

- bility and Statistics (E. Eberlein and M. S. Taqqu, eds.) 137–165. Birkhäuser, Boston.
- TAQQU, M. S. (1988). Self-similar processes. In *Encyclopedia of Statistical Sciences* 8 352–357. Wiley, New York.
- TERRIN, N. and TAQQU, M. S. (1991). Convergence in distribution of sums of bivariate Appell polynomials with long-range dependence. *Probab. Theory Related Fields* 90 57–81.
- VERWAAT, W. (1987). Properties of general self-similar processes. *Bull. Inst. Internat. Statist.* 52 199–216.
- WHITTLE, P. (1956). On the variation of yield variance with plot size. *Biometrika* 43 337–343.
- WHITTLE, P. (1962). Topographic correlation, power-law covariance functions, and diffusion. *Biometrika* 49 304–314.
- YAJIMA, Y. (1985). On estimation of long-memory time series models. *Austral. J. Statist.* 27 303–320.
- YAJIMA, Y. (1988). On estimation of a regression model with long-memory stationary errors. *Ann. Statist.* 16 791–807.
- YAJIMA, Y. (1989). A central limit theorem of Fourier transforms of strongly dependent stationary processes. *J. Time Ser. Anal.* 10 375–383.
- YAJIMA, Y. (1991). Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *Ann. Statist.* 19 158–177.

Comment: Short-Range Consequences of Long-Range Dependence

Arthur P. Dempster and Jing-Shiang Hwang

We welcome Jan Beran's informative sketch of the history of long-range dependence in many fields of applied statistical science, and likewise his review of the results of several decades of work by mathematical statisticians, mainly on asymptotic sampling theory of various robust as well as normality-based efficient estimators.

Our experience has been with applications of the models, most recently in Hwang (1992) and Dempster and Hwang (1992), to simultaneous estimation of employment time series of 51 U.S. states (including DC) given short input time series of $n = 48$ months. Since our data are fixed, we have emphasized issues related to modeling both time series of sampling error, which a priori have no long-range dependence (ignoring biases that cannot be assessed from our data), and underlying true time series, which appear empirically to have long-range dependence with parameter H close to 1 (but not greater than 1 because nonstationarity of unemployment and employment rate series is a priori implausible).

For inference about the true series, we have emphasized Bayesian thinking, and associated computational issues related to likelihoods of our fixed data, always under assumptions of normality, which appear generally to be reasonable in our case study. Although our theoretical approach to statistical inference is very different from that of Beran, we agree with his opening remarks about dangers from behaving as though tradi-

tional ways of thinking about level and variability of underlying short-memory stationary time series models continue to hold in the presence of stationary long-memory models. We direct our brief comments to exposing a few basic small n distinctions between inferences appropriate in situations characterized by short-range dependence and those with long-range dependence. We begin by exhibiting artificially generated pseudorandom "time series" that render in graphical form the main points about estimating the mean and variance of fractional Gaussian noise (fGn) data. We have found it convenient to use alternative notation τ^2 and d in place of σ^2 and H , where $d = 2H - 1$ and τ is chosen so that the spectral density

$$f(\lambda) \sim \tau^2 \lambda^{-d}$$

for λ close to zero. On this scale, $-1 < d < 1$ defines the range for fGn, but $0 < d < 1$ is the range of interest for long-range phenomena, with $d = 0$ corresponding to white noise and $d = 1$ marking the upper boundary where the spectral density first becomes nonintegrable at zero frequency. We use the same frequency domain conventions as Beran, namely, that $-\pi < \lambda < \pi$ and that $f(\lambda)$ is scaled so that σ^2 is its average value with respect to uniform measure. Appropriate roles for the alternative scale parameters τ^2 and σ^2 are elaborated below.

Figure 1 displays four series, each of length $n = 64$, simulated from four different fGn models with $d = 0.8, 0.9, 0.99, 0.999$. Part of the reason for the near coincidence of the curves apart from their levels is that all four were generated from innovations based on the same 64 normal pseudorandom values. In addition, however, the similarity implies covariance structures with remarkably similar forecast operators and resid-

Arthur P. Dempster is Professor and Jing-Shiang Hwang is Postdoctoral Fellow at Department of Statistics, Harvard University, Cambridge, Massachusetts 02138.

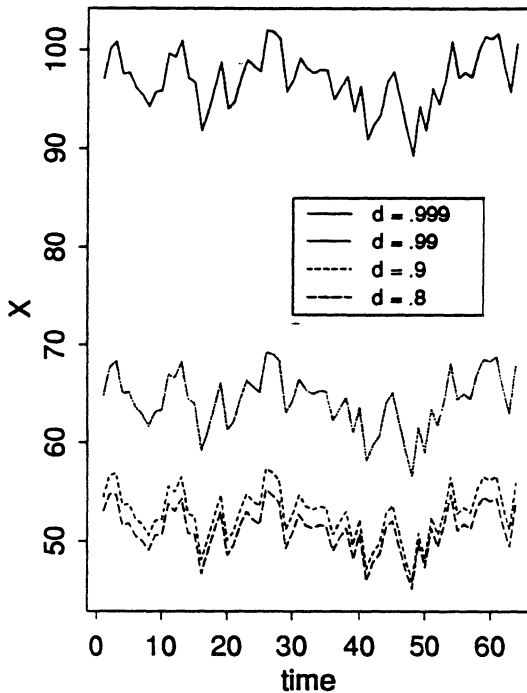


FIG. 1. Simulated time series of length $n = 64$ from 4 fGn models.

ual prediction variances. Figure 1 is typical of what we see in many repeated trials and suggests that discrimination among plausible values of d will be very difficult on the basis of a single short series.

Before turning to estimation of d we point out another strong feature of Figure 1 that may appear paradoxical. Whereas the empirical curves visually have different levels but common variability, exactly the opposite is true of the underlying models, which were in fact constructed to have a common $\mu = 50$, but very different $\sigma^2 = 8.3, 18.2, 198.0, 1998.0$. The key to resolving the aspect of paradox associated with variability is to be told that the four models were constructed to have a common $\tau^2 = 1$. From formulas given by Beran, it follows easily that τ^2 , as defined above through the small λ behavior of the spectrum, is related to σ^2 by

$$\tau^2 = \sigma^2 \frac{1}{2\pi} \Gamma(2+d) \cos\left(\frac{\pi d}{2}\right).$$

An approximation that is accurate to within 1% across $0 < d < 1$, and includes the correct limiting behavior as d approaches 1, is

$$\tau^2 \approx \sigma^2 \left(1 - \frac{d}{2}\right) (1-d).$$

Readers may note that the four values of σ^2 are roughly proportional to $1/(1-d)$.

Since the similar empirical variabilities shown in

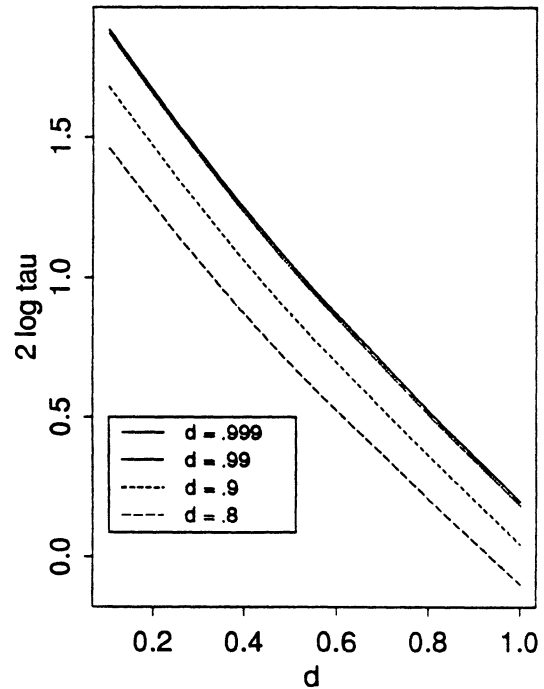


FIG. 2. Maximum likelihood estimates $\hat{\tau}^2$ (in log scale) for the four simulated series.

Figure 1 correspond to a very wide range of true σ^2 as d varies, practical estimation of σ^2 in the absence of accurate knowledge of d is impossible. Dependence on knowledge of d is much reduced in the case of estimating τ^2 , as illustrated in Figure 2 which shows $\log \hat{\tau}^2$ computed for a range of d values for each of the four simulated series, using maximum likelihood for a likelihood with μ integrated out and d assumed known.

The key to resolving the side of the paradox relating to μ is also implicit in Beran's discussion. His equation (3) is mathematically interesting because it holds for all n , and because it is an immediate consequence of the selfsimilarity property of fractional Gaussian motion. But he passes on without noting the unpleasant applied consequence for a statistician with fixed data. Not only is $\text{var}(\bar{X}_n)$ strongly dependent on d , but also it grows like $1/(1-d)$ as d approaches 1. It is mathematically interesting that \bar{X}_n is asymptotically at least 98% efficient for all positive d , but the virtues of \bar{X}_n are compromised for small n because \bar{X}_n is little better than any single observation, that is, $\text{var}(\bar{X}_n)$ is roughly σ^2 . Given only data like that shown in Figure 1, it would generally be foolhardy to attempt to estimate μ , at least without strong prior knowledge that d is not in the range illustrated in the simulations.

Again there is a silver lining. Prediction is likely to be what really matters to the time series data analyst. As noted above, the near coincidences in Figure 1 suggest that the forecast functions and residual variances are robust against variation in d close to $d = 1$.

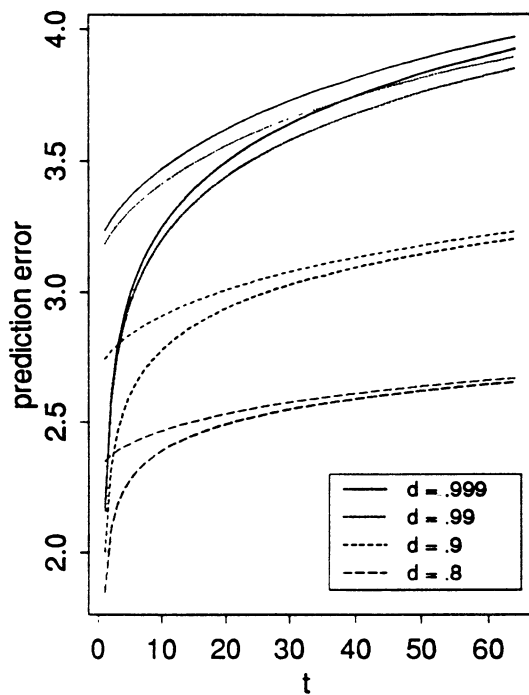


FIG. 3. Forecast standard errors for the four simulated series. In each pair, the upper one is the sample mean and the lower one is the optimum forecast.

Plots not shown here indicate that the optimum forecasts t steps ahead for $t = 1, 2, \dots, 64$ are indistinguishable from the optimum (BLUE) estimates of μ after $t = 5$ or 6, and these in turn differ only trivially from \bar{X}_{64} which does not depend on d . Figure 3 indicates that \bar{X}_{64} rapidly becomes a fully efficient forecast as t increases, but that prediction errors increase by about 50% as d moves from 0.8 to 0.999 across much of the forecast range $t = 1, 2, \dots, 64$. From Figure 3 it can be checked that prediction error variances for $d = 0.99$ and 0.999 grow roughly like $2 \log t$ which is the limiting rate at $d = 1$. Of course, for $d < 1$, stationarity ultimately restricts prediction error variance for large t to $\sigma^2 + \text{var}(\bar{X}_n)$.

Figure 1 supports a suspicion that d is hard to estimate from a single series. Figure 4 reinforces the suspicion by showing the log-log periodograms of the four series, along with the four model spectral densities on the same log-log scales. Figure 4 shows that estimation of d is, in effect, estimation of slopes of nearly linear functions in the log-log scale, and shows also that the (almost sufficient) periodograms are highly similar. Figure 5 shows plots of a marginal likelihood of d found by integrating out μ and σ^2 using uniform priors for μ and $\log \sigma^2$. These plots are perhaps the best visual way to show that the simulated series are really somewhat different, but also show that, while each series gives strong evidence of $d > 0$, there is little ability to discriminate among the larger values of d

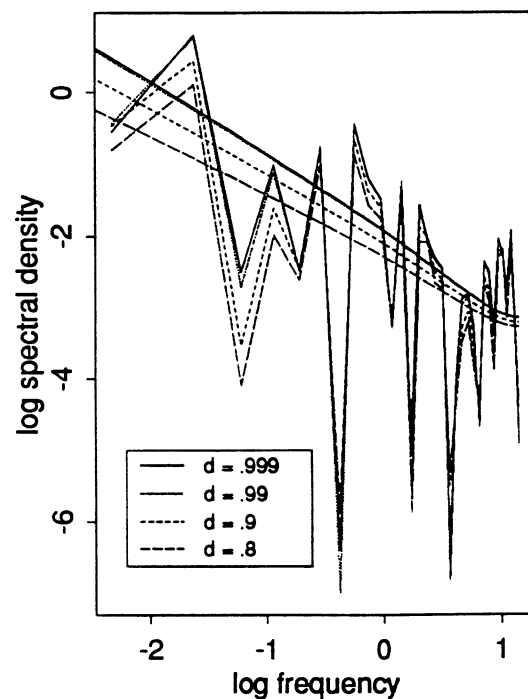


FIG. 4. Spectral densities and periodograms on a log-log scale.

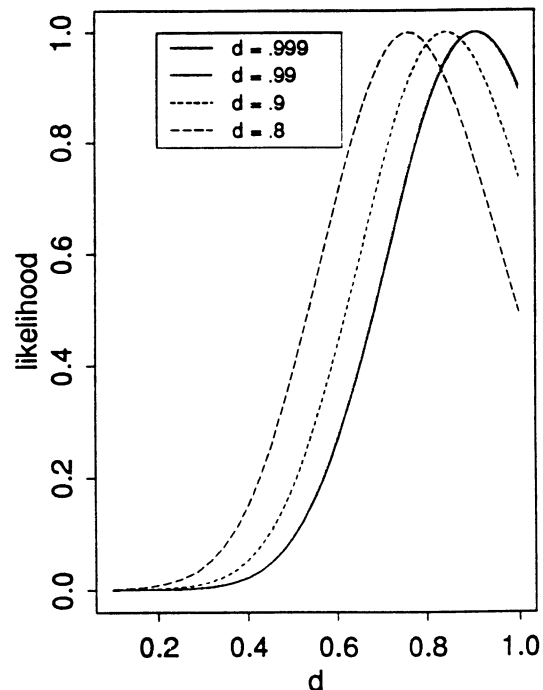


FIG. 5. Marginal likelihood function of d based on the above simulated series.

that critically affect estimability of μ and σ^2 . Figure 6 exhibits sampling variation of the likelihood functions of d based on 50 independent simulations from the same 4 models. The solid lines in Figure 6 show the products of the 50 likelihoods, simulating estimation

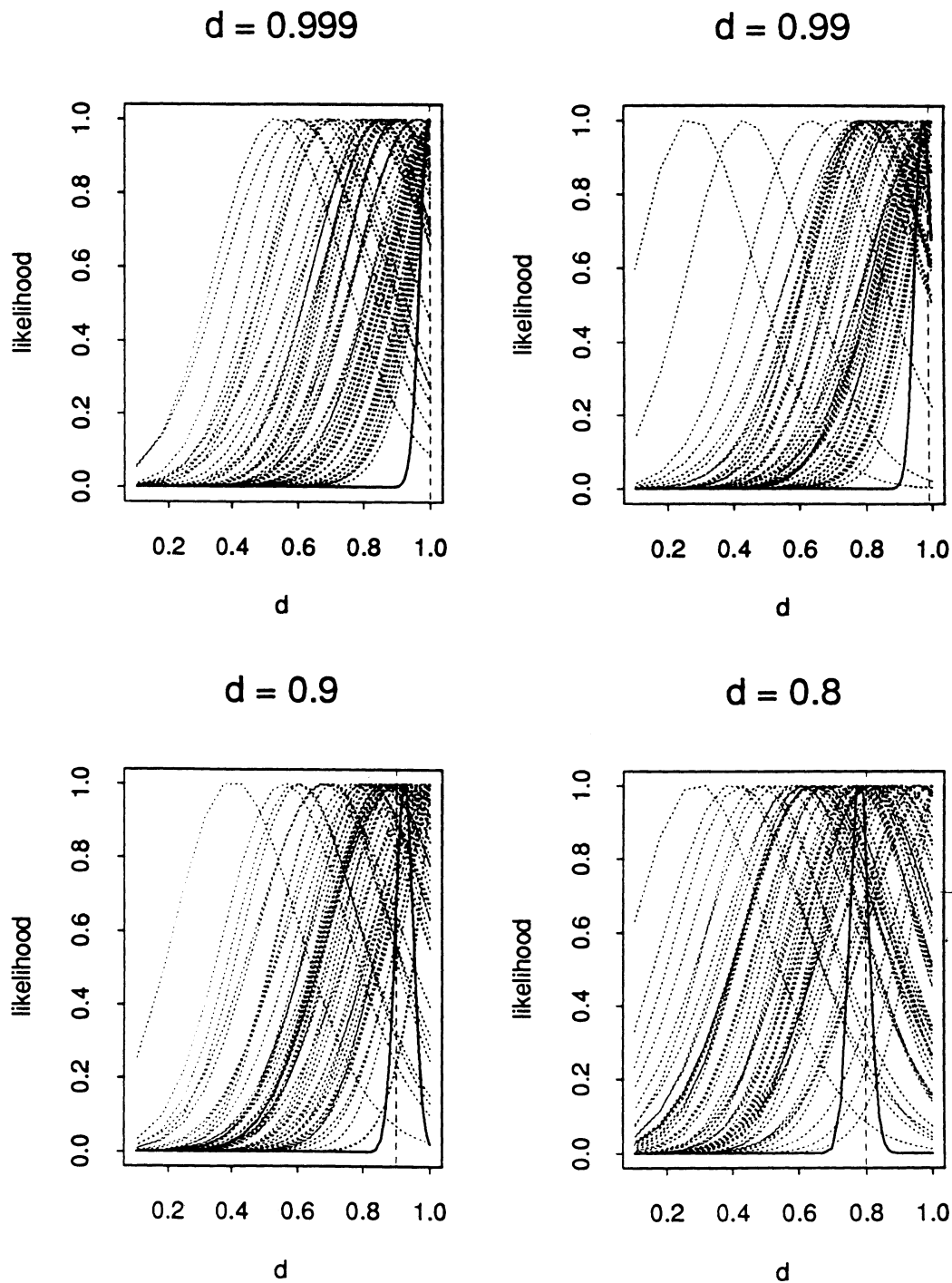


FIG. 6. Marginal likelihood functions of d based on 50 independent samples.

of d from 50 independent samples, allowed to have different μ and σ . From 50 samples, practical estimation of d is seen to be possible.

From the viewpoint of studies of the sampling behavior of probability models, the boundary $d = 1$ marks a radical change of behavior. But from the inverse inference viewpoint appropriate in applied statistics, it ap-

pears that one should be permitted to cross the barrier with scarcely a ripple affecting the machinery of inference. For example, the Figure 5 likelihood functions of d appear as though they should have natural extensions beyond $d = 1$. At present, it is easy to include the case $d = 1$, since the limiting stationary distribution of the process of the differences of fGn as d approaches

1 with fixed τ^2 is easily defined either through its covariance function or its spectral density, but we are not aware of studies of possible natural extensions beyond $d = 1$.

Comment

Emanuel Parzen

All statisticians should be made aware of the message of Jan Beran's comprehensive and stimulating paper, that the practice of statistics cannot be successful without applying awareness of long memory and long tail behavior in data. The data he discusses, especially "NBS precision measurements on the 1-kg check standard weight," demonstrate that statisticians who analyze data must be aware of the trichotomy that should be answered as one of the first steps in a data analysis: should the data be regarded as white noise (independent or zero memory), short memory (weakly dependent) or long memory (long range dependent). I would like to describe some heuristic concepts that I find useful for understanding, diagnosing, and modeling long range dependence.

Given a time series sample $Y(t)$, $t = 1, \dots, n$, I define the sample spectral density (or periodogram) as a function of ω , $0 \leq \omega < 1$:

$$\tilde{f}(\omega) = \frac{\sum_{t=1}^n |\tilde{Y}(t) \exp(-2\pi i t \omega)|^2}{\sum_{t=1}^n |Y(t)|^2}.$$

The spectral density $f(\omega)$, $0 \leq \omega < 1$, is defined (as a descriptor of the hypothetical population of sample paths in the probability model) as the limit as n tends to ∞ of $E[\tilde{f}(\omega)]$.

A time series is called short memory dependent if $f(\omega)$ is bounded above and below (white noise if the spectral density is constant).

A time series is defined to be long memory if $f(0)$ is infinity (more generally if $f(\omega)$ has zeroes or infinities at some frequency). An important role is played by the spectral density $f(1/n)$ at frequency $1/n$. It can be used to express the variability of the sample mean $\bar{Y}_n = (1/n) \sum_{t=1}^n Y(t)$; asymptotically $\text{VAR}[\bar{Y}_n] = (1/n) f(1/n) C$ for a suitable constant C depending on the index δ of the (self-similar) representation

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation and the Bureau of Labor Statistics through Westat, Inc.

$$f(\omega) = \omega^{-\delta} L(\omega)$$

where $L(\omega)$ is a slowly varying (log-like) function at $\omega = 0$, and $L(0) > 0$. The Hurst exponent H in Beran's formula (3) corresponds to long memory index $\delta = 1 - 2H$.

In addition to spectral techniques, we recommend changepoint and cusum analysis techniques to provide diagnostics of various types of long memory behavior, based on weak convergence theorems for cusum processes such as

$$\tilde{C}(\tau) = n^{\frac{1}{2}} \sum_{t=1}^{[n\tau]} \frac{Y(t) - \bar{Y}_n}{\tilde{\sigma}}, \quad 0 \leq \tau < 1,$$

where $\tilde{\sigma}$ is the sample standard deviation.

The cusum process is important for applications to quality control problems of identifying changepoints in the series under the null hypothesis that it is white noise.

To identify if there is long memory dependence and to detect changepoints in correlated data, one approach could be to estimate the exponent $\delta = 1 - 2H$, H the Hurst exponent, in the spectral density formula $f(1/n) = n^{\delta} L(1/n)$.

Values of the "fractal dimension" delta have received much public attention since they are used to describe music and how the brain works. *U.S. News and World Report*, June 11, 1990, writes (p. 62): "Surprisingly, the same mathematical formula that characterizes the ebb and flow of music has been discovered to exist widely in nature, from the flow of the Nile to the beating of the human heart to the wobbling of the earth's axis."

Estimation of δ (which can be considered estimating a "fractal dimension") is a central research problem of the analysis of long memory time series. Beran notes that it has analogies with the problem of estimating the tail index of a long tailed distribution. One expects to estimate how $f(1/n)$ depends on n essentially from the values of the sample spectral density in a band of low frequencies (omitting zero frequency) to be selected by suitable criteria.