LA-UR--91-477

DE91 008583

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

**TITLE: LIMITATIONS TO THE METHOD OF POWER SPECTRUM ANALYSIS:
NONSTATIONARITY, BIASED ESTIMATORS, AND WEAK
CONVERGENCE TO NORMALITY**

**AUTHOR(S): WILLIAM I. NEWMAN, MARTHA P. HAYNES AND
YERVANT TERZIAN**

**SUBMITTED TO: The Proceedings of "Statistical Challenges in Modern
Astronomy" held August 11-14, 1991 at Pennsylvania
State University**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos **MASTER**
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Limitations to the Method of Power Spectrum Analysis: Nonstationarity, Biased Estimators, and Weak Convergence to Normality

William I. Newman^{1 2}

Martha P. Haynes and Yervant Terzian^{3 4}

ABSTRACT The "Power Spectrum Analysis" method developed by Yu and Peebles (1969) has been widely employed as a technique for establishing the existence of periodicities. This method generates a sequence of random numbers from observational data which, it was claimed, is exponentially distributed with unit mean and variance, essentially *independent* of the distribution of the original data. We show that the derived random process preserves a subtle imprint of the original distribution, rendering the derived process nonstationary and producing a small but systematic bias in the usual estimate of the mean and variance. Although the derived variable may be reasonably described by an exponential distribution, the tail of the distribution is far removed from that of an exponential, thereby rendering statistical inference and confidence testing based on the tail of the distribution completely unreliable.

1.1 Introduction

Many phenomena in the physical sciences lend themselves to the conjecture that there exists some underlying structure or periodicity present in different kinds of data. Astronomy has had a long history of claims of pattern and form which ultimately proved to be false. Sheehan [17], a psychiatrist

¹Departments of Earth and Space Sciences, Astronomy, and Mathematics, University of California, Los Angeles, CA 90024

²Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

³Department of Astronomy, Cornell University, Ithaca, NY 14853

⁴National Astronomy and Ionosphere Center, Cornell University, Ithaca, NY 14853

and psychologist, has chronicled the evolution of planetary astronomy and how observational information of a largely qualitative sort became prone to misinterpretation, a consequence of the way the eye and brain function together. Schiaparelli and Lowell's claimed observation of "canals" on Mars is a quintessential example of this phenomena. Only recently have psychologists developed an appreciation for how the eye seemingly finds pattern where none exists. The pioneering work of researchers such as Julesz [10] have succeeded in *quantifying* how images, consisting of high densities of points (cf. astronomical photographic plates), could suggest to viewers the presence of pattern that high-order correlation statistics methods then showed to be nonexistent. The development of other quantitative measures of pattern and ultimately of hypothesis testing [6] has become a fundamental objective of the observational sciences.

Data samples are particularly prone to misidentification of pattern or of clustering. Recently, we [12] showed that the inappropriate application of statistical tests to *random noise* could provide an illusion of pattern, a phenomena we identified with the statistics of small numbers. A landmark development in astronomical investigations of clustering is the celebrated "Power Spectrum Analysis" (PSA) method of Yu and Peebles [22] which generates a sequence of random numbers from observational data that, it was claimed, is exponentially distributed with unit mean and variance, *essentially independent of the distribution of the original data*. The name the method, however, is somewhat misleading as its intent is to identify clustering phenomena in a set of discrete measurements. (We will see later how this name naturally emerged.) The purpose of the PSA technique is to determine whether observational data is "smoothly" distributed in some sense, or possesses a regular, possibly periodic clustering of samples.

The Power Spectrum Analysis method has been widely employed in the investigation of astronomical data for two decades. In their original paper [22], Yu and Peebles analyzed the distribution in the sky of the "rich," "compact" clusters of galaxies in an attempt to find an independent test for the existence of superclusters in these data. Yu and Peebles employed catalogued sequences of angular positions on the sky to *assess* the reality of superclustering, and concluded that there was a possibly significant component of nonrandomness in the data. Not long thereafter, Burbidge and O'Dell [3] employed the method to explore some controversial questions pertaining to the distribution of redshifts of quasi-stellar objects and related emission-line objects. Tift [19] in a problematic paper employed the PSA method, in addition to some unquantifiable *ad hoc* pattern recognition schemes, claiming that it validated his hypothesis that the Coma Cluster of galaxies displays evidence of some form of "quantization" with respect to redshift. More recently, Guthrie and Napier [9] examined data from the Virgo Cluster seeking evidence without success for redshift quantization. Broadhurst, *et al.* [2] used the method of Power Spectrum Analysis and claimed to find clustering of galaxies near the Galactic poles on extraor-

dinarily large distance scales. The appearance of clustering in some of the above situations seemingly contradicts long-accepted results in relativistic astrophysics and cosmology. If these claims are correct, then a profound change will become necessary in our picture of the large-scale structure of the universe. Before we abandon our contemporary view of physics, it is essential that we ascertain that we have not been led astray by some subtle *mathematical* artifice associated with the method of Power Spectrum Analysis.

In this paper, we will carefully examine the derivation of this widely used method of astronomical data analysis. In so doing, we will show that the derived random process preserves a subtle imprint of the original distribution, rendering the derived process nonstationary and producing a small but *systematic* bias in the usual estimate of the mean and variance. Further, as the method's derivation employs assumptions derived from the central limit theorem [4][5], the *asymptotic rate of convergence* of the random variables to an exponential distribution becomes critically important. Although the derived variable may be reasonably described by an exponential distribution over much of its range, the tail of the distribution is far removed from the behavior expected of a unit mean, unit variance exponential. Consequently, statistical inference and confidence testing based on the tail of the distribution, which was performed in some of the more speculative articles cited above, is completely unreliable. We believe that the Power Spectrum Analysis method incorporates some interesting ideas in producing a statistical test for clustering, but subtle mathematical details renders the method inapplicable to situations presented by available sample sizes. We will conclude by considering some analytic and simulation examples which illustrate these issues and provide a better measure of the utility and limitations of this useful method of statistical analysis.

1.2 Method of Power Spectrum Analysis

Following Yu and Peebles [22] (and temporarily employing their equation numbers), consider N points x_i distributed in the interval 0 to 2π , and let

$$z_n = N^{-1} \sum_{i=1}^N e^{inx_i}, \quad (1)$$

where n is an integer. We may regard the z_n as an ancillary series that has the appearance of a Fourier transform. We note that if the measured datum x_i were clustered, particularly around uniformly spaced points separated by $\Delta x < 2\pi$, then z_n would be large when $n \approx (\Delta x)^{-1}$. The variable n now has the role of a "frequency" and, in that sense, the power spectrum for the distribution of points is $|z_n|^2$, $n = 1, 2, \dots$

Yu and Peebles go on to say that,

"if the points x_i are distributed at random in the interval, the ensemble average of z_n (when $n \neq 0$) is

$$\langle z_n \rangle = N^{-1} \sum \langle e^{inx_i} \rangle = N^{-1} \sum \int_0^{2\pi} \frac{dx_i}{2\pi} e^{inx_i} = 0 \quad (2)$$

Similarly, the ensemble average value of the square of the absolute value of z_n is, for a random distribution,

$$\langle |z_n|^2 \rangle = \frac{1}{N} \sum_i \langle |e^{inx_i}|^2 \rangle + \frac{1}{N} \sum_{i \neq j} \langle e^{in(x_i - x_j)} \rangle = 1 \quad (3)$$

When N is large, and the points are distributed at random, the real and imaginary parts of z_n will be statistically independent Since $|z_n|^2$ is the sum of the squares of two independent variables, each normally distributed, $|z_n|^2$ must have an exponential distribution with the width fixed by equation (1.3). Thus we conclude that, when N is large, and the points x_i are distributed at random in the interval $0 \leq x \leq 2\pi$, the *a priori* probability for finding a value of $|z_n|^2$ greater than x is

$$P(|z_n|^2 > x) = e^{-x} \quad (\text{random}) \quad (5)$$

It is readily seen, as in equation (4), that the coefficients z_n are statistically independent in the sense that the ensemble average $\langle z_n z_m \rangle$ vanishes when $n \neq m$."

Yu and Peebles then go on to explore what might happen if the data manifested clustering and observed, for the case of "exact" clustering, that is each point is in a cluster and there N_c points in a cluster,

$$\langle |z_n|^2 \rangle = N_c \quad (6)$$

We will return later to the question of the precision of their equations (2), (3), and (5), since as we shall show the outcome of this question has a critical effect on the reliability of hypothesis testing.

1.3 A Real Example: The Coma Cluster

In order to illustrate how the method of Power Spectrum Analysis behaves, consider the example of the Coma cluster. During the years since the original claim by Tift (1976) [20] of quantization in the redshift distribution of galaxies in the Coma region, many additional redshifts have been measured for galaxies in the region. A thorough discussion and compilation of the redshift distribution in the Coma cluster has been presented by Kent

and Gunn (1982) [11]. The redshift database in their paper includes all of the redshift information available in the literature at that time as well as some new observations. Subsamples include (a) a complete sample with the magnitude $m < 15.7$ within a radius of 3° of the cluster center and (b) a more broadly distributed sample with $m < 15.0$ with a radius of 6° . It is relevant to note that both Tift and Gregory [21] and Kent and Gunn easily identify the Coma cluster out to a distance of 3° from the center because it is relatively isolated in redshift space but restrict analysis of the dynamics within the core to that distance because of contamination from galaxies in the surrounding supercluster. As an update to this earlier work, we have also compiled a sample of galaxies close to the Coma core, making use of a compilation of published optical redshifts kindly made available to one of us (MPH) by J. Huchra and supplemented by a small number of HI 21 cm line redshifts obtained by Haynes and collaborators. This "core" sample includes 192 galaxies with magnitudes of 15.7 or brighter and measured heliocentric redshifts less than $V < 15000 \text{ km s}^{-1}$ that lie within a projected radius of two degrees from the center of the Coma cluster, taken as R.A. = $12^{\text{h}}57^{\text{m}}23.5^{\text{s}}$ and Dec. = $+28^\circ15'10''$. Thus, our sample can be regarded as the most complete and accurate available to date. This data will be discussed in greater detail elsewhere [13].

In a preliminary effort to develop a sense of the statistical distribution of the data, we plotted a frequency histogram of the data over the entire velocity range, incorporating 7 class intervals ("bins") in accord with Sturges' rule [18]. Sturges (1926) developed an *ad hoc* formula for the class interval in applications to frequency distributions.

"It is based on the principle that the proper distribution into classes is given, for all numbers which are powers of 2, by a series of binomial coefficients. For example, 16 items would be divided normally into 5 classes, with class frequencies 1, 4, 6, 4, 1. Thus if a statistical series had 64 items ..., it should be divided into 6 plus 1 or 7 classes. ... If the formula gives 2, 10 may be chosen, but if the formula indicates 7 or 8, the one actually used should generally be the next lower convenient class interval, 5."

Accordingly, given N items, Sturges' rule would allocate $1 + \log_2 N$ class intervals, generally reduced to a multiple of 5. By employing Sturges' rule, we avoid the problem of introducing artificial fluctuations associated with excessive numbers of class intervals. Many other statisticians, notably including Pearson, Doane, Scott, Freedman, Diaconis among others, have explored refinements of Sturges' rule as well as alternative definitions for the appropriate number of class intervals. The point we wish to make here is that, as a preliminary descriptive statistic of a distribution, frequency histograms should contain only moderate numbers of intervals. In the case of the updated Coma cluster data, we should incorporate from 5-10 frequency

intervals.

In Figure 1, we provide the frequency histogram for the Coma cluster data set containing 192 points of which 9 samples have been excluded owing to their not being kinematically associated with the Coma cluster. By Sturges' rule, eight class intervals are suggested and is accordingly reduced to five class intervals in the range $4500 - 9500 \text{ km s}^{-1}$. No discernible irregularity or departure from smoothness may be observed here. On the other hand, astronomers have traditionally employed much larger numbers of class intervals. For example, Broadhurst *et al.* [2] employ more than one hundred class intervals, lending to the figure the appearance of clustering. We have shown previously [12] that the use of inordinately large numbers of class intervals can provide randomly selected data with a sense of clustering or of pattern when displayed this way. This point has been made by others, such as Gentleman [8], in the statistics and statistics education literature.

As an illustration of how pattern may seemingly be established, we plot in Figure 2 the same histogram, but where the class interval was selected to be 72.45 km s^{-1} in order to conform with Tift's claim of "quantization" based on a much smaller and less accurate data set. (We have restricted the velocity range of the samples to eliminate sources within the field that kinematically could be associated with a different population, thereby reducing the sample size by 5%.) Interestingly, the pattern observed here does not exhibit the characteristic pattern of alternation that Tift observed in his substantially more restrictive, yet less accurate data set. No imprint of a "periodic" feature at 72.45 km s^{-1} intervals is immediately evident. (Tift's original data yields a histogram that has the appearance of a "picket fence" superposed on a Gaussian curve.) The use of an inordinately large number of class intervals has provided an illusion of some kind of pattern, albeit not one that shows any verifiable evidence of clustering. The smooth, indeed, dull behavior of Figure 1 as well as its bell-shaped character suggests that the data might be drawn from a population that is in some sense close to a Gaussian.

As a more refined way to examine our data, now restricted to the range $4500 - 9500 \text{ km s}^{-1}$, we provide in Figure 3 a Normal probability plot, a descriptive way of establishing whether the numbers in a variable are approximately Normally distributed. Starting with the smallest value in the data set, we pose the question: "If this were a sample from the standard Normal distribution (that is, the Normal distribution with zero mean and unit variance), what would I expect the smallest value to be?". The answer is the first normal score which depends only on the number of cases in the variable, and can be directly evaluated. The probability plot graphs the observed smallest value against the value expected under the assumption of a Normal distribution. The question is repeated for the second smallest value: "If this were a sample from the standard Normal distribution what would I expect the second smallest value to be?". The observed second smallest value is plotted against the expected second smallest value. The

corresponding questions are posed and calculations performed for each subsequent data value, and the resulting plot graphs the observed value versus the calculated normal score for each case. The closer the resulting plot is to a straight line, the closer the data is to a Normal population. In Figure 3, we compare the Normal probability plot (solid) against an ideally Normal population (dashed line).

The remarkable apparent degree of normality in the Coma cluster data deserved some measure of quantification, and we employed the Kolmogorov-Smirnov (K-S) test. The list of data points is directly converted to an unbiased estimator $S_N(x)$ of the cumulative distribution function of the probability distribution from which it was drawn, which in our case we hypothesize to be Normal. One can construct any number of statistics to measure the overall difference between two cumulative distribution functions, i.e. that describing the population of Coma cluster galaxies and the Normal population. One might consider the absolute value of the area between them, or the integrated mean square difference, essentially L_1 and L_2 norms, respectively. The Kolmogorov-Smirnov statistic D is a particularly simple measure: It is defined as the *maximum value* of the absolute difference between two cumulative distribution functions, essentially an L_∞ norm. The K-S statistic D is useful in that its distribution in the case of the null hypothesis (i.e. that the data set were drawn from a population that is Normally distributed) can be *approximately* calculated. Within the context of this approximation, we calculate $Q_{KS} \approx \text{Probability}(D > \text{observed})$. For the Coma cluster data, we obtain that $D \approx 0.05$ and that $Q \approx 1.34 \times 10^{-40}$. There are limitations to the K-S statistic, and the approximations involved necessarily diminish from the reliability of this measure of the improbability of the null hypothesis. Nevertheless, we can be quite confident that the Coma cluster data, if not Normally distributed, is very smoothly distributed and shows no evidence of substantial clustering. Thus, the Coma cluster data is a superb, practical test of the viability of the Power Spectrum Analysis method.

1.4 Power Spectrum Analysis in Practice

In Figures 4 and 5, we plot the Power Spectral Amplitude for the Coma cluster data and for Gaussian random noise (generated by conventional algorithms [15]), where the two data sets were normalized to zero mean and unit variance. Accordingly, given 192 samples in each, none of the normalized data were outside of the range $(0, 2\pi)$, in keeping with equation (1) of Yu and Peebles given earlier. (We have mapped the interval from π to 2π to the interval from $-\pi$ to 0.) Nevertheless, we observe occurrences of $|z_n|^2$ in excess of 6. (No qualitative change in the Power Spectrum Analysis plot emerges if the reduced set of 183 samples are employed.) In light of Yu

and Peebles' equation (5), it seems at face value highly unlikely (≈ 0.002) that such a large excursion could take place. However, it is essential to note that we are looking at the outcome of a *set* of experiments, when we look at such plots, as we are identifying the extremum of a distribution of individual spectral estimates. What is required in looking at the maximum value of the Power Spectral Amplitude or, more conventionally, the maximum value of m is that we employ the "statistical theory of extremes" [7]. The maximum value that m can obtain from a set of "experiments," i.e. the variation with respect to the reciprocal period n in Yu and Peebles' equation [1], has a probability distribution that is very different from the statistic P of equation (5). In particular, if $P(x)$ is the probability that in a given experiment the observed value $\geq x$, then $1 - P(x)$ is the probability that in a given experiment the observed value $\leq x$. Thus, in a set of N independent experiments, the probability that none of the observed results exceeds x is $[1 - P(x)]^N$. Finally, the probability $P_N(x)$ that at least one of the observed N independent results exceeds x is given by

$$P_N(x) = 1 - [1 - P(x)]^N . \quad (1.1)$$

This result is rigorous and general; the definition of $P(x)$ need not be that of equation (5) as derived by Yu and Peebles.

As a simple illustration of how different extreme value statistics can be from Normal statistics, consider a Gaussian random process with zero mean and unit variance. We are well familiar with the fact that 95% of the values realized in this process should lie between -1.96 and 1.96 units, and that 99% of the values should lie between -2.58 and 2.58 units. Similarly, for a single realization of this process, 95% of all values should lie *below* 1.65 units and 99% of all values should lie *below* 2.33 units. (The numbers in this latter situation are of course different, since half of all realized values are negative and we are considering only an *upper* cut off or maximum value.) Suppose now that we looked at the maximum value of N experiments and that these experiments are statistically independent. For a set of 100 experiments, we would expect a largest value of 3.28 with 5% likelihood, and 3.72 with 1% likelihood [14]. Meanwhile, for a set of 1000 experiments, we would expect a largest value of 3.88 with 5% likelihood, and 4.26 with 1% likelihood. The purpose of this demonstration is to show just how different extreme value behavior is from Normal behavior.

It is difficult to assess the number of statistically independent realizations that are present in Figures 4 and 5, but in principle we might expect that there are at most 192 (the number of datum in the original set). Employing Yu and Peebles' equation (5) as though it were exact, we directly obtain that there is a 38% likelihood of seeing at least one excursion in excess of 6.0 in both Figures 4 and 5. Therefore, it should come as no surprise that the PSA plot provides excursions as large as those we have provided.

There are, however, examples in the literature of PSA plots where the excursions are very large indeed. As an illustration, Guthrie and Napier

[9] have power spectral amplitudes as great as 20. Could this be result of a subtle imprint of the approximations employed by Yu and Peebles in deriving the exponential distribution (5)?

1.5 Power Spectrum Analysis Approximations Revisited

In Yu and Peebles equation (2), reprinted earlier, we note that they had implicitly employed a distribution function for x , that is uniformly distributed over the interval $[0, 2\pi]$ and, as a consequence, their claim of vanishing mean in z_n for $n \neq 0$ is a consequence of this assumption. (We suspect that they believed their result to be generally correct as a consequence of a "random phase" approximation.) Suppose that x is distributed according to some distribution function $\mathcal{P}(x)$ on a doubly infinite domain. Then, it follows that

$$\begin{aligned} \langle z_n \rangle &= N^{-\frac{1}{2}} \sum \langle e^{inx_i} \rangle \\ &= N^{\frac{1}{2}} \int_{-\infty}^{\infty} e^{inx} d\mathcal{P}(x) \end{aligned} \quad (1.2)$$

$$= N^{\frac{1}{2}} \mathcal{F}(n) \quad (1.3)$$

where $\mathcal{F}(n)$ is the characteristic function of $\mathcal{P}(x)$ defined by

$$\mathcal{F}(n) = \langle e^{inx} \rangle = \int_{-\infty}^{\infty} e^{inx} d\mathcal{P}(x) \quad (1.4)$$

(In order to extend the range of integration from that in (2), it is useful to think of Yu and Peeble's distribution function as having compact support, i.e. it vanishes identically outside of the interval from 0 to 2π .) As a relevant example owing to the observed Normality of the Coma cluster data, suppose $\mathcal{P}(x)$ is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance of σ^2 , i.e.

$$\mathcal{P}(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x' - \mu)^2}{2\sigma^2}\right] dx' \quad (1.5)$$

Then, the characteristic function $\mathcal{F}(n)$ becomes

$$\mathcal{F}(n) = e^{in\mu} e^{-\frac{n^2\sigma^2}{2}} \quad (1.6)$$

and, therefore,

$$\langle z_n \rangle = N^{\frac{1}{2}} e^{in\mu} e^{-\frac{n^2\sigma^2}{2}} \neq 0 \quad (1.7)$$

for our Normally distributed example.

In a similar way, it follows that Yu and Peebles' equation (3) for the variance of z_n should generally be

$$\begin{aligned}\langle |z_n|^2 \rangle &= \frac{1}{N} \sum_i \langle |e^{inx_i}|^2 \rangle + \frac{1}{n} \sum_{i \neq j} \langle e^{in(x_i - x_j)} \rangle \\ &= 1 + (N-1) \mathcal{F}(n) \mathcal{F}^*(n) \geq 1\end{aligned}\quad (1.8)$$

and, for our Normally distributed example,

$$\langle |z_n|^2 \rangle = 1 + (N-1)e^{-n^2\sigma^2} \quad (1.9)$$

Equations (1.7) and (1.9) show that the random variables z_n are not stationary with respect to n . Moreover, we observe that, for $n \neq m$,

$$\begin{aligned}\langle z_n z_m^* \rangle &= \frac{1}{N} \sum_{i,j=1}^N \langle e^{i(nz_i - mz_j)} \rangle \\ &= \frac{1}{N} \sum_i \langle e^{i(n-m)x_i} \rangle + \frac{1}{N} \sum_{i \neq j} \langle e^{i(nz_i - mz_j)} \rangle \\ &= \mathcal{F}(n-m) + (N-1) \mathcal{F}(n) \mathcal{F}^*(m) \\ &\neq 0.\end{aligned}\quad (1.10)$$

This result reduces to (1.8) if the case $n=m$. By inspection we note that the correlation matrix is Toeplitz, and only for a set of measure zero does it reduce to a diagonal one. Thus, we see that the ancillary variables z_n are not individual independent deviates but are correlated. This underscores the difficulty in estimating the number of independent random variables N in a Power Spectrum Analysis plot. Now we must explore the role of the Central Limit Theorem in this discussion.

1.6 Central Limit Theorem

In particular, the Central Limit Theorem [4][5] states that if $s_i \in \mathbb{R}$, $i = 1, \dots, N$ are individual independent deviates drawn from the same population and if their mean μ and variance σ^2 exist, i.e. $|\mu|, \sigma < \infty$, then the variable

$$S \equiv \sum_{i=1}^N \left[\frac{s_i - \mu}{\sqrt{n}\sigma} \right] \quad (1.11)$$

describes, in the limit $N \rightarrow \infty$, a Gaussian zero mean, unit variance process $\mathcal{N}(0,1)$. Counterexamples to the Central Limit Theorem [16] are particularly rare, and generally involve some subtle violation of the assumed conditions.

Special attention must be paid to the fact that z_n , as defined in (1), is complex valued, and that the properties of both the real and the imaginary parts of z_n must be considered separately. It is easy to show that

$$\begin{aligned}\mu &\equiv \langle z_n \rangle \\ &= \langle \Re(z_n) \rangle + i \langle \Im(z_n) \rangle \\ &= \mu_R + i\mu_I\end{aligned}\tag{1.12}$$

where μ is the population mean of z_n and where μ_R and μ_I denote the real and imaginary parts of μ , respectively. Additionally, we obtain

$$\begin{aligned}\sigma^2 &\equiv \langle |z_n - \mu|^2 \rangle \\ &= \langle |\Re(z_n - \mu)|^2 \rangle + \langle |\Im(z_n - \mu)|^2 \rangle \\ &= \sigma_R^2 + \sigma_I^2\end{aligned}\tag{1.13}$$

where σ^2 is the variance of z_n defined by $\langle |z_n - \mu|^2 \rangle$ and where σ_R^2 and σ_I^2 are the variances of the real and imaginary parts of z_n , respectively. However, it is now easy to show that the Central Limit Theorem can be employed to each of the components of z_n . The preceding results (1.7) and (1.8) show that conditions *necessary* for the Central Limit Theorem to hold with respect to Yu and Peebles equation (1) are not satisfied until n is asymptotically large, and (1) results in a biased estimate for a Gaussian zero-mean, unit variance process. The one outstanding issue that remains is to determine the distribution function for the real, semi-positive definite variable $|z_n|^2$, as well as to consider issues of accuracy.

1.7 Distribution Function for $|z_n|^2$

Consider now, for any n , the decomposition of z_n into its real and imaginary parts X and Y , respectively. Let us now define a variable W' according to

$$W' = X^2 + Y^2\tag{1.14}$$

It should be clear that W' has the role of $|z_n|^2$. From the preceding discussion of the Central Limit Theorem, we know that the distribution functions for X and Y are approximately Gaussian but, for the present time, will assume that they are completely arbitrary, say $\mathcal{P}_x(X)$ and $\mathcal{P}_y(Y)$, respectively. It follows that the characteristic function for $\mathcal{P}_w(W')$ must satisfy

$$\int e^{i\lambda W'} d\mathcal{P}_w(W') = \int e^{i\lambda(X^2+Y^2)} d\mathcal{P}_x(X) d\mathcal{P}_y(Y)\tag{1.15}$$

We multiply both sides of the equation by $\frac{1}{2\pi}e^{-ikW''}$, and integrate over k from $-\infty$ to ∞ . We then integrate over W'' from W to ∞ to obtain

$$1 - \mathcal{P}_w(W) = \int \Theta(x^2 + y^2 - W) d\mathcal{P}_x(X) d\mathcal{P}_y(Y) \quad (1.16)$$

This equation has a direct intuitive interpretation. The left hand side represents the probability that $|z_n|^2$ is greater than W , namely $P(|z_n|^2 > W)$ as in Yu and Peebles' equation (5), while the right hand side represents the contribution to the probability from all points $X + iY$ such that $X^2 + Y^2 > W$. As an illustrative example, suppose that both X and Y are $\mathcal{N}(0, 1/2)$. The right hand side then becomes

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\pi} \Theta(x^2 + y^2 - W) e^{-(x^2 + y^2)} dx dy = \int_{r=\sqrt{W}}^{\infty} \int_{\theta=0}^{2\pi} \frac{1}{\pi} e^{-r^2} r dr d\theta \quad (1.17)$$

after making the usual transformation to polar coordinates and then becomes

$$\int_{\sqrt{W}}^{\infty} d(r^2) e^{-r^2} = e^{-W} ; \quad (1.18)$$

combining the left and right sides, we have verified equation (5) of Yu and Peebles. However, our general result (1.16) will become important later in the discussion.

1.8 Consequences of Slow Convergence Rates

Suppose that n is sufficiently great that these concerns regarding the biased estimates of the mean and variance in equations (1.7) and (1.8) are not realized. Then, the real and imaginary parts of $\lim_{n \rightarrow \infty} z_n$ must become Gaussian zero-mean one-half unit-variance random variables. Then, the mean and variance of the complex variable z_n , as defined in equations (1.12) and (1.13), become zero and unity, respectively. The operative question then becomes how rapidly does z_n converge to a Gaussian random variable as a function of large N ? The formal answer, in the sense of probability (see Dudley [4], for example), is that the convergence rate goes as $N^{-1/2}$. The practical answer is that convergence is reasonably rapid near the center of the distribution, but is manifestly slower in the tail, depending on the underlying distribution for s_i in equation (1.11).

As a concrete example of this phenomenon, consider the "Sum of Uniform Deviates" method described in Abramowitz and Stegun [1] for generating Gaussian random numbers. Quoting from Abramowitz and Stegun, let $U_1, U_2, \dots, U_n$ be a sequence of n uniform deviates. Then

$$X_n = \left(\sum_{i=1}^n U_i - \frac{n}{2} \right) \left(\frac{n}{12} \right)^{-1/2}$$

will be distributed asymptotically as a normal random deviate. When $n = 12$, the maximum errors made in the normal deviate are 0.009 for $|X| < 2$, 0.9 for $2 < |X| < 3$. In other words, random numbers generated in this way that are between two and three standard deviations from the mean can have errors nearly as large as one standard deviation! The error in the normal deviate will be greater still if it is further than three standard deviations from the mean. There is a possibly apocryphal story that IBM used exactly this scheme for generating Gaussian random numbers on their mainframe computers many years ago, but later abandoned it owing to the anomalous behavior in the tail of the generated distribution.

Therein lies the crux of the problem. Hypothesis testing using the method of Power Spectrum Analysis is of particular interest only as $|z_n|^2$ becomes large, i.e. when we are in the tail of the distribution. Owing to the form of equation (1.16), it follows that the calculation of $P(|z_n|^2 > W)$ depends critically and in a nonlinear way upon the tail of the distribution functions $\mathcal{P}_x(X)$ and $\mathcal{P}_y(Y)$, respectively. Empirically, we observe that the errors in the associated probability density functions for X and Y can be as much as an order of magnitude. This error is further amplified by equation (1.16) in the calculation of $P(|z_n|^2 > W)$, and the relative error between the true distribution and the assumed exponential one can readily exceed an order of magnitude. For example, the data displayed in Figure 5 are manifestly *not* exponentially distributed, and it is in fact easy to produce a random set of data with "spectral peaks" that are very large. Finally, when we apply the extreme value statistic $1 - [1 - P(x)]^N$ in equation (1.1) to estimate the likelihood that a peak exceeds some threshold, the typically large value of N then dramatically amplifies this error.

1.9 Conclusion

Observational astronomy has provided researchers for many decades with deep questions emerging from the illusive appearance of pattern, of which, issues of potential "clustering" have been of profound importance. In recent years, efforts have shifted to the application of quantitative schemes for assessing the likelihood of clustering. Of these, the Power Spectrum Analysis method of Yu and Peebles [22] has become the principal tool of astronomers examining issues of clustering. The method itself is very novel and is widely accepted as being robust. The question that naturally emerges is why should it suggest that Gaussian random data, both of natural origin (the Coma Cluster) as well as from computer simulations, have substantial clustering?

We have explored in a mathematically conservative way the underlying assumptions and approximations employed by this method. We found that a subtle bias in certain implicit estimates (the mean and the variance of

a derived statistical quantity) renders part of the range of a Power Spectrum Analysis plot unreliable. However, far from the origin, the plot could contain useful statistical information, depending upon convergence to normality of certain quantities derived in the context of the Central Limit Theorem. We have shown that, owing to slow convergence to Normality—particularly in the tail of the distribution, the Power Spectral Amplitude will be *far removed* from the exponential distribution approximation of Yu and Peebles' (5). The exponential distribution function is a good approximation for only the small amplitude events in the PSA plot. In the context then of examining the *extreme* value of the Power Spectral Amplitude, i.e. the maximal peak, the exponential statistic is unreliable and any inference made for clustering is invalid.

In a limited *sense*, the Power Spectral Analysis method can provide some evidence for the *lack* of clustering. What is necessary, and what we hope that this meeting will produce, is that astronomers and statisticians will focus collectively on developing a robust method for *assessing* the presence and significance of clustering.

1.10 Figure Captions

Figure 1. Frequency histogram for Coma cluster velocities. Satisfies Sturges' rule for number of class interval for ≤ 183 samples [13].

Figure 2. Frequency histogram for Coma cluster velocities excluding unrelated galaxies in field. Class interval selected to be 72.45 km s^{-1} in violation of Sturges' rule. Total sample size 183.

Figure 3. Normal probability plot for Coma cluster velocities excluding unrelated galaxies in field. Solid line represents Normal probability plot for observed data; dashed line represents plot for ideal Normal distribution.

Figure 4. Power spectrum analysis plot for Coma cluster data, normalized to zero mean and unit variance. Sample size 192.

Figure 5. Power spectrum analysis plot for simulated Gaussian random noise, normalized to zero mean and unit variance. Sample size 192.

Acknowledgements: We wish to thank G. Molchan for several stimulating discussions regarding the asymptotic character of the variables employed in the method of Power Spectral Analysis. W.I.N. wishes to thank the Department of Astronomy at Cornell University for its hospitality while this work was initiated; this research was partly supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University. W.I.N. was the Stanislaw Ulam Visiting Scholar at Los Alamos

National Laboratory where much of this work was done. M.P.H. and Y.T. are associated with the National Astronomy and Ionosphere Center which is operated by Cornell University under a management agreement with the National Science Foundation.

1.11 REFERENCES

- [1] M. Abramowitz and I.A. Stegun 1965. *Handbook of Mathematical Functions*, New York: Dover Publications, pp. 952-953.
- [2] T.J. Broadhurst, R.S. Ellis, D.C. Koo, and A.S. Szalay 1990. "Large-Scale Distribution of Galaxies at the Galactic Poles," *Nature*, **343**, 726-728.
- [3] G.R. Burbidge and S.L. O'Dell 1972. "The Distribution of Redshifts of Quasi-Stellar Objects and Related Emission-Line Objects," *Astrophys. J.*, **178**, 583-605.
- [4] R.M. Dudley 1989. *Real Analysis and Probability*, Pacific Grove, California: Wadsworth and Brooks/Cole, 436 pp.
- [5] W. Feller 1968. *An Introduction to Probability Theory and its Applications*, Third Edition, Volume 1, New York: John Wiley and Sons, 509 pp.
- [6] K. Fukunaga 1990. *Introduction to Statistical Pattern Recognition*, second edition, Academic Press Inc.: Boston, 591 pp.
- [7] J. Galambos 1978. *The Asymptotic Theory of Extreme Order Statistics*, New York: John Wiley & Sons.
- [8] J.F. Gentlemen 1977. "It's All a Plot (Using Interactive Computer Graphics in Teaching Statistics)," *Amer. Statistician*, **31**, 166-175.
- [9] B.N.G. Guthrie and W.M. Napier 1990. "The Virgo Cluster as a Test for Quantization of Extragalactic Redshifts," *Mon. Not. R. astr. Soc.*, **243**, 431-442.
- [10] B. Julesz 1981. "Textons, the Elements of Texture Perception, and their Interactions," *Nature*, **290**, 91-97.
- [11] S.M. Kent and J.E. Gunn 1982. "The Dynamics of Rich Clusters of Galaxies. I. The Coma Cluster," *Astron. J.*, **87**, 945-971.
- [12] W.I. Newman, M.P. Haynes, and Y. Terzian 1989. "Double Galaxy Redshifts and the Statistics of Small Numbers," *Astrophys. J.*, **344**, 111-114.

- [13] W.I. Newman, M.P. Haynes, and Y. Terzian 1991. "Pattern, Perception, and Probability: Statistical Inference and Astronomical Measurement," submitted to *Astrophys. J.*.
- [14] E.S. Pearson and H.O. Hartley 1962. *Biometrika Tables for Statisticians*, Cambridge: Cambridge University Press, Table 24.
- [15] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling 1986. *Numerical Recipes*, Cambridge: Cambridge University Press, 818pp.
- [16] J.P. Romano and A.F. Siegel 1986. *Counterexamples in Probability and Statistics*, Monterey, California: Wadsworth & BrooksCole Advanced Books & Software, 303 pp.
- [17] W. Sheehan 1988. *Planets and Perception Telescopic Views and Interpretation*, Tucson: University of Arizona Press, 324 pp.
- [18] H.A. Sturges 1926. "The Choice of a Class Interval: Case I. Computations Involving a Single Series," *J.A.S.A.*, 21, 65-66.
- [19] W.G. Tift 1973. "Properties of the Redshift-Magnitude Bands in the Coma Cluster," *Astrophys. J.*, 179, 29-44.
- [20] W.G. Tift 1976. "Discrete States of Redshift and Galaxy Dynamics. I. Internal Motions in Single Galaxies," *Astrophys. J.*, 206, 38-56.
- [21] W.G. Tift and S.A. Gregory 1976. "Direct Observation of the Large Scale Distribution of Galaxies," *Astrophys. J.*, 205, 696-708.
- [22] J.T. Yu and P.J.E. Peebles 1969. "Superclusters of Galaxies?", *Astrophys. J.*, 158, 103-113.