(2) provides an approximate formula that can be used to estimate the standard error of $\hat{\beta}$. As an example, we consider a series of 1632 average monthly temperatures over the Northern Hemisphere (land and sea) used by the IPCC (Intergovernmental Panel on Climate Change) for its global warming analyses. Figure 3a presents the data, smoothed by applying a 49-month moving average and centered around the 1950–1979 mean value for each month. It shows a pattern of steady behavior until about 1920, followed by a sharp rise between 1920 and 1940, then a gradual decrease until about 1975, followed by the sharp rise that has triggered the present alarm about global warming. Over the whole series, there is a clear rise in temperature, but whether it is due to the greenhouse effect is a matter of intense debate among climate scientists.

A linear trend was fitted to this data series (unsmoothed) and resulted in a estimated trend of 0.40°C per century, a figure consistent with several other estimates of global warming over the last century and a half. The first 120 periodogram ordinates of the residuals are plotted on log-log scales in Figure 3b. The pattern is quite similar to the two series quoted by Beran, and again seems to show evidence of long-range dependence. This is confirmed by the estimates $H = 0.90$ with standard error 0.05, based on $n_0 = 1$, $n_1 = 120$; also $\hat{b} = 0.0033$. When these figures are inserted into (2) (adjusted for the unit of trend) the standard error of the estimated trend is around 0.1, which is again consistent with earlier estimates of standard error including those quoted by Bloomfield (1992). My main doubt about this conclusion is whether the series can really be assumed stationary, given the obvious inconsistencies in methods of measurement over the last century and a half, but this would take us into other aspects beyond the scope of the present discussion.

I believe the message of all three examples is that the concept of long-range dependence must be taken seriously. At the same time, exactly how these examples are to be interpreted could be a matter of considerable debate. Jan Beran is to be congratulated on his very clear and comprehensive review, and I hope it will act as a springboard for much further research in this area.

ACKNOWLEDGMENTS

Example 3 is based on data provided by Dr. P. Jones of the Climate Research Unit, University of East Anglia. I thank Jan Beran for providing me with the Nile and NBS data, and Peter Bloomfield for copies of his references. This work was supported in part by NSF Contract DMS-91-15750.

Rejoinder

Jan Beran

I would like to thank the discussants for their stimulating comments and valuable suggestions. Their comments emphasize once more that long memory is an important issue to anybody who uses statistical inference, since it occurs rather frequently in real data and strongly influences the validity (and power) of standard tests and confidence intervals. Particularly interesting are the data examples analyzed by Smith (global warming–climatological data), Haslett and Raftery (wind speed–meteorological data) and Dempster and Hwang (employment series–economic data), since these are examples that concern everyone (and not just a selected group of scientists). Parzen summarizes the main message of the paper very clearly by saying that in data analysis, we always have to decide whether the data (either the original measurements or residuals, e.g., after subtracting a regression function) are white noise, a short-memory process or a long-memory process. The same view is expressed in a more general context by Mosteller and Tukey (1977, p. 119 ff): "even in dealing with so simple a statistic as the arithmetic mean, it is often vital to use as direct an assessment of its internal uncertainty as possible. Obtaining a valid measure of uncertainty is not just a matter of looking up a formula." In other words, no formula should be applied without checking its approximate validity. Naturally, this does not only refer to "classical" formulas, such as $\text{var}(\overline{X}) = \sigma^2 n^{-1}$, but also to the "new" formulas, such as $\text{var}(\overline{X}) = L(n)n^{2H-2}$ $(0 < H < 1)$, given in the present review paper.

One major reason why the question of long memory is usually not dealt with in daily statistical practice is the lack of statistical software packages. Haslett and Raftery's program (and its implementation in the next release of SPLUS) is therefore a welcome contribution. As already mentioned briefly after formula (12) and

described in more detail in Raftery's comment, an efficient algorithm for maximum likelihood estimation for fractional ARIMA processes can be found in Haslett and Raftery (1989). Another fast method of fitting parametric models with long memory to very long-time series is given in Beran and Terrin (1992). This method is especially suitable for computers where calculations can be done on several parallel processors simultaneously.

Most of the technical comments and questions in the discussion are challenging, and mostly unsolved, research problems.

Parzen suggests to consider the CUSUM statistic. This is certainly a direction worth pursuing. He also suggests to define the spectral density f as the asymptotic limit of the expected value of the periodogram. This has the advantage that the spectral density is also meaningful for $H \geq 1$, although it is not integrable there. Values of $H \geq 1$ are often observed in engineering and physics (some references are given, e.g., in Percival, 1985). Using the above definition of the spectral density, Solo (1989) derived the asymptotic distribution of the periodogram for a class of nonstationary processes where $f(x) \sim cx^{1-2H}$ (as $x \to 0$) with $H \geq 1$ and $c > 0$. This also partially answers the question by Dempster and Hwang on whether there are possible extensions beyond $H = 1$. However, it is not clear if the models considered by Solo indeed yield anything more than the correct spectrum. For example, the nonstationarity of those processes might contradict at least the visual impression of observed time series with $H \geq 1$. I am not aware of any comparisons of simulated sample paths of these processes with real data. Certainly, more research needs to be done to obtain useful stochastic processes with $H \geq 1$ and to derive suitable statistical methods.

Another intriguing point raised by Smith and Parzen is how to obtain a good estimate of H without knowing, or without estimating, the entire shape of the spectrum. One possibility is to bound the influence of periodogram ordinates at higher frequencies. This approach is taken by Graf (1983). Another possibility is to use periodogram ordinates at low frequencies only, such as proposed, for example, in Geweke and Porter-Hudak (1983). Since we are disregarding a part of the information, we might lose a considerable amount of efficiency. The difficulty is to choose a cut-off point such that, on one hand, $\hat{H}$ is not too biased and, on the other hand, the variance of $\hat{H}$ is not too large. The question of what the "best" cutoff point is, is the subject of current research, and no definite answer is known at the present time. Generally speaking, the hope is that one can exploit the analogy between the estimation of the tail of a distribution and estimation of the pole of the spectrum at zero. A complication arises from the fact pointed out by Smith—for very low frequencies the approximation of the distribution of the periodogram by the exponential distribution is very inaccurate. Results along this line can be found in Künsch (1986a) and Hurvich and Beltrao (1992). Künsch suggests to leave out a few of the lowest frequencies. How do we decide how many frequencies we have to leave out? Smith's calculations illustrate the difficulty.

Apart from the shape of the distribution, the expected value of the periodogram $I(x)$ at low frequencies can be far from the asymptotic value $f(x)$. This is taken into account in Graf's method by replacing $f(x)$ by the exact value of $E(I(x))$ which can be calculated exactly for each n.

Smith mentions two estimators, a least squares estimator (LS) and a maximum likelihood type estimator (ML). The ML estimator is in fact a simplified version of Graf's HUB00 estimator, in that it uses $f(x)$ instead of the exact value $E(I(x))$.

Smith's questions concerning Graf's HUBINC estimator can be answered as follows: a theoretical (or at least heuristic) justification of the HUBINC estimator is given in Graf's unpublished PhD thesis (Graf, 1983; also see Graf et al., 1984). A more thorough mathematical theory is currently being developed in an ongoing research project. Some major points are however already clear at the current stage—deviations from the model spectrum at high frequencies do not influence the estimate too much. The HUBINC estimator bounds the influence of both positive and negative outliers. Thus, the group of negative outliers in Figure 1 and the extreme negative outlier for the NBS data do not have any great effect on the result. (One should mention that Graf also proposed a so-called HUBa estimator where a chosen percentage a of the most extreme residuals [largest in absolute value] is huberized.) In contrast to that, LS- and ML-type methods are very much influenced by outliers. Thus, comparing the results of LS and ML estimates for different subsets of the data—with and without certain outliers—mainly reflects the instability of these estimators. It does not give us a measure of uncertainty for a robust estimator.

The strength of good robust methods is their stability with respect to inference—small bias and almost the same distribution (as under the ideal model) under deviations from the model. In particular, the latter point implies that confidence intervals for the HUBINC estimator derived under the ideal assumption of fractional Gaussian noise, should give us a realistic measure of variability, even if the spectral density is not exactly equal to the spectral density of fractional Gaussian noise, and most likely this measure of variability is much more accurate than comparisons of LS and ML estimates when leaving out certain outliers. This is at least the case, when we consider outliers at

frequencies that are not too close to zero. Near the origin the estimator is not robust, though it at least uses $E(I(x))$ instead of $f(x)$, so that deviations from the exponential distribution might cause a problem. This point will need to be addressed in future research. It is in particular a problem of model choice or, when we are not willing to assume any parametric model, of nonparametric estimation. For a given parametric model, one can easily check by simulations how far the different behaviour of the periodogram at the very low frequencies influences $\hat{H}$. For instance, in the case of fractional Gaussian noise, $\hat{H}$ is almost unbiased and its distribution is already very well approximated by the central limit theorem [see equation (13)] for fairly small sample sizes (about $n \geq 200$). However, which minimal sample size is required in order that the effect of the periodogram ordinates at the lowest frequencies is negligible, depends on the model.

Another difficulty with LS and ML estimators is worth mentioning—residuals from a nonrobust fit are not a reliable diagnostic tool for the detection of outliers. Often the actual outliers have small residuals, because the fitted curve tends to be pulled towards them. Therefore, apart from the obvious extreme negative outlier for the NBS data, it seems difficult to interpret the residual and probability plots in Figures 1 and 2.

Dempster and Hwang take a Bayesian view. Since for short series the variability of $\hat{H}$ is rather high, as is demonstrated in their comment, the Bayesian approach can be useful wherever it can be applied in a reasonable way. Dempster and Hwang show how repeated measurements can improve the precision of $\hat{H}$ dramatically. In real data, we might not expect each of the single time series (or individual) to have the same value of H. Instead, we might think of H as an individual characteristic. We draw a random sample from the population of individuals so that $\hat{H}$ becomes a random variable. It seems natural to consider empirical Bayes methods in this context.

Dempster and Hwang also illustrate some other interesting features of long-memory processes—the interrelationship between estimating the variance and estimating H, and a comparison of the sample mean and optimal prediction. There is nothing to be added to their illuminating comments.

ADDITIONAL REFERENCES

BERAN, J. and TERRIN, N. (1992) Testing for change points and estimation of the long-memory parameter, based on a multivariate central limit theorem. Unpublished manuscript.

BLOOMFIELD, P. (1992). Trends in global temperature. *Climatic Change* 21 1–16.

BLOOMFIELD, P. and NYCHKA, D. (1992). Climate spectra and detecting climate change. *Climatic Change* 21 275–287.

DEMPSTER, A. P. and HWANG, J.-S. (1992). Bayesian implementation of a complex hierarchical model. Presented at the International Symposium on Multivariate Analysis and Its Applications, Hong Kong.

HWANG, J.-S. (1992). Prototype Bayesian estimation of US employment and unemployment rates. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

RAMSEY, F. (1974). Characterization of the partial autocorrelation function. *Ann. Statist.* 2 1296–1303.

SMITH, R. L. (1993), Long-range dependence and global warming. In *Statistics in the Environment* (V. Barnett and F. Turkman, eds.). Wiley, Chichester. To appear.