Psychonomic Bulletin & Review
2005, 12 (3), 409-441

Provenance of correlations in psychological data

THOMAS L. THORNTON and DAVID L. GILDEN
University of Texas, Austin, Texas

Two distinct families of statistical processes are considered in the production of psychophysical time series data (Gilden, 1997, 2001; Gilden, Thornton, & Mallon, 1995). We inquire whether the spectral signatures of the underlying dynamics are better described in terms of short-range autoregressive moving-average (ARMA) processes or long-range fractal processes. A thorough presentation of both families is given so as to clarify the scope and generalizability of the models as descriptions of choice reaction time data. Analyses of data are supplemented by the construction of a spectral likelihood classifier that discriminates between the two families of processes. The classifier has sufficient sensitivity to ensure that fractals are correctly identified and that ARMA processes will rarely be misconstrued as belonging to the fractal family. Spectral likelihood classification illustrates an extremely general framework for testing competing spectral hypotheses and is offered for use in measuring the specific character of fluctuations in designed experiments.

There are few truisms in the field of psychology, but one of them is surely that measurement error is found in all experiments. Data are inevitably produced that do not perfectly reflect the logic imposed by the experimental design. To the extent that a psychological experiment succeeds in measuring something or in making some sort of distinction, the data will partially reflect the design, and this leads to a way of thinking about data that is found throughout all the experimental sciences: *data* = *signal* + *noise*. This innocent equation almost always contains an implicit but critical assumption: that the noise may be regarded as independent samples from some distribution—typically taken to be the Gaussian distribution. In this way, the residual error is conceived of as a featureless background of white noise in which the interesting part, the treatment means, are more or less buried.

Often this conception of data is justified. Whenever there is random assignment to cells and each participant contributes a single datum, errors may be expected to be uncorrelated. However, in all of sensory psychophysics and most of cognitive psychology, single individuals respond to entire blocks of trials in a given experimental session. Here, the residual error will develop correlations by virtue of the circumstance that the response history was laid down by a nervous system that has memory. In many situations, these correlations are little more than a nuisance created by explicit recollection of responses on previous trials. This is the case, for example, in magnitude estimation, where the fact that people tend to reiterate their chosen responses (a response, say, of *loud* is likely to engender another response of *loud*; Luce, Nosofsky, Green, & Smith, 1982; Staddon, King, & Lockhead, 1980; Verplanck, Collier, & Cotton, 1952) makes it difficult to extract a meaningful relation between distal signals and their subjective experience. More generic, and potentially much more interesting, are those situations in which explicit memory is not implicated, and then the correlations and the memory system that created them may become the focus of inquiry.

The study of sequential correlations has been particularly intensive in reaction time (RT) methodologies. This body of work has revealed numerous forms of response contingency that are collectively referred to as *priming*. Almost any aspect of a past stimulus or response may influence the speed at which subsequent decisions and discriminations are made (see Bertelson, 1963; Hale, 1967; Luce, 1986; Pashler & Baylis, 1991; Rabbitt, 1968; M. C. Smith, 1968). The issue of relevance here is the range of influence that a given trial has on the trials that follow. Typical ranges are 5–10 trials, and for this reason, priming is thought to be served by a short-term—in most cases, implicit—memory system that decays primarily through interference (Maljkovic & Nakayama, 2000). If priming effects were the only mechanism through which RT histories became internally correlated, then on all scales larger than a few trials, the latencies would form a white noise in any experimental design in which the stimuli were randomly interleaved. Because priming is local, it is unable to create large-scale structures in response history that are not already present in the stimulus sequence. There are, however, numerous demonstrations that repeated measurement generates response histories that are globally distinguished from white noise (Gilden,

Preparation of this article was supported by NIMH Grants R01-MH58606 and R01-MH065272. We thank Guy Van Orden and colleagues for sharing their word-reading data with us (Van Orden, Holden, & Turvey, 2003). We also thank Jerome Busemeyer, Robert Nosofsky, Guy Van Orden, Eric-Jan Wagenmakers, and several anonymous reviewers for their helpful comments, suggestions, and emendations. Finally, we acknowledge S. Spear and A. Stah for their continued and unwavering support. Correspondence regarding this article can be sent to D. L. Gilden or T. L. Thornton, Department of Psychology, University of Texas, 1 University Station A8000, Austin, TX 78712 (e-mail: gilden@psy.utexas.edu or t.thornton@mail.utexas.edu).

1997, 2001; Gilden, Thornton, & Mallon, 1995; Van Orden, Holden, & Turvey, 2003; Van Orden, Moreno, & Holden, 2003; Wagenmakers, Farrell, & Ratcliff, 2004), and it is apparent that there may be a serious gap in our understanding of what is being measured in psychophysics and, generally, in experimental psychology.

The processes that create global correlations in psychophysical methodologies are distinguished from response bias and priming effects in three important ways: They appear to be generic across task and stimuli, they have an enormous range extending over all resolvable trial scales, and they often appear to have an identifiable and lawful form as $1/f$ noise (the power spectrum denoted $S(f)$ goes as f^{-1}; for details, see the Short- and Long-Range Processes and the Statistical Background sections below). All three aspects of this form of memory have important implications for psychological theory. Its universality implies a dynamic common to choice and discrimination. That it is long range entails that this particular dynamic has no characteristic time scale. And finally, to the extent that episodes of repeated measurement are revealed to have power spectra that fall off inversely with frequency, cognitive psychology may find itself allied with modern theories of complexity in which these noises are actively being investigated (Bak, 1990, 1992; Bak, Tang, & Wiesenfeld, 1987, 1988; Handel & Chung, 1993; Li, 2003).

It is clear that the import and relevance of any of these claims rests upon the assertion that repeated measurement generates a structure that may be fairly described as fractal (Mandelbrot & Van Ness, 1968; Schroeder, 1991) and, in particular, a fractal that has the specific structure of $1/f$ noise. In several articles (Gilden, 1997, 2001; Gilden et al., 1995) we have demonstrated that data sets drawn widely from the corpus of psychophysical methodologies are well fit by a family of whitened $1/f$ noises. Yet agreement between model and data, however compelling, is not proof that the model is correct. There is an enormous difference between post hoc data fitting and the prior specification of a theoretical model that has been shown to agree with data (Roberts & Pashler, 2000). Insofar as there are no theoretical models of cognition that predict the correlations we observe, the interpretation of them as $1/f$ noise is vulnerable. There may be nonfractal processes that generate the observed data, and they may just happen to look like $1/f$ noises. This possibility is a central concern of this article (see also Wagenmakers, Farrell, & Ratcliff, 2004).

Here, we entertain the proposition that there are viable models of psychophysical time series that are nonfractal. The particular nonfractal family of models that we shall consider is the autoregressive moving-average (ARMA) process. It has been argued recently that the ARMA family provides a natural foil to $1/f$ noises and to the entire family of fractional Brownian motions (hereafter variously referred to simply as *fractals*) in terms of a key distinction relating to whether the internal correlations are of long or short range (Wagenmakers, Farrell, & Ratcliff, 2004). Aligned with this distinction is whether a process is scale free or governed by a controlling time scale. Fractal noises have a symmetry referred to as *self-affinity* (Mandelbrot, 1985), and this property entails both the absence of discernible scales and the presence of long-range correlations (see Gilden, Schmuckler, & Clayton, 1993; Mandelbrot, 1983; Maylor, Chater, & Brown, 2001; Schroeder, 1991, and the references therein). The exact sense in which time scales and the range of correlations intertwine will be discussed in detail below.

There are two central issues that are addressed in this article. The first concerns whether the ARMA process is a viable explanation for the correlational structure observed in psychophysical time series. Answering this question involves more than just fitting ARMA processes to data, because we must also consider what sort of statistical framework will be used in drawing conclusions. Wagenmakers, Farrell, and Ratcliff (2004), have recently posed this problem by nesting ARMA within a more general long-range model known as *ARFIMA* (autoregressive fractionally integrated moving average). Their method involves determining whether there is sufficient evidence to reject the short-range ARMA as the default model for any given time series. We will argue that this is not a good approach to analyzing psychophysical time series and that the true state of affairs is better illuminated if we allow fractals to compete with ARMAs as two independent families.

The second issue bears on the degree to which the ARMA family can be discriminated from fractal noises in practice. This concerns the possibility that select members of the ARMA family may effectively masquerade as fractals. In the absence of a theory of cognition that provides a principled account of the observed correlations, this threat is an ongoing concern. There is only one way to address this problem, and that is to construct a classification scheme that is accurate and unbiased. The latter part of this article will be devoted to a detailed description of optimal spectral classifiers that solve the discrimination problem through explicit calibration on the ARMA and fractal families. Spectral classification offers a powerful method for discriminating long- from short-range processes. It decides the classification problem for any two families, with no requirement that they bear any particular relation to each other. Furthermore, spectral classifiers of the sort presented in this article are entirely general and permit the testing of *any* potential hypothesis that might be of interest, provided that the hypothesis can be adequately formalized or simulated. Moreover, they can easily be extended to include information regarding the prior probability of the models, the prior probability of parameter values, and model complexity.

Short- and Long-Range Processes

The mathematical distinctions that make a process short or long range are of some subtlety, and it helps to have a set of concrete examples that illustrate the pivotal role of time scale. This is especially true in the present

psychological context, where we will contemplate statistical processes that are formally of short range, yet have no relationship to short-term memory. As we shall discuss in some detail below, the short-range ARMA processes that are candidate models for psychological data create correlations between scores of trials, not the 5 to 10 that might be attributed to some aspect of a short-term memory system.

Short-range processes are not peculiar to psychology and have relevance across a range of disciplines. The formal expression of a short-range process is given by the Langevin equation. This equation is the point of departure for most discussions of fluctuations in statistical physics, and it expresses quite generally the relaxation of a system to equilibrium following a perturbation. It is a first-order equation and involves only the first-order terms that arise when the perturbation is small, yet larger than the mean size of the random fluctuations that exist in the equilibrium state (for an introductory treatment, see Landau & Lifshitz, 1958). It is written

$$\frac{dX}{dt} = -\frac{X}{\tau} + \varepsilon, \tag{1}$$

where X is some state variable, τ is the relaxation time, and ε is a source of white noise with uncorrelated increments. The relaxation time τ acts here as the time scale over which the perturbation dies and the system returns to the equilibrium state. It reflects the various physical properties that mediate this return, and it will be different in different systems.

The relaxation time is essentially what defines a short-range process. It provides the ruler that measures the time over which the state variables are self-correlated. Where such a ruler exists, the autocorrelation function $\rho(k)$ will decay exponentially with e-folding time τ (the time it takes for the correlation to decay by a factor of e), and this determines the exact sense of *short-range*:

$$\rho(k) = \frac{\mathrm{cov}[X(t), X(t-k)]}{\mathrm{var}[X(t)]} = e^{-\frac{k}{\tau}} \tag{2}$$

(*cov* and *var* denote the covariance and variance, respectively; k is a specific choice of temporal lag). Again, we wish to stress that since τ is completely free to vary, there is no entailment for what kind of memory system would be relevant were exponentially decaying correlations observed in a psychological context.

What would it mean for a physical process that unfolds in time not to have a characteristic time scale? Essentially, this would mean that there is no information coming out of the process that reveals how long the process has been observed. There are no features present in a *scale-free* process that inform on the sampling rate and, hence, on how the number of samples relates to the overall observation interval. In the context of spatial fractals, the absence of scale entails that there is no information, say in a photograph, about the size of the objects in the image or of the camera distance. This kind of symmetry, termed *self-affinity*, is well understood, and it forms the basis of fractal geometry (Mandelbrot, 1983). The fractal that is of importance in long-range memory systems is known as *fractional Brownian motion* (fBm). These are random fractals that have autocorrelation functions that decay as power laws (see Peitgen & Saupe, 1988). Power laws generally arise in discussions of scale-free dynamics and are, in fact, the empirical evidence that a system is scale free (see Schroeder, 1991, for an excellent introduction to scale-free processes).

Scale-free processes may come about in any number of ways. The simplest scale-free process that is relevant to psychology is the random walk. Besides arising in the calculation of any mean quantity (the numerator viewed as a sequence of partial sums is a random walk), they are often used in theoretical models of speeded choice and discrimination where information acquisition is thought to occur incrementally (e.g., Link, 1975; Ratcliff, 1978; Ratcliff & Rouder, 1998; P. L. Smith, Ratcliff, & Wolfgang, 2004). The particular scale-free processes that generate the $1/f$ noise structure visible in the records of human performance are, however, neither simple nor well understood. The theory of $1/f$ noise is a current problem in biophysics, statistical mechanics, and the theory of complexity (see the articles in Handel & Chung, 1993; Li, 2003, and the references therein). It will suffice for our purposes here to sketch the kinds of processes that produce $1/f$ noise.

In a definite sense, $1/f$ noises are intermediate between white noises and the contours of random walks, also known as *Brown noises* (after their application in the kinetic theory of gases, where molecules describe random walk paths referred to as *Brownian motion*). The simplest way to demonstrate this relationship is via the power spectrum. The power spectrum of a time series can be computed by taking the Fourier transform of the autocorrelation function, and in this way, we go from a discussion of correlations at different temporal lags to a treatment of power at different spectral frequencies (denoted by f, or the inverse of wavelength). In the spectral domain, we see that Brown, $1/f$, and white noises have a very simple and organized relationship: Their power spectra are given by f^{-2}, f^{-1}, and f^{0}, respectively. The steep spectral falloff of a Brown noise (f^{-2}) is what makes it so predictable, the point-to-point variations induced by each step are small perturbations of slowly varying trends. In contrast, white noises (f^{0}) have no predictability, since the absence of spectral falloff implies the absence of any trend that might allow prediction. The $1/f$ noises (f^{-1}) are intermediate in predictability, and the kinds of processes that can create them must incorporate aspects of both order and disorder. An example of such a process is the random random walk, where the probability of a positive increment varies at each walk position (in contrast to the usual random walk, where this probability is stationary). Here, the underlying transition probabilities are themselves unpredictable in terms of walk position, and they modulate the otherwise highly correlated random walk to produce a true $1/f$ noise. Another example in this vein is

the tangent bifurcation observed in some iterated maps (Devaney, 1992)—the logistic map in particular (Keeler & Farmer, 1986; Pomeau & Manneville, 1980). In these maps, there are parameter values at which the output orbit generally bounces around chaotically but is occasionally trapped where it executes a highly predictable cycle. The overall signal is $1/f$. Perhaps the most significant development in the theory of $1/f$ noise is Bak's (1990, 1992; Bak et al., 1987, 1988) theory of self-organized criticality. In this theory, $1/f$ noises are thought to be the natural fluctuations emanating from metastable systems that have converged to a phase transition between order and chaos.

In addition to dynamical models, various schemes have been proposed that effectively average over the fluctuations arising from a number of subsystems with different fixed time scales.[1] Brute-force approaches to averaging procedures that produce $1/f$ noise are well known and have even been the subject of a "Mathematical Games" column by Martin Gardner (1978). Gardner shows how $1/f$ noises may be generated using only three spinners of the type popular in board games. Cognitive psychologists who have endeavored to account for $1/f$ noises have generally followed this logic by identifying three levels of cognitive activity and showing how their outputs may be summed to create $1/f$ noise (Pressing, 1999; Wagenmakers, Farrell, & Ratcliff, 2004; Ward, 2002). The latter accounts have no physical motivation and cannot be viewed as anything more than an elaboration of Gardner's three-spinner game. A greater problem for this approach is an observation made by Hausdorff and Peng (1996) that spinner models of $1/f$ noise require careful tuning to produce the desired $1/f$ spectrum, tuning that has never been addressed by any psychological theory.

Although the distinctions governing long- versus short-range statistical processes are formal and, indeed, foreign to current discussions of memory, they may nevertheless be highly relevant to how we view the memory processes that are implicated by correlations in temporally ordered data (Gilden, 2001). Most important, if it is the case that human performance generates an underlying $1/f$ noise signal, cognitive theory must eventually reckon with the problem of how it is that sequences of decisions manifest complex structure. Making this case requires careful analysis of the signals that derive from iterated performance, and the first question that must be addressed is whether these signals are indeed long-range—that is, fractal (for a similar logic, see Wagenmakers, Farrell, & Ratcliff, 2004). Deciding which fractal best exemplifies behavior is predicated on showing that fractal descriptions are necessary in the first place.

Statistical Background

The temporal properties of data sequences may be analyzed directly in the time domain, as correlations at different lags, or indirectly, in terms of spectral power at different frequencies. We prefer the spectral description for three reasons. First, as we have previously pointed out, fractional Brownian motions that include random walks and $1/f$ noises have extremely simple spectra; they are straight lines in the log-frequency–log-power plane. Short-range processes also generate a well-defined and coherent family of spectral shapes. Second, the properties of the Fourier transform are well understood, and the influences of averaging and windowing are easily handled here. Finally, we are interested in a realistic long-range model for choice RT in which pure fBms mix with additive white noise (Gilden, 2001); in the spectral domain, these kinds of processes enjoy a relatively straightforward and transparent formulation.

The basic statistical problem in deciding whether a given process is short or long range inevitably rests upon an analysis of the shape of the power spectrum. Often, this analysis focuses on the low-frequency part where the spectrum has the greatest variability but where it is also particularly diagnostic. Consider, for example, the Debye–Lorentzian, which is the spectrum associated with the Langevin equation and any process that has an exponentially decaying autocorrelation function (see Schroeder, 1991, p. 123). It is given by

$$S(f) = \frac{\tau}{1 + (2\pi f \tau)^2}, \tag{3}$$

where f is the frequency measured, say, in inverse trial number (the natural unit that replaces Hertz when data are received in discrete trials as opposed to a regular sampling interval). This spectrum is illustrated in Figure 1 for $\tau = 1$. Note the low-frequency white plateau. Because low frequencies in the spectral domain correspond to large time lags in the time domain, the plateau reiterates the fact that the process is uncorrelated over time scales greater than τ; that is, the process is short range. Much of this article concerns whether or not this plateau is discernible in data sets that are collected within the practical constraints set by human observers.

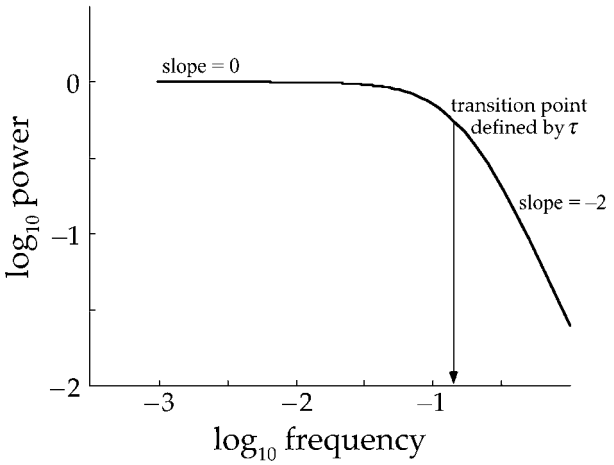

Figure 1. Debye–Lorentzian power spectrum characterizing the short-range process. Note the whitening at frequencies lower than the reciprocal of the characteristic time scale.

Autoregressive Processes

In this article, we analyze a class of short-range processes known as ARMAs. These processes combine single-parameter autoregressive and moving-average processes (i.e., two first-order processes) and are denoted by ARMA(ϕ, θ), where ϕ and θ denote the model's two parameters (Priestley, 1981). First-order ARMA models are the simplest descriptions of short-range behavior that have parameters capable of estimation from the kind of data received in typical psychophysical experiments. More important, they are characterized by correlations that decay over a single time scale, and this makes them useful as a statistical foil in assessments of long-range structure (Wagenmakers, Farrell, & Ratcliff, 2004). In contrast, higher order ARMAs have multiple time scales (Granger & Morris, 1976) and so become increasingly more like a true long-range process as time scales are added (Granger, 1980). In this article, we are interested solely in distinguishing long-range fractal processes from those processes that are governed by a single time scale. Although higher order ARMAs and AR-mixture models do arise occasionally in psychology (e.g., in some models of bipolar disorder, as in Benedetti, Barbini, Colombo, Campori, & Smeraldi, 1996; or in criterion learning models, as in Busemeyer & Myung, 1992), they are more applicable in such fields as electrical engineering, where they provide descriptive approximations to analog filter functions and various order differential equations (Karl, 1989; Priestley, 1981).

The short-range ARMA(ϕ, θ) process may be expressed recursively as a rule that prescribes how a current output is generated from random inputs and the previous output. The autoregressive part takes the form of a leaky integrator:

$$O_t = \phi O_{t-1} + \varepsilon_t, \tag{4}$$

where O_t is the current output, O_{t-1} is the previous output at time $(t - 1)$, and ε_t is the current random input. The parameter ϕ may take on both positive and negative values, but only the positive branch generates time series that look anything like psychophysical data and so we will restrict our discussion to $\phi > 0$. In the limit that $\phi \to 1$, the process described by Equation 4 becomes a nonstationary long-range Brownian motion [also called a *random walk* or *Brown noise*, $S(f) = f^{-2}$]. All first-order autoregressive processes defined on $|\phi| < 1$ are stationary and short range. Equation 4 may be expanded as a weighted sum over a sequence of random inputs where the weights form a geometric series in ϕ; the weight at time $(t - k)$ is equal to ϕ^k. When $|\phi| < 1$, ϕ^k decays geometrically with look-back time, and this leads to an autocorrelation function defined on the output sequence O_t that decays exponentially with increasing lag.[2] By definition, then, O_t is a short-range process. However, we wish to stress that when ϕ is near unity (as is often true for fits of the model to RT data), the decay of the autoregressive process will be quite slow, and there may be palpable correlations between current values and those in the remote past. Consequently, the sense of *short range* associated with these models should not be confused with the capacity limitations of working memory or with the short-term effects due to repetition priming (see Luce, 1986).

The power spectrum of a first-order autoregressive process is written

$$S(f) = \frac{1}{\left[1 - 2\phi\cos(2\pi f) + \phi^2\right]} \tag{5}$$

and is graphed in the left panel of Figure 2 for $\phi = .7$. As is shown, the spectrum is dominated by power at the low frequencies, a reflection of the fact that leaky integration generates discernible hills and valleys in a time series. Like the Debye–Lorentzian spectrum shown earlier in Figure 1, the autoregressive spectrum also has a brown region where power falls sharply [$S(f) \propto f^{-2}$]. The knee where the flat white noise region connects with the brown noise region occurs at a frequency that scales as $-\log(\phi)$. The similarity between the Debye–Lorentzian and the autoregressive spectral functions is no accident;

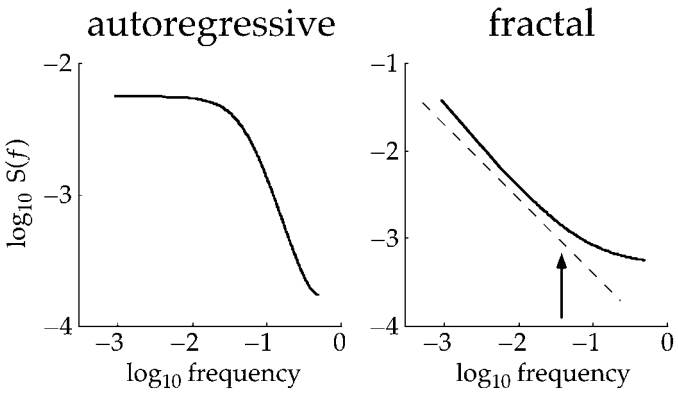

Figure 2. Comparison of power spectra associated with autoregressive and fractal processes. The autoregressive spectrum plots Equation 5 for $\phi = .7$. The whitened fractal spectrum plots Equation 9 for $\alpha = -1$ and $\beta = 1$. The inset dashed line has a slope of −1 for reference.

formally, the autoregressive process is a first-order difference equation that approximates the Langevin differential equation (Equation 1) in discrete time. Accordingly, the autoregressive spectral representation is simply the discrete parameter version of a continuous parameter Debye–Lorentzian where $\phi = e^{-1/\tau}$.

In applied settings, the single-parameter AR(ϕ) process is not particularly useful as a descriptive tool, simply because it does not have enough flexibility to adequately capture the shapes of typical psychophysical spectra (e.g., Pressing & Jolley-Rogers, 1997). For this reason, the autoregressive process must be augmented by a moving-average component to form a two-parameter ARMA model. The additional degree of freedom that defines the hybrid model is obtained by adding in a fraction θ of the previous random input at step $(t - 1)$ so that the full ARMA(ϕ, θ) process is written recursively as

$$O_t = \phi O_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}. \tag{6}$$

In practice, the parameter θ can take either positive or negative values, and this allows the moving-average component of the model to implement two qualitatively distinct types of filtering: *smoothing* and *differentiation*. For $\theta > 0$, the moving-average component averages the sequence of random inputs over a single time step—producing a spectrum that is flat at low frequencies with sharp attenuation of power at high frequency. For $\theta > 0$, the moving-average component implements a one-step differencing of temporally adjacent inputs. In this case, the moving-average power spectrum rises linearly with frequency, and in the limit of $\theta = -1$, the moving-average reduces to a pure derivative operator whose spectrum is linear with a slope of $+2$. In the description of psychophysical data, only the negative θ branch is relevant. The positive branch generates time series that are much smoother (by virtue of averaging) than is observed in psychophysics. Putting the two pieces together, the power spectrum of an ARMA(ϕ, θ) process is written as the product of its constituent parts:

$$S(f) = \frac{\left[1 + 2\theta\cos(2\pi f) + \theta^2\right]}{\left[1 - 2\phi\cos(2\pi f) + \phi^2\right]}. \tag{7}$$

The range of spectral shapes achievable through Equation 7 is quite remarkable. The generality of ARMA models corresponds, roughly speaking, to the class of *rational functions* (i.e., functions that are ratios of two polynomials; Priestley, 1981). To be precise, ARMA(ϕ, θ) models can reproduce *exactly* any spectral function that can be expressed as a ratio of first-order polynomials. This is the first indication that, as a class, the two-parameter ARMA(ϕ, θ) model is of very high complexity and so may not be of much use as a theoretically meaningful description of behavior. Given that typical spectra associated with psychophysical and cognitive time series are monotonic and have, at most, the structure of a quadratic (in log–log coordinates), we should expect the first-order ARMA to do a fair job in describing global trends in these data, provided its parameters are appropriately chosen. In the work described below, we will identify these parameter regimes in some detail, but we will also make it quite clear that the typical ARMA process generates spectra that look nothing like those observed in psychophysical data.

Fractal Processes

For the long-range process, we will consider a family of models in which white noise is added into a purely fractal process. White noise is used to modulate the rate of spectral descent in this family in much the same way that the moving average modulates autoregression. We denote this class of process as *fBmW*, to emphasize its hybrid structure, and write its time domain expression formally as

$$O_t = \text{fractal}(\alpha)_t + \text{white}(\beta)_t\,, \tag{8}$$

where the first term on the left denotes the current value of a fractal sequence (i.e., a fractional Brownian motion with exponent α), and the second term denotes the current value of an independent white noise sequence whose variance is β^2. We have had considerable success using this hybrid model to describe the fluctuations associated with choice RT tasks (Gilden 1997, 2001; Gilden et al., 1995). The fact that Equation 8 is written only in terms of the present time (t) should not be construed to imply that earlier times are not implicitly involved in the creation of the fractal dependence. Equation 8 is, in this sense, heuristic, and not constructive as in the ARMA definition. It is the case that the constructive time domain function for the fBmW(α, β) process, were it specified by a physical, biological, or statistical model, would involve feedback of the kind made explicit in the ARMA expression—although in this case, the recursion would necessarily be of infinite order, or at least on the order of the entire series (e.g., see fractional differencing; Hosking, 1981). This highlights one of the key reasons that we chose to frame the classification problem in the frequency domain. Here, long-range processes such as the fBmW, which have a complicated expression in the time domain, enjoy a relatively succinct and transparent description as spectral power laws plus white noise.

Accordingly, the power spectrum of the hybrid fBmW (α, β) has a straightforward expression built from its correlated and uncorrelated parts:

$$S(f) = N(\alpha) f^{\alpha} + \beta^2. \tag{9}$$

This expression is graphed in the right panel of Figure 2 for $\alpha = -1$, $\beta = 1$. The leading normalization constant $N(\alpha)$ on the right-hand side is necessary to ensure that the fBm component has unit variance in the time domain. Thus, $1 + \beta^2$ gives the total variance of the random process, and $\beta^2/(1 + \beta^2)$ and $1/(1 + \beta^2)$ are the representative fractions of total variance attributable to the white and fractal components of the process. The power spectrum of the fBmW(α, β) process has two regimes: a low-frequency linear region that terminates in a high-frequency

roll-off. The locus of the transition point (denoted by the arrow in the figure) is controlled by the white noise parameter β, and it occurs where the power associated with the white noise component is roughly commensurate with that of the fractal component. As β increases, this transition point migrates to lower frequencies, and the slope of the spectrum in the linear, low-frequency region decreases. Note that by virtue of its fractal lineage, the fBmW has spectral power that continues to rise even as the frequency goes to zero. Practically what this means is that fractal signals are correlated on all scales. Such asymptotic behavior is the signature of long-range processes, and it stands in direct contrast to the spectra of short-range processes governed by a characteristic temporal or spatial scale (τ). For any short-range process, the spectrum must turn over and flatten at frequencies below its correlation limit.

Global Properties

Figures 3 and 4 give detailed portraits of the spectra produced by the ARMA and fBmW processes. The parameter ranges in these figures were chosen on the basis of data relevance. The fBmW parameters were truncated at the boundaries suggested by our own experiments (Gilden, 2001): $\alpha \geq -1$ and $\beta \leq 2$. ARMA parameters were confined here to intervals that contain descending spectra that are not too steep: $\phi > 0$ and $\theta < 0$. Autoregression tends to generate spectra that drop off in frequency quite rapidly, as $1/f^2$ spectra. Signals with this rapid a spectral decay are relatively smooth and more typically

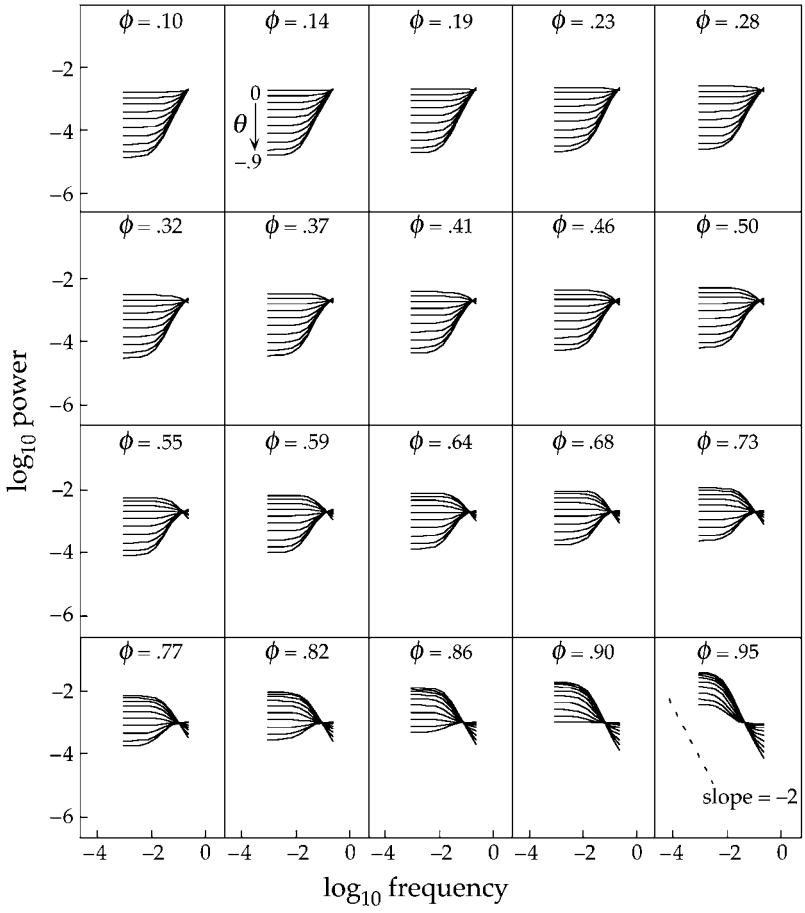

Figure 3. Power spectra associated with the family of first-order ARMA(ϕ, θ) processes. Each panel plots spectra for a single fixed level of the autoregressive parameter ϕ across a range of θ (the moving-average parameter θ decreases from top to bottom). Only a subset of the 400 spectra making up the ensemble used in classification are displayed. This subset was produced by the factorial combination of 20 evenly spaced levels of ϕ on the interval [.1, .95] with 10 staggered levels of θ defined by the quadratic $\theta_k = -.084k + .002k^2$ (for $k = 1, 3, 5, \ldots$) on the interval $[-.9, -.082]$. The two bottom right panels show the power spectra expected of ARMA processes having $\phi = .9$ and $\phi = .95$. These values of ϕ are of particular interest because they most resemble whitened fractals, given an appropriate choice of θ.

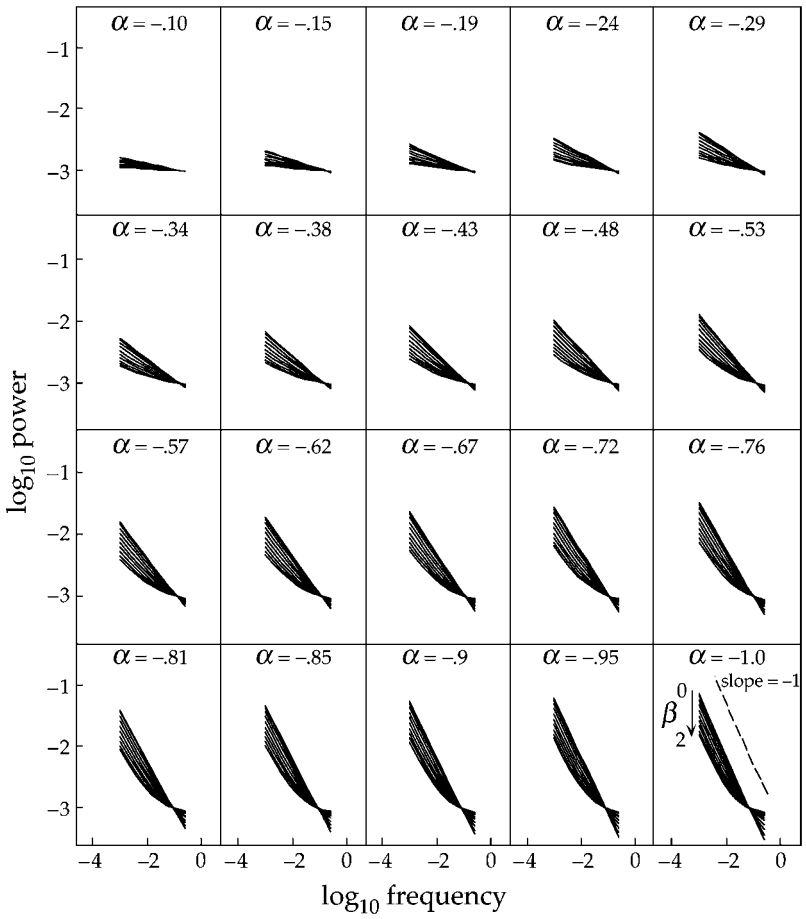

Figure 4. Power spectra associated with the family of whitened fractal processes. Each panel plots spectra for a single fixed level of α across a range of β (β increases from top to bottom). Only a subset of the 400 spectra making up the ensemble used in classification are displayed. This subset was produced by the factorial combination of 20 linearly spaced values of α on the interval $[-1, -.1]$ and 10 linearly spaced values of β on the interval [0, 2]. The bottom right panel shows the spectra generated by a pure $1/f$ signal ($\alpha = -1$) that has been contaminated by additive white noise. In general, the fBmW is linear at low and intermediate frequencies, bowing upward at higher frequencies where the uncorrelated noise source dominates the fluctuation.

associated with landform topography (Feder, 1988; Keller, Crownover, & Chen, 1987). Visually, psychophysical data contours are much rougher than landscape silhouettes, and we have never observed time series data to decay more quickly than $1/f$ (Gilden, 1997, 2001; Gilden et al., 1995). In order to be of psychological relevance, autoregression requires an additional process to add roughness that graduates the spectral falloff. The moving-average part of the ARMA achieves this roughening when $\theta < 0$ and random increments one step back are subtracted. In this case the so-called *moving average* is, in fact, a differencing operator. Differencing produces spectra that increase as f^2, permitting a judicious choice of ϕ and θ to delicately modulate the spectral decay so that the falloff is shallow enough to approximate behavioral data. Note how, in Figure 3, increasing spectra are gradually transformed into decreasing spectra as ϕ exceeds $|\theta|$. When $\phi + \theta = 0$, the balance between autoregression and differencing is perfect, and pure white noises with flat spectra are generated (see note 2). The degree of freedom that balancing affords permits the ARMA to closely approximate any monotone spectrum, so long as the spectrum has an asymptotic white region at low frequencies.

Figure 4 illustrates the family of hybrid fBmW(α, β) processes. Each panel of Figure 4 displays a subset of spectra having a particular value of the fBm parameter α. Within each panel, the various spectra are distinguished by the amplitude of additive white noise, β. In these plots, there are 20 linearly spaced values of α taken on the interval $[-1, -.1]$ and 10 linearly spaced values of β taken on the interval [0, 2]. These two sets of parameter values were crossed to yield 200 fBmW spectra. In

contrast to the family of ARMA spectra, the fBmW spectra display a general family resemblance. Spectral power falls linearly with increasing frequency before gradually flattening. This property greatly restricts the class of time series that may be fit by the fBmW.

The global properties of these two families of temporal correlations are evidently quite different, and this leads to their having different utilities. The ARMA process is a statistical jackknife and is used for a variety of purposes in applied time series analysis (e.g., forecasting); it is discussed in virtually all texts on the analysis of sequential structure. Unfortunately, the very breadth that allows the ARMA to fit a profusion of unrelated filter functions is the same feature that undermines it as a theory of psychological process; because it can *describe* so much, it must in turn *explain* very little (cf. Figure 1 in Roberts & Pashler, 2000). In contrast, whitened fractals are highly constrained in their spectral appearance, and this is what makes them useful as descriptions of behavior. It is meaningful and interesting to assert that human behavior is fractal, precisely because such a claim is positioned for falsification. General engineering models such as the ARMA are not designed for scientific conjecture, and fractal models describe well just those events that do, in fact, have fractal structure. The consequences of this observation will become increasingly obvious in the treatment of data that follows.

THE INTERPRETATION OF FLUCTUATIONS IN PSYCHOPHYSICAL DATA

Now that these two families of statistical processes have been introduced, we may meaningfully address what roles they might play in understanding psychophysical data. We will return to the two issues that guide this inquiry.

1. Is the short-range ARMA process an appropriate hypothesis for describing the structure of correlations found in repeated psychophysical measurement? We will present analyses of data from four experiments that provide motivation for discarding the ARMA as a viable account of psychological process. These analyses will contrast the poor fits of the ARMA model with the fits of a long-range fBmW model, which successfully captures the bowed shapes evident in all four power spectra. Importantly, we show that not only does the best-fit fractal model outperform the best-fit ARMA model in every case, but also that the data are embedded within the fractal and ARMA families in a categorically distinct and informative manner. To anticipate, our analyses show the data to be highly similar to a large fraction of the exemplars defining the fractal family and, at the same time, to be highly *dissimilar* to all but a vanishingly small subset of relatively poor-fitting ARMAs.

2. In laboratory practice, how reliably can short-range autoregressive time series be discriminated from long-range fractal time series? Here, we ignore the irrelevance of the ARMA as a theory of psychological process and focus instead on its utility as a short-range statistical foil to the fBmW. In this regard, we build on the early insights of Davies and Harte (1987) and Pressing and Jolley-Rogers (1997), who pointed out the potential for short- and long-range processes to be confused in certain regimes, as well as the more recent work by Wagenmakers and colleagues, who introduced the ARMA as a more stringent means to test for long-range structure (Wagenmakers, Farrell, & Ratcliff, 2004). Although this issue has received some preliminary quantitative treatment (Wagenmakers, Farrell, & Ratcliff, 2004), we go substantially beyond that work to examine the discriminability of two entire families of short- and long-range processes in the spectral domain. In so doing, we hope to provide a complete account of where, in each model's parameter space, there is the potential for confusion.

To attack this problem, we have constructed a spectral likelihood classifier that uses the shape of the power spectrum to decide among competing short- and long-range descriptions of data. We show this classifier to be both highly sensitive and unbiased in discriminating fBmW processes from their most confusable ARMA relatives. The construction of the classifier is proof that the provenance of correlations can be reliably decided using realistic length data sets and, more important, that the two families of process are not as confusable as analyses limited to first-order spectral slope measures might suggest. We illustrate the practical utility of the classifier by applying it to sequences of word-reading RTs from a study by Van Orden, Holden, and Turvey (2003) that has been at the center of a recent controversy concerning the nature of the observed fluctuations (Wagenmakers, Farrell, & Ratcliff, 2005; Van Orden, Holden, & Turvey, 2005).

A Framework for Deciding the Provenance of Fluctuations

There are at least two distinct logics for the categorization of time series data, and the occasions for their respective usage will depend critically on what we know about the categories and their relational structure. For example, suppose that we believe that all the objects to be classified have a particular attribute p that makes them instances of a Class A, unless they also have an attribute q, in which case they are in Class B. In such a world, Class A is nested within Class B, for we may regard the objects in A to have the null value of q. In this way, we can construct an inferential statistical framework for class membership; any given object is to be placed in A (the null hypothesis) unless there is evidence for presence of the defining attribute q. This approach to the interpretation of time series has been explored in some detail in a recent article by Wagenmakers, Farrell, and Ratcliff (2004). In their framework, there are two processes that generate time series: ARMA and ARFIMA. ARFIMA augments the ARMA process through an additional parameter d. When $d > 0$, the ARFIMA process

generates time series with long-range correlation functions. When $d = 0$, ARFIMA reduces to ARMA. Wagenmakers, Farrell, and Ratcliff (2004) conducted tests on experimental data to determine whether the hypothesis of $d = 0$ can be rejected. It is implicit in this framework that any given time series is to be regarded as short range (ARMA) unless there is evidence to the contrary that it is long-range (ARFIMA).

The nested models logic of ARMA $\subset$ ARFIMA does succeed in providing the necessary structure for setting up an inferential framework, but it does so with both theoretical and statistical costs. The theoretical costs are in fact quite numerous, the first being that a default hypothesis should be a plausible outcome, and we might ask whether we are ready to accept the ARMA process as the default interpretation of psychophysical time series. As this article will make quite clear, the ARMA family makes no substantial contact with psychophysical time series and, consequently, a default ARMA interpretation is not an inference one should draw in the absence of strong evidence to the contrary.[3] In an inferential framework based on a null hypothesis, there is generally a bias to accept the null. And since the classifier devised by Wagenmakers, Farrell, and Ratcliff (2004) presumes the ARMA as null, it has a much larger miss rate than false alarm rate in tests of long-range structure. Although this logic makes sense in situations in which the null has face validity (no effect, say), we see no reason to embrace the ARMA process in the absence of positive evidence for its descriptive relevance.

Second, the classification framework adopted by Wagenmakers, Farrell, and Ratcliff (2004) is theoretically beholden to autoregression. Autoregression is but one possible biological dynamic that might be considered in choice and discrimination, and just as there is more to physics and biology than diffusion, there may also be more to psychology. There are places where autoregression does appear in psychological theory (e.g., in models of criterion learning; Busemeyer & Myung, 1992, and the references therein; Dorfman & Biderman, 1971) and it is ironic, at least from our point of view, that it is primarily in the theoretical modeling of RT that autoregression has found its most profound psychological applications. Models of RT use autoregression to generate diffusion to a decision boundary (Ratcliff, 1978; Ratcliff & Rouder, 1998; P. L. Smith & Vickers, 1989; Usher & McClelland, 2001). This is an enormously elegant framework, and we have used aspects of it in our work on visual search (Thornton, 2002; Thornton & Gilden, 2005). However, and this is a key point, in diffusion models of RT it is bits of information or perceptual evidence that accumulate, whereas in autoregressive models of RT it is RT itself that accumulates. A model that describes how individual latencies are produced may have nothing to say about their correlations, and typically, diffusion models assume that latencies are uncorrelated. If there are reasons to believe that autoregressive processes are active in the formation of correlations, they have not been articulated within the corpus of psychophysics.

The third cost to the inferential approach as practiced in Wagenmakers, Farrell, and Ratcliff (2004) arises from within the logic of hypothesis testing with nested models. The more general class, ARFIMA, is virtually guaranteed to fit data better than ARMA does, because it has the extra parameter, d. Even if there were no long-range dependencies in the data, d would permit some ARFIMA configuration to fit the sampling error better than any ARMA does. Here badness-of-fit measures must be replaced by some criterion that takes into account the additional degree of freedom that ARFIMA possesses. Such a criterion would determine whether the additional parameter d was justified, given that ARFIMA has three parameters and ARMA only two. However, this problem is not as easily solved as might appear from the psychological literature. The AIC criterion (Akaike, 1974) used by Wagenmakers, Farrell, and Ratcliff (2005) is, in fact, one of a panoply of possibilities (Myung, 2000). This is not a situation in which a one-size-fits-all correction for degrees of freedom is prudent or judicious. When models have unequal degrees of freedom, the selection process ultimately involves answering questions centering on generalization and theoretical relevance (Forster, 2000; Myung, 2000). The theoretical relevance of the ARFIMA is an unavoidable issue, and it appears that its sole utility arises from its nesting relationship to the ARMA, an irrelevance in the practical matter of deciding the classification problem.[4]

In this article, we advocate a second route to classification that is unbiased, does not assume autoregression in the construction of long-range models, and does not create the statistical uncertainties that come from associating nested ARMA/ARFIMA models with short- and long-range processes. This route is Bayesian in spirit and is built around the notion of allowing models to compete on the basis of prior probability and likelihood. We formulate the classification problem in terms of Bayesian inference by replacing ARFIMA with the family of whitened fractional Brownian motions (fBmW). This class of fractals has long-range correlations and is entirely suitable for evaluating psychophysical time series in terms of the long-range/short-range distinction.

The benefits of this reformulation are immediate. First, the ARMA process no longer represents a default hypothesis through which all data are presumed to be short range unless there is strong evidence to the contrary. Rather, in our approach, the ARMA becomes simply a candidate description of data that competes with descriptions based on the fBmW. This manner of categorizing assigns provenance to the model with the highest single likelihood or the largest marginal likelihood (i.e., true Bayesian selection). Second, the connection to autoregression is replaced by a theoretically richer class, one that makes contact with the statistical literature in physics, biology, physiology, astronomy, and meteorology (see the references in the introduction). In particular, the specific fractals known as $1/f$ noises have generated a great deal of theoretical attention, as evidenced by the wide variety of nonautoregressive mechanisms that have

been considered—self-organized criticality, extremal dynamics, and tangent bifurcations, to name only a few. ARFIMA, in contrast, despite its popularity in the fields of econometrics and geology (see Tong, 2001), is arguably little more than a mathematical formalism that lacks a supporting theoretical literature. Finally, the whitened fBms are defined by two parameters and compete with the ARMA family on somewhat equal footing. The statistical subtleties of unequal degrees of freedom are, for the most part, avoided. Within a Bayesian perspective, the question of whether fractals or ARMAs provide a better description of psychophysical time series will be decided on the basis of likelihood, since we have no knowledge of the prior probabilities of the two classes. The arguments that derive from this kind of analysis are quite strong. Not only will we be able to evaluate which class has the member with the greatest likelihood, but also we will be able to obtain a clear picture of how the two classes as a whole embed the data. The latter distinction is important because it reflects model-specific differences in complexity (Navarro, Pitt, & Myung, 2004; Pitt, Myung, & Zhang, 2002), a dimension that may prove critical in assessing whether the ARMA or the fBmW is a better description of psychophysical data.

The data in contention derive from RT methods. This is the domain that Wagenmakers, Farrell, and Ratcliff (2004) consider and it is especially important in view of the global use of RT in experimental psychology. We have considered RT methods in earlier work (Gilden, 1997), and we will use those data to illuminate the properties of the ARMA and fractal families here. The relevant experiments involve speeded judgments in mental rotation, lexical decision, serial visual search, and parallel visual search. The experimental methods are given in Gilden (1997), and we will confine our present comments to the construction of spectra.

In each of these studies, there were 6 observers, and they responded to over 1,024 trials in an experimental session. The sequences of RTs were treated identically, independently of study. Cell means were removed to create sequences of residuals, and these were then linearly detrended and standardized to have zero mean and unit variance. Eight-point spectra for each residual sequence were computed, using the methods described in detail in Appendix A. These spectra were then averaged over the 6 observers. The effects of observer averaging are discussed in Appendix B. Models were evaluated by computing the summed χ^2 of the residuals (i.e., a *weighted least squares* cost function) between the observer-averaged spectra and the theoretical spectral expectations deriving from a subset of the ARMA and fBmW families. This subset was chosen to include all processes that have descending spectra and are no more correlated than a pure random walk. RT spectra never ascend, and pure random walks provide an effective upper bound on the observed correlations. In this way, we allow all potentially relevant members from each family to compete for the data. The results of these calculations are shown in Figures 5A and 5B.

Figure 5A illustrates the best-fitting models to the data. It is clear that, in each experiment, there is a whitened fractal that does indeed fit the data quite closely. The ARMA fits are not good, and were it not for rigorous error checking of the code, we might wonder whether these are, in fact, the best fits. The ARMA models are obviously hamstrung by the presence of a time scale that leads to flattening at low frequency. The ARMA processes are forced to snake an S-shape through data that continues to ascend through the low frequencies. Fractals have exactly the property required to fit every data set: scale freedom.

There is no question that the best-fitting fractal mirrors the observed spectra much better than the best-fitting ARMA does. Still, it must be recognized that these fits are made without the benefit of a prior theory. We are decidedly not comparing data with the predictions of a theory. There is no assurance that the best-fitting models are generic to their class. It could be the case that the ARMA family has greater likelihood (smaller χ^2) generally and that the best-fitting fractal is atypical of its class. Stronger statistical arguments are mandated, and we attempt to supply these through Figure 5B. In this figure, we have plotted the value of χ^2 from its minimum value up to 20 times the minimum value. The central ovals in each of the fractal plots mark out those models that were within 10% of the minimum χ^2. The upper triangle in the ARMA space is grayed out, because it contains only ascending spectra and none of these will fit the characteristic descending spectral data. We have found these plots to be instructive.

Figure 5B makes it clear that almost the entire ARMA family has low likelihood. The few ARMA processes that resemble data are confined to the edge of the parameter space, where the process is a combination of a pure random walk with a pure differencing operator. The confinement of the best ARMA fits to this region of the parameter space is noteworthy for two reasons. First, it indicates that the only way a short-range process can approximate the observed data is to incorporate near-perfect autoregression. Accordingly, we see that all the ARMA fits cluster near the boundary $\phi = 1$. Recall that this boundary represents a qualitative divide separating the class of short-range processes from the class of long-range random walks (i.e., nonstationary processes with $\phi \geq 1$). Apparently, the ARMA process is able to mimic data only when the parameter ϕ is very close to the limit of stationarity and the moving-average parameter θ is negative with a slightly lower absolute magnitude. In this configuration, the substantial correlations induced by an autoregressive process with $\phi \approx 1$ end up being attenuated with anticorrelated differencing via the moving-average ($\theta < 0$) process, and this allows the full ARMA process to attain a global approximation to data. When the two parameters are equal in absolute value, there is cancellation, and the ARMA generically reduces to white noise with a flat spectrum (this occurs for all ϕ, θ pairs along the major diagonal in Figure 5B). Thus, we see that the ARMA achieves its snakelike parody of smoothly

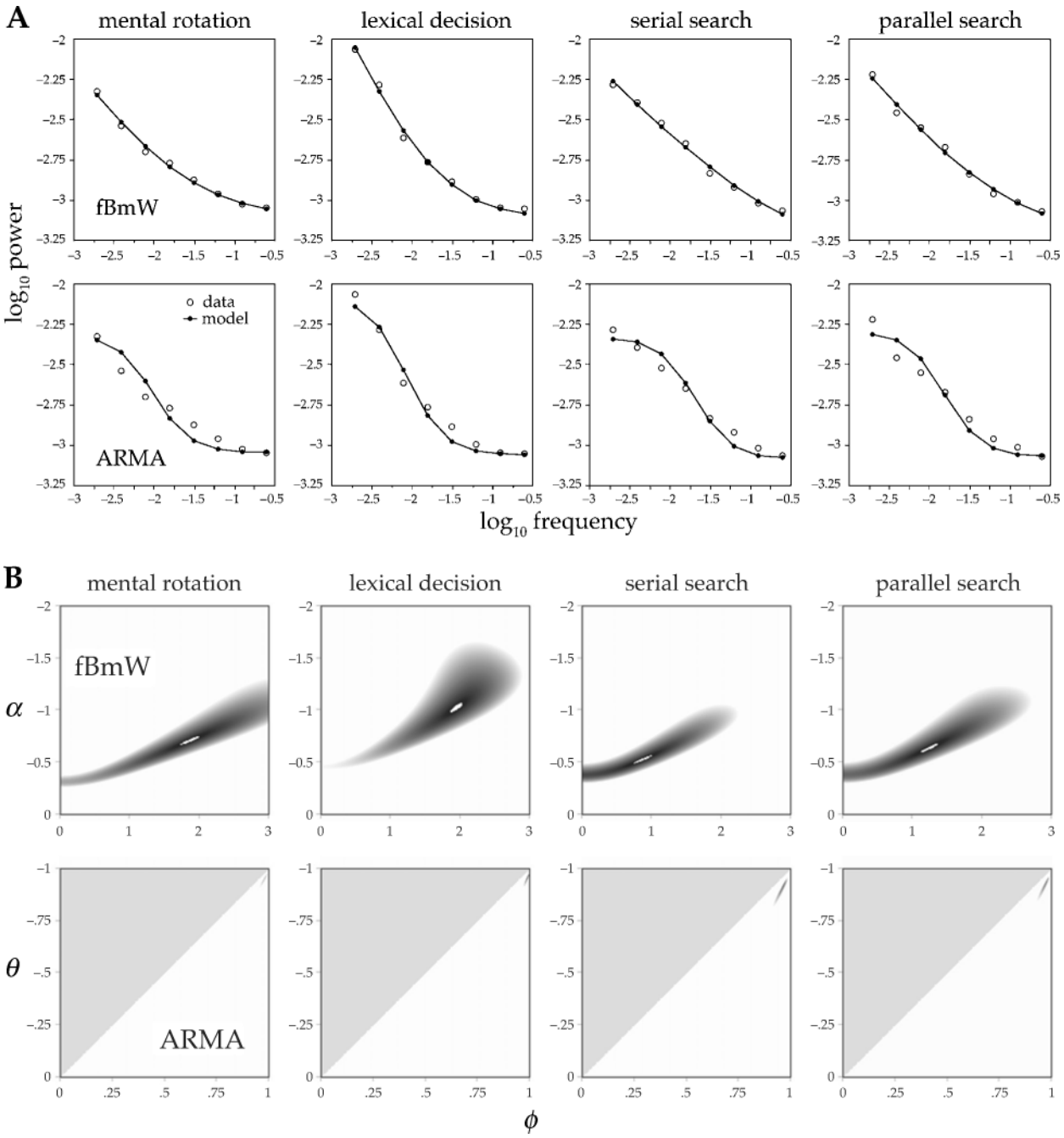

Figure 5. (A) Best-fitting ARMA and fractal models to data from experiments in mental rotation, lexical decision, and visual search. These experiments are described in Gilden (1997). (B) Portraits of χ^2 goodness of fit across the parameter spaces of the two families. Values of χ^2 are plotted from the minimum value up to 20 times the minimum. The white ovals in the top panels are the regions of fractal parameter space within 10% of the minimum value. The upper triangle in the ARMA space is grayed out because it contains ascending spectra that are not relevant to psychophysical data.

descending power law spectra only when its autoregressive and differencing components respect a very specific configuration. The isolation of the ARMA fits to such a small subset of the full parameter space means that the fits are not generic to the process. This reiterates the second key insight provided in Figure 5B: The probability of selecting an ARMA process that actually resembles data is close to zero. This result was presaged in Figure 3, where it is evident that the ARMA process generates a profusion of spectral shapes. Despite its diversity,

the ARMA family is unable to produce a single process that succeeds in fitting our RT data.

Now consider the portraits of χ^2 in the fBmW family; these serve to illustrate that the best-fitting fractal models are generic to the process. In contrast to the ARMA case, here the data literally light up a large portion of the parameter space, and we see that there are many exemplars from this family that resemble the data. This is not a demerit but the best evidence that the family embeds the data within a framework that is robust to perturbation. Figure 5B also shows that the fractal family is sensitive to the specific spectral shapes of the data. Each data set generates a different portrait of χ^2 with a different oval of best fits. Again, this is not a demerit but evidence that parameter estimation within this family is meaningful. The search data have lower values of α and β, whereas the lexical decision and mental rotation data have higher values of both parameters. This is just the sort of information we want from a model, because it suggests a distinction that might bear on the dynamics that create the correlations in the first place. The ARMA family, because its diversity is not an expression of the data, can produce only the same poor-fitting models, regardless of the experiment from which the data were derived.

The experiments that produced these data were selected only to typify experimental practice. They were not selected to fit fractal models. This is an unbiased sample of choice RT data, and Figures 5A and 5B make it clear that the short-range ARMA process is not a viable model of any member of the sample. We have identified a process that does naturally fit residual spectra, and the derived parameters are indeed within the range of what have been termed $1/f$ noises.

Spectral Tools for Deciding Provenance

The demonstration that a few experiments in choice RT are more likely fractal than ARMA does not establish the case that correlations are necessarily produced by a complex dynamic that emits fractal structure. In fact, no set of positive instances of fractals will constitute proof, although for an optimal decision maker it may change the prior probability that fractal descriptions are correct. In the absence of a psychological theory of correlation, we are faced with the problem that data from each and every experiment remain vulnerable to misinterpretation. In this theoretically uninformed state, the only recourse is to develop a statistical tool that allows the two classes to be discriminated.

The remainder of this article will present a set of analysis tools that we have developed that can be used to decide the nature of fluctuations in psychological data sets. First, we provide a low-variance method for estimating the form of the power spectrum. This method is greatly to be preferred over standard Fourier estimates, which are too noisy to be of much use in the issue of deciding the provenance of single sequences. We then present the details and calibration results of a powerful classification framework based on spectral likelihood and Bayesian inference. The spectral classifier allows us to rationally solve the data interpretation issue by assigning provenance to whichever description is most likely, given the data. This particular decision framework is known to be optimal, in that it maximizes long-run classification accuracy. With these tools, we can be fairly certain that short-range processes cannot successfully masquerade as fractals, and we will not be fooled into thinking that objects are fractal when they are not. From a practical standpoint, the spectral classifier provides a standard benchmark for research in this area.

The "Whole-Spectrum" Approach

Traditionally, long-range structures in empirical time series have been assessed in the frequency domain by fitting power laws to spectral estimates (i.e., *spectral tilt* analyses). In double-log coordinates, this reduces to simple linear regression, where the slope of the best-fit line is taken to estimate the fractal exponent of an fBm. When empirical spectra are linear—for example, in production experiments (Gilden, 2001; Gilden et al., 1995)—this practice makes some sense and can be used to distinguish serially correlated fluctuations from white noise. However, when our spectra are not linear, as is the case for all choice RT data, the fitting of lines has the potential to lead to confusion, especially when the models we are interested in testing cannot be distinguished in terms of slopes (Wagenmakers, Farrell, & Ratcliff, 2004). In this domain, the correct approach is to try and understand the underlying nonlinear forms in the data and to see whether our models are capable of reproducing these forms. This leads us to inquire whether there is, in fact, sufficient information available in the shapes of choice RT power spectra to discriminate the predictions of the fBmW and the ARMA models.

The Classification Problem

Let us consider the worst case scenario for classification by entertaining the following question: Is it possible to reliably discriminate long-range processes from an unconstrained class of short-range processes that have been configured so as to approximate the long-range processes? The worst case occurs when we consider a single time series of limited length (i.e., no more than 1,024 trials). This case arises when averaging over observers is not justified or is not prudent (Appendix B). For example, in exploratory data analysis, one may not have a sense yet for the range of processes at work, and in cases in which there is sufficient heterogeneity, averaging may actually distort the true state of affairs (e.g., computing a d' by first averaging hits and false alarms across observers with widely different biases). The potential for distortion via averaging is well recognized in the psychological literature—most notably, in assessments of the form of learning curves (Brown & Heathcote, 2003; Heathcote, Brown, & Mewhort, 2000; Wixted & Ebbesen, 1997), and in work on models of categorization (Maddox, 1999). Accordingly, there has been a growing

emphasis on supplementing standard observer-averaged analyses with analyses based on single-observer data. In the context of discriminating models, this inevitably leads us into a domain in which confusability is high. When spectral data are not averaged and, instead, relatively short single-observer time series form the unit of analysis (Van Orden, Holden, & Turvey, 2003; Wagenmakers, Farrell, & Ratcliff, 2004, 2005), it becomes increasingly possible for an ARMA foil to mimic the spectra of a long-range fBmW. Avoidance of this potential quagmire for interpretation begins with solid and reliable techniques that transform data into spectral representations. We will show that with a high-quality estimate of the power spectrum, there is sufficient information in a single time series to distinguish short- and long-range models.

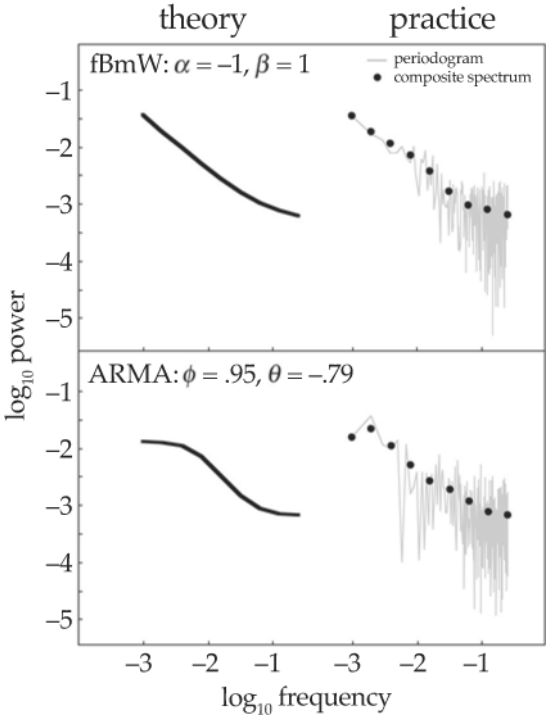

Figure 6. Theoretical expectations and empirical estimates of power spectra associated with the fBmW(−1, 1) process and its closest ARMA relative ($\phi = .95, \theta = -.79$). The smooth spectral forms on the left were derived using Equations 7 and 9. The functions shown in the upper and lower right are estimates of the spectral density based on a standard periodogram analysis (light gray) and a composite spectral analysis (black symbols) in which periodograms of overlapping segments are averaged to form the final estimate of the power spectrum (Welch, 1967). The estimates were obtained by analyzing a single exemplar of a simulated fBmW(−1, 1) process and a single exemplar of a simulated ARMA(.95, −.79) process (both exemplars consisted of 1,024 trials).

Properties of the Raw Periodogram

Estimates of the power spectrum that are numerically derived by a discrete Fourier transform have been historically referred to as the *periodogram*. The periodogram is computed by squaring the amplitudes of the constituent sine waves defined at the frequencies $f_k = k/N$, where $k = 1, 2, \ldots, N/2$ and N is the length of the sampled time series. Insofar as the periodogram is discrete and the power spectrum that it estimates is continuous, the estimates at each frequency should be thought of as an averaged power over a small window centered on each frequency (see Press, Flannery, Teukolsky, & Vetterling, 1992).

The periodogram has the unfortunate property that its reliability does not improve as sample length grows (Priestley, 1981). This is in contrast to standard estimators, such as the sample mean or variance, which do improve with N. The nonconvergent nature of the periodogram arises because increasing N generates a finer frequency resolution but does not decrease the variability in the estimates of power. This state of affairs arises from the fact that each point estimate of power derives from a sum of N random variables and even though every member of the sum (actually, a sample autocovariance; see Priestley, 1981, p. 432) is independent and has a variance that does decrease as $1/N$, the overall variance obtained in summing N such quantities, will not change with the length of the time series

$$\left[\mathrm{Var}_{\mathrm{Sum}} = \sum \mathrm{Var}_{\mathrm{Summands}} \propto N(1/N)\right].$$

The thin gray lines in the upper and lower right quadrants of Figure 6 are the respective periodograms of single examples of an fBmW(−1, 1) and ARMA(.95, −.79) process (each periodogram is based on a simulated time series with $N = 1{,}024$). These functions are typical of the quality of spectral estimates obtained in taking a straight Fourier transform of psychophysical data. As the figure makes clear, the periodograms bear little resemblance to their theoretical expectations in the left panels (the smooth black lines). The spectral estimates are extremely noisy functions that fluctuate wildly from frequency to frequency. This feature of the periodogram is a direct consequence of the fact that estimates of power are, by definition, uncorrelated across frequency (Priestley, 1981).

In distinguishing the fBmW from the ARMA process, we are inevitably led to consider the low-frequency behavior of the power spectrum. It is in this region that long- and short-range processes generate unique and salient spectral signatures. Unfortunately, periodogram estimates not only are unimproved by increases in sequence length, but also satisfy a Weber-like relation; the variability of spectral estimates is proportional to the magnitude of the estimated power. Because psychophysical spectra have their greatest power at a low frequency, this is also the regime of greatest variability. Practically, this means that the region of the spectrum that is most diagnostic for distinguishing short- and long-range descriptions of data is inherently also the most unreliable.

Computation of Low-Variance Spectra

The intrinsic fluctuations of spectral estimates can be reduced only by some sort of smoothing or averaging process. In typical experimental designs, variability is

reduced by observer averaging, but this technique is not available if the relevant data are individual time series.

Fortunately, low-variance estimates of spectral power based on a single time series are readily available, although they come at the cost of reduced frequency resolution. The method we have used is elementary and is based on the simple artifice of breaking up a given data sequence of size N into a series of overlapping windows of size m and averaging their raw spectra (Welch, 1967). Averaging over windows reduces variability, but now the lowest resolved frequency is $1/m$, instead of $1/N$. The method of computing window-averaged spectra has been in use for some time in engineering disciplines (see Press et al., 1992) and is preferred over the straight N-point periodogram because it reduces the spectral variance per datapoint by $9K/11$, where K is the number of windows used in the average. We have extended this technique by allowing m to vary, in order to obtain the best estimate available at each frequency. Appendix A gives a detailed algorithm for the computation of the eight-point least variance composite spectrum that is used throughout our work. In Figure 6, the solid black symbols that overlap the gray periodograms on the right are the estimates of power provided by the composite spectral method (both the standard periodogram and the composite estimates were derived using the same input sequence). Clearly, the estimates of power provided by the composite spectrum are much improved, relative to those of the standard periodogram, in approximating the true spectrum.

The second class of refinements we bring to the analysis of time series is to use classification tools that emphasize the *whole* spectrum. As Figure 6 makes clear, even when the ARMA and the fBmW are aligned so as to be maximally confusable, there will still be local features in the shapes of spectra that are unique to each model. These local features are critical for model testing, and they are not captured by statistics that reduce the spectrum to its overall slope. In fact, any scheme based on spectral tilt would fail to distinguish the ARMA and the fBmW expectations shown on the left in Figure 6. The whole spectrum approach to classification begins with a detailed computation of spectral sampling distributions defined at each frequency.

Sampling Distributions of Spectra

Provenance cannot reliably be assigned in the presence of sampling error without the aid of a statistical tool that incorporates both information about expected spectral shapes and the expected deviations from the mean at each frequency. The sampling distributions of spectra are not available in closed form, and so we have implemented a brute-force Monte Carlo simulation approach to calculating them across the entire ARMA and fBmW parameter spaces.[5] These distributions provide all of the necessary information required to derive the likelihoods of data, given each model.

Sampling distributions of power spectra were created at each point, using the following procedure. The ARMA sampling distributions were realized using a parameter grid that consisted of the factorial combination of 20 linearly spaced values of ϕ on the interval [.1, .95], along with 20 staggered values of θ on the interval [−.082, −.9]. The fBmW distributions were realized using a parameter grid consisting of the factorial combination of 20 linearly spaced values of α on the interval [−.1, −1], along with 20 linearly spaced values of β on the interval [0, 2]. Subsets of each of these grids were used to generate the spectral shapes shown in Figures 3 and 4.

At each of the 400 points in the 20×20 parameter spaces, 2,000 independent time series of length 1,024 were simulated. For example, at the point $[\alpha_i, \beta_j]$ we simulated the fBmW(α_i, β_j) process repeatedly to yield 2,000 exemplar time series of length 1,024. Each series was then normalized and transformed to yield a high quality 8-point composite spectral representation (see Appendix C). The resulting ensemble of 2,000 power spectra were then used to estimate the ensemble-averaged power spectrum at the point $[\alpha_i, \beta_j]$, as well as an 8×8 covariance matrix. Together, the average spectrum and covariance matrix suffices to characterize the underlying spectral sampling distribution of fBmW(α_i, β_j) processes. The above procedure (i.e., estimating the average spectrum and the ensemble covariation) was repeated for each unique parameter combination defining the ARMA and fBmW families to yield a complete set of 800 spectral sampling distributions.

Classification Using Maximum Likelihood

The spectral classification framework we have developed consists of two components (see Appendix C for a complete description). The first is a reference library based on the 800 sampling distributions generated by the two candidate processes. This library encapsulates a relatively complete range of spectral shapes that may be observed in the two models. The second component is a rational decision structure based on maximum likelihood. This component uses the library to find the most likely source of an input data spectrum—namely, it decides whether the given data are more consistent with an autoregressive or a fractal interpretation. Recognizing that the priors for ARMA and fBmW are unknown, categorization accuracy is maximized by deciding in favor of the model with the single largest likelihood. The sensitivity and bias of the classifier are determined by explicit calculation of the hits (correctly classifying a long-range fractal as such) and false alarms (incorrectly classifying an ARMA as a fractal) for all parameter inputs to ARMA and fBmW. This approach to the problem of spectral classification is nearly optimal, given the inherent limitations imposed by sampling error.

An example will clarify the details involved in classification. Consider an input time series of length 1,024 drawn from either ARMA or fBmW. The time series is standardized and then transformed into an 8-point composite spectrum, using methods described in Appendix A. The classification problem now reduces to determining which ARMA and fBmW processes are most likely to have produced the input data, and this problem

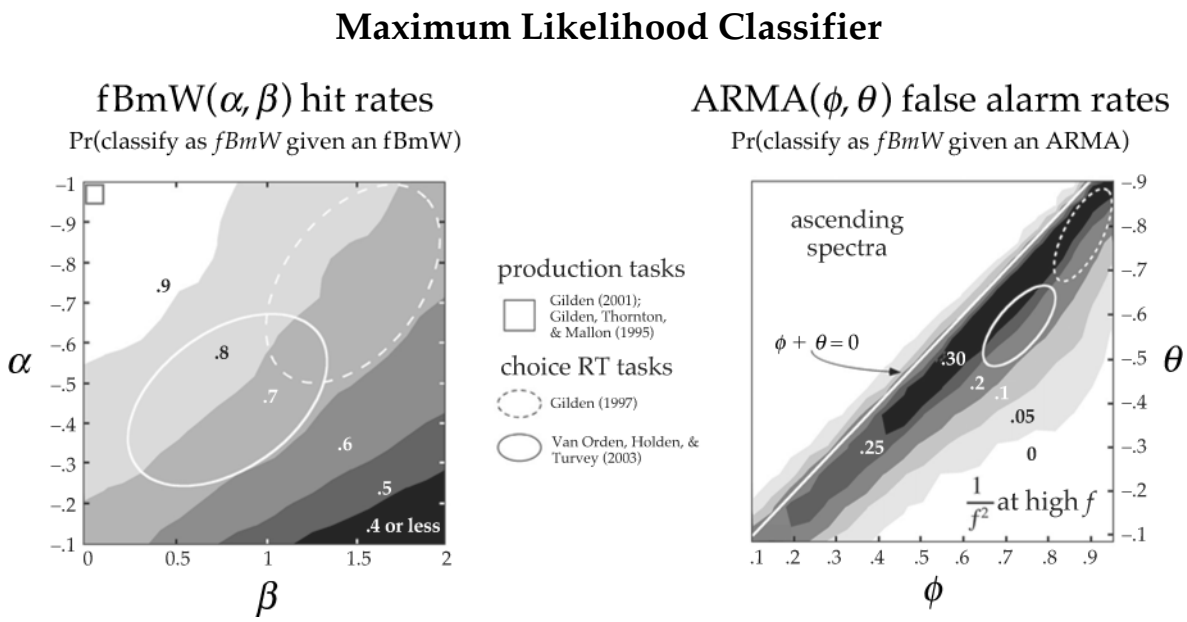

Figure 7. Accuracy of the spectral likelihood classifier in discriminating simulated exemplars of long- and short-range processes. The contour plot on the left shows classifier hit rates for correctly recognizing fBmW(α, β) time series. The hit rates associated with each contour are inset for reference and reflect the proportion of times the maximum likelihood classifier correctly assigned long-range provenance to fBmW time series as a function of the parameters α and β. The contour plot on the right shows the classifier's false alarm rate for incorrectly classifying time series generated by short-range ARMA(ϕ, θ) processes. The inset false alarm rates indicate the proportion of times the classifier mistook an ARMA time series for a fractal as a function of the parameters ϕ and θ. A total of 2,000 exemplar time series were simulated and classified at each of the 400 parameter combinations in each family.

is solved by a brute-force calculation of likelihood across the library of sampling distributions. There are 800 likelihoods, one at each of the 400 points defining the ARMA (ϕ, θ) space, and one at each of the 400 points defining the fBmW(α, β), space, that must be computed. There will be one likelihood that is greatest, and the family to which this belongs is assigned provenance. This procedure is fair, in the sense that each model is given its best opportunity to explain the data and the classifier always chooses blindly among competitors.

Calibration results. Deciding whether statements of provenance can be reliably made reduces to the question of how well the classifier performs on a test bed of decision problems. The test bed used in this work consisted of 2,000 independent time series at each of the 800 parameter points across the two process spaces (800,000 decisions in each space). This ensemble is sufficiently large to construct stable estimates of the hit rates (asserting *fBmW* when it is fBmW in fact) and the false alarm rates (asserting *fBmW* when it is ARMA in fact).

Figure 7 displays the calibration tests on the classifier. The left panel of Figure 7 shows a contour plot of the classifier's hit rate as a function of the fBmW parameters α and β (each line in the plot is an iso–hit rate contour). In general, the maximum likelihood classifier had little trouble in correctly recognizing a long-range spectrum. The median hit rate taken over the entire range of α and β is 78%. This number represents the overall power in correctly deciding that long-range structure is present. There are, however, portions of the parameter space where whitened fractals are confusable with short-range processes. Consider the regime where the fBmW is virtually white, where fractal exponents are small ($-.4 < \alpha < -.1$), and additive white noise is high ($\beta > 1.5$). Here, the classifier is correct in its fractal classifications only about 40% or less of the time, indicating that for those long-range processes that look *white* (i.e., nearly flat power spectra), there is a slight bias to categorize them as ARMA. This bias arises primarily because the ARMA processes that are confusable with these sorts of fractals actually have less spectral variability on the whole and so are generally preferred by the maximum likelihood classifier, all else being equal.

In the region of the fBmW space more characteristic of $1/f$ noise and psychophysical data (i.e., $\alpha < -.5$), the power of the classifier is observed to be much higher, exceeding 85% on average. These regions are depicted in the left panel of Figure 7 by a square for production data and by ellipses for RT data. Production data typically have linear spectra at low frequency and are easily discriminated from the whitening at the low frequency characteristic of the ARMA process. RT methods generate data that do have palpable amounts of white noise (Gilden, 1997, 2001; Van Orden, Holden, & Turvey, 2003), but even here the classifier misses only at a rate of about 20%.

In the right panel of Figure 7, we show a contour plot of the maximum likelihood classifier false alarm rates—

classifying a sequence as fBmW when ARMA is the source of data. Most apparent is that the false alarm rates are uniformly zero across the majority of the ARMA space. The lower triangle is dominated by spectra that decay as $1/f^2$ at high frequency, whereas the upper triangle contains spectra that increase with frequency. Neither of these regions can be modeled by a fractal process that is no more correlated than a $1/f$ noise, and this classifier has a library that is limited to $\alpha \in [-.1, -1]$. In fact, only a small subset of ARMA processes are even remotely confusable with fBmWs. The median false alarm rate taken across the subset of ARMA processes whose spectra decay with frequency is 8%. False alarm rates greater than this occur only along the relatively narrow ridge proximal to the major diagonal. This ridge contains sequences that are consistent with RT methods, as shown by the ellipses (there is no region in the ARMA plane that can produce the sort of data observed in production tasks—i.e., a pure $1/f$ noise—so there is no demarcation of production tasks here). Again, we observe that the classifier is about 80% accurate in its assessment of provenance for data that might be of practical concern. Thus, the classifier does not achieve high hit rates at the expense of a large false alarm rate.

Although in no other work has the confusability of fractal and autoregressive descriptions been examined with the detail attempted here, we can compare the performance of the spectral likelihood classifier with a related result from the work of Wagenmakers, Farrell, and Ratcliff (2004). There, long-range dependence was assessed at a single point in the fBmW space, using nested ARMA/ARFIMA models and exact maximum likelihood estimation. Specifically, a three-parameter ARFIMA was used to discriminate simulated exemplars of the fBmW(-1, 1) process from simulated exemplars of the ARMA(.896, $-.691$) nearest relative. They found a fractal hit rate of 74% and an ARMA false alarm rate of 8%. The spectral likelihood classifier produces a hit rate of 88% and a false alarm rate of 8% for these sequences. Although these differences in performance are not large, they do provide good evidence that assignments of provenance with spectral maximum likelihood is at least as powerful as, if not more so than, time domain techniques based on ARFIMA.

Bayesian Classification Based on Integrated Likelihood

A growing body of recent work on model selection has highlighted the need to augment local assessments of fit with more global metrics that take stock of the full predictive range of a model (see the articles in the special issue of the *Journal of Mathematical Psychology*, Myung, Forster, & Browne [Eds.], 2000; Navarro et al., 2004; Pitt et al., 2002; Rissanen, 1996; Roberts & Pashler, 2000). Here, a model's proximity to data, as well as its generalizability to new data sets, is used in selecting the best computational description.

This brings us to one of the principal strengths of the spectral classification approach we advocate here: It is general enough to incorporate virtually any other kind of information one might want to include in selecting among long- and short-range models, including global information about the generalizability of the models. In particular, with a relatively minor change in the decision metric, the maximum likelihood classifier becomes a proper Bayesian classifier that uses marginal likelihoods and parameter priors to estimate which model is more probable, given the data. This extension to Bayesian selection is possible because we have access not only to the single maximum likelihood estimate, but also to the full set of likelihoods of data across the entire fBmW and ARMA parameter spaces. This means that we can easily go from assignments of provenance based on the mode of the likelihood distributions (i.e., max-likelihood) to assignments based on the integrated distribution, with negligible change to the core framework.

In the Bayesian selection framework, we compute the integrated likelihood of the data, given each model, as

$$L(\text{data}|\text{model}) = \int_{x,y} \text{prior}(x, y) L(\text{data}|x, y), \quad (10)$$

where x and y denote particular values of the process parameters. The likelihood L of the data at each point $[x, y]$ in the model's parameter space is weighted by the corresponding prior probability of that parameter combination. The left-hand term is the integrated likelihood of the data, given the model, and is simply the prior-weighted sum of likelihood taken over the entire space of parameter values. In this method of selection, the *best* model for a data set is that which maximizes Equation 10 (under the assumption of equal model priors). The Bayesian metric differs from maximum likelihood classification, where decisions are made solely on the basis of each model's single maximum likelihood (taken over all x, y):

$$\arg\max_{x,y} [L(\text{data}|x, y)], \quad (11)$$

where "arg max" is just a shorthand description for an algorithm that finds the maximum likelihood across each model's parameter space. Quite simply, decisions in Bayesian classification are made via a comparison of weighted means, whereas decisions in maximum likelihood are made via a comparison of maxima.

We denote a classifier that computes Equation 10 as a *Bayesian classifier*. Bayesian classifiers are generally preferable to maximum likelihood classifiers because they are sensitive to differences in parameter number and model complexity (see Myung, 2000; Pitt et al., 2002; Wasserman, 2000). Although the short- and long-range models we examine here do not differ in their parameter number, they do differ in complexity, and recent work has emphasized the importance of using global estimates of complexity in selecting among competing models (Pitt et al., 2002; Rissanen, 1996; Wasserman, 2000).

The kind of complexity that is relevant here is related to *falsifiability* and measures the descriptive breadth and scope of a computational model (Cutting, Bruno, Brady, & Moore, 1992). Highly flexible models are hard to fal-

sify because they can take whichever shape a particular data set demands. Protean models that generate a large number of distinguishable predictions have higher intrinsic complexity than do models that produce a more constrained set of predictions. By this view, *good* models are ones with low complexity, because they are falsifiable, robust to perturbation, and tend to generally predict only the specific forms observed in data (Roberts & Pashler, 2000). This is exactly the point we attempted to make when we showed the full range of spectral shapes achievable by the two-parameter ARMA and fBmW families. In these plots, it is evident that the fBmW is not a complex model, whereas the ARMA is.

One attractive feature of a Bayesian classifier is that it allows us to adjust for differences in complexity (Myung & Pitt, 1997; Wasserman, 2000) in models that are intrinsically stochastic (e.g., simulation-dependent modeling). All else being equal, a Bayesian metric based on integrated likelihood (Equation 10) will tend to favor lower complexity models. This arises because low-complexity models generate broad distributions of likelihood (shape changes slowly with parameter variation), whereas complex models generate more peaked distributions of likelihood (shape changes quickly with parameter variation). Although highly peaked likelihood distributions may be advantageous from the point of view of parameter estimation (the peakedness of the distributions are related to the confidence region), they are the defining feature of an overly complex model (Hochreiter & Schmidhuber, 1997), especially when they occur in the context of a large number of diverse and distinguishable predictions (Pitt et al., 2002).

In the context of discriminating short- and long-range models of data, these points imply that Bayesian selection based on Equation 10 will effectively penalize the more complex ARMA model. Fits of the ARMA to typical data sets generate nonnegligible likelihood over only a very small region of the parameter space. In contrast, fits of the fBmW to data generate palpable likelihoods over a fair fraction of configurations (see Figure 5B). Because the ARMA fits are nongeneric to the process and lead to highly confined distributions of likelihood in the parameter space, a proper Bayesian classifier will exhibit a general bias to make long-range assignments. This bias is not a shortcoming; it is evidence that the Bayesian metric is sensitive to the intrinsic differences in complexity associated with the ARMA and the fBmW models.

Calibration results for the Bayesian classifier. In order to facilitate comparisons, we have calibrated the Bayesian classifier on the same families of process that were used to calibrate the maximum likelihood classifier. Rather than using the single largest likelihood to assign provenance, here we estimate the integrated likelihood given each power spectrum by taking a weighted sum over all likelihoods in each model's respective parameter spaces (see Equations C4 and C5 in Appendix C).

Calibration results using this metric are shown in Figure 8. The fBmW hit rates (correctly assigning long-range status to long-range sequences) and ARMA false alarm rates (incorrectly assigning long-range status to short-

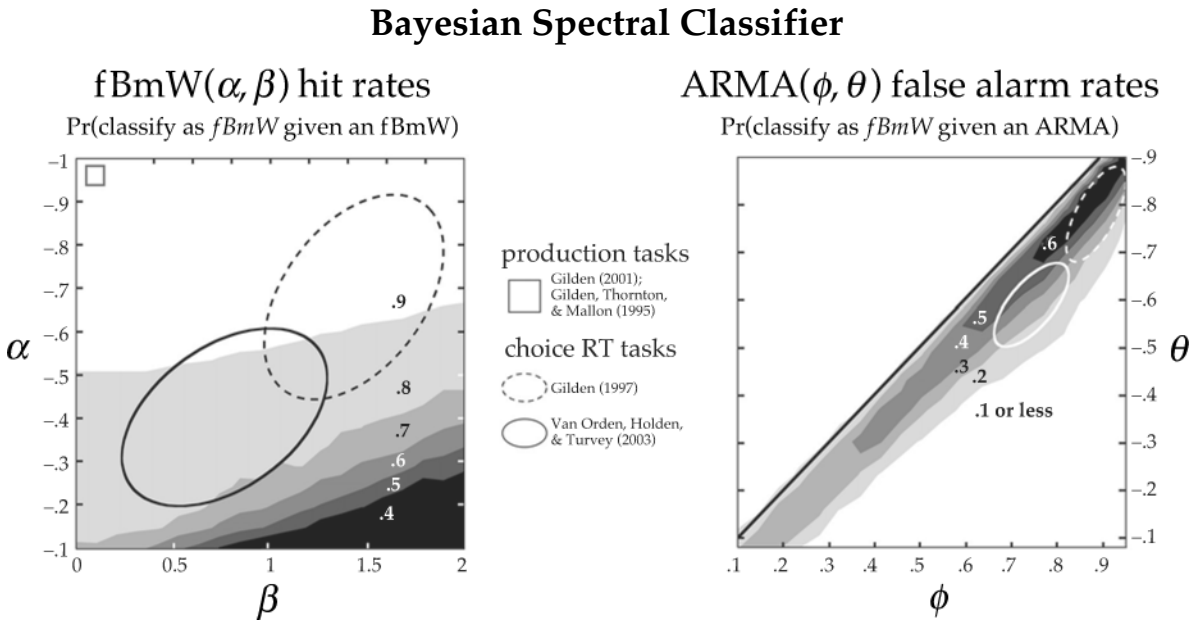

Figure 8. Accuracy of the Bayesian integrated likelihood classifier in discriminating simulated exemplars of long- and short-range processes. The contour plot on the left shows classifier hit rates for correctly recognizing fBmW(α, β) time series; the contour plot on the right shows the classifier's false alarm rates for incorrectly classifying time series generated by short-range ARMA(ϕ, θ) processes. A total of 1,000 exemplar time series were simulated and classified at each of the 400 parameter combinations in each family.

range sequences) are depicted in adjacent panels as iso-accuracy contour plots. The most striking feature of this figure is that the results under Bayesian classification are qualitatively indistinguishable from the maximum likelihood results in Figure 7. Both classifiers have no trouble discriminating the vast majority of short- and long-range processes. Performance is really seen to deviate from high accuracy only in the portion of the fBmW parameter space where white noise dominates the long-range fBm signal and in a limited region of the ARMA parameter space where autoregression and differencing variables are perfectly balanced.

Although the maximum likelihood and the Bayesian classifiers are largely congruent in sensitivity, the calibrations do indicate some quantitative differences in their respective patterns of hits and false alarms. As the inset accuracies in Figure 8 show, when integrated likelihood is the basis for classification, there are model-wide increases in both the fBmW hit rates and the ARMA false alarm rates. The rise in false alarms is particularly pronounced in the isolated region of the ARMA parameter space where short-range processes most closely mimic global aspects of long-range spectra. The classifier's increased tendency to assign a long-range status to these sequences reflects the fact that in those parameter regimes in which the ARMA process looks most like the fBmW (i.e., where ϕ is high and just a little bigger than $|\theta|$), there will naturally be a strong and broad distribution of likelihoods produced in the fBmW model. Under Bayesian classification, this signal swamps the integrated likelihood in the ARMA space, because these kinds of short-range spectra are more like the generic fBmW than the generic ARMA (the majority of ARMA configurations generate near-zero likelihood to such spectra). This state of affairs leads the classifier to increasingly reject the correct ARMA interpretation and to assign provenance to the model with the higher volume of likelihood. We argue that these kinds of errors do not represent failures but, rather, reflect an optimal strategy in which model complexity is being used to penalize the overly complex ARMA (Myung & Pitt, 1997). Because the maximum likelihood spectral classifier does not take into account complexity (i.e., the peakedness and breadth of the likelihood distribution in the parameter space), it does not show this bias, and its assignments are made primarily on the basis of goodness of fit. Although the Bayesian classifier does have a tendency to false alarm to certain short-range configurations, its overall accuracy is never compromised.

This is shown in Figure 9, where we compare d's under Bayesian and maximum likelihood classification for discriminating the family of fBmW processes from a corresponding set of ARMA mimics. The line along the major diagonal has unit slope and represents the point of d' equality. It is evident from the plot that even though the hit and false alarm rates differ across classifiers, there is little observable change in true discrimination sensitivity. Again, this is what we would expect if the Bayesian classifier was simply penalizing the more complex ARMA model by shifting its decision criterion so as to favor the low-complexity fBmW. Such a shift predicts a change in hit rates and false alarm rates with no reliable alteration of d', and this is precisely what we see in the figure.

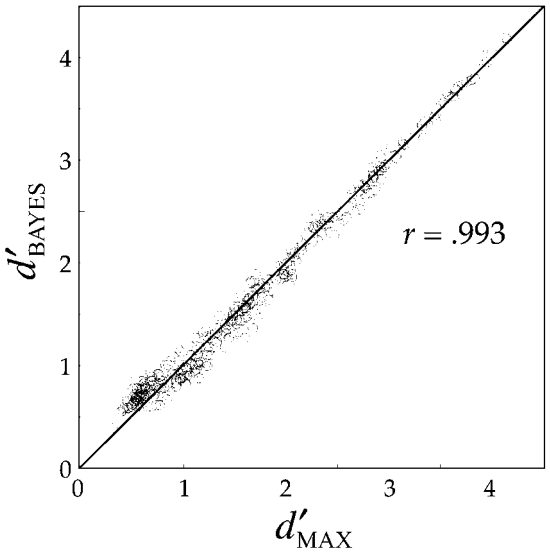

Figure 9. Comparison of classifier sensitivity for discriminating each of the 400 members of the fBmW(α, β) family from its own matched ARMA process (see the text for details). The sensitivity of the maximum likelihood classifier, d'_{MAX}, is plotted on the abscissa; the sensitivity of the integrated likelihood classifier, d'_{BAYES}, is plotted on the ordinate.

Which Spectral Classifier to Use?

Now that we have fully calibrated the maximum likelihood and Bayesian classifiers on simulated time series of known origin, it is natural to ask which of the two metrics is to be preferred in the analysis of empirical data sets. Given the nearly equivalent sensitivity of the two classifiers in discriminating short- from long-range processes, the short answer is that it depends on the goal at hand. In this regard, we provide the following heuristics simply as guidelines for the appropriate uses of each classifier.

Maximum Likelihood Spectral Classifier

Largely unbiased over the full ranges of short- and long-range processes; equates Type I and Type II errors.

Appropriate for estimation of parameter values.

Bayesian Spectral Classifier

Incorporates differences in complexity (and parameter number) and is, therefore, biased against protean, overly flexible models. Offers a natural counterpoint to current usage of ARFIMA (Wagenmakers, Farrell, & Ratcliff, 2004), where calibrations suggest a bias favoring short-range interpretations.

Is slightly more powerful in high-noise regimes (e.g., when α is near zero and β is high; see Figure 9, where $d'_{MAX} < 1$).

Admits other sources of prior information that might inform selection.

A Case Study: Analysis of Naming Data From Van Orden, Holden, and Turvey, 2003

The necessity for a solid data-analytic framework for time series analysis is most evident in the controversy generated by the work of Van Orden, Holden, and Turvey (2003). In that article, Van Orden and colleagues analyzed fluctuations in word-reading latency. Specifically, they were interested in the question of whether these latencies are complex—that is, whether they form $1/f$ noise. Their spectral analyses showed that naming latency power spectra were distinguishable from white noise in terms of slopes estimated from regressions of log power against log frequency. On the basis of this result and other converging analyses in the time domain (dispersion analyses of the fractal dimension; Caccia, Percival, Cannon, Raymond, & Bassingthwaighte, 1997; Eke, Herman, Kocsis, & Kozak, 2002), they concluded that they had found evidence for fractal structure.

Wagenmakers, Farrell, and Ratcliff (2004, 2005) had a radically different interpretation of these data. They reanalyzed Van Orden, Holden, and Turvey's (2003) data in terms of a variety of long- and short-range autoregressive models, including the second-order ARMA. These analyses indicated that although a majority of the sequences were best fit by one of the long-range ARFIMA models in their ensemble, the class of short-range ARMA models remained competitive. This led them to reject the earlier claims that reading latencies were unequivocally $1/f$ noise. In their own words: "[there was] some support for the existence of persistent serial correlations. However, this support [did] not appear to be very strong . . . . The analyses reported by [Van Orden, Holden, & Turvey, 2003] do not support their claim of $1/f^{\alpha}$ noise" (Wagenmakers et al., 2005). So which characterization of the Van Orden data is correct? Our methods can be used to inform the issue.

We present three analyses of Van Orden, Holden, and Turvey's (2003) data in order to illustrate the different ways spectral likelihood classification may be used to decide provenance and to settle the provenance issue, relative to whitened fractals and first-order ARMAs. We begin with an analysis based on classification of individual observer data. This involves a simple enumeration of the sequences in Van Orden, Holden, and Turvey's (2003) sample that are more likely to have been derived from a long-range fractal process. We then present a more powerful analysis in which classification is based on the observer-averaged spectrum. The final analysis offers an alternative method for combining evidence across observers.

Method. Van Orden, Holden, and Turvey's (2003) data were derived from 20 observers, each of whom provided a single uninterrupted sequence of 1,024 word-naming RTs (see Van Orden, Holden, & Turvey, 2003, for details). Prior to analysis we standardized and spectrally transformed each sequence, using the composite routines in Appendix A. In the following analyses, we report results using both raw and detrended versions of the data in which overall linear trends were removed prior to analysis (the results for the detrended series will be given within brackets). Detrending had little appreciable effect on any of the analysis outcomes.

Analysis 1: Classification of single-observer sequences. We visually examined the set of 20 power spectra in Van Orden, Holden, and Turvey's (2003) sample and found that a small fraction of the sequences were virtually indistinguishable from white noise (this was verified using linear regression). Recognizing that flat spectra provide no opportunity for deciding the issue of short- versus long-range correlation and that the spectral classifiers have a bias to assign a short-range status to such sequences, we excluded them from our single-sequence analyses. This left a total of 16 remaining sequences consistent with at least some type of serial correlation. These were submitted to spectral likelihood classification, using the algorithm detailed in the Maximum Likelihood Algorithm section in Appendix C. Of the 16 spectrally colored sequences, 14 [12] were described best as whitened fractals. These results provide preliminary evidence that Van Orden, Holden, and Turvey's (2003) word-reading data are generally long range in character.

An issue that reoccurs in these kinds of analyses is the appropriateness of detrending prior to classification. This issue is outside the purview of the classifier, because it depends critically on whether one believes large-scale fluctuations in data are the result of secular trends. It is standard practice in the physical sciences to remove linear trends as a caution. In domains in which there is theoretical cause or experience with similar data to suggest higher order contaminants in a signal (e.g., cyclical drift over time due to the environment, etc.), these trends should also be removed to ensure proper characterization of the fluctuations. Perhaps similar kinds of concerns motivated Van Orden, Holden, and Turvey (2003) to choose to quadratically detrend their naming latencies prior to analysis. Nonetheless, in the problem domain that we face of distinguishing highly confusable ARMA mimics from long-range fractals, quadratic detrending introduces clear dangers to interpetation (see the Detrending Time Series section in Appendix B for a treatment of the effects of detrending in this context). In particular, this practice is extremely ill suited for spectral likelihood classification, because it makes all sequences look short range. We know of no theory of latency that predicts such structures, and we have never observed any evidence of systematic quadratic contaminants in our laboratory. Unless there is good reason to believe that higher order secular effects are corrupting one's data, the best practice is to limit trend removal to first order.

Bayesian classification. In addition to classification by maximum likelihood, we also analyzed the 16 Van Orden,

Holden, and Turvey (2003) sequences by using the Bayesian integrated likelihood classifier. For this analysis, we used prior distributions as were used in the earlier calibrations of the classifier (we also explored a variety of alternative prior distributions, and our results were robust to these variations; for more here, see the Parameter Priors section in Appendix C). Of the 16 sequences showing some form of correlation, 15 [12] were assigned a long-range provenance under Bayesian classification.

Analysis 2: Observer average classification. Observer averaging offers a straightforward manner for reducing the variance of spectral estimates and so leads to improvements in the accuracy of our classifications (see Figure B1, Appendix B). Preliminary visual inspection of the average spectrum of Van Orden, Holden, and Turvey's (2003) data revealed the unmistakable signature of an fBmW process: a bowed spectrum that saturates in white noise at the high frequencies and rises continuously as the trial windows become ever larger. A subsequent quantitative analysis using the maximum likelihood classifier indicated that the average spectrum (over all 20 sequences) was some 20 [31] times more likely to have been generated by a whitened fractal than by a short-range ARMA. These results are consistent in magnitude and interpretation with data sets collected in our laboratory using similar methods (Gilden, 1997, 2001).

Analysis 3: Combining log-likelihoods. We supplement the previous two analyses with a final analysis in which model evidence is combined across the set of single-observer results (we thank J. Busemeyer for this suggestion). Because each observer provides logically independent evidence for one model over the other, a principled combination rule is to sum the individual log-likelihood ratios across observers (Jaynes, 2003), treating the ratios as approximate log Bayes factors (i.e., log odds favoring one model over the other; see Wasserman, 2000). The extent to which this sum is positive indicates that the combined evidence favors a fractal interpretation (similarly, when this sum is negative, the evidence will favor an ARMA interpretation). For the 20 time series collected by Van Orden, Holden, and Turvey (2003), the spectral likelihood classifier generates a combined log-likelihood ratio of 21 [15] (exponentiation gives the combined odds). This value clearly favors the whitened fractal model over the short-range ARMA.

Taken together, these three analyses provide converging evidence that the most likely provenance of Van Orden, Holden, and Turvey's (2003) data is whitened $1/f$ noise. The majority of observers who showed some form of serial correlation had data sequences classified as long range (14/16 with maximum likelihood, 15/16 with integrated likelihood); the averaged spectrum was approximately 20 times more likely to have been generated by a fractal process, and the combined log-likelihood ratio taken over all sequences provides clear and overwhelming evidence for long-range correlation in the data.

The verdict: How ARFIMA stacks up to spectral likelihood. There is little distinction at the level of the single observer. Single-participant analysis is intrinsically limited in terms of statistical power and thus represents the worst-case domain for distinguishing the nature of observed fluctuations in data. With spectral likelihood classification, we found that 14 of 16 sequences were assigned a long-range provenance, as compared with 12 of 20 sequences under ARFIMA testing. Insofar as the 4 *white noise* sequences excluded from our analysis were most likely given a null short-range assignment by ARFIMA (Wagenmakers et al., 2005), the argument comes down to 14 versus 12—that is, an immaterial distinction.

One of the principal advantages of our approach, however, is that it is not limited to the worst-case domain in which there is no averaging and single spectra are the unit of analysis. By virtue of spectral averaging, both within and across observers, our methods move substantially beyond current ARFIMA tests to reveal clear and strong evidence favoring the long-range interpretation of Van Orden, Holden, and Turvey's (2003) data (see Analysis 2). Although averaging over observers always carries the caveat that it may distort our view of the underlying processes, we will show in Appendix B that even under realistic levels of process heterogeneity, classification of averaged spectra remains robust and unbiased and offers a preferred means for increasing the statistical power of our classifications. We suspect that the primary reason observer averaging is downplayed in the application of ARFIMA methods is that the canned time domain algorithms required for these analyses simply do not permit such reduction (see the Ox and R routines in Doornik, 2001, and the references in Wagenmakers, Farrell, & Ratcliff, 2004, 2005).

The spectral likelihood classifier enjoys another practical advantage over previous ARFIMA tests, in that it has been thoroughly calibrated. We have constructed and tested both the maximum likelihood and the Bayesian classifiers, using two complete families of ARMA and fBmW processes. One goal of these calibrations is to provide a clear picture of where confusability among the competing models is greatest and where, as researchers, we should concentrate the bulk of our empirical efforts. In contrast, the ARFIMA method remains largely uncalibrated on the range of processes relevant to psychophysical data (it was calibrated on a single short- and long-range process in Wagenmakers, Farrell, & Ratcliff, 2004).

CONCLUSIONS

The evidence is that whitened fractals provide a better description of psychophysical data than do ARMA processes.

The evidence from a collection of studies incorporating RT measurement is that the ARMA process is not causal of the observed correlations. The volume of the ARMA parameter space that generates processes resembling data is almost zero. The best-fitting ARMA spectra are, in fact, poor fits, because they are S-shaped and the data are not. The data, rather, appear to be fractal, in

the sense that the power increases through the range of low frequencies, evidence for the absence of a temporal scale. We do not view the rejection of the ARMA as explanatory as the rejection of a null, but as the rejection of a model that is just the wrong model. Since there is little evidence for autoregression in the relevant data, we see little motivation to extend the ARMA into ARFIMA and conduct inference testing on the basis of the significance of an additional long-range parameter. Rather, we prefer to conduct the entire discussion in terms of fractional Brownian motions, which are long-range by design, are routinely used in a variety of scientific literatures, and have growing empirical support (Gilden, 1997, 2001; Gilden et al., 1995).

Although Wagenmakers, Farrell, and Ratcliff (2004, 2005) have provided the first published reports in which ARMA as a competing hypothesis has been seriously considered, we have been questioned informally on a number of occasions as to whether fractal descriptions are appropriate when a "simpler" first-order ARMA process might explain our data (Gilden, 2001) equally well. First-order ARMA processes may in some restricted sense be simple, but more to the point, they are familiar to practitioners within the field of time series analysis. ARMAs have the virtue that they can be written in closed form, they lend themselves to an elegant mathematical formalism, and they are invariably the first process encountered in the practical matter of learning about time series. We suspect that it is their familiarity and the ease with which they may be incorporated into existing models that underlie their appeal, notwithstanding the observation that they have no prior claim on psychophysics. We note that the ARMA apparently has no claim in the posterior sense either. ARMA/ARFIMA testing has replicated virtually all of the earlier research findings (Wagenmakers, Farrell, & Ratcliff, 2004), strengthening the claim that long-range fractal processes are the best explanation for the fluctuations that characterize psychological time series.

With good tools, competing short- and long-range descriptions of data are not that confusable.

We have created a powerful general purpose classifier, based on Bayesian inference, for tackling the difficult problem of discriminating long-range fractals from short-range ARMAs. The classifier capitalizes on a high-quality estimate of the power spectrum and an extensive set of spectral sampling distributions to compute the most likely provenance of time series data. By testing the classifier across a broad array of simulated ARMA and fractal time series, we demonstrated it to be highly sensitive in distinguishing short- from long-range data. This method is greatly preferred to analysis based only on spectral slope and is generally more flexible and transparent than using ARFIMA software and AIC penalties. Most important, in process regimes characteristic of typical production and choice RT data, the classifier achieved accuracies of 75%–95% correct in both hits and correct rejections. These numbers indicate that even specially tailored ARMA configurations that globally approximate fractal structure can be rejected when local features in the power spectrum are used in classification. We applied the classifier to a controversial set of word-reading data (Van Orden, Holden, & Turvey, 2003) and convincingly demonstrated that these time series are most likely fractal in origin.

We close by emphasizing the fact that there is a large and growing body of evidence that sequences of psychophysical data fluctuate as $1/f^{\alpha}$ noises. To date, long-range fractal fluctuations have been discovered in a diverse set of measurement domains, including signal detection and discrimination (Gilden & Wilson, 1995a), skilled motor performance (Chen, Ding, & Kelso, 1997; Gilden, 2001; Gilden & Wilson, 1995b), production of spatiotemporal intervals and force magnitudes (Gilden, 2001; Gilden et al., 1995), and both simple and choice RT (Gilden et al., 1995; Van Orden, Holden, & Turvey, 2003; Wagenmakers, Farrell, & Ratcliff, 2004). As the empirical evidence supporting long-range fluctuations continues to build, repeated verification of the phenomenon through statistical means will become increasingly unnecessary. The principal question that faces future research is not whether $1/f$ noise exists in cognitive activity but, rather, what its presence has to say about the structure of the mind.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Bak, P. (1990). Self-organized criticality. *Physica A*, **163**, 403-409.

Bak, P. (1992). Self-organized criticality in non-conservative models. *Physica A*, **191**, 41-46.

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of $1/f$ noise. *Physical Review Letters*, **59**, 381-384.

Bak, P., Tang, C., & Wiesenfeld, K. (1988). Self-organized criticality. *Physical Review A*, **38**, 364-374.

Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, **37**, 1-16.

Benedetti, F., Barbini, B., Colombo, C., Campori, E., & Smeraldi, E. (1996). Infradian mood fluctuations during a major depressive episode. *Journal of Affective Disorders*, **41**, 81-87.

Bertelson, P. (1963). S–R relationships and reaction time to new versus repeated signals in a serial task. *Journal of Experimental Psychology*, **65**, 478-484.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, **108**, 624-652.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, **35**, 11-21.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, **121**, 177-194.

Caccia, D. C., Percival, D. B., Cannon, M. J., Raymond, G. M., & Bassingthwaighte, J. B. (1997). Analyzing exact fractal time series: Evaluating dispersional analysis and rescaled range methods. *Physica A*, **246**, 609-632.

Cannon, M. J., Percival, D. B., Caccia, D. C., Raymond, G. M., & Bassingthwaighte, J. B. (1997). Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A*, **241**, 606-626.

Chen, Y., Ding, M., & Kelso, J. A. S. (1997). Long memory processes ($1/f^{\alpha}$ type) in human coordination. *Physical Review Letters*, **79**, 4501-4504.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364-381.

Davies, R. B., & Harte, D. S. (1987). Tests for the Hurst effect. *Biometrika*, **74**, 95-101.

Devaney, R. L. (1992). *A first course in chaotic dynamical systems: Theory and experiment*. Reading, MA: Addison-Wesley.

Diebold, F. X., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, **105**, 131-159.

Doornik, J. A. (2001). *Ox: An object-oriented matrix language.* London: Timberlake Consultants.

Dorfman, D. D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology*, **8**, 264-284.

Eke, A., Herman, P., Kocsis, L., & Kozak, L. R. (2002). Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement*, **23**, 1-38.

Feder, J. (1988). *Fractals*. New York: Plenum.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, **44**, 205-231.

Gardner, M. (1978). Mathematical games: White and brown music, fractal curves and one-over-*f* fluctuations. *Scientific American*, **238**, 16-32.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, **8**, 296-301.

Gilden, D. L. (2001). Cognitive emissions of 1/*f* noise. *Psychological Review*, **108**, 33-56.

Gilden, D. L., Schmuckler, M. A., & Clayton, K. (1993). The perception of natural contour. *Psychological Review*, **100**, 460-478.

Gilden, D. L., Thornton, T., & Mallon, M. (1995). 1/*f* noise in human cognition. *Science*, **267**, 1837-1839.

Gilden, D. L., & Wilson, S. G. (1995a). On the nature of streaks in signal detection. *Cognitive Psychology*, **28**, 17-64.

Gilden, D. L., & Wilson, S. G. (1995b). Streaks in skilled performance. *Psychonomic Bulletin & Review*, **2**, 260-265.

Giraitis, L., Kokoszka, P., & Leipus, R. (2001). Testing for long memory in the presence of a general trend. *Journal of Applied Probability*, **38**, 1033-1054.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, **14**, 227-238.

Granger, C. W. J., & Hatanaka, M. (1964). *Spectral analysis of economic time series*. Princeton, NJ: Princeton University Press.

Granger, C. W. J., & Morris, M. (1976). Time series modeling and interpretation. *Journal of the Royal Statistical Society A*, **139**, 246-257.

Hale, D. J. (1967). Sequential effects in a two-choice serial reaction task. *Quarterly Journal of Experimental Psychology*, **19**, 133-141.

Handel, P. H., & Chung, A. L. (Eds.) (1993). *Noise in physical systems and 1/f fluctuations*. New York: American Institute of Physics.

Hausdorff, J. M., & Peng, C.-K. (1996). Multi-scaled randomness: A possible source of 1/*f* noise in biology. *Physical Review E*, **54**, 2154-2157.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185-207.

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, **9**, 1-42.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, **68**, 165-176.

Jaynes, E. T. (2003). *Probability theory*. Cambridge: Cambridge University Press.

Karl, J. H. (1989). *An introduction to digital signal processing*. San Diego: Academic Press.

Keeler, J. D., & Farmer, J. D. (1986). Robust space–time intermittency and 1/*f* noise. *Physica D*, **23**, 413-435.

Keller, J. M., Crownover, R. M., & Chen, R. Y. (1987). Characteristics of natural scenes related to the fractal dimension. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **9**, 621-627.

Laming, D. (1968). *Information theory of choice reaction times.* New York: Academic Press.

Landau, L. D., & Lifshitz, E. M. (1958). *Statistical physics* (E. Peierls & R. F. Peierls, Trans.). London: Pergamon.

Li, W. (2003). *A bibliography on 1/f noise*. Manhasset, NY: Robert S. Boas Center for Genomics and Human Genetics. Available at http://www.nslij-genetics.org/wli/1fnoise/.

Link, S. W. (1975). The relative judgment theory of two-choice response time. *Journal of Mathematical Psychology*, **12**, 114-135.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental operations*. New York: Oxford University Press.

Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A. F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics*, **32**, 397-408.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, **61**, 354-374.

Maljkovic, V., & Nakayama, K. (2000). Priming of popout: III. A short-term implicit memory system beneficial for rapid target selection. *Visual Cognition*, **7**, 571-595.

Mandelbrot, B. B. (1983). *The fractal geometry of nature*. San Francisco: Freeman.

Mandelbrot, B. B. (1985). Self-affine fractals and fractal dimension. *Physica Scripta*, **32**, 257-260.

Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *Society for Industrial & Applied Mathematics Review*, **10**, 422-437.

Maylor, E. A., Chater, N., & Brown, G. D. A. (2001). Scale invariance in the retrieval of retrospective and prospective memories. *Psychonomic Bulletin & Review*, **8**, 162-167.

Miller, S. L., Miller, W. M., & McWhorter, P. J. (1993). Extremal dynamics: A unifying physical explanation of fractals, 1/*f* noise, and activated processes. *Journal of Applied Physics*, **73**, 2617-2628.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.

Myung, I. J., Forster, M. R., & Browne, M. W. (Eds.) (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, **44**.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, **49**, 47-84.

Pagano, M. (1974). Estimation of models of autoregressive signal plus white noise. *Annals of Statistics*, **2**, 99-108.

Pashler, H., & Baylis, G. C. (1991). Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 33-48.

Peitgen, H., & Saupe, D. (Eds.) (1988). *The science of fractal images.* New York: Springer-Verlag.

Peng, C.-K., Buldyrev, S., Havlin, S., Simons, M., Stanley, H., & Goldberger, A. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, **49**, 1685-1689.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

Pomeau, Y., & Manneville, P. (1980). Intermittent transition to turbulence in dissipative dynamical systems. *Communications in Mathematical Physics*, **74**, 189-197.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes* (2nd ed.). Cambridge: Cambridge University Press.

Pressing, J. (1999). Sources for 1/*f* noise effects in human cognition and performance. In *Proceedings of the 4th Conference of the Australian Cognitive Science Society*. Available at http://www.geocities.com/padeusis/n1ciss.html.jp.

Pressing, J., & Jolley-Rogers, G. (1997). Spectral properties of human cognition and skill. *Biological Cybernetics*, **76**, 339-347.

Priestley, M. B. (1981). *Spectral analysis and time series*. London: Academic Press.

Rabbitt, P. M. A. (1968). Repetition effects and signal classification strategies in serial choice-response tasks. *Quarterly Journal of Experimental Psychology*, **20**, 232-240.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59-108.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, **9**, 347-356.
Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40-47.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, **107**, 358-367.
Schroeder, M. (1991). *Fractals, chaos, power laws: Minutes from an infinite paradise*. New York: Freeman.
Smith, M. C. (1968). Repetition effect and short-term memory. *Journal of Experimental Psychology*, **77**, 435-439.
Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research*, **44**, 1297-1320.
Smith, P. L., & Vickers, D. (1989). Modeling evidence accumulation with partial loss in expanded judgment. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 797-815.
Staddon, J. E., King, M., & Lockhead, G. R. (1980). On sequential effects in absolute judgment experiments. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 290-301.
Thornton, T. (2002). *Attentional limitation and multiple target visual search*. Unpublished doctoral dissertation, University of Texas, Austin.
Thornton, T. L., & Gilden, D. L. (2005). *What can be seen in a glance?* Manuscript submitted for publication.
Tong, H. (2001). A personal journey through time series in Biometrika. *Biometrika*, **88**, 195-218.
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550-592.
Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, **132**, 331-350.
Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and 1/*f* scaling. *Journal of Experimental Psychology: General*, **134**, 117-123.
Van Orden, G. C., Moreno, M. A., & Holden, J. G. (2003). A proper metaphysics for cognitive performance. *Nonlinear Dynamics, Psychology, & Life Sciences*, **71**, 49-59.
Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, **44**, 273-282.
Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin & Review*, **11**, 579-615.
Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2005). Human cognition and a pile of sand: A discussion on serial correlations and self-organized criticality. *Journal of Experimental Psychology: General*, **134**, 108-116.
Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, **48**, 28-50.
Ward, L. M. (2002). *Dynamical cognitive science*. Cambridge, MA: MIT Press.
Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92-107.
Welch, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. In D. G. Childers (Ed.), *Modern spectrum analysis* (pp. 17-20). New York: IEEE Press.
West, B. J., & Shlesinger, M. F. (1989). On the ubiquity of 1/*f* noise. *International Journal of Modern Physics*, **B3**, 795-819.
Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, **25**, 731-739.

NOTES

1. In West and Shlesinger's (1989) theory, 1/*f* noises are derived from a hierarchy of short-range relaxation processes in which the time scales are distributed log-normally. The theoretical question here involves isolating the characteristics of systems that might be expected to generate the multiplicative probabilistic structures that lead to log-normals (the log of a product is a sum in the log quantities, and the central limit theorem entails the normal). Another form of averaging is based on the theory of extremal dynamics (Miller, Miller, & McWhorter, 1993). The latter theory is framed quite generally in terms of system templates that are relevant whenever the underlying assumptions concerning extreme value statistics are justified.

2. The first-order AR(ϕ) process (Equation 4) may be written recursively as a weighted sum of random inputs:

$$O_t = \varepsilon_t + \sum_{k=-\infty}^{t-1} \phi^{t-k} \varepsilon_k ,$$

where $|\phi| < 1$ is required for stationarity. There is a similar summation for the first-order ARMA(ϕ, θ) process (Equation 6):

$$O_t = \varepsilon_t + (\phi + \theta)\left(\varepsilon_{t-1} + \sum_{k=-\infty}^{t-2} \phi^{t-k-1} \varepsilon_k \right).$$

Two important points are revealed by this reexpression. First, we see that the ARMA(ϕ, θ) process can be recast as the sum of a first-order AR process and a source of uncorrelated noise (the right-hand term in parentheses describes an AR process; ε_t is the noise; see Pagano, 1974, and Granger & Morris, 1976, for a rigorous development of this equality). Second, whenever $\phi = -\theta$, the ARMA reduces to a random process with independent increments—a white noise.

3. Personal communications from several readers of this article and of our earlier work have been concerned with the existence of a number of RT studies in which latency sequences were fit by models of short-range process. These studies, it has been argued, provide the necessary rationale to justify the use of the ARMA as a null for psychological time series. A thorough search of the literature reveals only a single potentially relevant data set. In a series of experiments on local priming of RT (Laming, 1968; see also Botvinick, Braver, Barch, Carter, & Cohen, 2001), Laming made the informal observation that many of his sequences appeared to have exponentially decaying serial correlations. However, he did not conduct any tests on the form of this decay, and his time series were unusually short (100–200 trials) and so have little statistical power with which to distinguish exponential decay from power law decay. We have examined the relevant figures in some detail and have found, rather, that the decay appears to be more consistent with a long-range process. In particular, the exponential decay model offered by Botvinick et al. simply does not faithfully reproduce the observed autocorrelation function.

4. To be fair, we point out that there is nothing in the general approach of ARFIMA testing that requires a nested model relationship or the use of AIC selection. In fact, in subsequent related work, Wagenmakers et al. (2005) have relaxed the nesting relationship in order to examine comparisons between various order ARFIMA and various order ARMA parameterizations, using two different model selection metrics (i.e., AIC and BIC). Importantly, the removal of the nesting relationship means that the short-range ARMA models no longer have a default claim on the data.

5. In contrast to brute-force simulation, it is also possible to employ standard optimization routines to fit analytic expressions for the ARMA and fBmW to spectral data (i.e., using Equations 7 and 9; for an example, see the fits in Figure 5). We have used such an approach to verify that the solution that we derived from explicit simulation of the models did not depend on the particulars of the discrete parameter grids; in all cases, we found that the analytic solutions were consistent with our simulation-based results.

APPENDIX A
Spectral Methods

The discrete Fourier transform, Φ, of a time series $X(t)$ is written

$$\Phi(f) = \frac{1}{N}\sum_{t=0}^{N-1} X(t)e^{-i2\pi ft}, \tag{A1}$$

where f is the frequency (inverse trial number), defined as

$$\frac{n}{N}\text{, for } n = 1, 2, \ldots, \frac{N}{2}$$

and N is the length of the time series. The periodogram estimate of the underlying power spectral density is defined to be the squared modulus of this quantity:

$$\hat{S}(f) = \left|\Phi(f)\right|^2. \tag{A2}$$

All the analyses reported in this article are derived from a low-variance method of spectral estimation from Bartlett (1950) and Welch (1967). This method consists of portioning the series $X(t)$ into overlapping windows and then averaging the power spectra across these windows (see also the method in Press et al., 1992). Provided that the windows are chosen so as to overlap by half their length, the method turns out to be nearly optimal in terms of variance reduction (it provides a maximal number of windows to be averaged, while minimizing the interwindow correlation). Specifically, this choice of overlap reduces the error in the estimate of the power, $\hat{S}(f)$, by the factor

$$\frac{9K}{11},$$

where K represents the number of windows available in $X(t)$ (Welch, 1967). Window averaging leads to a smoother spectral estimate and is, in fact, equivalent to convolving the straight power spectrum (Equation A2) with a Gaussian-like function (i.e., a *Fejer* kernel of order m; see Priestley, 1981, pp. 439–440).

Although window averaging leads to a less variable estimate of the true power spectrum, it has the unfortunate consequence that estimates of power at neighboring frequencies are no longer uncorrelated

$$\left[\text{i.e., } \left\langle \hat{S}(f_i)\hat{S}(f_j)\right\rangle \neq 0\right].$$

Practically, this means that accurately fitting models to these spectra requires use of the full covariance matrix.

The Window-Averaged Spectrum

Let $X(t)$ represent the original intact time series of length N, where N is some power of 2 (throughout this article, N has been fixed at 1,024). We form the window-averaged spectrum by breaking $X(t)$ into

$$K = \frac{2N}{m} - 1$$

overlapping segments of size m:

$$x_1(j) = X(j),\ j = 0, \ldots, m-1$$
$$\vdots$$
$$x_k = X\left(j + \frac{(k-1)m}{2}\right),$$

where k is simply an index that runs over the K segments. For each segment, x_k, we form a straight m-point periodogram, using Equations A1 and A2:

$$\Phi_k(f) = \sum_{t=0}^{m-1} x_k(t)e^{-i2\pi ft}, \tag{A3}$$

where frequency is now defined over segments of length m. The power spectrum associated with the kth data segment is the squared modulus of the transform in Equation A3,

$$P_k(f) = \left|\Phi_k(f)\right|^2, \tag{A4}$$

and the final estimate of power in the window-averaged spectrum is simply the average over the K power spectra in Equation A4,

$$\left\langle P(f)\right\rangle_m = \frac{1}{K}\sum_{k=1}^{K} P_k(f). \tag{A5}$$

APPENDIX A (Continued)

For all the calculations in this article, power spectra were estimated using FFT routines that effectively compute Equation A3.

The Composite Spectrum

The window-averaged spectrum (Equation A5) is always defined relative to a particular choice of segment size, m. In practice, determining the best m is an art that depends partly on the degree of frequency resolution, as well as on the statistical power one requires. Choosing a large value of m enables one to estimate power at lower frequencies, because the window is longer, but with lower reliability, because a larger m implies fewer windows in the average. Similarly, smaller values of m lead to more stable averages at the cost of lower frequency resolution. In general, the lowest frequency that can be resolved within a window of size m is

$$f_1 = \frac{1}{m}$$

(excluding the DC point defined by f_0). If one wishes to estimate power at a frequency below f_1, m must be made larger.

For all the spectral estimates derived in this article, we form the *composite* spectrum, $C(f)$, using a combination of multiple window-averaged spectra defined over a set of m (cf. Gilden et al., 1995). Specifically, we combine spectral estimates of low-frequency power based on large windows with spectral estimates of intermediate- and high-frequency power based on smaller windows. In this way, we acquire highly reliable estimates of power at each frequency by always choosing the smallest m (i.e., the largest number of segments, K) that is available. Accordingly, all spectral estimates in this article consist of composite spectra defined at the following eight discrete frequencies:

$$F = \left\{ \frac{1}{m_j},\ j = 1,\ 2,\ \ldots,\ \left(\log_2 N\right) - 2 \right\}, \tag{A6}$$

corresponding to the window sizes

$$m_j = 2^{j+1},\ \text{for}\ j = \left\{1,\ 2,\ \ldots,\ \left(\log_2 N\right) - 2\right\}. \tag{A7}$$

To reiterate, we estimate the power at a given frequency with the least variability by choosing the smallest window that resolves it.[A1]

Finally, the composite spectrum requires that spectral estimates from different-sized windows be appropriately normalized by $1/m_j$. If we denote the power estimate at the lowest frequency in a window of size m_j as

$$\left\langle P\left(f_1\right)\right\rangle_{m_j},$$

then the complete eight-point composite spectrum is obtained by normalizing each of the window-averaged power estimates by its window size and concatenating the estimates into a single spectral vector:

$$\hat{C}(F) = \left\{ \frac{\left\langle P\left(f_1\right)\right\rangle_{m_j}}{N m_j} \right\}, \tag{A8}$$

for the set of m_j in Equation A7.

Normalization by m_j is necessary for the proper alignment of estimates from different window-averaged periodograms, because the overall power in any periodogram is proportional to sequence length (i.e., window size). In Equation A8, we also normalize by the total length of the time series (N), in order to bring the composite spectral estimates into accord with the theoretical expressions for the ARMA and fBmW spectral densities.

The combined estimates of power across the set F are highly reliable, because each estimate is based on the smallest possible choice of m and so the window average is taken over the largest possible number of segments.

NOTE

A1. The composite spectra analyzed in this work are based on eight points and do not include estimates of power at the lowest available frequency in a sequence of length N (i.e., $f_{\min} = 1/N$). The lowest frequency we use is defined at $2/N$. This choice was made for a number of reasons. First, estimates of power at $f_{\min}$ are highly unreliable, because power is generally greatest at this frequency, and spectral variability grows proportionally with power. Second, estimates of power at $f_{\min}$ are especially vulnerable to secular trends in a time series such as might occur from learning or fatigue (see Appendix B). A final reason to forgo the use of power estimates at $f_{\min}$ is that these estimates do not enjoy the variance reduction properties associated with window averaging; in a sequence of length N, there can be only one N-point window with which to estimate power at $f_{\min}$. Moreover, in the absence of window averaging, the estimates at $f_{\min}$ will be χ^2 distributed, and this runs counter to the Gaussian assumptions of the spectral classifier.

APPENDIX B
Analyses of Observer Averaging and Trend Removal

Averaging Spectra Over Observers

An alternative means of increasing reliability is to average power spectral estimates across sequences (observers). Provided that there are no systematic and/or large variations in the underlying process parameters across different observers, averaging presents a preferred means for improving statistical power. In the following figure, we show an example of how the discrimination of short- and long-range processes improves with the number of averaged spectra (n).

The figure plots the sensitivity of the maximum likelihood classifier in correctly classifying fBmW(-1, 2) and ARMA(.95, $-.86$) processes as a function of n. These particular long- and short-range processes were chosen because each of them generates time series that approximate the structure of choice RT data, and this fractal is especially susceptible to misclassification because 80% of its variability is due to pure white noise. Note that when n equals 1, the figure shows the classifier's hit rate and correct rejection rate for classifying single fBmW and ARMA sequences. The figure also shows how classifier performance changes in categorizing averaged spectra based on n independent exemplars of each process. It is evident that the power of the classifier improves dramatically with n, and by the time 8–16 spectra have been averaged, classification accuracy has improved from about 75% to near-perfect levels of discrimination. Classification accuracy improves with n, as was expected, because averaging even a small number of spectra tends to smooth out the sampling errors associated with periodogram estimation. As n increases, the averaged ARMA and fBmW spectra converge on their smooth theoretical forms and so become increasingly easier for the classifier to distinguish. We point out that the improvement in classification sensitivity with n is in no way limited to a specific parameter regime but holds across the entire model parameter spaces. The results shown in Figure B1 represent a worst-case situation, and we have verified that for other parameter regions in which the single sequence performance is higher to begin with, averaging only pushes classification performance to ceiling at a faster rate. For example, for simulations of the ARMA(.9, $-.69$) process, single-sequence spectra are classified as short range with an accuracy of approximately 81%, whereas for spectral averages formed over eight exemplar sequences, the classifier achieves 98% accuracy.

One possible reason not to use observer-averaged spectra is that the average may not reflect the structure of any single spectrum. For example, if we form an average of AR(ϕ) processes that differ widely in their parameters, the averaged spectrum may end up looking fractal. It is well known that fractal-like spectra may be approximated by taking an equally weighted sum of multiple AR(ϕ_K) processes with logarithmically spaced time scales [e.g., let $\phi_K = \exp(-1/\tau_K)$, $\tau_1 = 1$, $\tau_2 = 10$, and $\tau_3 = 100$]. Similarly, aggregation of a very large number of AR(ϕ_K) processes, each with a parameter drawn from a suitable beta distribution, can also generate spectral signatures that look long-range (Granger, 1980). Physical models have capitalized on this fact to account for the ubiquity of $1/f$ noise in natural systems (Miller et al., 1993; West & Shlesinger, 1989). These models do require physical motivation for the spacing of time scales, and the generic superposition is decidedly not fractal (Hausdorff & Peng, 1996). For psychologically realistic distributions of the autoregressive parameters ϕ and θ (e.g., skewed beta distributions that emphasize high levels of ϕ), averages based on four to eight ARMA spectra generate lower false alarm rates, relative to single exemplars. This indicates that even for ensembles of moderately heterogeneous ARMA(ϕ, θ) spectra, averaging does not generate long-range structure. We find, instead, that averages of short-range processes converge on short-range structure, whereas averages of long-range spectra converge on long-range structure. This point is confirmed by a general *landscaping* analysis that allows one to assess the overall discriminability of two competing models (Navarro et al., 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). We find that the ARMA and the fBmW models become increasingly more discriminable as more representative spectra are averaged.

Detrending Time Series

It is well recognized in the time series literature that assessments of serial correlation will be inaccurate when the data are corrupted by secular trends (Giraitis, Kokoszka, & Leipus, 2001). In psychological data, such unwanted trends might be identifiable with learning or fatigue over the course of a trial sequence, and these kinds of effects can, depending on their magnitude, introduce enough nonstationarity into a time series to inflate our estimates of autocorrelation and spectral power. For these and other reasons, many analysts typically suggest that raw time series data be detrended prior to analysis. One standard practice is to remove linear trends, using either least-squares regression or bridge detrending (i.e., subtracting a line connecting the end points of the series; Cannon, Percival, Caccia, Raymond, & Bassingthwaighte, 1997). In certain contexts, there may even be informed cause to eliminate higher order polynomial trends, so as to further increase the stationarity of the time series.

APPENDIX B (Continued)

Although detrending may have a number of desirable features in certain problem domains, it is important to understand that, in other contexts, it may not be justified. Consider the classification problem we are interested in here, that of distinguishing long-range processes from short-range ARMA mimics. In this domain in particular, it matters a great deal whether low-frequency *trends* in data are treated as nuisance variation or, instead, as reflections of the process of interest. Anyone who has synthesized a long-range $1/f^{\alpha}$ noise knows that characteristic *linear* and *quadratic* trends are discernable in the simulated record. In fact, arbitrarily removing linear or quadratic effects from a *pure* fractal signal is at odds with the scale-free property of these noises; detrending removes power only at the lowest scales, and so, the fractal can no longer be self-affine.

This effect is illustrated in Figure B2, where we show the expected power spectrum of a whitened fractal ($\alpha = -1, \beta = 2$) with its trends intact (filled points), with a linear trend removed (solid line), and with a quadratic trend removed (dashed line). The results are clear: As the order of trend removal increases, the low-frequency power of the fractal is increasingly attenuated. More problematic from our viewpoint is that detrending causes real fractals to look more like members of the ARMA family (note the flattening of power reminiscent of short-range power spectra).

We can quantify the relative increase in confusability caused by detrending fractal time series by using the spectral likelihood classifier. To do so, we simulate an ensemble of exemplars from the fBmW($-1, 2$) process and then submit the ensemble to maximum likelihood analysis, using either unprocessed or detrended versions of the same underlying sequences. This analysis indicates that the detrended fBmW exemplars are mistakenly classified as short range about 50% of the time, relative to 20% when the same time series are not detrended (similar patterns of increased miss rates with detrending were found across the fBmW family, although the absolute size of the increase was reduced with higher overall accuracy). This decrement in performance with detrending arises primarily because the most diagnostic region of the power spectrum for distinguishing short- from long-range processes is also the region most susceptible to the effects of trend removal (i.e., at low spectral frequency).

Now consider what happens when we detrend a short-range sequence. In comparing classification performance for raw and detrended ARMA sequences, we find an opposite pattern, in which detrending leads to a slight rise in accuracy. This is expected, given that a detrended ARMA process looks increasingly less like any long-range fBmW. From a practical standpoint,

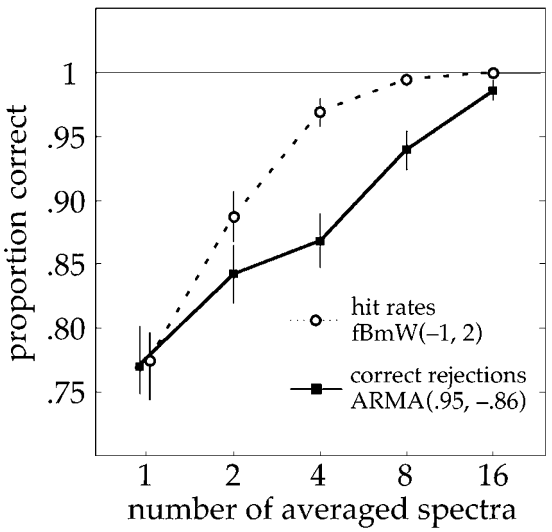

Figure B1. The effect that averaging power spectra has on classifier sensitivity in correctly categorizing long- and short-range time series. Hit rates and correct rejection rates of the maximum likelihood classifier are shown for discriminating averages of fBmW($-1, 2$) spectra from averages of its nearest ARMA relative ($\phi = .95$, $\theta = -.79$); n denotes the number of spectra used in the average. Significant increases in hit rates and correct rejections are realized for modest values of n. These simulations are based on ensembles of 1,000 averages for each n.

APPENDIX B (Continued)

the asymmetry in confusability that detrending introduces, whereby detrended fBmWs look more like ARMAs and detrended ARMAs still look like ARMAs, leads to both a marked reduction in discriminability and a significant increase in the classifier's bias to assign short-range provenance. Together, these effects indicate that detrending has the unfortunate property that it increases the likelihood of missing true long-range structures in data.

It is clear that the decision to detrend is a far more subtle issue than is often appreciated. The real conundrum here is that there is no principled way to distinguish trend-induced nonstationarity from low-frequency fluctuations in the signal itself (Diebold & Inoue, 2001; Granger & Hatanaka, 1964). We deal with the issue of *trends* using the following three guidelines. First, we adopt a composite spectral representation in which estimates of power at the lowest frequency ($f_1 = 1/N$) are excluded. As Figure B2 makes clear, estimates of power at f_1 show the greatest effects of trend removal. By excluding these estimates from our spectral representations, we greatly decrease the classifier's tendency to categorize detrended fractals as ARMAs. Second, we remove quadratic trends only when there are good reasons to do so—either because of a specific theory or through accumulated experience with similar data. A good rule of thumb, and one that appears to be shared by other independent assessments of long-range structure, is that trend removal should generally be limited to first-order analyses (e.g., *scaled windowed variance analyses*, Cannon et al., 1997; Mandelbrot, 1985; *detrended fluctuation analysis*, Eke et al., 2002; Peng et al., 1994). As an additional means of guarding against brief nonstationarity in the early epochs of data collection, we suggest that observers receive some preliminary practice trials (e.g., by including a warm-up block that is not subject to analysis). This may seem obvious, but it is critical.

Finally, we advocate routinely reporting all analysis results for both the raw and the linearly detrended cases (cf. the analyses in the Case Study section). In this way, it is possible to assess to what extent our relatively uninformed decisions regarding the need to detrend effect final assignments of provenance.

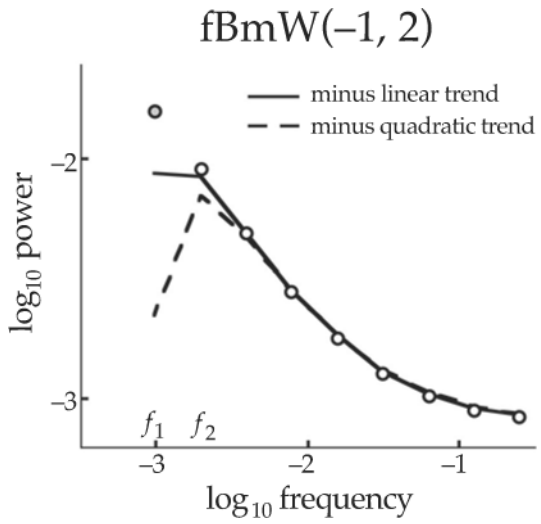

Figure B2. Spectral consequences of detrending a long-range time series. In the figure, the solid points represent the expected power spectrum of the unadulterated fBmW(−1,2) process based on averaging composite spectral estimates from 200 exemplars (the white points represent the eight-point composite spectra used in this article; the single gray point denotes the estimate of power at the lowest frequency [f_1] available in a series of length 1,024). The solid line is the nine-point composite spectral expectation for the same set of fBmW(−1, 2) exemplars with all linear trends removed; the dashed line is the expectation under quadratic trend removal.

APPENDIX C
The Spectral Likelihood Classifier

Normality of Spectral Sampling Distributions

For the purposes of classification, we treat the composite spectra as eight-element vectors, one element for each frequency. Our analysis assumes that the sampling distribution of vectors associated with a single fixed fBmW(α, β) or ARMA(ϕ, θ) process forms a multivariate Gaussian distribution, characterized by an 8-point mean vector, μ, and an 8×8 covariance matrix, C. Accordingly, the sampling distribution of vectors associated with a given fBmW(α, β) process is defined by a mean vector $\mu_{\alpha,\beta}$ and a covariance matrix $C_{\alpha,\beta}$. Similarly, the sampling distribution of each ARMA(ϕ, θ) process has a mean vector $\mu_{\phi,\theta}$ and a covariance matrix $C_{\phi,\theta}$. For the window-averaged spectra we form (Welch, 1967; see Appendix A), the distributions of power are, in fact, approximately Gaussian for most of the frequencies, and any departures from normality are small and arise only for estimates at the lowest frequencies. Normality of estimates occurs in the window-averaged periodogram via the central limit theorem; power at each frequency is estimated using an average over overlapping segments, each of which is χ^2 distributed.

We have investigated the extent to which departures from normality might affect classification performance. In general, the performance of the spectral likelihood classifiers appears to be quite robust to mild violations of the Gaussian assumption of the sort that arise in the use of window-averaged periodograms. The validity of this assumption is important because it allows us to use simple analytic expressions for the n-dimensional Gaussian in order to estimate the likelihoods of spectral data.

Simulation Details

For all the work reported in this article, sequences were explicitly simulated using either time domain expressions for the short-range ARMA(ϕ, θ) process or frequency domain expressions for the long-range fBmW(α, β) process. First-order ARMA processes were simulated recursively using Equation 6 over a set of Gaussian deviates with zero mean and unit variance. To ensure the stationarity of these processes, the first 500 values simulated for each time series were excluded from the final analysis.

The fBmW(α, β) process was simulated by adding white noise to spectrally synthesized fractional Brownian motions. Simulations of the fBm were realized in the frequency domain by creating a $1/f^{(\alpha/2)}$ amplitude spectrum with random phase angles (cf. the method of Peitgen & Saupe, 1988). The resulting Fourier representation was then inverse transformed to yield the appropriate time domain expression. To ensure sampling variability across exemplars, all fBm spectra were synthesized using 2,048 values, and only the first 1,024 of these were used in the fBmW composition (without this padding, the distributions of fBm spectra have zero variance). Although there are a number of explicit methods for creating spectral variability across simulated fBm exemplars (Davies & Harte, 1987; Peitgen & Saupe, 1988), we prefer the 2,048-point truncation method because it allows us to build and calibrate spectral classifiers that are largely unbiased in the regime of production and choice-RT data (see Figure 7).

The Maximum Likelihood Algorithm

A schematic of the structure of the spectral maximum likelihood algorithm is shown in Figure C1. The decision algorithm begins with an input time series ($N = 1{,}024$) that is transformed to z scores and converted into a composite spectral representation, $\vec{S}_K$ (here, the subscript K simply denotes that the time series could be the Kth exemplar of some generating process or, alternatively, data from the Kth hypothetical observer).

To categorize $\vec{S}_K$ as being of either ARMA or fBmW origin, the spectral likelihood classifier computes 800 likelihoods, 1 at each of the 400 points defining the ARMA(ϕ, θ) space and 1 at each of the 400 points defining the fBmW(α, β) space. The expression for the likelihood of $\vec{S}_K$ at the point $[\alpha_i, \beta_j]$ in the fBmW space is

$$L(\vec{S}_K, \alpha_i, \beta_j) = \frac{\exp\left(-\frac{1}{2}\left[\left(\vec{S}_K - \mu_{\alpha_i,\beta_j}\right) C^{-1}_{\alpha_i,\beta_j} \left(\vec{S}_K - \mu_{\alpha_i,\beta_j}\right)'\right]\right)}{\sqrt{2\pi^n \left|C_{\alpha_i,\beta_j}\right|}}. \tag{C1}$$

The quantities $|C_{\alpha_i,\beta_j}|$ and $C^{-1}_{\alpha_i,\beta_j}$ denote, respectively, the determinant and inverse of the covariance matrix at the point $[\alpha_i,\beta_j]$ (each covariance matrix was estimated via explicit simulation of 2,000 exemplars); the quantity n in the denominator corresponds to the dimension of the com-

APPENDIX C (Continued)

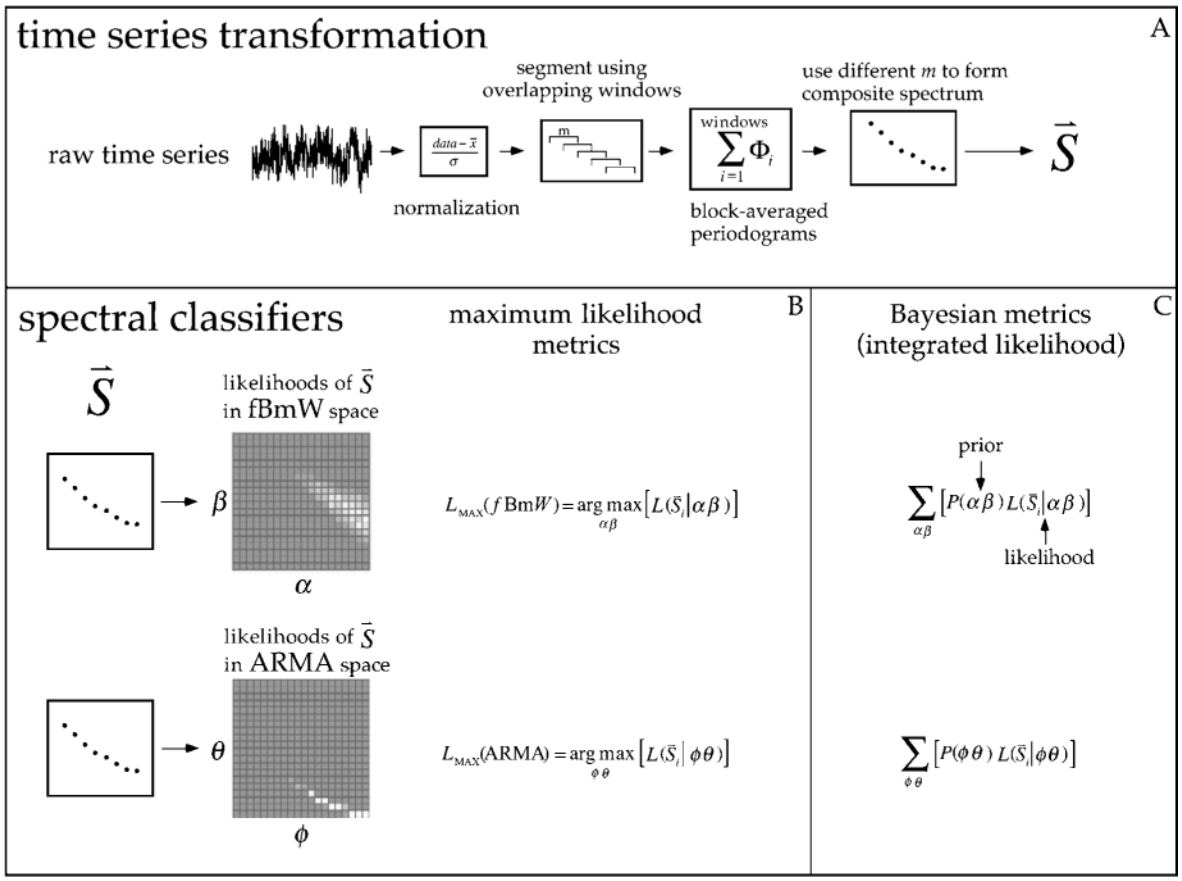

Figure C1. Schematic representation of the structure of the Bayesian spectral classifier. Panel A represents the algorithm that is used to transform raw time series data into a high-quality composite spectral estimate. The lower panels depict the two classification metrics that are used to decide whether the data vector $\vec{S}$ was generated by a short- or a long-range process. The maximum likelihood classifier (panel B) computes a map of the likelihood of $\vec{S}$ for both the fBmW and the ARMA families, assigning provenance to the family with the maximum likelihood; the Bayesian classifier (panel C) computes an integrated posterior probability for both the fBmW and the ARMA models and assigns provenance to whichever model generates the higher probability.

posite spectral representation ($n = 8$ for all analyses reported here). Equation C1 is just the multidimensional analogue of the more familiar expression for the density of a Gaussian random variable (the term in brackets reduces to a z^2). Using Equation C1, the classifier computes a separate likelihood of the spectral data $\vec{S}_K$ at every point in the fBmW parameter space. The classifier then assumes that the best fractal description of the data occurs at the point in the space with the greatest likelihood. We denote the maximum likelihood of the data in the fBmW(α, β) space as

$$L_{\mathrm{MAX}}(\mathrm{fBmW}) = \arg\max_{\alpha\beta}\left[L\left(\vec{S}_K|\alpha\beta\right)\right]. \tag{C2}$$

In similar fashion, the classifier computes 400 likelihoods of $\vec{S}_K$ across the ARMA(ϕ, θ) space to find the best autoregressive description of the data based on the quantity

$$L_{\mathrm{MAX}}(\mathrm{ARMA}) = \arg\max_{\phi\theta}\left[L\left(\vec{S}_K|\phi\theta\right)\right]. \tag{C3}$$

The final decision metric is based on a comparison of Equations C2 and C3 and reduces to assigning provenance to whichever model has the single greatest maximum likelihood.

APPENDIX C (Continued)

The Bayesian Algorithm

The Bayesian integrated likelihood classifier assigns provenance on the basis of whichever model generates the greatest integrated likelihood to an input data spectrum. Specifically, for each spectral vector $\vec{S}_K$, we compute the following two quantities:

$$P\left(\vec{S}_K \,|\, \text{fBmW}\right) \cong G\sum_{i,j} p\left(\alpha_i\beta_j\right) L\left[\vec{S}_K \,\middle|\, \alpha_i\beta_j\right] \tag{C4}$$

and

$$P\left(\vec{S}_K \,|\, \text{ARMA}\right) \cong G\sum_{i,j} p\left(\phi_i\theta_j\right) L\left[\vec{S}_K \,\middle|\, \phi_i\theta_j\right], \tag{C5}$$

where the left-hand terms are proportional to the marginal likelihood of the data spectrum given each model, $p(\alpha_i\beta_j)$ and $p(\phi_i\theta_j)$ are the respective prior probabilities of specific parameter choices (a uniform constant for the fBmW and a truncated uniform for the ARMA; see the Parameter Priors section below), the right-hand summations represent the weighted average of likelihood taken over all 400 parameter combinations defining each model, and G is simply a normalizing constant that falls out in the final decision ratio. In the state of ignorance regarding the prior probability of the two model classes themselves, we assume equality, and this allows us to use Equations C4 and C5 as estimates of the posterior probabilities of the models, given the spectral data. We assign provenance to each spectral vector, $\vec{S}_K$, using a rational strategy based on the sign of the following ratio:

$$\ln(\text{Bayes}) = \log_e\left(\frac{P\left(\vec{S}_K \,\middle|\, \text{fBmW}\right)}{P\left(\vec{S}_K \,\middle|\, \text{ARMA}\right)}\right), \tag{C6}$$

where the left-hand term stands for the log Bayes factor.

Parameter Priors

In using a Bayesian classifier to assign provenance across the entire families of fBmW and ARMA processes, we have assumed either a uniform or a truncated uniform distribution on the prior probability of various parameter combinations. This is surely incorrect in the context of analyzing choice RT data, because previous work (Gilden, 1997, 2001; Van Orden, Holden, & Turvey, 2003) has indicated that most sequences occupy a limited region of the model parameter spaces (cf. Figures 7 and 8). However, in the interest of calibrating the classifier over the entire fBmW and ARMA families, the uniform assumption is justified for two reasons. First, it is the only distribution that treats each process equally. For any other prior that has a peak value in the parameter space, classification performance will be reduced when the peak does not coincide with the particular process that is being tested. For example, if we assume some sort of fixed two-dimensional Gaussian or beta prior distribution that is centered on a specific point in the fBmW and ARMA parameter spaces—say $[\alpha_i, \beta_j]$ and $[\phi_k, \theta_l]$—then in calibrating the classifier on other processes whose parameters are different from these values, the likelihoods will be arbitrarily attenuated, depending on the particulars of how fast the prior decays from its peak (e.g., if the distribution has low variance). The second reason that we use a uniform prior during classifier calibration is that this is the actual distribution of processes; in the context of calibration, each and every ARMA and fBmW process is, by definition, as likely as any other.

The truncated uniform prior. The ARMA is a highly complex family, and fully half of its predicted spectra are inconsistent with psychophysical data; they ascend with increasing frequency. It is the case that the hit and false alarm rates of the Bayesian classifier are sensitive to the ARMA parameter integration limits and so to the range of ARMA processes that we wish to consider as viable candidates. In this work, we adopt the conservative policy of integrating only over descending spectra. This artificially reduces the complexity of the ARMA family and makes those ARMA processes with descending spectra more probable.

For the calibration results reported in the Bayesian Classification Based on Integrated Likelihood section, we enforced a uniform prior on the fBmW parameters and a truncated uniform on the ARMA parameters. The truncated uniform was constructed to be 0 for all ARMA processes with ascending spectra (i.e., parameter combinations for which $\phi \leq |\theta|$); the remaining subset of processes with descending spectra have priors set to a constant value (in all cases, the priors were normalized to form proper probability distributions).

APPENDIX C (Continued)

Priors for choice RT analyses. In the analysis of actual data, we do not know the proper priors. Although the uniform assumption is correct for calibration purposes, it does not reflect the experiential knowledge that we acquire as researchers regarding the range of parameters relevant to actual psychophysical data. Recognizing that only a subset of the family of fBmW and ARMA processes are consistent with the form of choice RT spectra, we are best advised to construct prior distributions on the parameters that reflect this knowledge (Jaynes, 2003). A conservative approach here is to explore classification performance, using a variety of priors that respect the basic qualitative constraints imposed by our empirical experience. In this way, we can determine the extent to which our results are robust in the face of what is surely a misspecified prior distribution.

In using the Bayesian classifier to categorize Van Orden, Holden, and Turvey's (2003) data, we parameterized the priors within a family of general purpose two-dimensional beta distributions. These distributions were constrained to peak in the parameter regions previously associated with our own choice RT data (i.e., regions where power spectra descend smoothly with frequency and are not linear; see the dashed ovals in Figures 7 and 8). In separate analyses the beta parameters were manipulated to create a variety of prior distributions that differed in central tendency, variability, and skew (in all cases, the priors were normalized to form proper probability distributions). So long as the prior distributions did not overly penalize parameter regions consistent with choice RT data (i.e., modest α and high β for the fBmW, high ϕ for the ARMA), the final assignments were changed little from the uniform case reported in the Case Study section, indicating robustness to the specific form of the priors.

(Manuscript received May 19, 2004;
revision accepted for publication January 16, 2005.)