Walsh–Fourier Analysis and Its Statistical Applications

DAVID S. STOFFER*

The aim of this article is to acquaint statisticians and practitioners, whose research activities require statistical methodology, with the statistical theory and applications of Walsh–Fourier analysis. It has been suggested that Walsh spectral analysis is suited to (albeit not restricted to) the analysis of discrete-valued and categorical-valued time series, and of time series that contain sharp discontinuities. I explain the need for Walsh–Fourier analysis, review the history and properties of Walsh functions, and outline the existing Walsh–Fourier theory for real-time stationary time series. I discuss various statistical applications based on the Walsh–Fourier transform and provide an annotated bibliography.

KEY WORDS: Analysis of variance; Classification and prediction; Coherency; Discrete- and categorical-valued time series; Multivariate time series; Scaling; Sequency; Spectral analysis; Walsh–Fourier transform; Walsh functions.

1. INTRODUCTION

Many of us were introduced to the scientific world of *sinusoids* (sine and cosine waves—see Figure 1) through elementary exhibitions and discussions of light waves. When we were young, we saw rainbows and were told that sunlight (white light) was made up of the different colors of the rainbow. Teachers gave us prisms to look through and asked us to memorize the array of colors that emerged from the prism (Roy G. Biv: red, orange, yellow, green, blue, indigo, violet) and we called the array of colors the *spectrum*. Later, we learned that light was a wave (sinusoid). We were told that if a beam of light is passed through a prism, different wavelengths (colors) would be refracted through different angles. The long wavelengths (red) are refracted least, and the short wavelengths (violet) are refracted most. We were introduced to the concept of the *frequency* of oscillation of a wave, measured in cycles per second (cps). For the visible spectrum, the color red had the slowest frequency of oscillation (about 4×10^{14} cps) while violet had the fastest frequency of oscillation (about 7.5×10^{14} cps). Although we could see the colors, we could not see the waveforms. Moreover, there were soundwaves and invisible electromagnetic waves, such as infrared or x-rays, that we envisioned to be sinusoidal. Perhaps we did so because some of the waveforms that we did see, such as the wave motion caused by plucking a guitar string or the wave motion observed at the surface of a body of water produced by the wind or some other disturbance, were sinusoidal.

But not all waveforms are sinusoidal. In the latter part of the nineteenth century, Alexander Graham Bell tried to develop telegraphy equipment using sinusoidal functions but failed because he coud not produce sinusoidal voltages. Instead, his voltages were square waves (block pulses—see Figure 2) while his receivers resonated with sine waves. Bell's contribution, of course, was the discovery of voice transmission by electricity, but his telegraphy transmitter decomposed voice into square waves and the receiver recomposed it from the square waves. Although the trigonometric functions (sine and cosine) were well known (as was Fourier analysis which I shall discuss briefly in a moment) almost no practical use could be made of this knowledge with the technology available at the time. Evidently, the decomposition of voice into square waves predates the decomposition into sine waves by many decades. See Harmuth (1977, pp. 3–7) for further discussion and references on the history of communications.

In 1822, the French mathematician J. L. Fourier offered a solution to a heat conduction problem by a trigonometric series representation (Fourier 1822). Since then, Fourier analysis has played a major role in applications to problems in science and engineering. Fourier showed that (almost) any periodic (repeating) function can be represented, to any desired degree of approximation, by a series consisting of a sum of sine and cosine functions (sinusoids)—this notion will be discussed in more detail in Section 3. In the statistical analysis of time series, Fourier methods are used to discover and analyze the regularity or periodicity in data; this technique formalizes the concept of dependence or correlation between adjacent time points that one typically encounters in the collection of data over time. For example, Figure 3 shows a plot of the average monthly temperatures recorded in Dubuque, Iowa, from January 1964 to December 1975 (this data set was taken from Cryer 1986). The data clearly show sinusoidal behavior with (quite understandably) a yearly cycle. Using linear regression techniques, we can approximate the behavior of the data using sines and cosines; in particular, Cryer (1986, pp. 34–35) fit the regression model

$$X(t) = \beta_0 + \beta_1 \cos(2\pi t/12) + \beta_2 \sin(2\pi t/12) + \epsilon(t)$$

to the temperature data $X(t)$, where $\epsilon(t)$ is the usual regression error term and t denotes time in terms of months ($t = 1, \ldots, 144$; 12 years of monthly data). In this regression

* David S. Stoffer is Associate Professor, Department of Mathematics and Statistics, and Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260. This work was supported in part by National Science Foundation Grant DMS-9000522, by a grant from the Centers for Disease Control through a cooperative agreement with the Association of Schools of Public Health, National Institute on Alcohol Abuse and Alcoholism Grant AA06390, and National Institute on Drug Abuse Grant DA03209. The author thanks Nancy L. Day and Mark S. Scher, School of Medicine, University of Pittsburgh, for providing the neonatal EEG sleep data, Allan R. Sampson and David E. Tyler for their helpful discussions on the problem of scaling categorical variables, an associate editor and the editor, Donald Guthrie for detailed comments that substantially improved the presentation of this article. The article was partially written while the author was visiting the Department of Statistics, Carnegie Mellon University, and the Institute of Mathematics and Statistics, University of São Paulo, Brazil.

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Review Paper

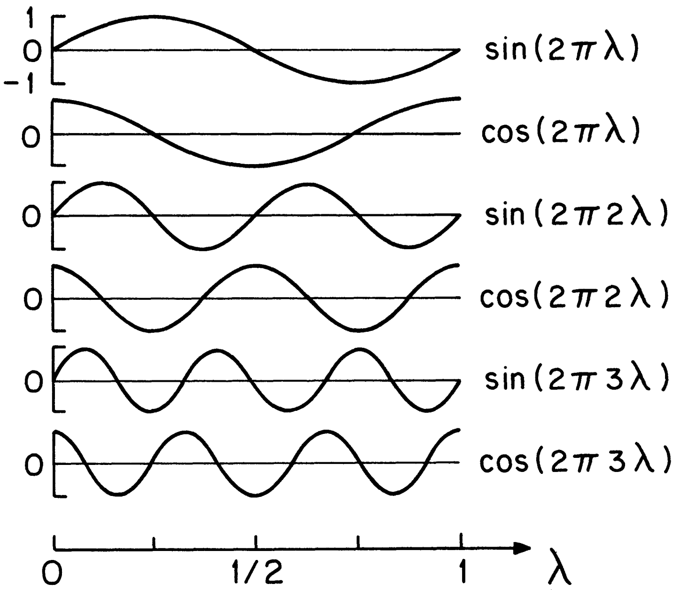

Figure 1. Sine and Cosine Waves.

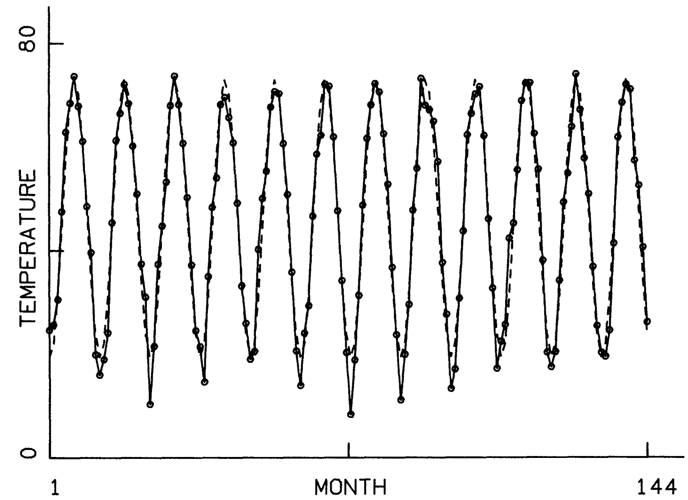

Figure 3. Average Monthly Temperature, Dubuque, Iowa, January 1964 to December 1975; Observed (Solid Line–Circles) and Predicted (Dashed Line).

model, $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$ are the independent or regressor variables; note that $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$ each make a complete cycle every 12 time points (months) and are said to have a frequency of $1/12$ cycles per month [that is, for example, the graph $\cos(2\pi t/12)$ vs. t ($t = 1, 2, \ldots$) repeats itself every 12 time points]. Cryer illustrated the Minitab commands needed to fit the above regression model as well as the output from the commands applied to

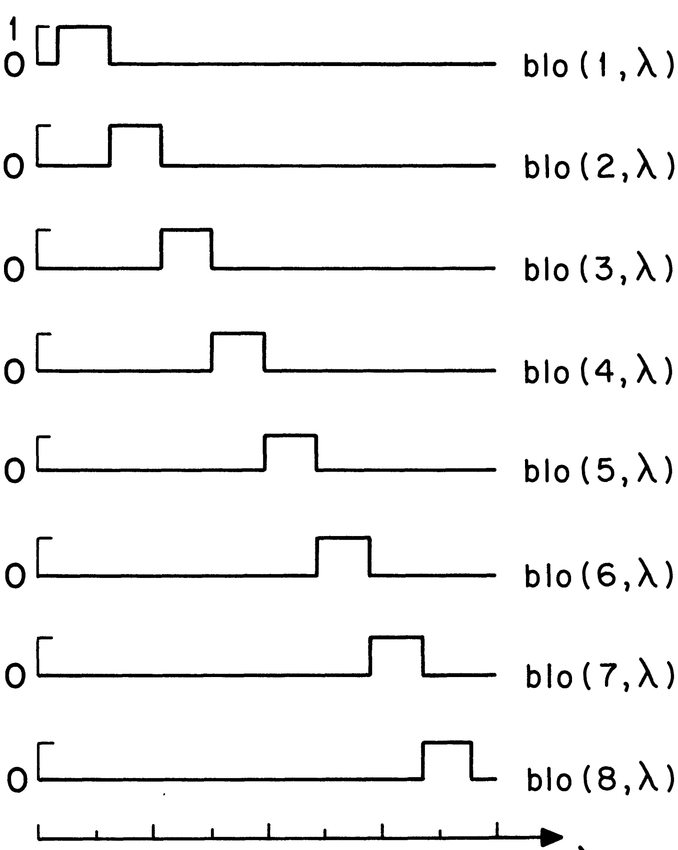

Figure 2. Block Pulses.

the temperature data. For this data set the fitted model is

$$\hat{X}(t) = 46.3 - 22.0 \cos(2\pi t/12) - 15.2 \sin(2\pi t/12),$$

which is also plotted in Figure 3; the R-square value in this case is 96.4%. This is a simple example, where the cyclic or periodic variation in the data is discovered through visual analysis. In more complex problems, we need a more precise method of identification, as well as an assessment of the statistical significance of the cyclic (harmonic) components of variation exhibited in the data. This need is best answered by using the techniques of *spectral analysis* (also referred to as *frequency domain analysis* or *Fourier analysis*) for time series.

But we have seen that not all waveforms are sinusoidal. For example, neurologists are interested in the cyclic behavior of electroencephalographic (EEG) sleep patterns; such patterns have been used in a variety of ways, from the assessment of the cerebral maturation and neurophysiological organization of the central nervous system in premature children to the treatment of psychiatric disorders (such as depression) in adults. There are basically two kinds of sleep: non-rapid-eye-movement (non-REM) or quiet sleep and rapid-eye-movement (REM) or active sleep, and there are various stages within each sleep state. REM sleep alternates with non-REM sleep in about 90-minute intervals in adults and at about 45–60 minute intervals in children. Figure 4 shows the per minute EEG sleep record (for 120 minutes) of a normal full-term infant (ID 465) approximately 24–36 hours after birth (the corresponding total number of body movements per minute is also shown—this will be used in Section 6.3). Here, sleep-state is categorized (per minute) into one of six possible states using the labels: State 1 quiet sleep—trace alternant; State 2 quiet sleep—high voltage; State 3 indeterminate sleep; State 4 active sleep—low voltage; State 5 active sleep—mixed; State 6 awake. Clearly, the EEG sleep-state waveform exhibited in Figure 4 is nonsinusoidal; in fact, it is a square waveform. We could, using Fourier (trigonometric) techniques, synthesize the sleep

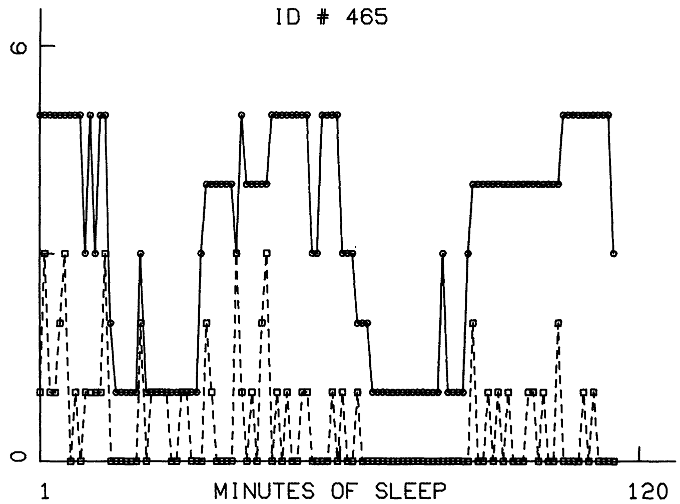

Figure 4. Sleep State (Solid Line–Circles) and Total Number of Body Movements (Dashed Line–Squares) of the Unexposed Infant.

pattern in Figure 4 to any desired degree of accuracy by the sum of sine and cosine functions. It is obvious, however, that we could never realistically recover the abrupt switching pattern between the EEG sleep states using the smooth sinusoidal waveforms. In fact, it makes more sense to describe square waveforms with abrupt switches, such as the sleep pattern in Figure 4, in terms of square waveforms such as the Walsh functions shown in Figure 5. It is simply a matter of using the right tool for the job.

Although the analysis of time series data using Fourier methods is well established in statistics, the statistical analysis of data (such as the sleep state data) based on the Walsh functions is virtually nonexistent; the aim of this article is to partially fill this gap. The remainder of this article is arranged as follows. I will briefly discuss the history of the Walsh functions in Section 2. In Section 3, I will present a brief discussion and comparison of Fourier-based and

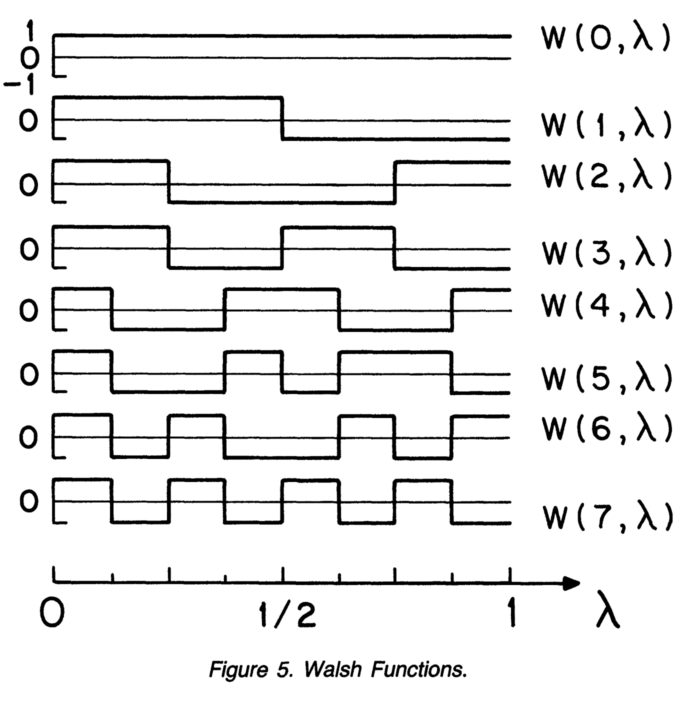

Figure 5. Walsh Functions.

Walsh–Fourier based analysis. In Section 4 I will offer a more detailed account of the Walsh functions, their generation, their behavior, and the role of *dyadic time* in Walsh analysis. In Section 5, I will present a summary of the existing Walsh–Fourier theory for real-time stationary time series. This will include the definition of the Walsh–Fourier transform, the Walsh–Fourier spectrum, and asymptotic results (such as central limit theorems) for the Walsh–Fourier transform; multivariate extensions and coherency will also be discussed briefly. In Section 6 I will present various applications of Walsh–Fourier analysis in statistics. First, Fourier and Walsh–Fourier analyses of the sleep data shown in Figure 4 are compared. Next, an analysis of variance where categorical time series are collected as part of the experimental design (Stoffer 1987; Stoffer, Scher, Richardson, Day, and Coble 1988) is presented. This is followed by a brief discussion of a Walsh–Fourier analysis of coherency (Stoffer 1990) and some preliminary results on scaling techniques for categorical time series based on the Walsh–Fourier transform. Finally, I will summarize the statistical application of the Walsh–Fourier transform in the classification of multivariate binary data (Ott and Kronmal 1976). A brief discussion is given in Section 7, and then a fairly complete set of references and an annotated bibliography is included.

Before proceeding, I would like to stress the following points. Walsh–Fourier (square wave) analysis is not a substitute for Fourier (trigonometric) analysis, nor is it a replicate of Fourier analysis. Each technique will have its own advantages in certain situations. Moreover, it is not always necessary to choose between the two approaches—both may be applied in some problems.

2. HISTORY OF WALSH FUNCTIONS

By the start of the twentieth century, scientists were well aware of the existence of many useful orthogonal systems of continuous functions, such as the set of orthogonal trigonometric functions that occur in Fourier analysis. The field was developed further as mathematicians constructed orthogonal systems with functions that were not continuous. In 1923, J. L. Walsh published a complete set of orthogonal functions that take on only two values, ± 1 ("on" and "off"), and are similar in oscillation and many other properties to the trigonometric functions (Walsh 1923). Although other binary-valued, sets of discontinuous orthogonal functions were constructed (Haar 1910; Rademacher 1922), the system due to Walsh is more prominent in terms of recent studies and applications. Paley (1932) reintroduced Walsh functions to the scientific community by defining them as the product of Rademacher functions. Walsh's definition, in terms analogous to the behavior of trigonometric functions, is more appealing in applications, nevertheless, Paley's definition was better suited for mathematical considerations. Fine published about eight articles in the 1950s (see Bramhall 1974, for a listing) dealing with some of the mathematical properties of Walsh functions as defined by Paley; the most important of these are Fine (1949, 1950, 1957). In particular, Fine generalized the Walsh functions (Fine 1950). Further advances in generalized Walsh functions were developed by Chrestenson (1955), and Selfridge (1955) combined these generalizations to develop a theory of Walsh transforms.

During the 1960s, H. F. Harmuth applied the Walsh functions to various problems in engineering; those applications included problems in signal detection, discrimination and classification, image coding and transmission, logical circuitry, and multiplexing (multiplexing deals with the simultaneous transmission of a group of independent messages over a shared channel). In 1969, Harmuth published an article on the applications of Walsh functions in the field of communications that appeared in the *IEEE Spectrum* (Harmuth 1969a); this article, as well as Harmuth's text on the subject (Harmuth 1969b; a second edition was published in 1972) inspired a tremendous amount of work in the area and led to a great surge in the research of the theory and applications of Walsh functions in engineering and computer science that lasted well into the next decade. Beginning in 1969, there were annual symposia on Walsh functions and their applications in Washington, D.C., and biennial symposia in Britain (in 1971, 1973, 1975). The *Proceedings of the Symposia on the Applications of Walsh Functions* held in Washington, D.C., were published for the years 1970–1974; however, the 1974 proceedings volume contained articles and recollections of various participants of the 1969 symposia. (The Washington, D.C. proceedings are published and made available by the National Technical Information Service, U.S. Department of Commerce, Springfield, VA 22151. The British *Proceedings of the Symposia on the Theory and Applications of Walsh Functions* were published by the Hatfield Polytechnic Institute, Hatfield, Hertfordshire, England.) Bramhall (1974) published a bibliography on Walsh and Walsh-related functions containing roughly 800 entries, most of which were published in the late 1960s and early 1970s. As an example of the Walsh fever during this period, N. M. Blachman of the General Telephone and Electronics Corporation (GTE), California, explained how he began his eminent research on the applications of Walsh functions: "In about 1971, because there seemed to be Important People among our customers who'd been impressed by rumors they'd heard . . . concerning the wonderful things that only Walsh functions could do, our Chief Engineer asked me whether we should be making use of Walsh spectral analysis in our work here on signal analysis . . ." (personal communication, 1989). But by the late 1970s, interest in Walsh functions began to wane and there were no more symposia (it is interesting to note that J. L. Walsh never did anything more with the Walsh functions after his 1923 article, however, he did appear at at least one of the symposia in Washington, D. C., where he was presented a pair of argyle socks patterned in the form of Walsh functions). H. D. Frankel, in his address at the 1971 Symposium on the Applications of Walsh Functions as Chairman of the Panel on the Applications of Walsh Functions, opened with the following remarks: "The purpose of the panel is to discuss . . . the possible uses of Walsh functions . . . what makes such a discussion timely is that the theory is coming along nicely, some interesting [equipment has] been built, and so many more people and agencies are involved. What makes the discussion especially attractive to anticipate, is that the field has matured to the point of having bred counter-revolutionaries; there exist former adherents who now strongly oppose the use of Walsh functions!" (Frankel 1971, p. 134). Evidently, many of these researchers realized that, contrary to their initial beliefs, Walsh functions were not the answer to all of their engineering problems.

The first mention of Walsh analysis in the statistics literature was Good (1958), who described a "mod 2 three dimensional discrete Fourier transform," which is the discrete Walsh–Fourier transform for a sample of size 2^3, although he did not give any references to the Walsh literature of the time. In that article, Good pointed out the fact that the effects in a 2^n factorial experiment could be regarded as an n-dimensional mod 2 discrete Fourier (Walsh–Fourier) transform of the data; he related this to Yates's algorithm for calculating the effects in a 2^n factorial design and used these ideas to develop a fast Fourier transform (also see Good 1971). In 1972, P. A. Morettin began developing a statistical theory for the analysis of time series based on Walsh functions (Morettin 1972). As in the engineering literature of the time, however, Morettin primarily based his work on *dyadic time*—dyadic time uses a strange clock where, for example, if it is 5 p.m. now, in 3 hours it will be 6 p.m. (I will discuss the concept of dyadic time in more detail in Section 4). Morettin (1973, 1974a) was the first to discuss results for the statistical analysis of *real-time* stationary time series; nevertheless, many of Morettin's articles on this subject concentrated on the development of a statistical theory for dyadically stationary time series (Morettin 1974b, for example). Morettin (1981) published an excellent review article of Walsh–Fourier analysis for dyadic- and real-time stationary processes. I will not discuss the dyadic-time references in detail, but interested readers can see Morettin (1981) for references and discussions. A statistical application based on the Walsh functions appeared in Ott and Kronmal (1976), where the Walsh–Fourier transform was used in a classification problem for multivariate binary data (a similar technique was described in the engineering literature in Ito 1970); this technique will be discussed at the end of Section 6. Texter and Ord (1989) used the Walsh functions to develop a procedure to choose the order of differencing (a technique used to remove trends in time series data) in automatic forecasting procedures.

Kohn (1980a,b) established theoretical results concerning the statistical application of Walsh functions to real-time stationary processes. In these articles, Kohn laid the groundwork for the Walsh–Fourier analysis of real-time stationary processes by showing that many of the results concerning the decomposition of real-time stationary time series using trigonometric functions have their Walsh functions analogs—Kohn's approach to the problem was based on results from the engineering literature, in particular, Robinson (1972a,b). Since Kohn (1980a,b), other articles in the statistical literature that established Walsh–Fourier theory for real-time stationary time series were Morettin (1981, 1983) and Stoffer (1985, 1987, 1990). And recently (Stoffer et al. 1988), an analysis of variance based on the

Walsh–Fourier transform was used to assess the effect of maternal alcohol consumption on neonatal EEG sleep-state cycling (this was an analysis of categorical time series based on the Walsh functions that will be described Section 6).

3. FREQUENCY AND SEQUENCY DOMAIN ANALYSIS: SOME BASIC CONCEPTS

The basis of frequency domain (Fourier) analysis for time series is the *spectral representation theorem* for stationary processes (see Brockwell and Davis 1987, th. 4.8.2, for example). By (second-order) *stationarity* (in real time), we mean a time series, $\{X(t), t = 0, \pm 1, \pm 2, \ldots\}$, that has a constant mean level $E\{X(t)\} = \mu$, for all t, and for which the covariance between observations at times t and s, $\text{cov}\{X(t), X(s)\} = \gamma(t - s)$, is a function only of the time difference or lag, $t - s$ for all times t and s (γ is called the *autocovariance function*). Roughly, the representation theorem says that we may think of a (mean-zero) stationary time series as being formed by the random superposition of sine and cosine waveforms,

$$X(t) = \sum_{j=1}^{q} [A(j) \cos(2\pi\lambda_j t) + B(j) \sin(2\pi\lambda_j t)], \quad (3.1)$$

where $\lambda_1, \ldots, \lambda_q$ are different frequencies measured in cycles per unit time and the $A(j)$'s and $B(j)$'s are mutually uncorrelated, mean-zero random variables with $\text{var}\{A(j)\} = \text{var}\{B(j)\} = \sigma_j^2$. This implies that the total variance in the time series is $\text{var}\{X(t)\} = \Sigma_{j=1}^{q}\sigma_j^2$, and that the total variance in the time series can be decomposed (as in an analysis of variance) into components σ_j^2 corresponding to sinusoidal waveforms at various frequencies of oscillation. Ideally, q is small and the λ's are well separated in (3.1). Figure 6 shows a particular example of a time series that would be produced by (3.1) in the case where $q = 4$, with frequencies $\lambda_1 = 2$, $\lambda_2 = 5$, $\lambda_3 = 20$, and $\lambda_4 = 50$ cycles per unit of time and with $[A(j), B(j)] = (2, 1), (1, 2), (1, -2), (2, -2)$ $(j = 1, 2, 3, 4)$.

In the temperature-data example presented in Section 1 and shown in Figure 3, it was easy to visually (and intellectually) establish the fact that the data had a yearly cycle, but in more complicated problems we would not be able to rely solely on our sight and intellect to discover the harmonic components of the data. Instead, the idea of correlating the data with sinusoids at various frequencies seems appealing. For example, the temperature data shown in Figure 3 is highly correlated with $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$, that is, the sinusoids that make one cycle every 12 months.

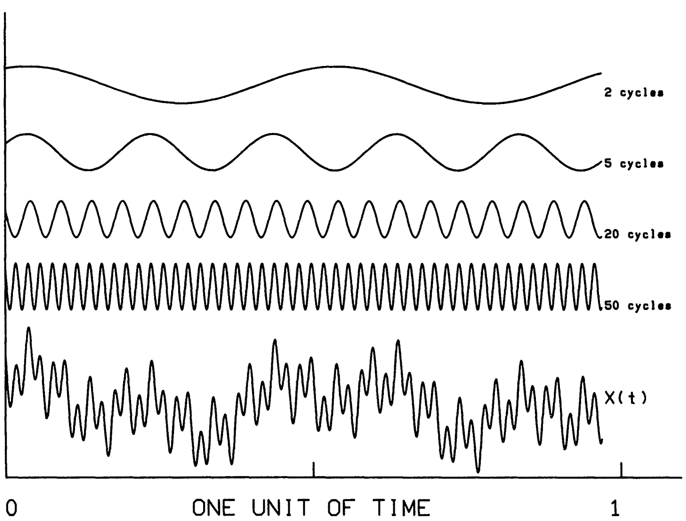

Figure 6. A Time Series X(t) as the Sum of Four Sinusoids.

If $X(0), X(1), \ldots, X(N - 1)$ represents time series data that we suspect have various periodic components, we may try to discover these harmonic components by computing the *cosine transform* (which is essentially the correlation of the data with cosines)

$$C(\lambda_j) = N^{-1/2} \sum_{t=0}^{N-1} X(t) \cos(2\pi\lambda_j t) \quad (3.2)$$

and the *sine transform* (which is essentially the correlation of the data with sines)

$$S(\lambda_j) = N^{-1/2} \sum_{t=0}^{N-1} X(t) \sin(2\pi\lambda_j t) \quad (3.3)$$

where $\lambda_j = j/N$ (that is, j cycles per N time points; $1 \le j \le N/2$). Typically the (Fourier) *periodogram* of the data

$$I_F(\lambda_j) = C^2(\lambda_j) + S^2(\lambda_j) \quad (3.4)$$

is computed, and a plot of $I_F(\lambda_j)$ versus λ_j is inspected for peaks. Again, the idea here is that $I_F(\lambda_j)$ will be large when the time series $X(t)$ contains harmonic components near the frequency λ_j—the periodogram $I_F(\lambda_j)$ is essentially the squared correlation of the data with the sine and cosine waves that oscillate at frequency λ_j.

A good introduction to the spectral (Fourier) analysis of time series can be found in the texts by Bloomfield (1976) and Shumway (1988); Priestley (1981) provided a comprehensive treatment of spectral methods at an intermediate level. Brillinger (1981) and Hannan (1970) are the classic works written at an advanced level, whereas a mathematical treatment of spectral analysis at the graduate level can be found in the texts by Brockwell and Davis (1987) and Fuller (1976).

The same basic ideas are employed in Walsh–Fourier analysis. As previously mentioned, the Walsh functions are similar in some respects to the system of sines and cosines used in Fourier analysis, however, unlike their sinusoidal counterparts, the Walsh functions are square waveforms that take on only two values, +1 and −1 ("on" and "off"). The sinusoids in Fourier analysis, $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$ $(n = 1, 2, \ldots)$, are distinguished by their frequency of oscillation n in terms of the number of complete cycles they make in the interval $0 \le \lambda < 1$. For example, for the frequency $n = 3$, $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$ each complete three cycles in the unit interval. The Walsh functions exhibited in Figure 7, denoted by $W(n, \lambda)$ $(n = 0, 1, 2, \ldots, 0 \le \lambda < 1)$, are distinguished by the number of times n that they switch signs in the unit interval. For example, $W(3, \lambda)$ switches signs three times in the unit interval $0 \le \lambda < 1$, from +1 to −1 at $\lambda = 1/4$, then from −1 to +1 at $\lambda =$

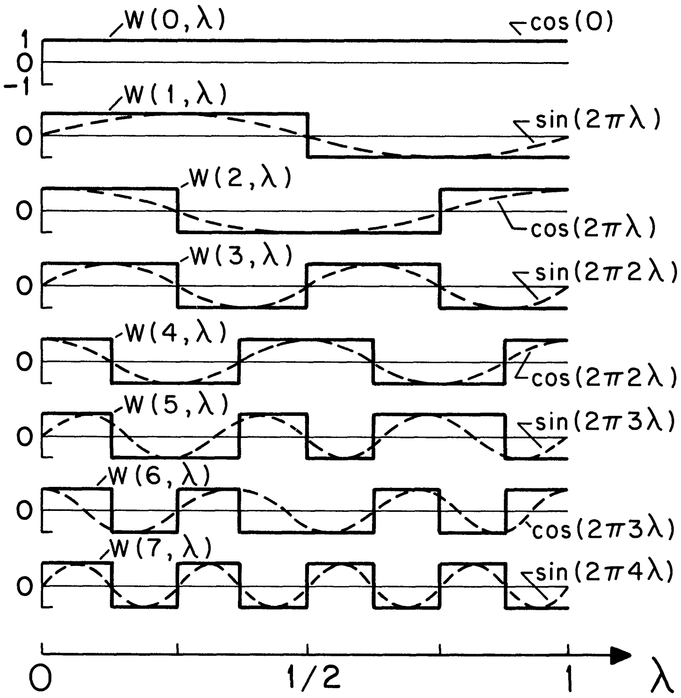

Figure 7. Walsh and Fourier Harmonics.

1/2, and finally from $+1$ to -1 at $\lambda = 3/4$. In order to compare the Fourier and Walsh harmonics, Figure 7 shows a plot of sinusoids superimposed on the Walsh functions. Of course the domain of the Walsh functions, like the sinusoids, is extended to the entire real line ($-\infty < \lambda < \infty$) in an obvious way.

Since the Walsh functions are aperiodic, the value n in the notation $W(n, \lambda)$ cannot be called frequency as in the case of the periodic sinusoids $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$. Harmuth (1969b) introduced the term *sequency* to describe generalized frequency to distinguish functions, however, such as the Walsh functions, that are not necessarily periodic. Harmuth noted that the frequency parameter n in $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$ may also be interpreted as one half the number of zero crossings or sign changes per unit time (the zero crossing at $\lambda = 0$, but not the one at $\lambda = 1$ is counted for sine functions). For example, in Figure 7, when $n = 3$, $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$ each cross zero (or change signs) six times. In analogy to the relationship of frequency to the number of zero crossings or sign changes in periodic functions, *Harmuth-sequency* is defined to be one half the average number of zero crossings or sign changes that a function makes per unit time—this concept can be applied to aperiodic as well as periodic functions and the definition of Harmuth-sequency coincides with that of frequency when applied to sinusoidal functions. I find Harmuth's definition of sequency to be forced in an attempt to get Walsh-based analysis to behave like Fourier-based analysis [Harmuth also defined "cosine Walsh" and "sine Walsh" or *cal* and *sal* functions to further the analogy of Walsh to Fourier analysis. This idea permeates the engineering literature on Walsh functions and interested readers can refer to Harmuth (1969b), Ahmed and Rao (1975), or Beauchamp (1975, 1984)]. But this brings me back to a point I made earlier: Walsh–Fourier and Fourier analysis are not replicates of each other—each has its own advantages and disadvantages, and there is no need to replace one with the other. Henceforth, I will adopt the term sequency in reference to the Walsh functions, however, the term will simply denote the number of zero crossings or sign switches that a Walsh function makes in the unit interval. For example, $W(3, \lambda)$ is the third sequency-ordered Walsh function since $W(3, \lambda)$ makes three sign switches or zero crossings in the unit interval.

In the sequency domain (Walsh–Fourier) analysis of time series, we will be interested in the discrete version of the Walsh functions since time series typically are sampled at discrete, equidistant time points. The first eight discrete, sequency-ordered Walsh functions, $W(n, m/N)$ ($m, n = 0, 1, \ldots, 7$), corresponding to a sample of length $N = 2^3$ are shown below as the rows (columns) of a symmetric matrix called the Walsh-ordered Hadamard matrix, $\mathbf{H}_W(3)$:

$$\mathbf{H}_W(3) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}$$

For example, the fourth row or column of $\mathbf{H}_W(3)$ corresponds to the discrete Walsh function $W(3, m/8)$ ($m = 0, 1, \ldots, 7$), which makes three sign changes (this is its sequency value) as m varies from 0 to 7. The discrete Walsh functions are simply the continuous Walsh functions evaluated at equidistant points along the unit interval. For example, $W(3, m/8)$ is the third Walsh function $W(3, \lambda)$ pictured in Figure 7, evaluated at $\lambda = 0/8, 1/8, \ldots, 7/8$.

By analogy to Fourier analysis, the *Walsh periodogram* of the data, $X(0), \ldots, X(N - 1)$, is

$$I_W(\lambda_j) = \left[N^{-1/2} \sum_{t=0}^{N-1} X(t)W(t, \lambda_j) \right]^2, \tag{3.5}$$

where λ_j is a sequency of the form $\lambda_j = j/N$ (j switches per N time points, $1 \leq j \leq N - 1$). The Walsh periodogram is essentially the squared correlation of the data with the Walsh functions at various rates of switching. One can plot $I_W(\lambda_j)$ versus λ_j to inspect for peaks, as in the periodic case. The term being squared in (3.5) is the *Walsh–Fourier transform* of the data, in analogy to the cosine and sine transforms in (3.2) and (3.3).

Unfortunately, we cannot obtain the analog of the trigonometric representation of $X(t)$ given in (3.1). That is, if we propose that a time series is the superposition of Walsh functions at various sequencies,

$$X(t) = \sum_{j=1}^{q} A(j)W(t, \lambda_j), \tag{3.6}$$

where $\lambda_1, \ldots, \lambda_q$ are q distinct sequencies and the $A(j)$'s are mutually uncorrelated mean-zero random variables with $\operatorname{var}\{A(j)\} = \sigma_j^2$, then the time series $X(t)$ is not stationary, but is *dyadically stationary* (see Section 4 for a definition of dyadic stationarity). Although this problem makes the Walsh–Fourier decomposition more difficult to interpret than the Fourier (trigonometric) decomposition, it does not totally negate the practical utility of the Walsh–Fourier decomposition.

As previously mentioned, there are many physical situations in which time series cannot be thought of as the superposition of well-separated sinusoids. For example, if the process of interest is discrete or categorical-valued with values in some finite set (e.g., square waveforms), then it makes little sense to correlate the data with smooth sines and cosines. As an alternative, it has been suggested that the spectral analysis of time series that contain sharp discontinuities be conducted in the sequency domain via the Walsh–Fourier transform. This seems to be a natural alternative to the usual Fourier analysis since the Walsh–Fourier transform is based on square-wave Walsh functions. This approach enables investigators of square-wave phenomena to analyze their data in terms of square waves and sequency (switches per unit time) rather than sine waves and frequency (cycles per unit time). Beauchamp (1975, chap. V, sec. F; 1984, sec. 3.3.4) empirically demonstrated that the respective roles of Walsh and Fourier spectral analysis for discontinuous and smooth varying time series, respectively, are clear. He concluded that where a signal (time series) is derived from a sinusoidally based waveform, Fourier analysis is relevant, and where the signal contains sharp discontinuities and a limited number of levels, Walsh analysis is appropriate. Beauchamp (1975, 1984) provided an excellent introduction to Walsh analysis with many examples and numerous comparisons between Walsh-based and Fourier-based spectral analysis. As a point of interest, Beauchamp (1984, sec. 3.5.2) presents two examples of synthesizing waveforms, that is, reconstructing an entire time series from a few sinusoids or Walsh functions in the sense of (3.1) and (3.6). In the first example he shows that in reconstructing a simulated continuous (smooth) seismic waveform [say $X(t)$], approximately twice as many Walsh terms are required to give about the same accuracy as may be obtained in the Fourier case [the number of terms corresponds to q in equations (3.1) and (3.6)]. The second example shows the reconstruction of a rectangular waveform in which there is efficient reconstruction with considerably fewer Walsh terms than Fourier terms. Although the results of these examples may be obvious, they reinforce the idea that analysis of smooth, continuous types of waveforms favors Fourier analysis, whereas a rectangular or discontinuous waveform favors Walsh analysis.

In addition to Beauchamp (1975, 1984), readers interested in the engineering perspective of Walsh–Fourier spectral analysis could refer to Ahmed and Rao (1975), Harmuth (1969b, 1977), Maqusi (1981), Tzafestas (1985), or the *Proceedings of the Symposia on the Applications of Walsh Functions,* to mention a few. Be forewarned that in examining the literature concerning this subject one must keep in mind that there are two modes of development (one based on the real clock and one based on the dyadic clock), even though the particular mode is not always apparent. This can be quite confusing since the results are considerably different, and results from one mode do not typically apply to the other.

4. WALSH FUNCTIONS AND THEIR GENERATION

There are a number of formal definitions of Walsh functions (e.g. Ahmed and Rao 1975; Beauchamp 1975, 1984; Kohn 1980a; Morettin 1981; or Ott and Kronmal 1976), but none give much insight. The Walsh functions form a complete orthonormal sequence on $[0, 1)$—roughly, this means that almost any waveform can be uniquely synthesized to any desired degree of accuracy by a linear combination of Walsh functions—and take only two values $+1$ and -1. Although the various definitions of the Walsh functions lead to different orderings, I will be interested primarily in the Walsh or sequency ordering since this ordering is comparable to the frequency ordering of sines and cosines. The sequency-ordered Walsh functions will be denoted by $W(n, \lambda)$, where $n = 0, 1, 2, \ldots; 0 \le \lambda < 1$. The first eight continuous sequency-ordered Walsh functions, $W(n, \lambda)$ $(n = 0, 1, \ldots, 7)$, were displayed in Figure 7. On further inspection of Figure 7, one notes that $W(0, \lambda)$ always takes the value $+1$, and makes no sign changes or zero crossings; the first Walsh function $W(1, \lambda)$, makes one sign switch from $+1$ to -1 as λ varies over its range $(0 \le \lambda < 1)$, the second Walsh function $W(2, \lambda)$ makes two sign switches, from $+1$ to -1, and then from -1 to $+1$, and so on until the seventh Walsh function $W(7, \lambda)$, which oscillates the fastest making seven sign switches. The corresponding eight discrete sequency-ordered Walsh functions, $W(n, m/N)$ $(n, m = 0, 1, \ldots, 7)$, corresponding to a sample of length $N = 2^3$ were shown in Section 3 as the rows (or columns) of the symmetric Walsh-ordered Hadamard matrix $\mathbf{H}_W(3)$. In the statistical analysis of time series using Walsh functions, we shall be interested primarily in the discrete Walsh functions.

The Hadamard matrix can be generated recursively as follows: Begin with $H(0) = +1$, and then let

$$\mathbf{H}(k+1) = \begin{bmatrix} \mathbf{H}(k) & \mathbf{H}(k) \\ \mathbf{H}(k) & -\mathbf{H}(k) \end{bmatrix}, \qquad k = 0, 1, 2, \ldots,$$

so, for example,

$$\mathbf{H}(1) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{H}(2) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

and

$$\mathbf{H}(3) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

The Hadamard matrix contains the Walsh functions as rows (or columns, noting the symmetry) in what is called natural or Hadamard ordering. For example, if we write the sequency-ordered or Walsh-ordered Hadamard matrix $\mathbf{H}_W(3)$ in terms of its column vectors, $\mathbf{H}_W(3) = [\mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_4, \mathbf{h}_5, \mathbf{h}_6, \mathbf{h}_7]$, where $\mathbf{h}_i$ is the ith sequency-ordered Walsh function, then the natural-ordered Hadamard matrix is $\mathbf{H}(3) = [\mathbf{h}_0, \mathbf{h}_7, \mathbf{h}_3, \mathbf{h}_4, \mathbf{h}_1, \mathbf{h}_6, \mathbf{h}_2, \mathbf{h}_5]$. In general, for $N = 2^p$ where p is a positive integer, the sequency-ordered Walsh functions can be obtained from the $N \times N$ Hadamard matrix $\mathbf{H}(p)$ by a bit-reversal Gray code (Ahmed and Rao 1975, pp. 92–95), which essentially counts the number of sign changes in each row or column of the $\mathbf{H}(p)$ and then reorders the rows or columns to obtain $\mathbf{H}_W(p)$. As shown in Good (1958) and others, it is possible to simplify the task of computing $\mathbf{H}(p)$, and in turn $\mathbf{H}_W(p)$, through a matrix factorization technique which reveals a large amount of redundancy in the repeated calculations implied by the aforementioned recursive generation technique. This is the essence of the *fast* Walsh–Fourier transform that will be discussed briefly in the next section.

The concept of dyadic addition must be introduced to describe some of the properties of the Walsh functions. If m and n are nonnegative integers, their binary expansions are

$$m = \sum_{j=0}^{q} m_j 2^j \quad \text{and} \quad n = \sum_{j=0}^{q} n_j 2^j,$$

where m_j and n_j are either 0 or 1. The *dyadic sum*, $m \oplus n$ (also called bitwise-exclusive-or addition and bit-by-bit mod 2 addition), of m and n is

$$m \oplus n = \sum_{j=0}^{q} |m_j - n_j| 2^j;$$

note that $1 \oplus 1 = 0 = 0 \oplus 0$, and $1 \oplus 0 = 1 = 0 \oplus 1$. For example, using binary expansion, $5 = 101$, $3 = 011$, so that $5 \oplus 3 = 101 \oplus 011 = 110 = 6$; this is why in dyadic time, 3 hours from 5 p.m. is 6 p.m.

Discrete Walsh functions $W(n, m/N)$ ($n, m = 0, 1, 2, \ldots, N-1$), have the following properties:

$$W(n_1, m/N)W(n_2, m/N) = W(n_1 \oplus n_2, m/N), \quad (4.1)$$

$$W(n, m_1/N)W(n, m_2/N) = W[n, (m_1 \oplus m_2)/N], \quad (4.2)$$

$$N^{-1} \sum_{n=0}^{N-1} W\left(n, \frac{m}{N}\right) = N^{-1} \sum_{m=0}^{N-1} W\left(n, \frac{m}{N}\right) = 0$$

$$\text{for } m, n \neq 0 = 1 \quad \text{for } m \text{ or } n = 0 \quad (4.3)$$

for $N = 2^p$ and $0 \leq n, m \leq N - 1$.

Relationships (4.1) and (4.2) are the Walsh analogies of the multiplicative/additive relationship of the trigonometric functions, for example, $2\cos(\alpha t)\cos(\beta t) = \cos(\alpha - \beta)t + \cos(\alpha + \beta)t$. These results reveal the connection between the Walsh functions and *dyadic-time stationarity,* and the trigonometric functions and *real-time stationarity*. That is, if (3.1) is true, then by using the multiplicative/additive relationships of sines and cosines, one can show that a time series generated by (3.1) has the properties $E\{X(t)\} = 0$ and $\text{cov}\{X(t + h), X(t)\} = \Sigma_{j=1}^{q} \sigma_j^2 \cos(2\pi\lambda_j h)$, independent of the time t, and hence the time series is real-time stationary (for example, see Shumway 1988, sec. 2.2). If (3.6) is true, then by using results (4.1) and (4.2) one can show that a time series generated by (3.6) has the properties $E\{X(t)\} = 0$ and $\text{cov}\{X(t \oplus h), X(t)\} = \Sigma_{j=1}^{q} \sigma_j^2 W(h, \lambda_j)$. In this case the time series is not real-time stationary since the covariance function is not a function of the lag $(t \oplus h) - t$; however, the time series $X(t)$ is dyadically stationary since the covariance function depends only on h, the dyadic difference between $t \oplus h$ and t, (see Morettin 1981 for details). This natural relationship between the Walsh functions and dyadic addition is the reason that most of the Walsh literature focused on the assumption of dyadic stationarity. Clearly, dyadic time has theoretical appeal in the Walsh domain; nevertheless, due to its strange behavior, dyadic time is of little practical use. Henceforth, the term stationary will refer only to real-time stationarity.

5. THE WALSH–FOURIER TRANSFORM AND SPECTRUM

The Walsh–Fourier transform of $X(0), X(1), \ldots, X(N-1)$, is

$$d_N(\lambda) = N^{-1/2} \sum_{t=0}^{N-1} X(t)W(t, \lambda), \qquad 0 \leq \lambda < 1, \quad (5.1)$$

in analogy to the cosine and sine transforms given in (3.2) and (3.3). This transform is essentially the correlation of the data with the Walsh functions. Assuming $X(t)$ is stationary, with autocovariance function $\gamma(h) = \text{cov}\{X(t + h), X(t)\}$, the variance of $d_N(\lambda)$ is

$$\text{var}\{d_N(\lambda)\} = \sum_{j=0}^{N-1} \tau(j)W(j, \lambda), \quad (5.2)$$

where $\tau(j)$ is the *logical covariance function* (Robinson 1972a,b; Kohn 1980a) of $X(t)$ given by

$$\tau(j) = 2^{-q} \sum_{k=0}^{2^q - 1} \gamma(j \oplus k - k), \qquad 2^q \leq j < 2^{q+1}.$$

If the autocovariance function of $X(t)$ is absolutely summable [that is, $\Sigma_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, which roughly means that observations taken far apart in time are nearly uncorrelated], then $\text{var}[d_N(\lambda)] \to f(\lambda)$ as $N \to \infty$ where

$$f(\lambda) = \sum_{j=0}^{\infty} \tau(j)W(j, \lambda), \qquad 0 \leq \lambda < 1, \quad (5.3)$$

is the *Walsh–Fourier spectrum* (or *spectral density*) of $X(t)$.

If the length of the sample is a power of two, say $N = 2^p$, then the transform (5.1) may be calculated for $\lambda_N = m/N$ ($m = 0, 1, 2, \ldots, N - 1$) using a fast Walsh–Fourier transform. The fast transform is computed from the matrix product $\mathbf{H}_W(p)\mathbf{X}$, where $\mathbf{X}$ is the $N \times 1$ vector of observations, $\mathbf{X} = (X(0), X(1), \ldots, X(N-1))'$. This calculation can be done quickly, with only a relatively small number of additions and subtractions of the data, using the matrix redundancy techniques discussed briefly in Section 3 (Ahmed and Rao 1975, chap. 6). If the sample length N is not a power of 2, dummy observations equal to the sample mean may be appended to the series in order to make the new sample length a power of 2. Similarly, observations may be deleted to obtain a new sample length that is a power of 2. Provided that not too many observations are appended or deleted, the transform will not be noticeably different than that of the original data, although the sequencies will be slightly different. That is, the new sequencies will be $\lambda' = m/N'$ where $N' = 2^p$ is the new sample length, whereas the sequencies associated with the original series are of the form $\lambda = m/N$. These considerations are similar for the fast Fourier transform (Bloomfield 1976, pp. 72–74; Shumway 1988, pp. 66–67). It is also possible to calculate the Walsh–Fourier transform directly from the definition of the Walsh functions—in this case, the computation of the transform could be slow.

If $N = 2^p$ and the mean, μ, of the series is unknown, the zero ($m = 0$) sequency is the only sequency of the form $\lambda_N = m/N$ for which the transform of the mean-centered series $\{X(t) - \mu\}$ cannot be evaluated since, by (4.3),

$$N^{-1/2} \sum_{t=0}^{N-1} \mu W(t, m/N) = N^{1/2}\mu \qquad m = 0$$
$$= 0 \qquad m \neq 0.$$

Thus, a mean-centered transform [that is, the transform of $X(t) - \mu$, or of $X(t) - \bar{X}$, where $\bar{X}$ denotes the sample mean] will be the uncentered transform except at the 0 sequency.

Various results relating the convergence of $d_N(\lambda)$ to $f(\lambda)$ exist. For example, Kohn (1980a) showed that if λ_N is dyadically rational (that is, its binary representation is finite) and $\lambda_N \oplus \lambda \to 0$ as $N \to \infty$ then $E\{d_N^2(\lambda_N)\} \to f(\lambda)$ as $N \to \infty$. Although the asymptotic ($N \to \infty$) covariance of the Walsh–Fourier transform at two distinct sequencies is not zero in general (this is in contrast to the trigonometric case, wherein the Fourier transforms of the data at two distinct frequencies are, under mild conditions, asymptotically independent), Kohn (1980a) shows that if $\lambda_{1,N}$ and $\lambda_{2,N}$ are dyadically rational, $|\lambda_{1,N} - \lambda_{2,N}| \geq N^{-1}$ and $\lambda_{j,N} \oplus \lambda \to 0$ ($j = 1, 2$) as $N \to \infty$, then $E\{d_N(\lambda_{1,N})d_N(\lambda_{2,N})\} \to 0$ as $N \to \infty$. These results are useful in obtaining consistent estimators of $f(\lambda)$. Kohn (1980a,b), Morettin (1973, 1974a, 1983), and Stoffer (1985, 1987, 1990) established central limit theorems (under the assumption of stationarity) for the Walsh–Fourier transform under a wide range of conditions. The basic result is that, under appropriate conditions (that typically include some type of *mixing condition*—loosely, events that occur far apart in time are nearly independent), $d_N(\lambda_N) \xrightarrow{D} N(0, f(\lambda))$ as $N \to \infty$; that is, the large sample distribution of the transform $d_N(\lambda_N)$ is Gaussian with mean zero and variance $f(\lambda)$ given in (5.3). Under these same conditions and using the aforementioned results, if $\{\lambda_{1,N}, \ldots, \lambda_{M,N}\}$ is a collection of M sequencies that are all chosen close to a sequency of interest, λ, such that $|\lambda_{j,N} - \lambda_{k,N}| \geq N^{-1}$ for $j \neq k$, and $\lambda_{j,N} \oplus \lambda \to 0$ for $j = 1, \ldots, M$, then as $N \to \infty$,

$$\sum_{j=1}^{M} d_N^2(\lambda_{j,N}) \xrightarrow{D} f(\lambda)\chi_M^2 \tag{5.4}$$

where χ_M^2 denotes a chi-squared distribution with M degrees of freedom. From this we deduce that $M^{-1}\Sigma_{j=1}^{M} d_N^2(\lambda_{i,N})$ is an estimate of $f(\lambda)$ having variance $2f^2(\lambda)/M$. If we let $M \to \infty$ as $N \to \infty$ with $M/N \to 0$, then the estimate is a mean square consistent estimate of $f(\lambda)$ ($0 < \lambda < 1$). Note that $d_N^2(\lambda_{j,N})$ is the Walsh periodogram of the data at sequency $\lambda_{j,N}$, as discussed in (3.5); hence, *a consistent estimate of the Walsh–Fourier spectrum, $f(\lambda)$, is simply the average of the Walsh periodogram at sequencies near the sequency of interest.*

A useful measure of the degree of association (at sequency λ) between two time series $X_A(t)$ and $X_B(t)$ is *coherency,*

$$K_{AB}(\lambda) = f_{AB}(\lambda)/[f_{AA}(\lambda)f_{BB}(\lambda)]^{1/2},$$

where $f_{AA}(\lambda)$ and $f_{BB}(\lambda)$ are the Walsh–Fourier spectra of series A and B, respectively, and $f_{AB}(\lambda)$ is the *cross-spectrum* of the two series. The cross-spectrum is related to the covariance of the Walsh–Fourier transforms of series A and B; it is a sequency-dependent measure of covariance between the series in much the same way that the Walsh–Fourier spectrum is a sequency-dependent measure of variance. Thus, coherency is a sequency-dependent measure of correlation and analogous to the usual correlation inequality, $-1 \leq K_{AB}(\lambda) \leq 1$. In the trigonometric (Fourier) case, cross-spectra are complex-valued and hence squared-coherency rather than coherency is measured; this is one advantage of working in the real-valued Walsh–Fourier domain, as we shall see in the application presented in Section 6.3. For an introduction to cross-spectral analysis in the trigonometric case, see Bloomfield (1976, chap. 9) or Shumway (1988, chap. 2). For a detailed account of cross-spectral Walsh–Fourier analysis and its applications, see Stoffer (1990).

6. STATISTICAL APPLICATIONS

6.1 A Comparison of Fourier and Walsh–Fourier Data Analysis

To examine some of the differences between Fourier and Walsh–Fourier analysis, the EEG sleep state data shown in Figure 4 is analyzed using the fast Fourier transform (FFT) and the fast Walsh–Fourier transform (FWT); see Ahmed and Rao 1975, chapters 4 and 6, for example. In order to make the two techniques comparable, Harmuth's definition of sequency is used. The discrete Walsh function $W(n, \lambda)$ makes n zero crossings (or sign switches) per unit time, and thus its corresponding sequency value is n; the Harmuth-sequency of that function is $n/2$ if n is even, and $(n + 1)/2$ if n is odd (recall that Harmuth-sequency and frequency coincide for sinusoids). To differentiate between the two definitions of sequency, I will denote Harmuth-sequency by H-sequency. Using this definition one will note that two Walsh functions have the same H-sequency, that is, $W(2n, \lambda)$ and $W(2n - 1, \lambda)$ ($n = 1, 2, \ldots$) each have an H-sequency of n; this is also true for the sinusoids, that is, $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$ each have a frequency of n. Thus, in this approach, the Walsh–Fourier periodogram given in (3.5) is modified to look like the Fourier periodogram (3.4) by setting $I_H(\lambda_j) = I_W(2j - 1) + I_W(2j)$ [$j = 1, 2, \ldots, (N - 2)/2$)], where λ_j represents H-sequency.

The duration of the sleep study exhibited in Figure 4 was 116 minutes, and hence to utilize the FFT and the FWT, the data were padded to $N = 128 = 2^7$. The Fourier periodogram and the Walsh periodogram are compared in Figure 8, but since $\Sigma_j I_H(\lambda_j) = 2\ \Sigma_j I_F(\lambda_j)$, excluding $\lambda_j = 0$ and $1/2$, $I_H(\lambda_j)/2$ and $I_F(\lambda_j)$ are superimposed to facilitate the comparison. Moreover, the values of the periodograms are an order of magnitude larger in the slower H-sequency/frequency range than in the moderate and fast ranges, and hence, in order to get the best visual comparison, Figure 8 shows $10 \log_{10}\{I_H(\lambda_j)/2 + 1\}$ and $10 \log_{10}\{I_F(\lambda_j) + 1\}$, with $\lambda_j = j/N$, for $j = 1, \ldots, 7$, and $I_H(\lambda_j)/2$ and $I_F(\lambda_j)$ for $j = 8, \ldots, 63$.

The two periodograms are very similar, and both suggest a strong peak in the neighborhood of $\lambda = 3/128$. In the frequency domain, λ is measured in cycles per minute; this would mean that the dominant harmonic component in the data is about one cycle every 45 minutes ($128/3 = 43$ minutes per cycle). In the H-sequency domain, Harmuth (1977, p. 9) would interpret this as 45 minutes being the dominant "*average* period of oscillation" [that is, the average separation in time of the zero (mean value) crossings multiplied by two]. In both cases, we are led to the same conclusion that this normal neonate alternates between non-REM and REM sleep in approximately 45 minute intervals. This component of normal infant sleep is well documented (Hauri 1982).

The most striking difference between the two periodograms is the peak in the Walsh periodogram at $\lambda = 14/128$ that is not present in the Fourier periodogram; this H-sequency corresponds to an average period of about 9 minutes (128/14). To explain this, note that the average length of time that the infant stays in any one state is approximately 4.52 minutes (for example, in Figure 4, the infant is in State 5 for nine minutes, State 3 for one minute, State 5 for one minute, etc.). Hence, the average separation in time of the state changes multiplied by two (which is Harmuth's definition of *generalized period*) is nine minutes. This behavior was consistent enough throughout the record of this infant to be sufficiently correlated with the Walsh functions corresponding to an average period of nine minutes. This component, of course, was missed in the frequency domain.

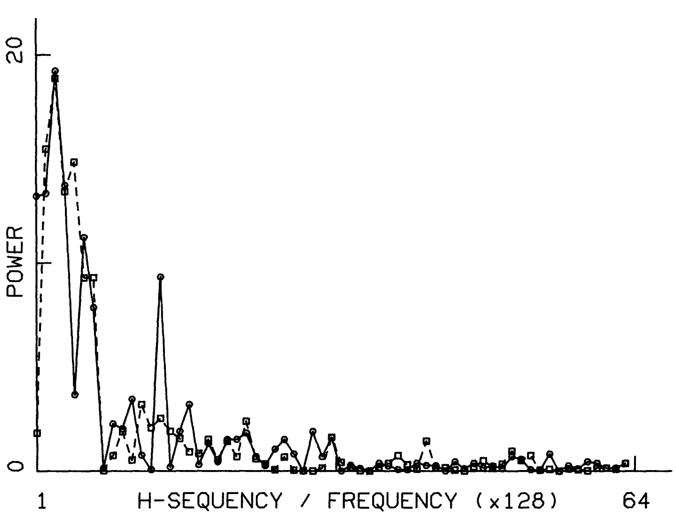

Figure 8. Comparison of the Walsh Periodogram (Solid Line–Circles) and the Fourier Periodogram (Dashed Line–Squares) for the Sleep-State Data Shown in Figure 4.

It is interesting to note that in an analysis of the effects of maternal alcohol consumption on the EEG sleep-state cycling of neonates (Stoffer et al. 1988), the nine minute generalized period was present in the unexposed infants but was absent in the exposed infants. This will be explained in more detail in the next section.

In the sections that follow, I will use sequency rather than H-sequency. Sequency will be measured in switches per unit time, and generalized period will be measured in terms of the average length of time between switches. It is easy to convert back and forth between the two measures since H-sequency is approximately one half of sequency. For example, in Figure 8 the strong peak at H-sequency 3/128 would occur at a sequency of either 5/128 or 6/128 switches per minute, and the peak at H-sequency 14/128 would occur at a sequency of either 27/128 or 28/128 switches per minute. The actual peak sequency values were 5/128 and 27/128, and hence, in the sequency domain, we would conclude that the average length of time per sleep-state change is approximately 25.6 (128/5) minutes with a secondary component of 4.74 (128/27) minutes.

6.2 A Walsh–Fourier Analysis of Variance

Data were collected for a study of the effects of moderate maternal alcohol consumption on neonatal EEG sleep patterns. A detailed description of the study design, the methods used for measuring alcohol use, and the scoring of neonatal EEG sleep records can be found in Day, Wagener, and Taylor (1985) and Scher, Richardson, Coble, Day, and Stoffer (1988). Briefly, an EEG sleep recording of approximately two hours duration is obtained on a full-term infant 24 to 36 hours after birth, and recordings are scored—by an electroencephalographer who is not aware of the prenatal substance exposure of the infant—for EEG sleep-state, REM's, arousals, and body movements using scoring epochs of 1 minute. Sleep-state is categorized (per minute) into one of six possible states: State 1 quiet sleep—trace alternant; State 2 quiet sleep—high voltage; State 3 indeterminate sleep; State 4 active sleep—low voltage; State 5 active sleep—mixed; State 6 awake, as described in the introduction.

Two groups of infants were compared. The first group contained 12 intants whose mothers abstained from using alcohol throughout pregnancy, while the second group contained 12 infants whose mothers used alcohol throughout pregnancy on a regular basis at a rate of at least .5 drinks per day. The actual sleep-state records (that is, per minute EEG sleep-state for 120 minutes for each infant in the study) can be found in Stoffer et al. (1988). The sleep-state record of a typical infant from the abstainer group (ID 465) is exhibited in Figure 4, and a similar plot for a typical infant in the exposed group (ID 223) is displayed in Figure 9. Recall that the corresponding total numbers of body movements per minute (discounting sucking) for each infant are

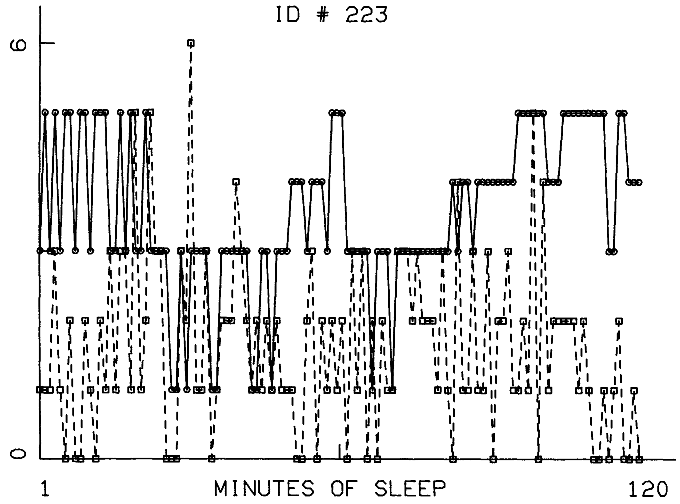

Figure 9. Sleep-State (Solid Line–Circles) and Total Number of Body Movements (Dashed Line–Squares) of the Exposed Infant.

also shown; they will be used in the next section in an analysis of coherency.

The problem of detecting whether a common sleep pattern existed among the 12 infants in each group was addressed using a nonparametric signal-plus-noise model for the sleep-state data in each group. Let $X_q(t)$ designate the sleep state of infant q at time t ($q = 1, \ldots, 12$; $t = 0, \ldots, 127$); the sleep studies were approximately 120 minutes long—to use the FWT given in Ahmed and Rao (1975), the data were padded to $N = 128 = 2^7$ minutes as described in Sections 4 and 6.1. Each sleep-state record in the group is decomposed into the sum of three components—an individual mean value θ_q, a common stochastic and stationary sleep pattern $S(t)$, and stationary error $\epsilon_q(t)$,

$$X_q(t) = \theta_q + S(t) + \epsilon_q(t). \tag{6.1}$$

$S(t)$ has mean zero and Walsh–Fourier spectrum $f_S(\lambda)$, and $\epsilon_q(t)$, uncorrelated with $S(t)$, has mean zero and Walsh–Fourier spectrum $f_\epsilon(\lambda)$. Testing whether there was a common sleep pattern among individuals in the group meant testing the null hypothesis H_0: $f_S(\lambda) = 0$ $(0 < \lambda < 1)$. To perform the test, the Walsh–Fourier transform was computed for each infant q in the group. Stoffer (1987) showed that, under appropriate conditions, the Walsh–Fourier transform of $X_q(t)$, $d_{N,q}(m/N)$, can be represented approximately as the sum of two independent random components, say,

$$d_{N,q}(m/N) \simeq U(m) + Z_q(m), \tag{6.2}$$

where $U(m)$ is normally distributed with mean zero and variance $f_S(m/N)$, and the $Z_q(m)$ are independent and normally distributed with mean zero and variance $f_\epsilon(m/N)$; moreover, $U(m)$ and $Z_q(m)$ are independent. Thus the problem of testing H_0: $f_S(\lambda) = 0$ is reduced to a random effects analysis of variance (ANOVA) problem where the data are the Walsh–Fourier transforms of the sleep-state time series.

Let $d_{N,\cdot}(m/N) = [\sum_{q=1}^{12} d_{N,q}(m/N)]/12$ be the average transform of the group. Then following Scheffé (1959, pp. 225–227), the hypothesis mean-square and the error mean-square at any particular sequency $\lambda = m/N$ were computed as $\mathrm{MSH}(m) = 12\, d^2_{N,\cdot}(m/N)$ and $\mathrm{MSE}(m) = \sum_{q=1}^{12}[d_{N,q}(m/N) - d_{N,\cdot}(m/N)]^2/11$, respectively. The test of H_0 was performed at sequencies of the form $\lambda = m/N$ by comparing the ratios

$$F(m) = \mathrm{MSH}(m)/\mathrm{MSE}(m) \tag{6.3}$$

with the F distribution with 1 and 11 df. See Stoffer (1987) for further details. Also, refer to the Brillinger (1980) for analysis of variance problems under continuous-valued time series models.

Figure 10 shows a plot of the square root of $F(m)$ for each group at all sequencies except the zero sequency ($m = 0$) and indicates the null significance levels of .01 and .001 (for ease of comparison, the square root of the F ratios are shown in Figure 10). Although it is difficult to get precise information from Figure 10, a table was provided in Stoffer et al. (1988, tab. 1) to facilitate the examination of the spectral components. As an example, for the abstainer group, the largest peak is $F(29) = 62.1$. In terms of generalized periodicity, this means that the component corresponding to 4.4 (128/29) minutes per state change, on the average, was a strong common component of sleep for the unexposed group. This is consistent with the individual analysis presented in Section 6.1. As previously mentioned, this component is absent from the exposed group of infants. From Figure 10 we concluded that a common sleep pattern does exist in the abstainer group, however, since there were only a few significant sequencies associated with the exposed group, we were hesitant to conclude that a common sleep pattern existed in the exposed group.

To extend the preceding analysis to comparisons between the two groups of infants (see Stoffer et al. 1988 for details), add a subscript for group membership and test the null hypothesis H_0: $f_{S_1}(\lambda) = f_{S_2}(\lambda)$ $(0 < \lambda < 1)$. Let

$$d_{N,\cdot\cdot}(m/N) = [d_{N,\cdot 1}(m/N) + d_{N,\cdot 2}(m/N)]/2$$

be the grand average transform and define the hypothesis mean-square and the error mean-square as

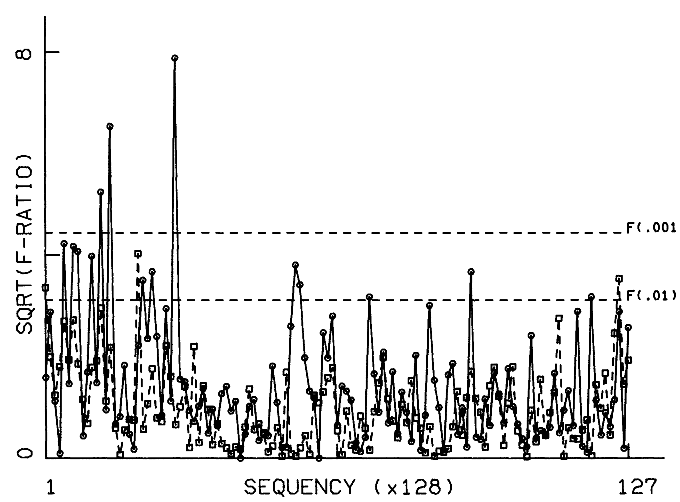

Figure 10. F Ratio (6.3) for Testing for a Common Signal in the Unexposed Group of Infants (Solid Line–Circles) and in the Exposed Group of Infants (Dashed Line–Squares).

$$\mathrm{MSH}(m) = 12 \sum_{k=1}^{2} [d_{N,\cdot k}(m/N) - d_{N,\cdot\cdot}(m/N)]^2$$

and

$$\mathrm{MSE}(m) = \sum_{k=1}^{2} \sum_{q=1}^{12} [d_{N,qk}(m/N) - d_{N,\cdot k}(m/N)]^2/2(12 - 1),$$

respectively. The statistic $F(m)$ was computed as in (6.3) and the hypothesis was tested by comparing $F(m)$ with the F distribution with 1 and 22 df. Figure 11 shows a plot of $F(m)$, in this case at all sequencies except the zero sequency, and indicates the null significance thresholds of .05 and .01. Again, refer to Stoffer et al. (1988, tab. 1 and fig. 5) for detailed information about Figure 11. Figure 11 also includes another plot for comparison—this other plot will be explained in Section 6.4.

Figure 11 shows significant differences between the groups. For example, for the sequency of 29/128, which corresponds to the generalized period of 4.4 (128/29) minutes per sleep state switch, $F(29) = 7.05$, exceeding the .02 significance threshold. Hence we conclude that (among other components) the generalized period of approximately 4.4 minutes per sleep state change is a component in the sleep architecture of normal neonates, and furthermore, this component is absent in neonates that are exposed to alcohol. From this analysis, the neurologist suggested that differences between the ultradian sleep cycles in the two groups may reflect differences in the central nervous system maturation. In particular, these differences indicated an alteration in the development and expression of diverse neurotransmitter sections within the brain stem and forebrain connections that are associated with the expression of active and quiet sleep.

The tests based on $F(m)$ are only appropriate at one predetermined sequency. Since these tests were used for an exploratory analysis, special consideration was given to the overall error rate. It has been suggested (Shumway 1988, p. 70) that the way to overcome the problem of the overall error rate involved in making simultaneous statements about

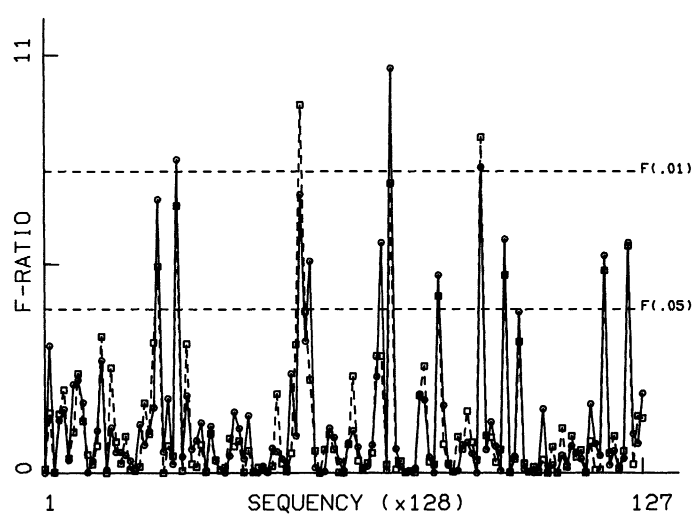

Figure 11. F Ratio for the Between-Group Comparison of the Unexposed and Exposed Groups of Infants Using the Techniques of Section 6.2 (Dashed Line–Squares) and the Scaling Techniques of Section 6.4 (Solid Line–Circles).

the value of the spectrum is to use a Bonferroni correction. In the context of the EEG sleep-state analysis, a bound on the overall error rate of the tests is the sum of the individual α levels at each sequency value. Thus in the investigation of a common group sleep pattern, we used the levels of .001 and .01 to indicate the strong and moderate spectral components. Subsequently, in the between-group analysis we were able to focus on a few particular sequencies of interest [although all the values of $F(m)$ are evaluated for computational and graphical convenience]; in this case we used the level of .01 to indicate group differences, however, we felt that since this was an exploratory analysis, it was appropriate to indicate crossings of the .05-level threshold.

6.3 A Walsh–Fourier Analysis of Coherency

The analysis presented here is from a continuation of the study of the effects of moderate alcohol exposure on neonatal cerebral and central nervous system maturation as assessed through EEG sleep studies and involved the 24 infants discussed in the previous section. Details can be found in Stoffer (1990). In particular, the coherency between the per minute EEG sleep-state time series and the corresponding per minute total number of body movements time series was used in detecting disturbances in neonatal sleep cycling as a result of alcohol exposure. Recall that coherency is a sequency dependent measure of correlation between the two time series. Figures 4 and 9 present typical data plots from the unexposed group and from the exposed group of infants, respectively. I will concentrate on the analysis of these infants before presenting the group analysis.

Let $d_A(\lambda)$ and $d_B(\lambda)$ denote the Walsh–Fourier transforms of the per minute EEG sleep state and the corresponding total number of body movements of an infant, respectively. The *sample coherency* between sleep state and number of body movements is

$$\hat{K}_{AB}(\lambda) = \hat{f}_{AB}(\lambda)/[\hat{f}_{AA}(\lambda)\hat{f}_{BB}(\lambda)]^{1/2}, \quad (6.4)$$

where $\hat{f}_{AB}(\lambda)$ is obtained by averaging the values of the product $d_A(m/N)d_B(m/N)$ over values of m in a neighborhood of λ. Similarly, $\hat{f}_{AA}(\lambda)$ and $\hat{f}_{BB}(\lambda)$ are obtained by averaging the Walsh periodograms $d_A^2(m/N)$ and $d_B^2(m/N)$ over m in a neighborhood of λ. These ideas were previously discussed in Section 5, after equation (5.4). As in the case of ordinary sample correlation, $-1 \le \hat{K}_{AB}(\lambda) \le 1$, with values close to -1 or $+1$ corresponding to strong association.

The coherencies of the unexposed infant and of the exposed infant are compared in Figure 12. Since the sample Walsh–Fourier cross-spectrum $\hat{f}_{AB}(\lambda)$ is real, unlike the Fourier coherency (which is absolute correlation relative to frequency), the Walsh–Fourier coherency can be negative; this was a considerable advantage over the trigonometric case in this analysis. It was evident from Figure 12 that there are small isolated ranges of sequencies at which the coherency between the sleep-state data and the body-movement data for the unexposed and the exposed infants differ markedly in sign, most notably at the very fast or high end of the sequency range: 120 to 127 switches per 128 time points. Here, the coherency between the two time series is

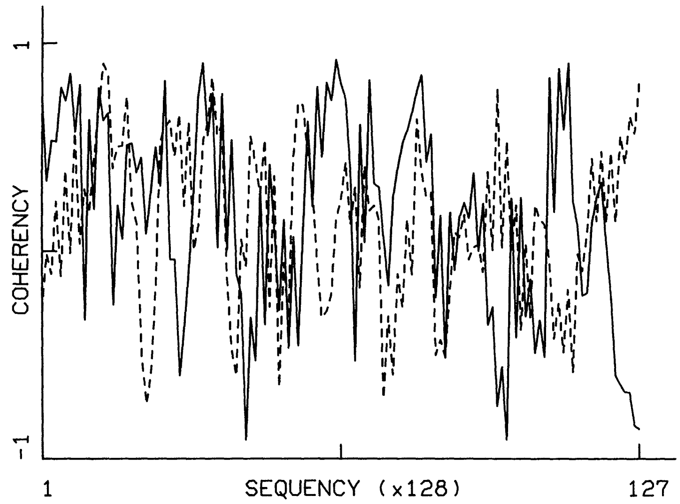

Figure 12. Coherency Between Sleep State and Body Movements for the Unexposed Infant (Solid Line) and the Exposed Infant (Dashed Line).

negative for the unexposed infant and positive for the exposed infant. It is believed that this difference will be an aid in the identification of a disturbance in the sleep cycle due to alcohol exposure. This distinction might have been missed in a trigonometric analysis, where an investigator must work with absolute or squared coherency—a positive-valued measure of association.

Finally, the average coherency of the group of 12 infants whose mothers abstained from alcohol during pregnancy was compared to the group of 12 infants whose mothers used alcohol on a regular basis throughout pregnancy (these groups were described in the previous section). Although there were differences between the average coherency of the two groups at various isolated ranges of sequencies (that were consistent with the differences between ID 465 and ID 223 pictured in Figure 12), differences in the sequency range 120–127 (per 128) will be illustrated here. The average coherency, with its standard error, for each group of 12 infants and the normalized difference between the groups at sequencies 120–127 (per 128) are listed in Table 1. While the difference between the exposed group and the unexposed group did not remain as marked as the individual analysis, a trend prevailed. It appeared that, for the unexposed group, the sleep-state series and the body-movement series have zero coherency on average at the fast sequencies (although there was a tendency to be on the negative side of zero as in the case of ID 465), while the average coherency between the two time series for the exposed group was positive (which was consistent with the individual analysis of ID 223). Although the neurophysiological aspects of these results are difficult to explain at this time, we believe that these coherency results will act as markers in establishing EEG abnormalities.

Table 1. Comparison of the Average Coherency Between Sleep State and Number of Body Movements for a Group of 12 Infants Whose Mothers Used Alcohol Regularly During Pregnancy (Exposed) and a Group of 12 Infants Whose Mothers Abstained From Alcohol Throughout Pregnancy (Unexposed)

Sequency	*Exposed group mean (se)*	*Unexposed group mean (se)*	*Normalized difference*
120	.07 (.14)	−.10 (.13)	.91
121	.21 (.13)	−.04 (.11)	1.46
122	.17 (.13)	−.08 (.17)	1.16
123	.20 (.14)	.01 (.12)	1.08
124	.30 (.10)	−.03 (.15)	1.84*
125	.29 (.10)	−.00 (.12)	1.91*
126	.23 (.14)	.00 (.17)	.99
127	.29 (.12)	−.06 (.16)	1.77*

NOTE: *Exceeds .05-level critical value for a one-sided t test.

6.4 Scaling Categorical Time Series

The scaling of categorical-valued time series is in the same vein as scaling in the analysis of contingency tables and regression with qualitative variables (e.g., Breiman and Friedman 1985; Greenacre 1984; Nishisato 1980). That is, in order to use quantitative statistical techniques on qualitative variables, scaling provides a method to optimally assign numerical values to the categories according to some specified reasonable criteria. For example, in Section 6.2, the six EEG sleep states were arbitrarily assigned the numerical values of 1 through 6 so that an analysis of variance could be performed on the data as if the numerical values were actually the outcomes. Although the results of Section 6.2 will not change if the data are transformed by a linear transformation, they are not invariant to even monotone nonlinear transformations of the values. Hence it is necessary to quantify the categories for use in spectral analysis according to reasonable optimization criteria.

Let $X^*(t)$ be the categorical-valued EEG sleep-state series of an infant in the study described in Section 6.2, and note that there are six possible states or categories. For example, the sleep-state series of infant ID 223 shown in Figure 9 is (indeterminate, mixed active, indeterminate, mixed active, ...). Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_6(t))'$ be a 6×1 binary-valued random vector such that

$$Y_j(t) = 1 \quad \text{if } X^*(t) \text{ is in state } j \text{ at minute } t$$
$$= 0 \quad \text{otherwise.}$$

From Figure 9, $\mathbf{Y}(0) = (0, 0, 1, 0, 0, 0)'$, $\mathbf{Y}(1) = (0, 0, 0, 0, 1, 0)'$, $\mathbf{Y}(2) = (0, 0, 1, 0, 0, 0)'$, $\mathbf{Y}(3) = (0, 0, 0, 0, 1, 0)'$, ... Applying the results of Stoffer (1990) for the Walsh–Fourier analysis of the vector observation $\mathbf{Y}(t)$, there is a consistent estimate $\hat{\mathbf{f}}_Y(\lambda)$ of the 6×6 spectral density matrix $\mathbf{f}_Y(\lambda)$ of $\mathbf{Y}(t)$, based on the 6×1 Walsh–Fourier transform of $\mathbf{Y}(t)$. This spectral matrix would be of some interest in the analysis of this data, however, by definition of the $\mathbf{Y}(t)$ time series, most coherencies [off diagonal elements of $\mathbf{f}_Y(\lambda)$] would be negative [since, for example, if one component of $\mathbf{Y}(t)$ increases from minute 1 to minute 2, one of the other components must decrease in that same minute].

Scaling consists of assigning numerical values (scales or scores) to the six categories of $X^*(t)$ according to some optimal criterion. This problem can be viewed as optimally selecting a 6×1 vector of real numbers, say $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)'$, to produce a numerical univariate time series $X(t) = \boldsymbol{\beta}'\mathbf{Y}(t) = \sum_{j=1}^{6} \beta_j Y_j(t)$ that allows quantitative analysis of the qualitative time series $X^*(t)$. Note that by definition of the $Y_j(t)$'s, if the categorical process is in state j at time t, then $X(t) = \beta_j$. The problem is how to select an appropriate criterion for determining the vector of scales $\boldsymbol{\beta}$.

I will first focus on the initial analysis of Section 6.2, that is, the detection of a common sleep signal in each group of infants, the alcohol exposed group and the unexposed group. As in Section 6.2, let $X_q^*(t)$ be the qualitative sleep state of infant q in the group ($q = 1, \ldots, 12$; $t = 0, 1, \ldots, 127$). Corresponding to each qualitative EEG sleep-state time series, $X_q^*(t)$, there is the quantitative 6×1 vector series $\mathbf{Y}_q(t)$, which we may model in an analogous manner to the model presented in Section 6.2: $\mathbf{Y}_q(t) = \boldsymbol{\theta}_q + \mathbf{S}(t) + \boldsymbol{\epsilon}_q(t)$, where $\mathbf{S}(t)$ represents the hypothesized common sleep pattern of the particular group with Walsh–Fourier spectrum $\mathbf{f}_S(\lambda)$ and $\boldsymbol{\epsilon}_q(t)$ is uncorrelated with $\mathbf{S}(t)$ and has Walsh–Fourier spectrum $\mathbf{f}_\epsilon(\lambda)$. The validity of models such as this for discrete-valued (in particular binary-valued for this discussion) time series can be found in Stoffer (1987). Under the model assumptions, the 6×6 spectral matrix of $\mathbf{Y}_q(t)$ is $\mathbf{f}_{Y_q}(\lambda) = \mathbf{f}_S(\lambda) + \mathbf{f}_\epsilon(\lambda)$ $(0 < \lambda < 1)$, and hence the spectrum of the univariate scaled time series $X_q(t) = \boldsymbol{\beta}'\mathbf{Y}_q(t)$ is $f_{X_q}(\lambda) = \boldsymbol{\beta}'\mathbf{f}_S(\lambda)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{f}_\epsilon(\lambda)\boldsymbol{\beta}$ $(0 < \lambda < 1)$; this result is similar to the usual variance result, $\operatorname{var}\{\boldsymbol{\beta}'\mathbf{Y}_q(t)\} = \boldsymbol{\beta}' \operatorname{var}\{\mathbf{Y}_q(t)\}\boldsymbol{\beta}$.

It can be shown that the noncentrality parameter of the test given in (6.3) of Section 6.2 is essentially the *signal-to-noise ratio* of the $X_q(t)$ process, namely, $\{\boldsymbol{\beta}'\mathbf{f}_S(\lambda)\boldsymbol{\beta}\}/\{\boldsymbol{\beta}'\mathbf{f}_\epsilon(\lambda)\boldsymbol{\beta}\}$; see Stoffer (1987) for details. Hence it is apparent that the score vector $\boldsymbol{\beta}$ should be chosen to maximize, in some sense, this signal-to-noise ratio. Of course, $\mathbf{f}_S(\lambda)$ and $\mathbf{f}_\epsilon(\lambda)$ are unknown and must be estimated; let $\hat{\mathbf{f}}_S(\lambda)$ and $\hat{\mathbf{f}}_\epsilon(\lambda)$ be consistent estimates of $\mathbf{f}_S(\lambda)$ and $\mathbf{f}_\epsilon(\lambda)$. The scaling problem, then, is to choose the score vector $\boldsymbol{\beta}$ to maximize $\{\boldsymbol{\beta}'\hat{\mathbf{f}}_S(\lambda)\boldsymbol{\beta}\}/\{\boldsymbol{\beta}'\hat{\mathbf{f}}_\epsilon(\lambda)\boldsymbol{\beta}\}$ in some sense. For example, one could choose $\boldsymbol{\beta}$ to maximize the signal-to-noise ratio at one particular sequency of interest, or to maximize an integrated signal-to-noise ratio, $\{\boldsymbol{\beta}'\bar{\mathbf{f}}_S\boldsymbol{\beta}\}/\{\boldsymbol{\beta}'\bar{\mathbf{f}}_\epsilon\boldsymbol{\beta}\}$, where

$$\bar{\mathbf{f}}_S = \sum_{m=1}^{N-1} \hat{\mathbf{f}}_S(m/N) \quad \text{and} \quad \bar{\mathbf{f}}_\epsilon = \sum_{m=1}^{N-1} \hat{\mathbf{f}}_\epsilon(m/N);$$

a weighted scheme could also be used, say

$$\bar{\mathbf{f}}_S = \sum_{m=1}^{N-1} \operatorname{tr}\left\{\hat{\mathbf{f}}_S\left(\frac{m}{N}\right)\right\}\hat{\mathbf{f}}_S\left(\frac{m}{N}\right)$$

and

$$\bar{\mathbf{f}}_\epsilon = \sum_{m=1}^{N-1} \operatorname{tr}\left\{\hat{\mathbf{f}}_\epsilon\left(\frac{m}{N}\right)\right\}\hat{\mathbf{f}}_\epsilon\left(\frac{m}{N}\right),$$

where tr denotes trace. In these cases the optimal score vector $\boldsymbol{\beta}$ can be obtained by the solution of a matrix eigenvalue problem (see Rao 1973, sec. 1).

In this particular problem, the categories associated with EEG sleep states are ordered, and hence the scales, or elements, of $\boldsymbol{\beta}$ should be ordered; for example we seek $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)'$ such that $\beta_1 \le \cdots \le \beta_6$. Nishisato (1980 chap. 8) discusses a few methodologies that could be adjusted to this problem. Techniques to obtain monotone transformations such as the one discussed in Kimeldorf, May, and Sampson (1982), or as is used in the ACE algorithm (Breiman and Friedman 1985), or suitable isotonic regression algorithms (Robertson, Wright, and Dykstra 1988) could also be adapted for this particular problem.

To apply these techniques to the EEG sleep-state data for the unexposed group and the exposed group of infants, the optimal score for each group was obtained as follows: (a) the 6×1 $\mathbf{Y}_q(t)$ vectors for each infant in the group were formed; (b) the Walsh–Fourier transform $\mathbf{d}_{N,q}(m/N)$ of the $\mathbf{Y}_q(t)$ data vectors were calculated using the fast transform; (c) the average transform was calculated, $\mathbf{d}_{N,\cdot}(m/N) = [\Sigma_{j=1}^{12}\mathbf{d}_{N,q}(m/N)]/12$; (d) the 6×6 error spectrum $\mathbf{f}_\epsilon(\lambda)$ was estimated by

$$\hat{\mathbf{f}}_\epsilon(m/N) = \sum_{q=1}^{12} [\mathbf{d}_{N,q}(m/N) - \mathbf{d}_{N,\cdot}(m/N)] \times [\mathbf{d}_{N,q}(m/N) - \mathbf{d}_{N,\cdot}(m/N)]'/11,$$

and a 6×6 hypothesis spectrum $\mathbf{f}_H(\lambda)$ was estimated by

$$\hat{\mathbf{f}}_H(m/N) = 12\, \mathbf{d}_{N,\cdot}(m/N)\mathbf{d}'_{N,\cdot}(m/N)$$

in analogy to the hypothesis mean-square defined in (6.3); the 6×6 signal spectrum $\mathbf{f}_S(\lambda)$ can be estimated by $\hat{\mathbf{f}}_S(m/N) = \hat{\mathbf{f}}_H(m/N) - [\hat{\mathbf{f}}_\epsilon(m/N)/12]$ as described in Stoffer (1987); (e) the integrated error spectral estimate was calculated as previously described, that is, $\bar{\mathbf{f}}_\epsilon = \Sigma_{m=1}^{N-1}\hat{\mathbf{f}}_\epsilon(m/N)$, however, the integrated signal spectrum $\bar{\mathbf{f}}_S$ was replaced by $\bar{\mathbf{f}}_H = \Sigma_{m=1}^{N-1}\hat{\mathbf{f}}_H(m/N)$; (f) the optimal score vector $\boldsymbol{\beta}$ was computed to maximize $\{\boldsymbol{\beta}'\bar{\mathbf{f}}_H\boldsymbol{\beta}\}/\{\boldsymbol{\beta}'\bar{\mathbf{f}}_\epsilon\boldsymbol{\beta}\}$ as the solution to a matrix eigenvalue problem. The reason for changing strategies from $\bar{\mathbf{f}}_S$ to $\bar{\mathbf{f}}_H$ was primarily based on a matrix conditioning problem.

The score vector obtained for the unexposed group was $\boldsymbol{\beta}_u = (8.60, 8.50, 8.41, 8.59, 8.43, 8.44)'$ and for the alcohol exposed group was $\boldsymbol{\beta}_e = (5.80, 5.81, 5.78, 5.86, 5.76, 6.13)'$; these scales correspond to the sleep-state labels in the order presented in Section 6.2; for example, in the unexposed group, the value 8.60 is given to quiet sleep—trace alternant and the value of 8.44 is given to the awake state. These vectors are unique up to a scaling constant, so for example, the largest scale value could also be the smallest scale value by multiplying the vector by -1. Since these optimal scores were not ordered (states 3, 4, and 5 are out of order in the unexposed group; states 3 and 5 are out of order in the exposed group) I ordered them by simply averaging neighbors to obtain the ordered score vector for the unexposed infants, $\boldsymbol{\beta}_{uo} = (8.60, 8.50, 8.49, 8.47, 8.45, 8.44)'$, where the scales for states 3, 4, 5 are computed as $[p \times 8.50 + (1 - p) \times 8.44]$ ($p = .75, .50, .25$), and the ordered score vector for the alcohol exposed infants, $\boldsymbol{\beta}_{eo} = (5.80, 5.81, 5.84, 5.86, 6.00, 6.13)'$, where 5.84 is the simple average of 5.81 and 5.86, and 6.00 is the simple average of 5.86 and 6.13.

The test described in (6.3), Section 6.2, for determining whether a group has a common sleep pattern was performed using the optimal and the ordered scales, however, the F ratios differed only slightly and the plots do not differ visually; hence only the ordered cases are presented here. Figure 13 shows the F ratios (as square roots) for the unex-

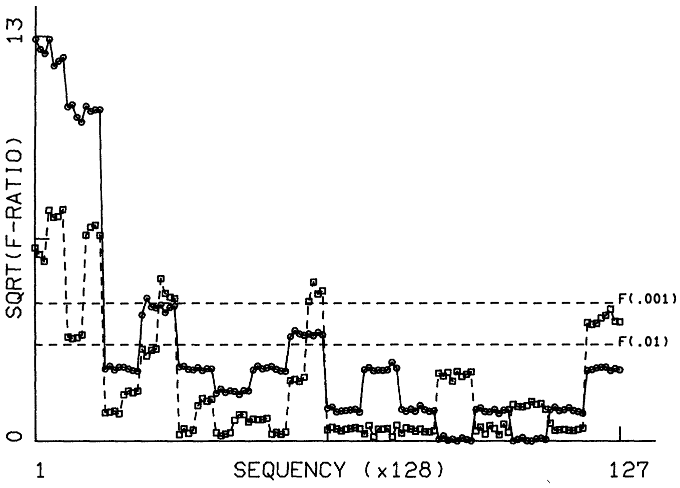

Figure 13. F Ratio (6.3) for Testing for a Common Signal in the Unexposed Group of Infants (Solid Line–Circles) and in the Exposed Group of Infants (Dashed Line–Squares) Using the Scaled Sleep States.

posed and the exposed groups of infants. First, note that the F ratios are considerably larger in general than they were in Section 6.2, and it is easier in this case to pick out the peak values. Second, notice that there is now strong evidence of a common signal in the exposed group. Third, although it is difficult to see in the figures, we are led to basically the same conclusions as to where the peak periods are in each group. For example, using these scaling results, the peaks for the unexposed group occur at the sequencies 1–15, 21–31, 56–63, with possible peaks in the ranges 72–79, and 120–127; these are approximately the same conclusions that were made previously in Section 6.2 and in Stoffer et al. (1988, tab. 1). For the exposed group of infants, these scaling results suggest peaks in the ranges 1–7, 11–14, 28–31, 60–63, and 120–127; this is fairly consistent with the previous (although not overwhelming) results, except for the inclusion of the 60–63 sequency range in this analysis. Finally, note that the two groups have peaks at roughly the same range of sequencies and hence it appears that the sleep behaviors of the two groups of infants have more in common than was previously believed.

The next step was to obtain optimal scales for the between-group comparison of the unexposed and exposed groups of infants. Following the steps (a)–(f) outlined above, with the 6×6 hypothesis and error spectral matrices defined as

$$\hat{\mathbf{f}}_H(m/N) = 12 \sum_{k=1}^{2} [\mathbf{d}_{N,\cdot k}(M/N) - \mathbf{d}_{N,\cdot\cdot}(m/N)] \times [\mathbf{d}_{N,\cdot k}(m/N) - \mathbf{d}_{N,\cdot\cdot}(m/N)]'$$

and

$$\hat{\mathbf{f}}_\epsilon(m/N) = \sum_{k=1}^{2} \sum_{q=1}^{12} [\mathbf{d}_{N,qk}(m/N) - \mathbf{d}_{N,\cdot k}(m/N)] \times [\mathbf{d}_{N,qk}(m/N) - \mathbf{d}_{N,\cdot k}(m/N)]'/22,$$

where $\mathbf{d}_{N,qk}(\lambda)$ is the Walsh–Fourier transform of the $\mathbf{Y}_{qk}(t)$ data vector for infant q ($q = 1, \ldots, 12$) in group k ($k = 1, 2$); the optimal score vector for comparing the two groups was computed to be $\boldsymbol{\beta} = (-.43, .03, .42, 2.16, -.14, 3.53)'$. Although only State 5 is out of order, the difference between the unordered scales $\boldsymbol{\beta}$ and an ordered version of $\boldsymbol{\beta}$—say where State 5 has a scale value between 2.16 and 3.53—was considerable. Nevertheless, an ordered set of scales was sought, and after trying scales for State 5 of the form $p \times 2.16 + (1 - p) \times 3.53$, where $0 \leq p \leq 1$, it was discovered that choosing $p = 1$ was the best in the sense that it gave results closest to the optimal but unordered scale vector $\boldsymbol{\beta}$. Hence the ordered score vector used was $\boldsymbol{\beta}_0 = (-.43, .03, .42, 2.16, 2.16, 3.53)'$ in which case the two active sleep states, low voltage and mixed, are given the same scale. Figure 11 shows a plot of the F ratios for the between-group comparison using the ordered scales $\boldsymbol{\beta}_0$—recall that Figure 11 also displays the results of this analysis from Section 6.2 using suboptimal scales. Note that the results of the two analyses do not differ by much, and one could draw the same conclusions that were drawn in Section 6.2. On the basis of the results given in Figure 13 and the fact that using the optimal scales for distinguishing between the two groups did not establish overwhelming differences, however, I would hesitate to make the conclusion of obvious sleep-cycling differences between the unexposed and the exposed groups of infants. Evidently, it is imperative that more infant sleep studies be performed before any definitive conclusions about the effects of moderate maternal alcohol consumption on infant sleep-state cycling are drawn.

The results on scaling for time series discussed in this section are preliminary at best. Clearly there are many approaches to the problem discussed here, and there are problems that were not discussed here, for example, how would one scale only one time series of interest or how one would scale multivariate categorical time series. Moreover, techniques must be developed to obtain optimal ordered scales.

6.5 Classification for Multivariate Binary Data Using Walsh Functions

Binary variables serve, in many instances, as good models for observations, such as the presence or absence of a disease, the favorable or unfavorable outcome of an experiment or treatment, et cetera. Often, one of these variables is a response that depends in some unknown way on a number of binary predictor variables. In a time series context, we may observe a binary-valued time series; for example, in the sleep studies described in Section 6.2, we may be interested in whether or not an infant is in REM sleep at minute t. In this case, the predictors would be the past, and the response would be the present state of the time series.

Let $(\mathbf{X}, Y) = (X_1, \ldots, X_p, Y)$ be a $(p + 1)$-dimensional vector of Bernoulli variables with a joint density $f(\mathbf{x}, y)$ (x_i, $y = 0, 1$). These variables may be different attributes, such as X_1: sex (male/female), ..., X_p: smoker (yes/no), and Y: myocardial infarction (yes/no), or the same attributes evolving over time, such as X_1: infant in REM sleep at minute 1 (yes/no), ..., X_p: infant in REM sleep at minute p (yes/no), and $Y = X_{p+1}$: infant in REM sleep at minute $p + 1$ (yes/no). The problem is to predict the outcome y of Y from a knowledge of the outcome $\mathbf{x}$. Many approaches to this problem exist. While some procedures require the estimation of the density f, others require only the estimation of suitable functions of the probabilities of the points in the sample space S of $(\mathbf{X}, Y)$; see Ott and Kronmal (1976) for a discussion.

It is clear—as well as optimal in the sense that the probability of misclassification is minimized—that one should predict

$$\begin{aligned} y = 1 &\quad \text{if } h(\mathbf{x}) = f(\mathbf{x}, 0) - f(\mathbf{x}, 1) < 0 \\ y = 0 &\quad \text{if } h(\mathbf{x}) = f(\mathbf{x}, 0) - f(\mathbf{x}, 1) > 0, \end{aligned} \tag{6.5}$$

and to make no prediction if $f(\mathbf{x}, 0) = f(\mathbf{x}, 1)$.

In practice, f is not known. Instead, a classification or prediction procedure can be based on estimates of $f(\mathbf{x}, y)$ or $h(\mathbf{x})$. An estimate based on the Walsh–Fourier transform is discussed here.

The sample space S of $\mathbf{X}$ (where $Y = X_{p+1}$) contains 2^{p+1} points each of which is a vector of zeros and ones. Let these points be numbered by a binary index vector $\mathbf{m}$. The discrete Walsh functions in *natural order* can be written as $W(\mathbf{x}, \mathbf{m}) = (-1)^{\mathbf{m}'\mathbf{x}}$; $\mathbf{m}'\mathbf{x} = \Sigma_{i=1}^{p+1} m_i x_i$ where m_i and x_i are either 0 or 1 (the notation throughout this section will be slightly different than the rest of this article to facilitate this discussion). For example, if $p + 1 = 3$, then S contains 2^3 points, $W[(011), (101)] = (-1)^{0+0+1} = -1 = W(3, 5)$ and $W[(111), (101)] = (-1)^{1+0+1} = 1 = W(7, 5)$; these are the (4, 6) and (8, 6) elements, respectively, of the natural-ordered Hadamard matrix $\mathbf{H}(3)$ displayed in Section 4. Since the Walsh functions form a complete orthonormal set, the probability of a point $\mathbf{x} \in S$ can be written as the orthogonal polynomial

$$f(\mathbf{x}) = 2^{-(p+1)} \sum_{\mathbf{m} \in S} d(\mathbf{m}) W(\mathbf{x}, \mathbf{m}). \tag{6.6}$$

The coefficients $d(\mathbf{m})$ are the Walsh–Fourier transforms of the density $f(\mathbf{x})$ at sequency $\mathbf{m}$,

$$d(\mathbf{m}) = \sum_{\mathbf{x} \in S} f(\mathbf{x}) W(\mathbf{x}, \mathbf{m}).$$

For a general discussion of the Walsh expansion of univariate and multivariate probability density functions, see Maqusi (1981, chap. 7).

The model free maximum likelihood estimate (MLE) of $f(\mathbf{x})$ is $\hat{f}(\mathbf{x}) = n(\mathbf{x})/n$, where n is the sample size and $n(\mathbf{x})$ is the number of times the outcome $\mathbf{x}$ is observed. Hence put

$$\hat{d}(\mathbf{m}) = \sum_{\mathbf{x} \in S} n(\mathbf{x}) W(\mathbf{x}, \mathbf{m})/n, \tag{6.7}$$

so that the estimate of $f(\mathbf{x})$ can be rewritten as

$$\hat{f}(\mathbf{x}) = 2^{-(p+1)} \sum_{\mathbf{m} \in S} \hat{d}(\mathbf{m}) W(\mathbf{x}, \mathbf{m}). \tag{6.8}$$

For example, in the case with $p + 1 = 3$ variables, consider the following random sample of size $n = 25$: $n(000) = 1$, $n(001) = 4$, $n(010) = 2$, $n(011) = 5$, $n(100) = 2$, $n(101) = 3$, $n(110) = 7$, $n(111) = 1$. The Walsh functions for all $\mathbf{x}$ and $\mathbf{m}$ needed for this example are given in the natural-ordered Hadamard matrix $\mathbf{H}(3)$ displayed in Section 4. The Walsh–Fourier transforms of the data given by (6.7) are $\hat{d}(000) = 1$, $\hat{d}(001) = (1 - 4 + 2 - 5 + 2 - 3 + 7 - 1)/25 = -.04$, $\hat{d}(010) = (1 + 4 - 2 - 5 + 2 + 3 - 7 - 1)/25 = -.20$, $\hat{d}(011) = -.28$, $\hat{d}(100) = -.04$, $\hat{d}(101) = -.44$, $\hat{d}(110) = .04$, $\hat{d}(111) = .28$. The representation (6.8) can be verified using these calculations; for example, $\hat{f}(001) = n(001)/n = 4/25 = 2^{-3}(1 + .04 - .20 + .28 - .04 + .44 + .04 - .28)$.

So far, nothing more has been done than rewriting $\hat{f}(\mathbf{x}) = n(\mathbf{x})/n$ in terms of the Walsh functions. However, Ott and Kronmal (1976) determined which coefficients in the series (6.8) are to be set equal to zero for the estimates to have certain desirable properties. This approach is in the same vein as the methods of classification for binary variables where the density is written as

$$f(\mathbf{x}) = \exp\left[a_0 + \sum_i a_i x_i + \sum_{i<j}\sum a_{ij} x_i x_j + \sum_{i<j<k}\sum\sum a_{ijk} x_i x_j x_k + \cdots\right], \tag{6.9}$$

$\mathbf{x} = (x_1, \ldots, x_{p+1})$ (the a_i's are the main effects, the a_{ij}'s are the first-order effects, etc.), and the classification procedure is constructed on the basis of this model assuming that interaction terms above a certain order are equal to zero. Ott and Kronmal (1976) discussed three classification methods based on (6.9) and compared them to four other methods based on the representation (6.6) [the representation (6.9) bears no particular relationship to (6.6)].

Ott and Kronmal (1976) based their term selection rules by requiring that the mean summed square error $J_S = E\{\Sigma_{\mathbf{x},y}[f(\mathbf{x}, y) - \hat{f}(\mathbf{x}, y)]^2\}$ be minimized. They showed that the increase in J_S due to the inclusion of the $\mathbf{m}$th term in the series (6.6) is

$$\begin{aligned} \Delta J_{\mathbf{m}} &= 2^{-(p+1)}\{\mathrm{var}[\hat{d}(\mathbf{m})] - d^2(\mathbf{m})\} \\ &= 2^{-(p+1)}[1 - (n + 1)d^2(\mathbf{m})]/n. \end{aligned}$$

Replacing $d^2(\mathbf{m})$ by an unbiased estimate, an unbiased estimate of the increase in the error J_S due to the inclusion of the $\mathbf{m}$th term is $\Delta\hat{J}_{\mathbf{m}} = 2^{-(p+1)}[2 - (n + 1)\hat{d}^2(\mathbf{m})]/(n - 1)$. Thus it is estimated that inclusion of $\hat{d}(\mathbf{m})$ in (6.8) leads to a decrease in the error J_S if $\Delta\hat{J}_{\mathbf{m}} < 0$, that is if

$$\hat{d}^2(\mathbf{m}) > 2/(n + 1). \tag{6.10}$$

Whenever $\Delta\hat{J}_{\mathbf{m}} \geq 0$, that is, if $\hat{d}(\mathbf{m})$ does not satisfy the term selection rule (6.10), $\hat{d}(\mathbf{m})$ is set equal to zero in (6.8). For the preceding example with $p + 1 = 3$ and $n = 25$, there are four terms whose absolute values are less than .277 [the square root of $2/(n + 1)$]. The density estimates resulting from the deletion of the four terms are $\hat{f}(000) = .07$, $\hat{f}(001) = .18$, $\hat{f}(010) = .07$, $\hat{f}(011) = .18$, $\hat{f}(100) = .11$, $\hat{f}(101) = .14$, $\hat{f}(110) = .25$, $\hat{f}(111) = 0$; recall that the MLE's $n(\mathbf{x})/n$ were .04, .16, .08, .20, .08, .12, .28, and .04 in the same order. The density estimates based on the deletion rule differ somewhat from the usual MLE's but not enough to change the predictions by applying (6.5). For example, $\hat{h}(01) = \hat{f}(010) - \hat{f}(011)$ equals $-.11$ $(.07 - .18)$

in the deleted terms case and equals $-.12$ ($.08 - .20$) in the MLE case; hence in both cases we would predict the outcome y to be 1 if we observed $\mathbf{x} = (0, 1)$.

For predictions of the outcome y at $\mathbf{x}$, only the estimate of the differences between the densities at $(\mathbf{x}, 0)$ and $(\mathbf{x}, 1)$ need to be estimated to apply (6.5). The relevant difference can be developed into the series

$$h(\mathbf{x}) = f(\mathbf{x}, 0) - f(\mathbf{x}, 1) = 2^{-p} \sum_{\mathbf{m}\in S(x)} e(\mathbf{m})W(\mathbf{x}, \mathbf{m}), \quad (6.11)$$

where $S(x)$ is the sample space of the variables X_i ($i = 1, \ldots, p$). The term $e(\mathbf{m})$ in (6.11) is the Walsh–Fourier transform of the difference of the densities, and the MLE of $e(\mathbf{m})$ is given by $\hat{e}(\mathbf{m}) = \Sigma_{\mathbf{x}\in S(x)} W(\mathbf{x}, \mathbf{m})[n(\mathbf{x}0) - n(\mathbf{x}1)]/n$, so that the MLE if $h(\mathbf{x})$ can be written as

$$\hat{h}(\mathbf{x}) = 2^{-p} \sum_{\mathbf{m}\in S(x)} \hat{e}(\mathbf{m})W(\mathbf{x}, \mathbf{m}). \quad (6.12)$$

If the mean summed square error

$$J_S = E\left\{ \sum_{\mathbf{x}\in S(x)} [h(\mathbf{x}) - \hat{h}(\mathbf{x})]^2 \right\} \quad (6.13)$$

is to be minimized, then following the preceding arguments, the selection rule for omitting $\hat{e}(\mathbf{m})$ terms in (6.12) has exactly the same form as (6.10) with $\hat{e}(\mathbf{m})$ replacing $\hat{d}(\mathbf{m})$.

Applying the exclusion rule can be thought of as giving a weight of zero or one to each of the coefficients in (6.12). Ott and Kronmal (1976) suggest that it may be more satisfactory to replace such an abrupt term selection rule by determining weights $w(\mathbf{m}) \in [0, 1]$ for each of the coefficients $\hat{e}(\mathbf{m})$ in (6.12). Thus write (6.12) in a more general way as

$$\hat{h}(\mathbf{x}) = 2^{-p} \sum_{\mathbf{m}\in S(x)} w(\mathbf{m})\hat{e}(\mathbf{m})W(\mathbf{x}, \mathbf{m}), \quad (6.14)$$

where the weights remain to be chosen. To minimize the mean summed square error J_S in (6.13), $\partial J_S/\partial w(\mathbf{m})$ was set equal to zero and $w(\mathbf{m})$ was solved for, obtaining the optimal weights

$$w(\mathbf{m}) = e^2(\mathbf{m})/E\{\hat{e}^2(\mathbf{m})\}.$$

Using the MLE for $e^2(\mathbf{m})$ the estimated weights were

$$\hat{w}(\mathbf{m}) = n\hat{e}^2(\mathbf{m})/[1 + (n-1)\hat{e}^2(\mathbf{m})].$$

To measure the performance of various classification procedures—three based on the decomposition (6.9) and four based on the decomposition (6.6)—Ott and Kronmal drew random samples from 11 populations containing lower to higher order interaction terms in the sense of (6.9) and applied each of the seven methods to each sample. The number of variables in each case was taken to be $p + 1 = 6$, while the number of random samples and the sample size varied. For each simulation, the mean error of misclassification and its standard error was computed, where the probability of misclassification is given by

$$P(\mathbf{x}) = f(\mathbf{x}, 1-y) \qquad \text{if the prediction is } Y = y$$
$$= \frac{[f(\mathbf{x}, 0) + f(\mathbf{x}, 1)]}{2} \qquad \text{if no prediction is possible.}$$

Their conclusion was that the weighted Walsh–Fourier procedure (6.14) with $w(\mathbf{m})$ replaced by $\hat{w}(\mathbf{m})$ dominated when they ranked the mean errors of misclassification of the seven procedures in each population. See Ott and Kronmal (1976) for details.

7. DISCUSSION

It is evident that statisticians have, for the most part, been absent from the development of Walsh–Fourier analysis. In their 1976 article, Ott and Kronmal wrote that "There is a vast literature about the theory and applications of discrete [Walsh] Fourier series, which has seemingly gone unnoticed by statisticians," (Ott and Kronmal, 1976, p. 397). Fifteen years later, this statement is still true (You won't find Walsh–Fourier anaysis mentioned in the *Encyclopedia of Statistical Sciences*). The mathematical statisticians that did work in this area were content to build a theory around the absurd but convenient assumption that the data were *dyadically* stationary. Of course they knew that such a theory was not applicable, but the mathematics was nice. I was once asked by a colleague who was working in the area if I had any data that were dyadically stationary, after a moment of silence we simply laughed at the thought. It is evident why Ott and Kronmal's statement remains true today. In engineering, Walsh–Fourier analysis went from "hot topic" to "nobody's interested" in less than a decade. N. M. Blachman wrote, "Thinking Walsh functions a fascinating case study for historians of science, [I] tried to get Gina Kolata to [write] the story of their rise and fall for the American Association for the Advancement of Science's weekly *Science* (for which she was working) five or ten years ago, but I was unsuccessful; she wrote that, since interest had died out, the subject was no longer of sufficient interest." (personal communication, 1989). K. G. Beauchamp explained that "An original idea behind most of the applications in the 1970s was that [Walsh–Fourier] analysis would prove a quicker, cheaper, and easier way of deriving spectral information. This may have the case when computers were slow and expensive to use, but the bulk of these applications were overtaken by the very much increased speed and power of the much cheaper machines which enabled Fourier methods to be more effectively applied. I think this was probably the reason for the demise of the annual Symposia on Walsh functions." (personal communication, 1990).

Walsh–Fourier analysis was nearly dead and buried because of the failure of engineers and statisticians to see the benefits of the technique in its proper perspective. Engineers tried to *replace* the elegant, general, and practical theory of Fourier (trigonometric) analysis with the idiosyncratic theory of Walsh–Fourier analysis solely on the basis that the computations are more suited to computers. Statisticians were willing to build a statistical theory of Walsh–Fourier analysis based on dyadic time that they knew good and well would never be of any practical importance. Fourier (trigonometric) analysis is like the Gaussian distribution—everything works, the theory is elegant, and it is applicable in a wide variety of problems—but there are more phenomena than the Gaussian distribution can explain and we need and use other distributions. Walsh–Fourier analysis is like a non-Gaussian distribution; it is useful but its use is more restricted and the theory is not quite so elegant.

I have made the point that Walsh–Fourier analysis is not a replacement for Fourier analysis; it is another tool with which to help answer some questions that arise in the statistical analysis of data. In this article I have strived to address this problem by exposing the mostly impractical side of Walsh-based analysis, by presenting various ways in which Walsh–Fourier analysis has been used successfully and realistically in statistics, and by suggesting problems for the future. There is clearly room for Walsh–Fourier analysis in our statistical bag of tricks, but there is still much to be done.

[*Received July 1989. Revised December 1990.*]

REFERENCES AND ANNOTATED BIBLIOGRAPHY

Ahmed, N., and Rao, K. R. (1975), *Orthogonal Transforms for Digital Signal Processing,* New York: Springer-Verlag.

Excellent introduction to Fourier and Walsh–Fourier analysis from the engineering perspective. Discusses the fast Fourier and fast Walsh–Fourier transforms in a fair amount of detail, and FORTRAN computer subroutines for both transforms are provided.

Beauchamp, K. G. (1975), *Walsh Functions and Their Applications,* London: Academic Press.

——— (1984), *Applications of Walsh and Related Functions,* London: Academic Press.

Excellent introductions to Walsh and related functions, also, and to the spectral decomposition of time series via the Walsh–Fourier transform from the engineering perspective. Comparisons between Walsh–Fourier and Fourier decompositions of simulated data, as well as waveform synthesis using sinusoids and Walsh functions, are discussed in detail. Provides a perspective on situations when Walsh based analysis, as opposed to trigonometric-based analysis, is relevant. The 1984 edition is essentially an updated version of the 1975 text and contains examples of many engineering applications of Walsh functions as well as a fairly up to date list of references.

Bloomfield, P. (1976), *Fourier Analysis of Time Series: An Introduction,* New York: John Wiley.

An introduction to statistical Fourier analysis that requires knowledge of calculus and an introductory course in mathematical statistics. Aimed at practitioners as well as students. Contains FORTRAN computer subroutines for analysis of data based on the introduced techniques.

Bramhall, J. N. (1974), "Bibliography on Walsh and Walsh Related Functions," *Proceedings of the Symposium on the Applications of Walsh Functions,* Springfield, VA: U.S. Department of Commerce, pp. 416–460.

Extensive bibliography containing roughly 800 entries on Walsh and related functions published between 1900 and 1974. Entries are categorized into five main categories and appropriate subcategories. The major categories are: mathematical treatment, generation of Walsh functions, applications, comparison with other types of analysis, miscellaneous.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformation for Multiple Regression and Correlation" (with comments), *Journal of the American Statistical Association,* 80, 580–619.

Brillinger, D. R. (1980), "Analysis of Variance Problems Under Time Series Models," in *Handbook of Statistics* (Vol. 1), ed. P. R. Krishnaiah, Amsterdam: North Holland, pp. 237–278.

The definitive article on analysis of variance for time series in the spectral domain.

Brillinger, D. R. (1981), *Time Series: Data Analysis and Theory* (2nd ed.), San Francisco: Holden-Day.

Brockwell, P. J., and Davis, R. A. (1987), *Time Series: Theory and Methods,* New York: Springer-Verlag.

Chrestenson, N. E. (1955), "A Class of Generalized Walsh Functions," *Pacific Journal of Mathematics,* 5, 17–31.

Extends Fine's work on generalized Walsh functions. See Fine (1950).

Cryer, J. D. (1986), *Time Series Analysis,* Boston: Duxbury Press.

Day, N. L., Wagener, D. K., and Taylor, P. M. (1985), "Measurement of Substance Use During Pregnancy: Methodological Issues," in *Current Research on the Consequences of Maternal Drug Abuse* (National Institute on Drug Abuse Research Monograph 59), ed. T. M. Pinkert, Washington, D.C.: U.S. Government Printing Office, pp. 36–47.

Fine, N. J. (1949), "On Walsh Functions," *Transactions of the American Mathematical Society,* 65, 372–414.

——— (1950), "The Generalized Walsh Functions," *Transactions of the American Mathematical Society,* 69, 66–77.

——— (1957), "Fourier–Stieltjes Series of Walsh Functions," *Transactions of the American Mathematical Society,* 86, 246–255.

These articles deal with some of the mathematical properties of Walsh functions. The 1950 paper establishes the generalized Walsh functions by extending the definition of Walsh functions to $W(x, y)$, where $x, y \in \boldsymbol{R}_+$. The 1957 paper essentially establishes a Walsh–Fourier spectral distribution function—see Kohn (1980a, Sect. 4) for details.

Fourier, J. L. (1822), *Theorie Analytique de la Chaleur.* English ed., *The Analytical Theory of Heat,* (1878); Reprinted New York: Dover, 1955.

Frankel, H. D. (1971), "Applications of Walsh Functions," *Proceedings of the Symposium on the Applications of Walsh Functions,* Springfield, VA: U.S. Department of Commerce, pp. 134–136.

Fuller, W. A. (1976), *Introduction to Statistical Time Series,* New York: John Wiley.

Good, I. J. (1958), "The Interaction Algorithm and Practical Fourier Analysis," *Journal of the Royal Statistical Society,* Ser. B, 20, 361–372.

The first appearance of the Walsh–Fourier transform in the statistics literature, but it makes no reference to the Walsh literature at the time. This work forms the basis of the sequency-order Walsh–Fourier transform by relating the calculations involved to Yates's algorithm for calculating effects in a 2^n factorial design.

——— (1971), "The Relationship Between Two Fast Fourier Transforms," *IEEE Transactions on Computers,* 20, 310–317.

Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analysis,* London: Academic Press.

Haar, A. (1910), "Zur Theorie der Orthogonalen Funktionensysteme," *Mathematische Annalen,* 69, 331–371.

Hauri, P. (1982), *The Sleep Disorders,* Kalamazoo, MI: The Upjohn Company.

Hannan, E. J. (1970), *Multiple Time Series,* New York: John Wiley.

Harmuth, H. F. (1969a), "Applications of Walsh Functions in Communications," *IEEE Spectrum,* 6, 82–91.

——— (1969b), *Transmission of Information by Orthogonal Functions,* (2nd Ed. 1972), New York: Springer-Verlag.

——— (1977), *Sequency Theory: Foundations and Applications,* New York: Academic Press.

The engineering approach to the study and use of the Walsh functions was originated in the first two works. These were the seeds that begot the hundreds of articles listed in Bramhall's bibliography. Each work contains interesting applications of Walsh functions as well as a history of the uses of nonsinusoidal functions. The introductions are interesting to read, but the main texts will be difficult for statisticians to read.

Ito, T. (1970), "Applications of Walsh Functions to Pattern Recognition and Switching Theory," in *Proceedings of the Symposium on the Applications of Walsh Functions,* Springfield, VA: U.S. Department of Commerce, pp. 128–137.

Outlines a statistical theory of pattern recognition for multivariate binary variables based on the Walsh expansion of the density functions along the same lines as the techniques presented in Ott and Kronmal (1976) and discussed in Section 6.5.

Kimeldorf, G., May, J. H., and Sampson, A. R. (1982), "Concordant and Discordant Monotone Correlations and Their Evaluations by Nonlinear Optimization," in *Optimization in Statistics,* (Studies in the Management Sciences Vol. 19), eds. S. H. Zanakis and J. S. Rustagi, Amsterdam: North Holland, pp. 117–130.

Kohn, R. (1980a), "On the Spectral Decomposition of Stationary Time Series Using Walsh Functions I," *Advances in Applied Probability,* 12, 183–199.

——— (1980b), "On the Spectral Decomposition of Stationary Time Series Using Walsh Functions II," *Advances in Applied Probability,* 12, 462–474.

A rather complete set of articles establishing the Walsh–Fourier theory for real-time stationary time series summarized in Section 3. Shows that many of the results concerning the decomposition of stationary time series using trigonometric functions have their Walsh function analogies. Part I focuses on the asymptotic properties of the Walsh–Fourier transform and introduces a Walsh–Fourier spectral density function. Part II discusses estimation of the Walsh–Fourier spectrum.

Maqusi, M. (1981), *Applied Walsh Analysis,* London: Heyden.

Concentrates on theory based on the generalized Walsh functions. Contains some engineering applications and has an entire chapter devoted to Walsh series expansions of univariate and multivariate probability density functions and their uses in the computations of general moments of nonlinear transformations.

Morettin, P. A. (1972), "Walsh–Fourier Analysis of Time Series," unpublished Ph.D. dissertation, University of California, Berkeley, Dept. of Statistics.

Establishes a statistical theory for the analysis of dyadically stationary time series using the Walsh–Fourier transform.

——— (1973), "A Note on a Central Limit Theorem for Dependent Random Variables," *Bulletin of the Brazilian Mathematical Society,* 4, 47–49.

——— (1974a), "Limit Theorems for Stationary and Dyadic-Stationary Processes," *Bulletin of the Brazilian Mathematical Society,* 5, 97–104.

These articles were the first to establish results for the analysis of real-time stationary time series based on the Walsh–Fourier transform.

——— (1974b), "Walsh-Function Analysis of a Certain Class of Time Series," *Stochastic Processes and Their Applications,* 4, 183–194.

——— (1981), "Walsh Spectral Analysis," *SIAM Review,* 23, 279–291.

Excellent review article on the statistical approach of Walsh spectral analysis. Discusses statistical results for dyadic- and real-time stationary time series. Provides a good opportunity to compare and contrast the dyadic- and real-time approaches to the problem. Includes many references.

——— (1983), "A Note on a Central Limit Theorem for Stationary Processes," *Journal of Time Series Analysis,* 4, 49–52.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and its Applications,* Toronto: University of Toronto Press.

Ott, J., and Kronmal, R. A. (1976), "Some Classification Procedures for Multivariate Binary Data Using Orthogonal Functions," *Journal of the American Statistical Association,* 71, 391–399.

The first application of Walsh functions in the statistics literature. Presents methods for prediction or classification for multivariate binary data based on a Walsh–Fourier expansion of the density. Good summary of Walsh functions and their generation in appendixes.

Paley, R. E. A. C. (1932), "A Remarkable Series of Orthogonal Functions," *Proceedings of the London Mathematical Society,* 34, 241–279.

Reintroduces the Walsh functions to the scientific community by defining them as the product of Rademacher functions. This definition is better suited for analytical manipulations where one uses the Walsh functions in Paley order.

Priestley, M. B. (1981), *Spectral Analysis and Time Series* (Vols. 1 and 2), New York: Academic Press.

Rademacher, H. (1922), "Einige Satze von Allgemeinen Orthogonalfunktionen," *Mathematische Annalen,* 38, 122–138.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: John Wiley.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference,* New York: John Wiley.

Robinson, G. S. (1972a), "Discrete Walsh and Fourier Power Spectra," in *Proceedings of the Symposium on the Applications of Walsh Functions,* Springfield, VA: U.S. Department of Commerce, pp. 298–309.

——— (1972b), "Logical Convolution and Discrete Walsh and Fourier Power Spectra," *IEEE Transactions on Audio and Electroacoustics,* 20, 271–280.

Introduces the concept of the logical covariance function, which was the basis for establishing a statistical Walsh–Fourier theory for real-time stationary processes. Compares and contrasts Walsh-based and trigonometric-based spectral analysis.

Scheffé, H. (1959), *The Analysis of Variance,* New York: John Wiley.

Scher, M. S., Richardson, G. A., Coble, P. A., Day, N. L., and Stoffer, D. S. (1988), "The Effects of Prenatal Alcohol and Marijuana Exposure: Disturbances in the Neonatal Sleep Cycling and Arousal," *Pediatric Research,* 24, 101–105.

Selfridge, R. G. (1955), "Generalized Walsh Transform," *Pacific Journal of Mathematics,* 451–480.

Shumway, R. H. (1988), *Applied Statistical Time Series Analysis,* Englewood Cliffs, NJ: Prentice-Hall.

An excellent text that provides an introduction to the spectral analysis of time series. Written for students in the physical, biological, and social sciences, and for applied statisticians. Requires knowledge of one year of calculus and a course on elementary mathematical statistics. Only text to discuss analysis of variance and discriminant analysis for time series data. Comes with software.

Stoffer, D. S. (1985), "Central Limit Theorems for Finite Walsh–Fourier Transforms of Weakly Stationary Time Series," *Journal of Time Series Analysis,* 6, 261–267.

——— (1987), "Walsh–Fourier Analysis of Discrete-Valued Time Series," *Journal of Time Series Analysis,* 8, 449–467.

An approach to the analysis of discrete-valued time series based on the Walsh–Fourier transform is presented. A general signal-plus-noise model for discrete-valued processes is developed, and solutions to problems involving signal detection and time series regression for discrete-valued processes are discussed.

——— (1990), "Multivariate Walsh–Fourier Analysis," *Journal of Time Series Analysis,* 11, 57–73.

Establishes statistical methodology for the spectral analysis of stationary multivariate time series via the Walsh–Fourier transform with an emphasis on cross-spectral analysis. Assumptions are fairly general so that techniques can apply to discrete-valued time series.

Stoffer, D. S., Scher, M. S., Richardson, G. A., Day, N. L., and Coble, P. A. (1988), "A Walsh–Fourier Analysis of the Effects of Moderate Maternal Alcohol Consumption on Neonatal Sleep-State Cycling," *Journal of the American Statistical Association,* 83, 954–963.

The Walsh–Fourier transform is used in an analysis of variance of categorical time series to investigate the spectral components of EEG sleep-state patterns of infants whose mothers abstained from alcohol during pregnancy and infants whose mothers used moderate amounts of alcohol throughout pregnancy.

Texter, P. A., and Ord, J. K. (1989), "Forecasting Using Automatic Identification Procedures: A Comparative Analysis," *International Journal of Forecasting,* 5, 209–215.

Uses Walsh functions to develop a procedure to decide the order of differencing in automatic, time domain, forecasting procedures. Shows the relationship between the Walsh–Fourier transform and Yates's algorithm.

Tzafestas, S. G., ed. (1985), *Walsh Functions in Signal and Systems Analysis and Design,* New York: Van Nostrand.

A diverse collection of articles—with editorial comments—that brings together major results of the engineering literature on Walsh functions.

Walsh, J. L. (1923), "A Closed Set of Orthogonal Functions," *American Journal of Mathematics,* 45, 5–24.

The article that first introduced Walsh functions to the scientific community.

Comment

K. G. BEAUCHAMP*

Stoffer has produced a very useful introduction to nonsinusoidal analysis for statisticians who would not normally consider (or even be aware of) such methods.

The summary given in the first half of the article is adequate, although speaking as an engineer more used to considering the Walsh function as a tool to assist the process of signal processing, I found several gaps in his presentation—examples being power spectral density analysis, waveform filtering, and correlation applications. In addition, two areas in this introductory section are rather misleading. The first area concerns the references to Harmuth's definition of sequency. This was not "forced in an attempt to get Walsh-based analysis to behave like Fourier-based analysis." About the time when this definition was conceived, computer time was expensive and Fourier analysis (even using the fast Fourier transform) for any useful accuracy *did* require a lot of computing time. Many scientists and engineers saw Walsh function analysis as a means of reducing computer costs (Beauchamp 1984; Temel and Linkens 1978). Often this led to a preliminary analysis using Walsh functions, with Fourier analysis applied later to selected items of the data that looked promising. Later, as computers became more powerful, the saving of time became less important and attention turned to those areas where there were significant differences in the *quality* of analysis using Walsh functions and where the special characteristics of these waveforms would be applicable.

The second area with which I take issue is the treatment of dyadic time. This is more than imposing scaling differences in real/dyadic time as inferred in Section 3. I considered this in some detail in Beauchamp (1984, chaps. 5, 6) and gave an illustration of the linear and dyadic time domain (Beauchamp 1984, fig. 3.1). In this diagram, consecutive data samples for a process are shown equally spaced (that is in real time) so that, whereas (for example) in arithmetic correlation, the lag increased uniformly with time, in the dyadic case, the time lag varies in a nonlinear way and can actually go backward. Consecutive data samples in real time are seen as taking place in a series of jumps unequal in length over both forward and backward time intervals due, of course, to the modulo-2 nature of the time additions. It is this that makes processing involving correlation using Walsh series so difficult to interpret and relate to Fourier methods. The problem is dealt with very thoroughly by Gibbs (1969, 1970) and Gibbs and Pichler (1971), who suggested a whole new mathematics to deal with the situation.

Turning to the statistical applications research section of Stoffer's article (Sec. 6), it is interesting to note that the appearance of a peak in the Walsh function periodogram, not present in the equivalent Fourier periodogram, has been noted by other workers, particularly those concerned with seismic waveform analysis (Kennett 1975). It has been suggested that this could indicate a specific difference between certain types of seismic data, not apparent with conventional Fourier analysis, leading to a better identification of earthquake characteristics, but this has not yet been adequately proven.

The statistical application section is limited by the presentation of only one real application, namely, the methods used for measuring alcohol use and the scoring of neonatal EEG sleep records. The conclusions reached in this case are interesting but would benefit from more extensive measured data, since a clinician may be chary of basing a treatment on the indications shown in the results of the analysis due to their sparse nature. Nevertheless, the stimulus provided by Stoffer's discussion to submit data to nonsinusoidal analysis is valuable, and I hope that in the near future it encourages more application of these methods to statistical data.

ADDITIONAL REFERENCES

Gibbs, J. E. (1969), "Walsh Functions as Solutions of a Logical Differential Equation," National Physical Laboratory Report, DES 1, Teddington, Middlesex, U.K.

——— (1970), "Discrete Complex Walsh Transforms," in *Applications of Walsh Functions: 1970 Proceedings, 31 March–3 April, Symposium and Workshop Held at Naval Research Laboratory, Washington, DC,* ed. C. A. Bass, Washington, DC: U.S. Government Printing Office, pp. 106–122.

Gibbs, J. E., and Pichler, F. R. (1971), Comment on "The Transformation of Fourier Power Spectra Into Walsh Spectra," *IEEE Transactions on Electromagnetic Compatibility,* 13, 51–54.

Kennett, B. L. N. (1975), "Short Term Spectral Analysis and Sequency Filtering of Seismic Data," in *Exploitation of Seismograph Networks,* ed. K. G. Beauchamp, Leiden, NL: Noordhoff, pp. 283–296.

Temel, Z. B., and Linkens, D. A. (1978), "Medical Data Analysis Using Microprocessor-Based Walsh Transformations," *IEEE Transactions on Biomedical Engineering,* 16, 188–194.

* K. G. Beauchamp is a private consultant, 35 Belle Vue Terrace, Greaves Road, Lancaster LA1 4TY, United Kingdom. He retired from the post of Director of Computing Services, University of Lancaster, in 1985.

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Review Paper

Comment

DAVID R. BRILLINGER*

Professor Stoffer is to be congratulated for having prepared such a fine review and for having lifted the Fourier–Walsh analysis of time series data to a higher statistical level. While many of the computations will be so very familiar to those knowledgeable in the design of experiments and analysis of variance, we are forced to view things in a new light and this has to be a good thing.

In my discussion I will focus on a naive technique to examine time series data for level changes. Suppose that time is continuous and runs over the unit interval, [0, 1]. By a dyadic subinterval of [0, 1] will be meant an interval of the form $I = (k2^{-p}, (k + 1)2^{-p})$ $(p > 0, k = 0, \ldots, 2^p - 1)$. One of the very special things about the Walsh functions in dyadic (and, as Professor Stoffer has pointed out to me, also in sequency) order is that if $S_n(t)$ denotes the

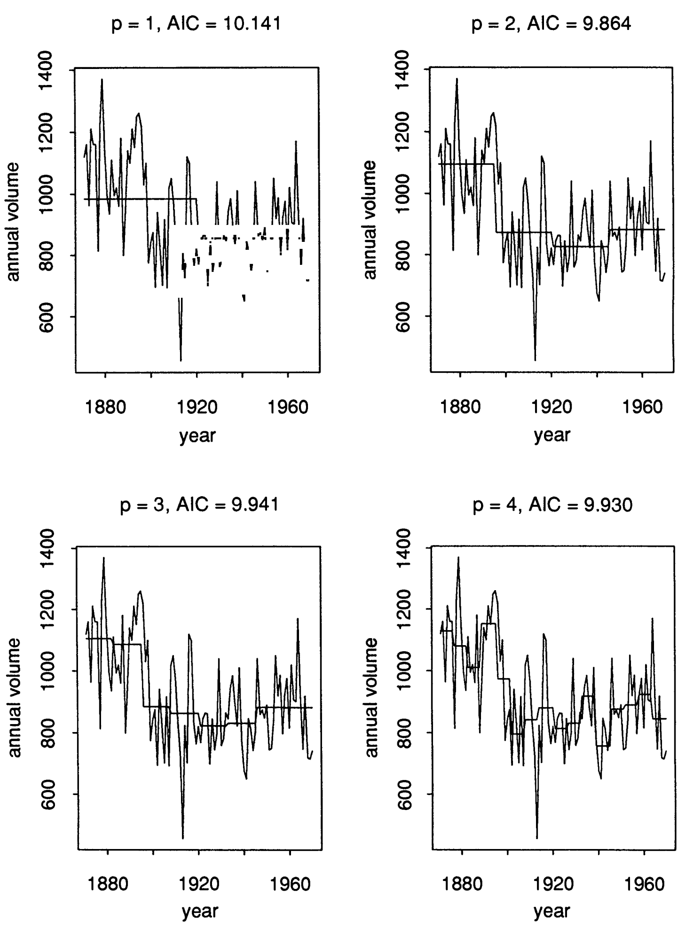

Figure 1. Nile River Discharge at Aswan. Data are from Cobb (1978). The series has been broken into 2^p approximately equal segments, and the values within each segment averaged to obtain a level for the segment.

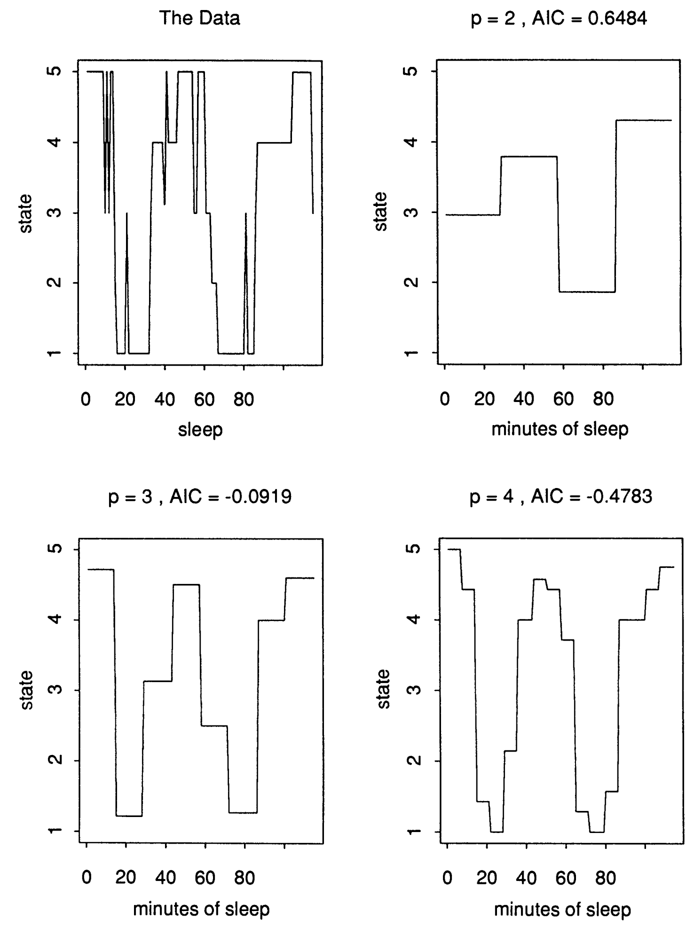

Figure 2. Sleep State of the Infant. Data are from Figure 4 of the article. The series has been broken into 2^p segments, and the values within each segment averaged to obtain a level for the segment.

partial sum of the first $n = 2^p$ terms of the expansion for $f(t)$ then

$$S_n(t) = 2^p \int_I f(u)\, du, \tag{1}$$

where $I = I_t$ is the dyadic interval containing t (see Fine 1949). Among other things, this result immediately gives the convergence of $S_n(t)$ to $f(t)$ for suitable $f(\cdot)$ as n tends to infinity. The function (1) strikes me as quite natural to consider in problems concerning functions that might be constant over intervals. One simply breaks the domain of observation into intervals and averages the measurements within the interval for a level for an interval. Figure 1 provides the results of such computations, for $p = 1, 2, 3, 4$, for the series of annual volumes of the Nile River discharge

* David R. Brillinger is Professor, Statistics Department, University of California, Berkeley, CA 94720. This work was supported in part by the NSF Grant DMS-8900613.

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Review Paper

at Aswan from 1871 to 1970. (The data may be found in Cobb 1978.) One of the issues with this data has been whether or not the mean level experienced an abrupt change at some point in time. In the computations here, the integral of (1) has been approximated by a discrete sum. To guide assessment of an appropriate number of Walsh function terms to include in the sum, the Akaike Information Criterion (AIC) has been computed for the cases, proceeding as if the errors were independent normals. The criterion is least for $p = 2$, and the corresponding graph of Figure 1 is highly suggestive.

I was led to think through the above approach because of a concern about the behavior of the "sequency" ordered functions when there is a shift of time origin. It seems to me that sequency statistics can well vary substantially if several of the initial points of a time series happen to be dropped, or if more early data points happen to turn up. I have corresponding concerns regarding vector time series analyses. How is one to align the components of series of this type? If one has an evoked response experiment, then there is a specific and crucial time origin, but is that the case here? I guess I am asking how robust are the analyses given here to dropping some of the data points from the beginning, as could well have happened had one simply commenced the measurements at a slightly later time? In the case of the analysis for my Figure 1, the effect seems easily understood. If the data values excluded are not extreme, then the "fitted" function itself should not vary a lot. It would seem that the amplitudes of the individual Walsh function components could well vary noticeably, however, and hence be the source of difficulty of interpretation.

Figure 2 presents the results of applying the preceding computations to the data in the article. Stepwise functions of the desired type have been produced. If the true phenomenon is actually fluctuating rapidly, these "smooth" functions will tend to miss that however.

Of course there can be no unique solution to problems of the type discussed in this article. Setting down specific models is one way that we focus our attention. The sort of model that occurs to me in the present context is a jump process in which the signal $S(t)$ is constant in segments filling up the time interval of observation. I wonder whether Professor Stoffer has some such formal model in mind? One for which a likelihood analysis is available?

ADDITIONAL REFERENCE

Cobb, G. W. (1978), "The Problem of the Nile: Conditional Solution to a Change Point Problem," *Biometrika*, 65, 243–251.

Comment

PEDRO A. MORETTIN*

In his comments on Morettin (1981), J. Durbin concluded by saying that "as regards the applications of Walsh functions to series which are stationary in the normal sense I am frankly sceptical about this." The current article by David Stoffer is therefore timely and stimulating.

When I was myself faced with the task of writing my thesis in the Walsh field I was intrigued by its novelty, as well as by the difficulties of interpretation which are due to the strange behavior of dyadic time. As properly remarked by Stoffer, even if it is quite natural to think about the intensity associated with each term of (3.6) leading to the concept of the *Walsh spectrum*

$$g(\lambda) = \sum_{j=0}^{\infty} B(j)W(j, \lambda), \qquad 0 \leq \lambda < 1, \tag{1}$$

it is less evident how to interpret the *dyadic covariance function*

$$B(j) = E\{x(t)x(t \oplus j)\}, \tag{2}$$

assumed to be absolutely summable. It follows that $X(t)$ given by (3.6) has to be *dyadically stationary* and not stationary in the usual sense.

The mathematical setup that is behind each case is as follows. When we analyze stationary processes, with continuous time, for example, we are considering a locally compact Abelian group (LCAG), namely, the additive group $\boldsymbol{R}$ of the real numbers. The characters of this group are the exponentials e^{iyx} (x, y in $\boldsymbol{R}$), hence the importance of sines and cosines. In this sense we can say that we live in a "sinusoidal world." In almost any field, as in communications engineering for example, especially in the theory of linear, time-invariant networks, Fourier analysis is perhaps the main mathematical tool.

If the Walsh functions are considered, the LCAG that is behind the scene is the so-called dyadic group D, which is the set of all sequences $\bar{x} = \{x_n\}$, where $x_n = 0$ or 1, with group operation given by $\bar{z} = \bar{x} + \bar{y}$ ($\bar{x}$, $\bar{y} \in D$) and $z_n = x_n + y_n$ (mod 2). Fine (1949) showed that the characters of

* Pedro A. Morettin is Professor, Department of Statistics, and Dean, Institute of Mathematics and Statistics, University of São Paulo, C.P. 20570, 01498, São Paulo, Brazil.

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Review Paper

this group can be identified with the full set of the Walsh functions. The appropriate class of stochastic processes to entertain, therefore, must have probabilistic structure unchanged under dyadic shifts. This can be generalized to processes defined on general LCAG's. See Morettin (1980) and Brillinger (1982) for further details. As a consequence of this general theory, spectral representations for $B(j)$ and $X(t)$, dyadically stationary, exist, analogous respectively, to those for the normal stationary processes.

If $X(t)$ is stationary and $\gamma(j)$ is the covariance function, then it follows from (2) that $B(j)$ is a function of $\gamma(t \oplus j - t)$, if it would make sense to define (2) for a stationary process. An estimate of (2) is

$$\hat{B}(j) = N^{-1} \sum_{t=0}^{N-j} X(t)X(t \oplus j), \tag{3}$$

while an estimate of $\tau(j)$, defined in (5.2), is proposed by Kohn (1980b) to be

$$\hat{\tau}(j) = N^{-1} \sum_{t=0}^{N-1} X(t)X(t \oplus j). \tag{4}$$

Robinson (1972a) defined the *logical covariance* as the average of possible $\hat{\tau}(j)$'s computed over an ensemble of "windows of data" and showed that there is a relationship between this logical covariance and $\gamma(j)$, which is mentioned by Stoffer and also defined by Kohn (1980a) in his Lemma 2. It would be nice if Stoffer could give a pedestrian account of these facts and tie them together. Other references are Yuen (1973), Gibbs (1967) and Pichler (1970).

The problem here is that although a Wiener–Khintchine theorem holds for $\tau(j)$, namely,

$$\tau(j) = \int_0^1 W(j, \lambda)\, dG(\lambda), \qquad j \geq 0, \tag{5}$$

a Cramér-type representation *does not hold for $X(t)$ stationary,* except in the trivial case $\gamma(j) = 0$ $(j \neq 0)$, as shown by Kohn (1980a). For dyadically stationary processes, this type of representation holds, as we just mentioned, but it seems that practical examples of these processes are scarce in nature; they tend to be man-made, as the response of a sequence filter to a white-noise input.

I also would like to have seen some mention of alternative estimators of the Walsh spectrum, besides the smoothed periodogram. A smoothed estimator obtained from (5.3) and (4) is computationally comparable with the first approach, as shown by Yuen (1973), for example.

REFERENCES

Brillinger, D. R. (1982), "Asymptotic Normality of Finite Fourier Transform of Stationary Generalized Processes," *Journal of Multivariate Analysis,* 12, 64–71.

Gibbs, J. E. (1967), "Walsh Spectrometry, a Form of Spectral Analysis Suited to Binary Digital Computation," National Physical Laboratory Report, Teddington, Middlesex, U.K.

Morettin, P. A. (1980), "Homogeneous Random Processes on Locally Compact Abelian Groups," *Annals of the Brazilian Academy of Sciences,* 52, 1–6.

——— (1981), "Walsh–Fourier Transforms," *Bulletin of the International Statistical Institute,* 49, 1211–1230.

Pichler, F. R. (1970), "Some Aspects of a Theory of Correlation With Respect to Walsh Harmonic Analysis," Technical Report R-70-11, University of Maryland, Dept. of Electrical Engineering.

Yuen, C. K. (1973), "A Fast Algorithm for Computing Walsh Power Spectrum," in *Applications of Walsh Functions: 1973 Proceedings, 16–18 April, Symposium Held at the Catholic University of America, Washington, DC,* eds. R. W. Zeek and A. E. Showalter, Springfield, VA: National Technical Information Service, pp. 279–283.

Rejoinder

DAVID S. STOFFER

I thank Beauchamp, Brillinger, and Morettin for their comments on the applications of Walsh–Fourier analysis in statistics. Much of my work in this area is based on their research, and their comments are a welcome complement to this exposition. I will address the comments of the discussants in lexicographic order (although I did give some thought to dyadic order).

1. RESPONSE TO BEAUCHAMP

I must begin by saying that anyone who is interested in using Walsh function analysis should first refer to Beauchamp's books (1975, 1984); they would certainly fill the gaps noted in the presentation. About Harmuth's definition of sequency being forced to make Walsh analysis behave like Fourier analysis, I believe that Beauchamp's explanation agrees with my interpretation. That is, at the time that Harmuth developed his theory of sequency, Walsh analysis was used as a quick-and-dirty Fourier analysis (this use of Walsh–Fourier analysis nearly brought about the demise of the technique once fast and cheap computers became available). There is a natural pairing of $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$, in that one is just a shift of the other; there is not, however, a natural pairing of $W(2n - 1, \lambda)$ and $W(2n, \lambda)$. My concern is that in Harmuth's definition of sequency $W(2n - 1, \lambda)$ and $W(2n, \lambda)$ are paired simply because in Fourier analysis $\cos(2\pi n\lambda)$ and $\sin(2\pi n\lambda)$ are paired.

Dyadic relationships are an unfortunate but necessary

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Review Paper

companion of the Walsh functions and must be dealt with. Results pertaining to the dyadic domain are entirely artificial, however. I have never seen dyadically stationary time series data (except for the trivial white noise case) or a real-world use for the notion of dyadic time.

I first encountered the appearance of a peak in a Walsh spectrum that was absent from the corresponding Fourier spectrum in Beauchamp (1975, pp. 101–102) where an analysis of a simulated seismic waveform was performed. I was intrigued by this occurrence, and it had a major influence in my pursuing the statistical applications of Walsh spectral analysis. That is, this occurrence was clear evidence that Walsh-based methodology was not merely a replicate of sinusoidal based methodology. That this occurred in the analysis of the sleep data in Section 6.1 adds to the evidence of the unique quality of Walsh–Fourier analysis. But, will that peak always be there whether or not it is meaningful? The answer is no; as discussed in Sections 6.1 and 6.2, this extra peak in the Walsh spectrum was present and meaningful in the unexposed infants and absent in the alcohol-exposed infants, even though there were similarities in the Walsh spectra of the two groups of infants at the lower sequencies (see Stoffer et al. 1988, Figure 3).

Finally, the clinical implications of some of this work were discussed in Stoffer et al. (1988) and were made available to clinicians in Scher et al. (1988).

2. RESPONSE TO BRILLINGER

Brillinger begins his comment with an excellent demonstration of analyzing level changes in time series via Walsh functions. This is an example of the applicability of Walsh analysis where sinusoidal analysis would not be considered. This type of analysis is precisely the type of useful statistical methodology that will be generated once researchers move away from trying to use Walsh–Fourier analysis to mimic sinusoidal analysis, and I thank Brillinger for this contribution. I would add that one might want to fine tune the procedure by dropping certain insignificant terms (according to some optimal procedure) in the expansion $S_n(t)$ in much the same way that coefficients were dropped in the analysis described in Section 6.5.

Brillinger next focuses on two concerns, one dealing with general results of the effect of phase or shift in Walsh–Fourier analysis and the other dealing with how these results pertain to the analysis of the sleep-state data in Section 6. The general problem concerns the difference between the transform of the data $X(t)$ and the transform of the phase-shifted data, say $Y(t) = X(t + \tau)$ $(t = 0, 1, 2, \ldots)$. The Fourier transform of $X(t)$ and $Y(t)$ are the same (certainly a desirable property); however, this is not the case for the Walsh–Fourier transform. This matter was discussed in further detail in Beauchamp (1975, pp. 42–45, pp. 89–94). Beauchamp (1975) noted, however, that phase shift yields comparatively small changes in the shape of Walsh power spectra. He further argued that phase shift is unimportant when the start of a signal can be readily defined (for example, a seismic disturbance or an evoked response).

Brillinger's second concern is that the start of the sleep signals might not be well defined and that the occurrence of phase or shift differences among the infant sleep signals will affect the individual and group analyses. Let me start with the individual analysis of the unexposed infant. Brillinger asks what will happen if several of the initial points of the sleep data happen to be dropped. Figure R.1 compares the Walsh periodogram of the original sleep data shown in Figure 8 (see Section 6.1 for details) and the Walsh periodogram of the same data but with the first 10 observations (10 minutes) removed (the altered data were padded to $N = 128$). As expected, there are differences in the periodograms, but as Beauchamp found in his work, essentially the same information is obtained (the peaks are shifted slightly and the amplitudes, as predicted by Brillinger, have decreased).

For the group analyses, the assumption that the sleep signals have been aligned (or are in phase) is critical, and Brillinger wonders how one can align the components of these sleep state series. The sleep series are aligned because of the protocol that was employed in these sleep studies. Briefly, only a select group of neonates participate in a sleep study—vaginal or cesarean section birth with no general anesthesia or Apgars less than six at five minutes, 38–42 weeks' gestation, male infants prior to circumcision, infant not on antibiotics, and infant 24–36 hours of age. Preparation for a sleep study takes about 30 minutes: the head is measured and marked, the scalp is cleansed, approximately 20 flat-disc electrodes are applied to the head with cream and tape, and the baby is swaddled and placed into a bed after being fed (so that the infant is asleep). The sleep study begins and records for two hours in a quiet, environmentally controlled room designated for neonatal sleep studies. Thus there is a fixed reference for the sleep signals analyzed due to the amount of time spent on preparation, the extensive preparation itself (that is, the neonate has been aroused for quite some time), and the feeding and swaddling to initialize the sleep session.

As far as setting down a parametric model, I would say

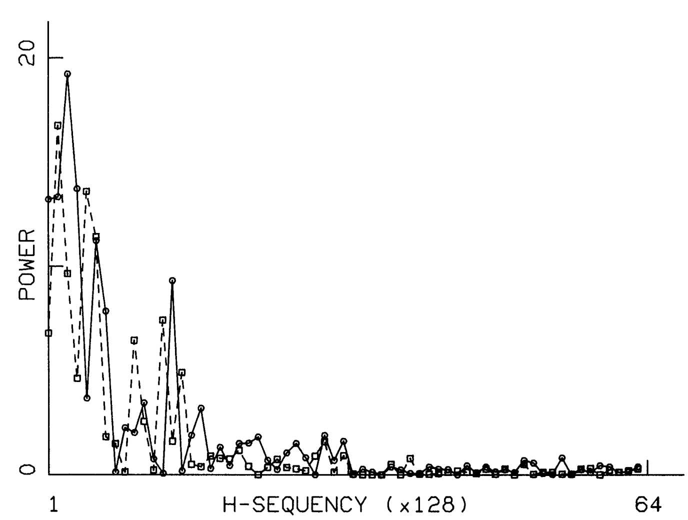

Figure R.1. Comparison of the Walsh Periodogram of the Sleep State Data for the Unexposed Infant (Solid Lines–Circles) and the Walsh Periodogram of the Same Data With the First 10 Minutes of Observations Deleted (Dashed Line–Squares).

that this would be difficult at this point since so little is known about infant sleep (very few infant sleep studies have been performed). However, parametric modeling might be applied to adult sleep since there have been extensive sleep studies on adults. In the studies presented in this article, I focused on one particular aspect of the data, that of the cyclic behavior of infant sleep—an investigation best approached by spectral techniques.

3. RESPONSE TO MORETTIN

Morettin gives a brief account of the Walsh theory for dyadically stationary time series, an area in which he has published extensively. I was reluctant to include any of this material due in part to space constraints, but primarily to the fact that, although the theory is convenient in the realm of Walsh analysis (as pointed out in Morettin's discussion of the dyadic group), it has no practical application. Morettin's comments partially fill this gap and interested readers should refer to Morettin (1981) for a more detailed discussion. I believe that, except perhaps for purely mathematical interests, it is time for the theory of dyadically stationary time series to go to its final resting place.

Morettin reiterates the fact that there is no spectral representation theorem for real-time stationary time series in terms of the Walsh functions. As he states, this makes interpretation of Walsh spectra difficult since we would like to think of a time series as being decomposed in terms of a linear combination of Walsh functions as in (3.6). I have often worried about this and have hope that further research will lead to an interpretation result for Walsh-based analysis. But again, this complication does not negate the practical utility of Walsh–Fourier analysis; it simply means that we must be careful in the interpretation of peaks in the Walsh spectrum—interpretation here is not as automatic as it is in Fourier analysis.

Finally, Morettin mentions smooth estimation of the Walsh spectrum. I have discussed this as a practical matter in Stoffer (1990); basically, other than the smoothed periodogram, the smoothed estimators that mimic the estimators of the Fourier spectrum (see Kohn 1980b or Morettin 1981 for details) will be extremely difficult to interpret due to the way that these estimators depend on dyadic addition.