not difficult. The difficulty lies in the rate of false detection. In this study, we compute upcrossing rates for Gaussian white noise spectra of high significance levels and determine the width of these excursions. These results give rise to a new approach to the justification of peaks in processes which appear to have "many lines". We give an example of a natural process which contains hundreds of line components whose frequencies correspond approximately with the normal modes of the sun. We show that the distribution of this spectrum is best described by a mixture model of approximately 68% noncentral and 32% central chi-squared components (lines and noise, respectively), and furthermore that, upon random permutation of the data samples, the detection rate falls to that expected for a Gaussian noise process. These experiments suggest that the peaks are genuine, and exhibit extraordinary correlation structure.
Key Words: Multitaper spectrum, Spectrum Estimation, Time Series, Crossings, Periodicities, Solar modes

1 Introduction

A major goal of spectrum estimation is to identify deterministic periodicities in time series, and implicit in this task involves assigning significance to a peak or peaks in a spectrum estimate under the null hypothesis that no harmonic component is actually present. When the frequencies of possible line components are known, this is not difficult as spectrum estimates are χ^2 distributed. Unfortunately, given very long datasets (perhaps in excess of $100,000$ samples), significant spurious detections can appear as many very narrow peaks or fewer wide peaks, and their simultaneous assessment becomes a very large multiple-comparisons problem. The assessment of significance of peaks in spectrum estimates is an old problem first considered by Schuster [23] and later by Fisher [8]. We give in this paper the average rate at which false detections occur, as well the width and shape of these excursions, as a new solution to this problem.

Beginning with the investigations of Rice [21] into upcrossing rates of Gaussian processes, expressions for the upcrossing rates of chi-square processes were derived by various authors [25, 2]. These upcrossing formulae are dependent upon the variance of the derivative of the process, or equivalently the serial correlations of the process. These crossing expressions are applied here to spectrum estimates. Fortunately, spectrum estimates have derivatives of all orders, so the main concern in the crossing literature, i.e. differentiability of the process, is not an issue.

The theoretical results in this paper were motivated by the discovery [34], of sharp peaks in power spectra of magnetic fields and fluxes of charged particles in interplanetary space. These frequencies corresponded approximately with those expected for the normal modes of the Sun. Agreement with measured frequencies of p-modes is excellent, and notably, over ten million solar p- and f- modes are expected to exist, [11]. When these findings were announced, they were disputed for various reasons. Most notably it was argued

that turbulence in the solar wind would destroy coherent modal structures in the interplanetary medium [22]. However, coexistence of discrete modes with interplanetary turbulence has been more recently shown in magnetohydrodynamic simulations using all three spatial dimensions [9].

These data present a challenge to conventional spectrum analysis methods. Under the null hypothesis, no modal structure is present in interplanetary space data, all perceived spectral lines are spurious, and the spectrum follows a "power law", characteristic of turbulent processes. Alternatively, there is a myriad of peaks throughout the spectrum, and this structure coexists with the power law background. In this study, we give evidence that data collected by neutron monitors on Earth show the presence of modal signatures in the cosmic ray background. This discovery may have profound implications, as modal structures may then manifest themselves in many natural processes on Earth, [33]. The present paper is part of a continuing effort to improve understanding of the statistics of these spectra.

1.1 Organization

Section 2 gives a brief introduction to spectrum estimation, especially the multitaper spectrum estimator. Notation is established and the distribution of these estimators is given. Section 3 gives an introduction to crossing literature in the time domain, and these results are applied in the frequency domain to obtain upcrossing rates for single taper and multitaper estimators. Section 4 gives some results concerning the width and shape of spurious and genuine detections. In §5, a real data example is given and a permutation test is proposed. Conclusions and some extensions are discussed in §6.

2 Spectrum Estimation

The power spectrum of a continuous time stationary process contains information about the temporal correlations in the data as a function of frequency. Delta functions in the power spectrum correspond to strong sinusoidal components in the data, and spectrum estimates are often computed for the purpose of determining where sinusoidal components contribute significantly to the total signal power. We give here a concise summary of nonparametric spectrum estimation.

2.1 Direct Estimators

2.1.1 Definition

Techniques for nonparametric estimation of power spectra began with the introduction of the periodogram by Schuster [23]. The periodogram is part of a larger class of *direct* spectrum estimators, which are computed as follows. Given equally-spaced samples of a stationary, zero-mean time series x_t, $t = 0,\ldots,N-1$, the direct spectrum is

$$\hat{S}^D(f) = \left| \sum_{t=0}^{N-1} x_t D_t e^{-i2\pi ft} \right|^2 \tag{1}$$

where D_t is a data window, often chosen to be one of those compared in [10]. The periodogram estimator results when $D_t = \sqrt{1/N}$ is substituted above. The nonuniform data window was introduced to reduce bias inherent in the periodogram. When a data window is used, $\mathrm{E}\{\hat{S}^D(f)\} = |D(f)|^2 * S(f)$, where $D(f)$ is the Fourier transform of the data window D_t, $S(f)$ is the true power spectrum, and $*$ denotes convolution.

Direct estimators are not only biased, but *inconsistent*, i.e. the variance of the estimator does not decrease as the number of samples is increased. The variance of this estimator can be improved using Welch's method, that is, by dividing the data into overlapped sections, computing a direct spectrum on each section and averaging the estimates. For more information on the Welch technique, see [36]. It can be shown that

the Welch estimator is consistent when the block length is fixed. The Welch method is often applied to multitaper spectra, described in §2.2.

2.1.2 Distribution

It can be shown that windowed Fourier transformed data is approximately complex-Gaussian distributed [15] and is certainly asymptotically so using a weak central limit argument. As such, the direct spectrum at a given frequency is χ^2_2 distributed, where uncorrelated real and imaginary parts each contribute a degree of freedom to the estimate, except at zero and the Nyquist frequencies, where the spectrum is real.

2.2 Multitaper Estimators

The multitaper estimator, [29], was developed as an approximate solution to the integral equation

$$y(f) = \sum_{t=0}^{N-1} x_t e^{-i2\pi(t-\frac{N-1}{2})} = \int_{-1/2}^{1/2} \frac{\sin N\pi(f-\nu)}{\sin \pi(f-\nu)} dZ(\nu) \tag{2}$$

obtained when the Fourier transform of the time series, $y(f)$, is substituted into the Cramér spectral representation theorem [6]. The spectrum of the process is then defined as $S(f) = \mathrm{E}\{|dZ(f)|^2\}$, where $dZ(f)$ is a stationary, orthogonal increments process. The method makes use of the discrete prolate spheroidal sequences, $\{v_n^{(k)}(N,W)\}_{k,n=0}^{N-1}$ or *Slepian* sequences which are a set of finite-length, orthogonal sequences $\mathscr{L}^2$-optimally concentrated on the band $(-W,W)$ in frequency [26]. They satisfy the eigenvalue equation

$$\sum_{m=0}^{N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)} v_m^{(k)}(N,W) = \lambda_k(N,W) \cdot v_n^{(k)}(N,W) \tag{3}$$

where the eigenvalue $\lambda_k(N,W)$ also denotes the fraction of energy in the band $(-W,W)$, and the index k sorts the sequences in descending order of energy concentration, $1 > \lambda_0 > \lambda_1 > \ldots > \lambda_{N-1} > 0$.[1] The first $K \approx 2NW$ of these are close to one, while the others rapidly drop to zero. Slepian sequences are best computed as the eigenvectors of a tridiagonal matrix, see [30].

2.2.1 Definition

The multitaper spectrum is constructed as the local least-squares solution to (2) in the band $(f-W, f+W)$, and as such, is a weighted average of magnitude squared Fourier transformed tapered data sequences. Denoting the *eigencoefficients* as

$$y^{(k)}(f) = \sum_{t=0}^{N-1} x_t v_t^{(k)} e^{-i2\pi ft}, \tag{4}$$

and the *eigenspectra* as $\hat{S}^{(k)}(f) = |y^{(k)}(f)|^2$, one forms the following weighted average

$$\hat{S}(f) = \sum_{k=0}^{K-1} d_k^2(f) \hat{S}^{(k)}(f) \tag{5}$$

where the optimum frequency-dependent weights, $d_k(f)$, are determined using an adaptive scheme, see [29] §V. For Gaussian white noise processes with $K = 2NW$ it can be shown that the expected weight is $\lambda_k/2NW$.

The multitaper usage of approximately bandlimited windows limits leakage from out-of-band frequencies, reducing bias, while the orthogonality of the windows guarantees a consistent estimate when W is

[1] The explicit dependence on (N,W) will be dropped for notational convenience.

fixed. The bias of the kth eigenspectrum, using the Cauchy-Schwarz inequality alone, can be bounded by $\sigma^2(1-\lambda_k)$, where σ^2 is the average of the spectrum across frequency. Additionally, W can be decreased to produce a less biased, better resolved estimate at the expense of variance, or increased to benefit the variance at the expense of bias, thus providing a compromise to the bias/variance tradeoff, see [30, 17]. Bronez [5] showed that between direct Welch and multitaper estimators with two of bandwidth, bias, and variance fixed, the third was always smallest in the case of the multitaper estimator. Stoica and Sundin [27] have shown that multitaper spectra are approximately maximum-likelihood spectral estimators.

2.2.2 Distribution

Multitaper estimators have $\chi^2_{2\alpha}$ distributions, where the effective degrees of freedom, 2α, can be approximated in terms of the weights

$$\alpha \approx \sum_{k=0}^{K-1} d_k^2(f) \approx K. \tag{6}$$

A particularly convenient *standardized* form of the spectrum

$$z(f) = \frac{\hat{S}(f)}{\mathrm{E}\{\hat{S}(f)\}} \tag{7}$$

is distributed as a scaled central $\chi^2_{2\alpha}$ random variable, and takes a $\Gamma(\alpha, 1/\alpha)$ probability distribution. This distribution has mean 1 and variance $1/\alpha$. The standardized spectrum will be used to derive all of the theoretical results of this paper.

Note that when the data contains a line component at frequency f_0, the distribution of the spectrum at f_0 changes from the null assumption of a central $\chi^2_{2\alpha}$, to a noncentral $\chi^2_{2\alpha}$ distribution, [13, 29]. In this case $\mathrm{E}\{S(f_0)\} = \lambda + 2\alpha$, where $\lambda > 0$ is the noncentrality parameter. This observation means, in practice, that spectra containing numerous line components may be inappropriately scaled if the spectrum is standardized using a local mean value. The 5% points of the central and noncentral distributions, however, agree approximately for reasonable values of λ, so standardization is better done by scaling the empirical 5% point of the spectrum so it matches the theoretical. Needless to say, careful prewhitening of the original spectrum must be done to ensure sufficient "flatness" in the residual spectrum has been achieved, see [32] Appendix D.

2.3 A Note on Rayleigh Resolution

The *Rayleigh resolution*, $\mathscr{R}$, of a time series is the original lower bound on the spacing of peaks in the spectrum which can be resolved given the length of time over which the data sequence is observed. Hence, $\mathscr{R} = 1/T$Hz where T is the length of the series (in seconds) or $\mathscr{R} = 1/N\delta t$Hz when the series is sampled at intervals of length δt seconds (we assume unit sampling throughout this paper). Note that the Rayleigh resolution does *not* depend on the sampling rate of the series, hence $\mathscr{R}$ does not change with zero-padding or downsampling. In this paper, frequency is standardized in units of the Rayleigh resolution. Standardizing both the spectrum and the units of frequency allows us to construct standard reference tables of false detection rates, looking ahead to table 2.

3 Crossing Problems

Rice [21] derived the expected number of zeros per second of a zero-mean, stationary, continuous-time Gaussian process as

$$\bar{N}_z = \frac{1}{\pi}\left[-\frac{\gamma''(0)}{\gamma(0)}\right]^{1/2} = 2\left[\frac{\int f^2 S(f)df}{\int S(f)df}\right]^{1/2} \tag{8}$$

where $\gamma(\tau)$ is the autocovariance function of the process at lag τ and γ'' is the curvature of the autocovariance with respect to τ. For history of Rice's formula [20] and an overview of the literature that has expanded from this result, consult [4, 1].

Miller [16] and Silverman [25] independently extended this result to χ^2 processes, see also [12, 24, 2, 35, 14]. Using the standardized spectrum (7), define

$$\Upsilon(\Delta) = \mathrm{Cov}\{z(f+\Delta), z(f)\}, \tag{9}$$

as the *antecovariance* of the spectrum estimate, where the antevariance is $\Upsilon(0) = 1/\alpha$. Denoting $U(z;\alpha)$ as the *upcrossing rate* of the level z of a $\Gamma(\alpha, 1/\alpha)$ process per Rayleigh resolution, we have

$$U(z;\alpha) = \psi\sqrt{\frac{z}{2\pi\alpha}}p(z;\alpha) \tag{10}$$

where $p(z;\alpha)$ is the $\Gamma(\alpha, 1/\alpha)$ probability distribution function and

$$\psi = \frac{1}{N}\left[-\frac{\frac{d^2}{d\Delta^2}\Upsilon(\Delta)|_0}{\Upsilon(0)}\right]^{1/2}. \tag{11}$$

The factor of $\Upsilon(0)$ in the denominator converts from covariance to correlation, and the $1/N$ converts to frequency in units of Rayleigh resolution. The term $-\frac{d^2}{d\Delta^2}\Upsilon(\Delta)|_0$ is the variance of the derivative $z' = \frac{d}{df}z(f)$.

The purpose of this paper is to apply crossing theory to spectra to obtain expressions for false detection rates above given significance levels. Under the null assumption that the data is white, Gaussian, the upcrossing rate $U(z_{95\%};\alpha)$, where $0.95 = \int_0^{z_{95\%}} p(z;\alpha)dz$ gives the rate of false detection of periodic components in the spectrum which have over 95% significance. This false detection rate depends on the serial correlations $\Upsilon(\Delta)$ of the spectrum estimates, and we derive these for both classes of estimators discussed in §2. Application of the above theory is applied with some care. It is usually assumed that the individual Gaussian processes making up the χ^2 process are independent with identical autocovariances, but when applied to multitaper eigencoefficients neither of these assumptions is true.

3.1 Direct Spectrum Upcrossing Rate

Beginning with equations (10) & (11), the *antecovariance*, or covariance of the spectrum estimate, $\Upsilon(\Delta)$, of the direct spectrum (1) is, [28] or [31] §4.4,

$$\Upsilon_D(\Delta) = |(D * D^*)(\Delta)|^2 = \left|\sum_{t=0}^{N-1} D_t^2 e^{-i2\pi t\Delta}\right|^2 \tag{12}$$

where $*$ used as a superscript denotes complex conjugate. The second equality follows when the process is white, zero-mean, and complex valued (as it is under the null hypothesis here). To obtain the upcrossing rate, substitute $\alpha = 1$ so the standardized spectrum has an exponential distribution, $p(z;\alpha) = e^{-z}$. Then (11) becomes, upon application of Isserlis's formula,

$$\psi_D = \frac{2\pi}{N}\left[\sum_{s,t=0}^{N-1}(s-t)^2 D_t^2 D_s^2\right]^{1/2} \tag{13}$$

and the upcrossing rate (10) is

$$U_D(z;1) = \frac{e^{-z}}{N}\left[2\pi z\sum_{s,t=0}^{N-1}(s-t)^2 D_t^2 D_s^2\right]^{1/2}. \tag{14}$$

For example, straightforward substitution of the periodogram estimate into the antecorrelation formula (12) results in

$$\Upsilon_P(\Delta) = \left| \frac{\sin N\pi\Delta}{N \sin \pi\Delta} \right|^2 \tag{15}$$

And the mean number of upcrossings of a level z for a periodogram per Rayleigh resolution becomes simply

$$U_P(z) = \left[\frac{\pi z}{3}\right]^{1/2} \frac{(N^2-1)^{1/2}}{N} e^{-z} \approx \left[\frac{\pi z}{3}\right] e^{-z}. \tag{16}$$

Note that for complex Gaussian white data, $\mathrm{E}\{z(f)\} = 1$ and all peaks are spurious, so the upcrossing rate should be as small as possible for high levels z. The upcrossing rate, among direct estimates and for N fixed, only changes in the constant ψ_D, which depends only on N and the data window D_t. It is possible to have $\psi_D = 0$ if one chooses, for example $D_t = 1$ when $t = 0$ and 0 everywhere else, but this choice gives a terrible spectrum estimate. In fact, for any "good" choice of D_t, see [10], one finds that the reduction in $U_D(z;1)$ as compared to $U_P(z;1)$ is marginal.

3.2 Multitaper Antecorrelation

The antecovariance, $\Upsilon(\Delta)$, for a multitaper estimate is evaluated in [31]. Beginning with a white, Gaussian, complex-valued series, as before, the covariance between the eigencoefficients separated in frequency by Δ, [29], is

$$\mathrm{Cov}\{y_j(f+\Delta), y_k^*(f)\} = \sum_{t=0}^{N-1} v_t^{(j)} v_t^{(k)} e^{i2\pi t\Delta} \tag{17}$$

where * denotes complex conjugate. When $\Delta = 0$, this expression collapses to $\delta_{j,k}$, because the Slepian sequences are orthonormal. Now, substituting the expected weights $d_k^2(f) = \lambda_k$, the antecovariance becomes

$$\Upsilon(\Delta) = \mathrm{Cov}\{z(f), z^*(f+\Delta)\} = \frac{1}{(2NW)^2} \sum_{n,m=0}^{N-1} \left[\sum_{k=0}^{K-1} \lambda_k v_n^{(k)} v_m^{(k)} \right]^2 e^{-i2\pi\Delta(n-m)} \tag{18}$$

When $K \lessapprox 2NW$, substituting (3) into the bracketed term and using orthogonality of the Slepian sequences, gives the approximation

$$\Upsilon(\Delta) \asymp \frac{1}{(2NW)^2} \sum_{n,m=0}^{N-1} \left[\frac{\sin 2\pi W(n-m)}{\pi(n-m)} \right]^2 e^{-i2\pi\Delta(n-m)} \tag{19}$$

$$= \frac{1}{(2NW)^2} \sum_{\tau=-(N-1)}^{N-1} (N-|\tau|) \left[\frac{\sin 2\pi W\tau}{\pi\tau} \right]^2 e^{-i2\pi\Delta\tau} \tag{20}$$

where the second equality uses the change of variables $\tau = n - m$. Remarks: (i) This expression is approximate because the bracketed term in (18) is a truncated spectral decomposition of (3). The approximation is accurate because eigenvalues become very small for $k > 2NW$. (ii) $\Upsilon(\Delta)$ drops rapidly to almost zero at $\pm 2W$, [31] Fig. 3. (iii) Real valued processes have half the degrees of freedom within $\pm 2W$ of the origin and the Nyquist frequency than they have at other frequencies. (iv) The standardizing factor $2NW$ is the trace of the sinc matrix in (3) and is the sum of the squared weights, λ_k here. (v) Eqn. (20) is the Fourier transform of a product of a Fejér kernel and a triangular window ($(N-|\tau|)$ for $|\tau| \le N$ and zero otherwise).

	K						
NW	$2NW$	$2NW-1$	$2NW-2$	$2NW-3$	$2NW-4$	$2NW-5$	$2NW-6$
3.5	0.7740	0.9514	0.9717	0.9612	0.9540	0.9502	0.9484
4	0.7230	0.8895	0.9118	0.9022	0.8947	0.8904	0.8881
4.5	0.6811	0.8383	0.8620	0.8533	0.8458	0.8412	0.8385
5	0.6458	0.7951	0.8199	0.8120	0.8045	0.7997	0.7967
5.5	0.6155	0.7580	0.7835	0.7764	0.7690	0.7641	0.7609
6	0.5892	0.7257	0.7518	0.7453	0.7381	0.7331	0.7298
7	0.5455	0.6718	0.6986	0.6934	0.6865	0.6815	0.6781
8	0.5103	0.6285	0.6557	0.6514	0.6449	0.6400	0.6364
10	0.4568	0.5622	0.5896	0.5870	0.5811	0.5764	0.5729

Table 1: Table of ψ, Eqn. (11), computed exactly for various choices of K and NW. For example, the choice of $NW = 5$, $K = 8 = 2NW - 2$ gives $\psi = 0.8199$.

Because the crossing rate (10) depends on the curvature of the antecorrelation, or equivalently, the variance of the derivative of the standardized spectrum, we take the second derivative of (20) at the origin,

$$\sigma_d^2 = \frac{1}{N^2}\,\mathbf{Var}\left\{\frac{d}{df}z(f)\right\} = -\frac{1}{N^2}\,\frac{d^2}{d\Delta^2}\Upsilon(\Delta)\Big|_0 \tag{21}$$

$$= \frac{1}{(N^2W)^2}\sum_{\tau=-N}^{N}(N-|\tau|)\sin^2 2\pi W\tau \tag{22}$$

$$= \frac{1}{2(NW)^2}\left\{1-\left[\frac{\sin 2\pi NW}{N\sin 2\pi W}\right]^2\right\} \tag{23}$$

$$\approx \frac{1}{2(NW)^2} \approx \frac{2}{\alpha^2} \tag{24}$$

where the scale factor $1/N$ is added to keep frequency in units of Rayleigh resolution. The Fejér kernel is used to simplify (21). The second term in (23) is zero for nonzero positive integer and half-integer values of NW, and is otherwise approximately $[\sin(2\pi NW)/(2\pi NW)]^2$. The second approximation in (24) holds exactly when $\alpha = 2NW$.

3.3 Multitaper Spectrum Upcrossing Rate

Beginning with (10) and (11), direct substitution of the expressions (20) and (24), and letting $K = 2NW$ gives $\psi \approx (2/\alpha)^{1/2}$, so that the multitaper upcrossing rate is

$$U(z;\alpha) \approx \sqrt{\frac{z}{\pi}}\frac{1}{\alpha}p(z;\alpha) \tag{25}$$

in Rayleighs assuming a white complex-vauled Gaussian process. This formula is deceptively simple as it involves the approximation (20). Exact evaluation of (11) can be computationally inconvenient when N is large. Table 1 gives values of ψ for various choices of NW, where $N = 10^5$ was used in the computation. Table 2 shows $U(z;\alpha)$ for common choices of NW, K.

It is common to average multitaper spectrum estimates from overlapped data blocks, as in the Welch method. Assuming J non-overlapping segments, each with K degrees of freedom, the crossing rate becomes

$$U_J(z;\alpha) \approx \psi\sqrt{\frac{z}{2\pi\alpha J}}p(z;\alpha J) \asymp \frac{1}{\alpha}\sqrt{\frac{z}{\pi J}}p(z;\alpha J) \tag{26}$$

$NW = 3.5$	$K = 4$		$K = 5$		$K = 6$		$K = 7$	
P %	z	$U(z;4)$	z	$U(z;5)$	z	$U(z;6)$	z	$U(z;7)$
50.0000	0.918	154,172	0.934	155,596	0.945	152,176	0.953	123,703
90.0000	1.670	61,808	1.599	62,952	1.546	61,986	1.505	50,655
99.0000	2.511	8,910	2.321	9,105	2.185	8,989	2.082	7,362
99.9000	3.266	1,093	2.959	1,118	2.742	1,105	2.580	906
99.9900	3.978	126	3.556	129	3.261	128	3.041	105
99.9990	4.666	14	4.130	14	3.756	14	3.480	12
99.9999	5.338	2	4.686	2	4.235	2	3.903	1
$NW = 4$	$K = 5$		$K = 6$		$K = 7$		$K = 8$	
P %	z	$U(z;5)$	z	$U(z;6)$	z	$U(z;7)$	z	$U(z;8)$
50.0000	0.934	144,458	0.945	145,837	0.953	142,168	0.959	115,494
90.0000	1.599	58,446	1.546	59,404	1.505	58,216	1.471	47,496
99.0000	2.321	8,453	2.185	8,614	2.082	8,460	2.000	6,915
99.9000	2.959	1,038	2.742	1,059	2.580	1,041	2.453	852
99.9900	3.556	120	3.261	122	3.041	120	2.870	98
99.9990	4.130	13	3.756	14	3.480	13	3.265	11
99.9999	4.686	1	4.235	1	3.903	1	3.645	1
$NW = 5$	$K = 7$		$K = 8$		$K = 9$		$K = 10$	
P %	z	$U(z;7)$	z	$U(z;8)$	z	$U(z;9)$	z	$U(z;10)$
50.0000	0.953	129,774	0.959	130,965	0.963	126,962	0.967	103,075
90.0000	1.505	53,141	1.471	53,859	1.444	52,399	1.421	42,669
99.0000	2.082	7,723	2.000	7,842	1.934	7,642	1.878	6,232
99.9000	2.580	951	2.453	966	2.351	942	2.266	769
99.9900	3.041	110	2.870	112	2.733	109	2.619	89
99.9990	3.480	12	3.265	12	3.093	12	2.952	10
99.9999	3.903	1	3.645	1	3.440	1	3.271	1
$NW = 6$	$K = 9$		$K = 10$		$K = 11$		$K = 12$	
P %	z	$U(z;9)$	z	$U(z;10)$	z	$U(z;11)$	z	$U(z;12)$
50.0000	0.963	119,005	0.967	119,995	0.970	115,801	0.972	93,999
90.0000	1.444	49,115	1.421	49,674	1.401	48,063	1.383	39,104
99.0000	1.934	7,163	1.878	7,255	1.831	7,028	1.791	5,725
99.9000	2.351	883	2.266	895	2.194	867	2.132	707
99.9900	2.733	102	2.619	103	2.524	100	2.442	82
99.9990	3.093	11	2.952	12	2.834	11	2.733	9
99.9999	3.440	1	3.271	1	3.130	1	3.010	1

Table 2: For quick reference, this table gives expected upcrossing rates, $U(z;K)$, per million Rayleigh resolutions for typical choices of NW and K. P denotes the probability of reaching the level z and $U(z;\alpha)$ is given in Eqn. (10). For example, the expected number of upcrossings in 10^6 Rayleighs of the 99% level in a multitaper spectrum of Gaussian white noise with $NW = 5, K = 8$ is 7842. If the number of peaks in a real spectrum greatly exceeds this, it is likely due to the process's correlation structure. To get the significance level for a χ^2_{2K} estimate, as opposed to a $\Gamma(K, 1/K)$ estimate, multiply z by $2K$.

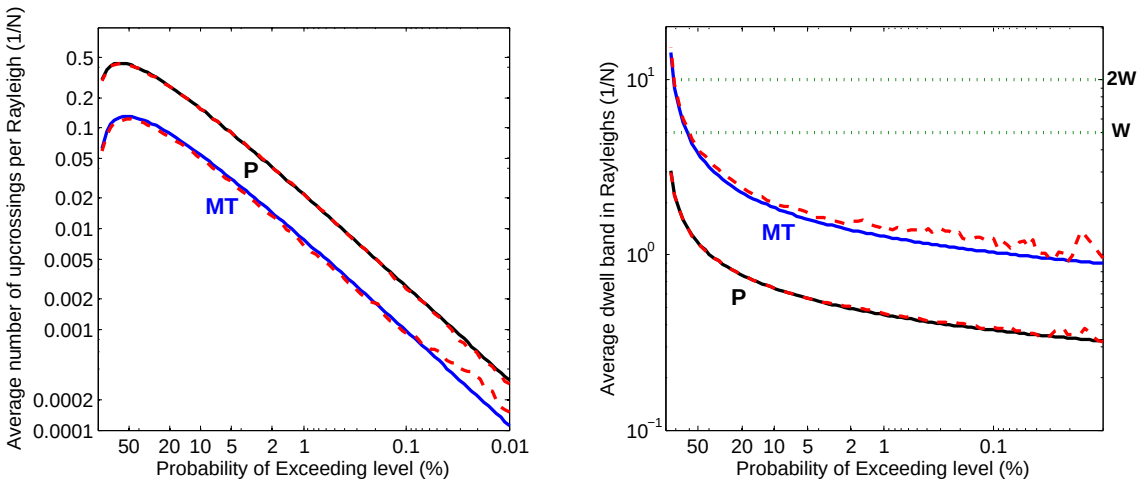

Figure 1: The left panel shows the multitaper $U(z_p;K)$ (blue, marked MT) versus the periodogram $U_P(z_p;1)$ (black, marked P) estimates' crossing rates per Rayleigh resolution with white complex-valued noise as input. Red dashed lines give upcrossing rates for simulated data ($N = 100,000$) for the respective estimates. The y-axis gives crossing rates for the level z_p which is obtained by inverting the cumulative probability distribution at the level $100(1-p)\%$, p is the p-value for the test for a harmonic component at a given frequency and is plotted on the x-axis. For the multitaper, parameters $NW = 5, K = 8$ were chosen. The right panel shows the average dwell bandwidth of crossings for these estimators.

where the last expression uses the asymptotic value of ψ.

3.4 Comparison

Combining the results of this section, the left panel of Fig. 1 shows the periodogram estimate's upcrossing rates against that of the multitaper, along with the result of a white Gaussian noise simulation where $N = 10^5$. Multitaper parameters were chosen as $NW = 5, K = 8$. Note that the multitaper curve is everywhere below that of the periodogram, indicating that at every level of significance, the multitaper estimate will give, given Gaussian white noise (where only spurious peaks are present), fewer detects. The reason for this lies in the expression (10), where the dominant term for the shape of $U(z;\alpha)$ is in the probability distribution $p(z;\alpha)$. The level z is implicit in this figure as the inverse complimentary (1-cdf) cumulative distribution of the values on the $x-$axis.

4 Shape and Width of Spurious Peaks

The upcrossing rate formulae have a particularly simple relationship with the width of these peaks, [18] equation (3.41). Define the *dwell bandwidth*, $w(z)$ as the average frequency difference from the point where the estimate crosses z with a positive slope and the next frequency where the estimate returns to a level below z. For small z, this dwell distance can correspond to more than one peak, but for large z the dwell band usually contains a single maximum, so *peak-bandwidth* is almost synonymous. The average number of such excursions times their mean bandwidth gives the total frequency range above z, the complimentary cdf (CCDF), $Q(z;\alpha)$, so $Q(z;\alpha) = U(z;\alpha)w(z)$.

For example, using (16) and noting that the CCDF is just e^{-z}, the average dwell bandwidth for spurious

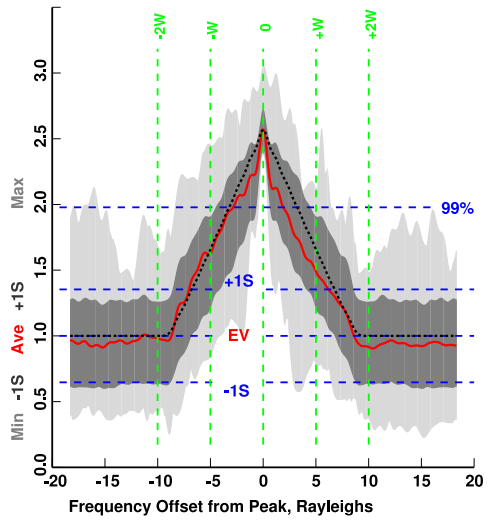

Figure 2: The red curve gives the average shape of the 50 largest (spurious) peaks in a multitaper white-noise spectrum. The light grey band shows the range, while the dark grey band shows one standard deviation away from the mean. The dotted black line shows the shape of the theoretical antecorrelation. Here, $N = 100,000, NW = 5$, and $K = 8$ was used. The dashed vertical lines give the $\pm W$ and $\pm 2W$ bands, and the horizontal dashed lines give the mean $z = 1 = \pm 1\sigma = \pm\sqrt{1/K}$ and the 99% significance level. At 99% significance, the average peak width is about W Rayleighs wide.

peaks in a periodogram is

$$w_P(z) \approx \sqrt{\frac{3}{\pi}} z^{-1/2} \tag{27}$$

Rayleighs. Similarly, combining (10) with the gamma CCDF gives the average dwell bandwidth for spurious peaks in the multitaper estimate, in Rayleighs,

$$w_{mt}(z) \approx \alpha \sqrt{\frac{\pi}{z}} \frac{Q(z;\alpha)}{p(z;\alpha)}. \tag{28}$$

The right panel of Fig. 1 gives a comparison of the average dwell bands for the two estimates.

Related to the dwell bandwidth problem is the question of the *shape* of random peaks in spectrum estimates. In the time-domain crossing literaure, [7], continuous-time Gaussian processes are assumed to be well-approximated by a second-order Taylor series approximation during an excursion. Fig. 2 shows the average of the 50 largest peaks in a multitaper spectrum estimate, and while a quadratic approximation may be convenient, it is apparently not appropriate for spectrum estimation. Blachman [3] showed that for Gaussian processes, large excursions take the shape of the autocovariance function. Adapting this result to spectra is not strictly valid as spectra have χ^2 distributions, but is asymptotically valid if one increases the degrees of freedom with N, and for spectra with > 20 df, the result applies reasonably well. Reconsidering Eqn (20), the antecorrelation $\Upsilon(\Delta)$ is dominated by a triangular window of width $2W$. Additionally, the antecorrelation is an entire function of frequency so its curvature is well-defined. We conclude, then, that the shape of random peaks in multitaper spectra is triangular with width $\pm 2W$. On the other hand, if the spectrum genuinely contains a δ-function at a given frequency, the spectrum estimate takes on the shape of the spectral window - which, for a multitaper estimate is roughly rectangular with width $\pm W$ centered about the frequency of interest.

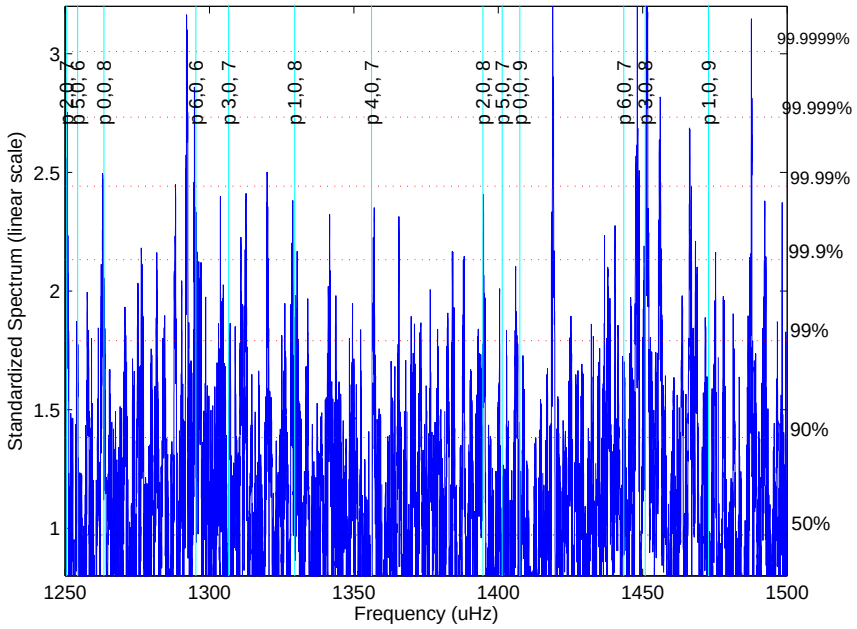

Figure 3: Standardized Newark neutron monitor spectrum between June 30-December 31, 1250-1500 μHz band shown, spectrum is an average of two nonoverlapping segments. Horizontal lines denote significance levels for a χ^2_{24} standardized estimate. Vertical lines denote low-l p-mode frequencies in that band, [19].

5 A Real Data Example

Returning to the problem mentioned in the introduction, namely, the discovery of solar modal structure in space [34, 32] and in the engineering environment on Earth [33], we examine neutron monitor data. These data present an excellent example of a process which appears to contain a pathological number of line components. We question whether the perceived line components could correspond to frequencies for solar p-modes, and by permuting the original data we test whether the lines are due to the temporal order of the data, i.e. the data's autocorrelation structure, or to the data's distribution. Neutron monitor data are given as counts per 10 seconds, so the process is expected to be Poisson distributed, however $\bar{x} = 99.2$ and $\hat{\sigma}^2 = 62.5$, suggesting that the data are underdispersed, so that the events are not uniformly distributed in time but "bunched".

Figure 3 shows a standardized spectrum estimate of neutron monitoring data obtained from the Newark neutron monitor of the Bartol Research Institute. Data were sampled from June 30, 2005 to December 31st, 2005, separated into two non-overlapping segments ($N = 799,197$ samples, 10s apart) and were interpolated, low-pass filtered and decimated to $\delta t = 30$s. The data was of high quality; less than 3% of the data was missing over this period. The Rayleigh resolution of the resultant spectrum was 62.6nHz. Because fine scale frequency resolution is desired, the time bandwidth $NW = 3.5$ was chosen with $K = 6$ tapers, giving a bandwidth of 438nHz, and yielding an estimate with ≈ 24 degrees of freedom.

Superimposed on this spectrum are $l \leq 6$ p-mode frequencies obtained from [19] (vertical lines in Fig. 3), labeled by the triplet (l,m,n), which indexes the modes. We perceive agreement between the 13 frequencies pictured on this band with highly significant peaks in the spectrum. Note that the standardization is for a $\Gamma(12,1/12)$ process, and that significance levels (horizontal lines in the figure) are labeled under the assumption that no modal structure is present. Abundant modal structure, such as would appear under our alternative hypothesis would cause this estimate to have a larger mean than expected.

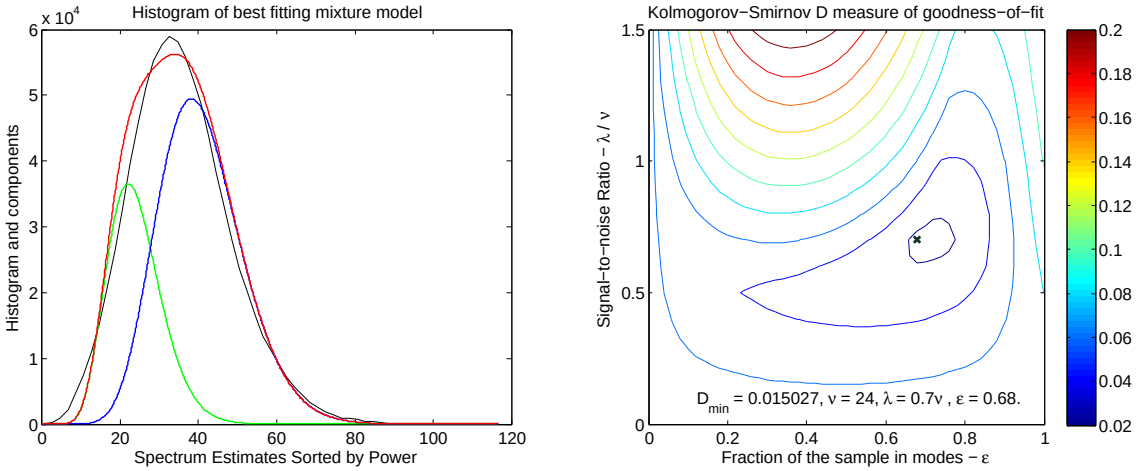

Figure 4: The left panel shows the best-fitting mixture (red) of noncentral (blue) and central (green) χ^2_{24} pdfs to the empirical distribution (in black) of the multitaper spectrum $NW = 3.5, K = 6, J = 2$ estimate on the band $(250-5000)\mu$Hz. This mixture was found by minimizing the Kolmogorov-Smirnov D statistic (contour plot, right panel) over a grid of choices for the mixing fraction ε (shown on the x-axis) and the noncentrality parameter λ (y-axis), expressed in units of the degrees-of-freedom parameter $\nu = 24$. The minimum of the D statistic (marked with an 'x') occurs on this grid at $\lambda = 0.7\nu, \varepsilon = 0.68$, and takes the value $D_{min} = 0.0150$ which indicates that the data are mainly modal on this band. Note that the central chi-square distribution ($\varepsilon = 0$) gives $D = 0.0663$.

Proceeding in the spirit of Thomson *et al.* [32], we decompose the empirical distribution of the spectrum in the p-mode band $(250-5000\mu$Hz) into a mixture of central and non-central χ^2_{24} components. Here we assume that the spectrum consists of a mixture of "noise" and "lines", the former having a central distribution and the latter having a noncentral distribution. The fit was done assuming that the noise power is uniform across frequency and constraining the mean of the two distributions to match the average across frequency. Denote by λ the noncentrality parameter, and ε the fraction of the spectrum in modes, both which are to be estimated. Expressing the noncentrality parameter in standard units gives a signal-to-noise ratio $\rho = \lambda/\nu$ where $\nu = 24$, the degrees of freedom, is the mean of the central component. The probability distribution of the mixture model, then, is

$$P_\varepsilon(s|\nu,\varepsilon,\lambda) = (1-\varepsilon)P_c(s\gamma|\nu) + \varepsilon P_{nc}(s\gamma|\nu,\lambda) \tag{29}$$

where the scale factor γ is chosen so that the expected value of the mixture distribution matches the sample average. Here s denotes a value of the spectrum at some frequency and P_c and P_{nc} are the central and noncentral χ^2_ν distributions, respectively. Creating a grid of choices for λ, ε, the Kolmogorov-Smirnov D statistic for goodness-of-fit between the empirical and mixture distributions was made at each point, see Fig. 4. The minimum D-statistic was $D_{min} = 0.015$ with $\lambda = 0.7\nu = 16.8$, and $\varepsilon = 0.68$. The right panel of Fig. 4 shows the best fitting pdf for the mixture model superimposed upon a histogram of the empirical distribution. Note that the central χ^2_{24} distribution gave $D = 0.0663$. This analysis suggests that the data are overwhelmingly modal on the $250-5000\mu$Hz band, where about 68% of the frequency bins indicate nonzero noncentrality.

Increasing the bandwidth and number of tapers to $NW = 5, K = 8$, to give $\nu = 2\alpha \approx 16$ for a single segment, we consider whether or not the peaks in the spectrum could be due to the unexplained underdispersion of the original Poisson process. Note that the data are otherwise somewhat non-Gaussian with skewness

P %	z	$U(z;8)$	$\sqrt{U(z;8)}$	Gauss	RP	Obs
50.0000	0.959	13,096	114	12,840	12,041	12,368
90.000	1.471	5,386	73	5,297	5,018	6,256
99.000	2.000	784	28	781	566	1,665
99.900	2.453	97	10	101	71	392
99.990	2.870	11	3	10	3	90
99.999	3.265	1	1	1	0	45

Table 3: Expected, $U(z;16)$, and observed counts of upcrossings per $100,000$ Rayleigh resolutions of the Newark neutron monitor multitaper $NW = 5, K = 8$ spectrum over the band $250 - 5000\mu$Hz. The column headed $\sqrt{U}$ is the standard deviation expected with Poisson statistics. The three right-hand columns give observed counts for the observed data (Obs), the same data after random permutation (RP) and white Gaussian (Gauss) data, for reference.

0.5159 and kurtosis 3.835. Using the data sequence of length $N = 799,197$, between September 30th and December 31st, 2005, sampled at $\delta t = 10$ seconds, the Rayleigh resolution is 125nHz. The p-mode band is about 4750μHz wide, ($\sim 250 - 5000\mu$Hz) which gives a band of width $37,962$ Rayleighs. Downsampling, computing the spectrum, and prewhitening as before, we give in Table 3 the result of a permutation experiment on these data. Upcrossings are given per 10^5 Rayleigh resolutions.

Because large peaks are rare events, counting "rare" upcrossings over an interval corresponds to a Poisson process, thus the column marked $\sqrt{U(z;\alpha)}$ is the expected standard deviation of the number of upcrossings of the level z, and gives rough error bounds associated with these counts. The table shows that there is an overwhelmingly large number of peaks in the observed spectrum, and that this number of peaks decreases to approximately that expected from a Gaussian process when the data is randomly permuted. For high significance levels, one finds many more upcrossings than expected with white Gaussian data. For reference, the column marked Gauss gives the number of upcrossings of a simulation of a white Gaussian process. We conclude that the pathological number of peaks observed in this spectrum is a result of the temporal order of the data, i.e. the process's autocorrelation structure and is *not* a result of any departure from Gaussianity.

6 Conclusions

We have derived expressions for the upcrossing rate of the spectrum of a Gaussian white noise process above a given significance level. This expression involves the autocorrelation of the spectrum, or antecorrelation, as well as the estimator's probability distribution. We showed that among nonparametric spectrum estimates, fewer spurious detects are always given by the multitaper estimate. From the upcrossing expression we also can derive the expected width of excursions above a given significance level. We found that the shape of the spectrum near a spurious peak corresponds to the shape of the antecorrelation there. Because the multitaper antecorrelation is roughly triangular in shape with width $\pm 2W$, spurious peaks take on that shape, while genuine peaks take the shape of the spectral window, which is rectangular with width $\pm W$.

Galactic cosmic radiation is known to undergo solar modulation in interplanetary space, and thus carries the signature of the 11 year solar cycle. In this study we consider solar modal structure as a possible driver for the more minor variations in cosmic ray intensity on Earth. Because the interplanetary medium is mainly turbulent, some have suggested that the perceived modal structure in these data could be due to its non-Gaussian distribution. We used a permutation test to justify the appearance of these modes. After random permutation, the large number of upcrossings of the spectrum in the high significance levels reduces

to the number expected for a Gaussian process. This evidence, combined with the fact that the spectrum appears to be about 68% modal, supports the argument that these peaks are genuine. We conclude that these data contain peaks which are indeed due to the correlation structure of the data and not its distribution. The agreement in Fig. 3 of these modal structures with known modal frequencies in the p-mode band supports these claims.

In conclusion, the present approach is part of an effort to better characterize spectra of this sort, and though we give here a single motivating example, the implications of these results are of widespread interest in the space physics and engineering communities.

Acknowledgment

The authors would like to thank Dr Roger Pyle and colleagues from the University of Delaware and the Bartol Research Institute for providing the data for this study.

References

[1] J. Abrahams. A survey of recent progress on level–crossing problems for random processes. In I. F. Blake and H. V. Poor, editors, *Communications and Networks*, pages 1–23. Springer–Verlag, New York, 1986.

[2] R. Barakat. Level–crossings of aperture–integrated isotropic speckle. *J. Opt. Soc. Am A*, 5:1244–1247, 1988.

[3] N. M. Blachman. Gaussian noise—Part I: the shape of large excursions. *IEEE Trans. on Information Theory*, 34:1396–1400, 1988.

[4] I. F. Blake and W. C. Lindsey. Level–crossing problems for random processes. *IEEE Trans. on Information Theory*, 19:295–315, 1973.

[5] T. P. Bronez. On the performance advantage of multitaper spectral analysis. *IEEE Trans. on Signal Processing*, 40:2941–2946, 1992.

[6] H. Cramér. On the theory of stationary random processes. *Ann. of Math.*, 41:215–230, 1940.

[7] H. Cramér and M. R. Leadbetter. *Stationary and related stochastic processes: Sample function properties and their applications*. Wiley, New York, 1967.

[8] R. A. Fisher. Tests of significance in harmonic analysis. *Proc. Royal Soc. London*, 125 A:54–59, 1929.

[9] S. Ghosh, D. J. Thomson, P. Dmitruk, W. H. Matthaeus, and L. J. Lanzerotti. Co–existence of discrete modes and turbulence in direct MHD simulations. In *AGU 2006 Fall Meeting, Supplement to EOS*. American Geophysical Union, 2006. Abstract, Session SH10.

[10] F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66:51–83, 1978.

[11] J. W. Harvey. Helioseismology. *Physics Today*, 48:32–38, October 1995.

[12] C. W. Helstrom. *Statistical theory of signal detection*. Pergamon, Oxford, second edition, 1968.

[13] H. O. Lancaster. *The Chi-squared Distribution*. John Wiley and Sons, New York, 1969.

[14] G. Lindgren. Slepian models for χ^2–processes with dependent components with application to envelope upcrossings. *J. Appl. Prob.*, 26:36–49, 1989.

[15] C. L. Mallows. Linear processes are nearly Gaussian. *J. Appl. Prob.*, 4:313–329, 1967.

[16] K. S. Miller, R. I. Bernstein, and L. E. Blumenson. Generalized Rayleigh processes. *Quart. Jour. Math.*, 16:137–145, 1958.

[17] D. B. Percival and A. T. Walden. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge Univ. Press, Cambridge, 1993.

[18] A. R. Pratt. Some theoretical considerations concerning time statistics in signal detection. In J. W. R. Griffiths, P. L. Stocklin, and C. van Schooneveld, editors, *Signal Processing*, pages 375–399. Academic Press, London, 1973.

[19] J. Provost, G. Berthomieu, and P. Morel. Low-frequency p– and g–mode solar oscillations. *Astronomy & Astrophysics*, 353:775–785, 2000.

[20] A. J. Rainal. Origin of Rice's formula. *IEEE Trans. on Information Theory*, 34:1383–1387, 1988.

[21] S. O. Rice. Mathematical analysis of random noise. In N. Wax, editor, *Selected papers on noise and stochastic processes (1954)*, pages 133–294. Dover, New York, 1944. Original in Bell System Tech. J. Vol 23 and 24 (1944) Available from http://bstj.bell-labs.com/.

[22] D. A. Roberts, K. W. Ogilvie, and M. L. Goldstein. The nature of the solar wind. *Nature*, 381:31–32, 1996.

[23] A. Schuster. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenonema. *Terrestrial Magnetism*, 3:13–41, 1898.

[24] K. Sharpe. Some properties of the crossings process generated by a stationary χ^2 process. *Adv. Appl. Prob.*, 10:373–391, 1978.

[25] R. A. Silverman. The fluctuation rate of the chi process. *IRE Trans. on Information Theory*, 4:30–34, 1958.

[26] David Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: the discrete case. *Bell System Tech. J.*, 57:1371–1429, 1978.

[27] P. Stoica and T. Sundin. On nonparametric spectral estimation. *Circuits Systems Signal Process.*, 18:169–181, 1999.

[28] D. J. Thomson. Spectrum estimation techniques for characterization and development of WT4 waveguide, *Part I. Bell System Tech. J.*, 56:1769–1815, 1977.

[29] D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70:1055–1096, 1982.

[30] D. J. Thomson. Quadratic-inverse spectrum estimates: applications to paleoclimatology. *Phil. Trans. R. Soc. Lond. A*, 332:539–597, 1990.

[31] D. J. Thomson. Multitaper analysis of nonstationary and nonlinear time series data. In W. Fitzgerald, R. Smith, A. Walden, and P. Young, editors, *Nonlinear and Nonstationary Signal Processing*, pages 317–394. Cambridge Univ. Press, Cambridge, 2001.

[32] D. J. Thomson, L. J. Lanzerotti, and C. G. Maclennan. The interplanetary magnetic field: Statistical properties and discrete modes. *J Geophys. Res.*, 106:15941–15962, 2001.

[33] D. J. Thomson, L. J. Lanzerotti, F. L. Vernon, III, M. R. Lessard, and L. T. P. Smith. Solar modal structure of the engineering environment. *Proceedings of the IEEE*, 95:1085–1132, 2007.

[34] D. J. Thomson, C. G. Maclennan, and L. J. Lanzerotti. Propagation of solar oscillations through the interplanetary medium. *Nature*, 376:139–144, 1995.

[35] D. Veneziano. Envelopes of vector random processes and their crossing rates. *Ann. Prob.*, 7:62–74, 1979.

[36] P. D. Welch. The use of the fast Fourier transform for estimation of spectra: A method based on time averaging over short, modified periodogram. *IEEE Trans. on Audio and Acoustics*, 15:70–74, 1967.