Comment: Computational Aspects of Fractionally Differenced ARIMA Modeling for Long-Memory Time Series

Adrian E. Raftery

Congratulations to Jan Beran on an excellent survey of statistical methods for long-memory time series! This is an important phenomenon in practice because it arises frequently and, in spite of being hard to detect, can lead to completely invalid inferences if it is ignored.

In our study of Irish wind speeds (Haslett and Raftery, 1989), we found $\hat{d} = 0.328$ (i.e., $\hat{H} = 0.828$), which seems to be a typical value for meteorological time series. (This data set can be obtained by sending a message to *statlib@stat.cmu.edu* consisting of the single line "send wind from data." It belongs to the Irish Meteorological Service, who have agreed to release it on condition that it be used only for research into statistical methods.)

We were working on the project for 4 years before we noticed the long-memory dependence. Yet its effect is enormous: for estimating the mean wind speed at a site, 20 years of data contains about the same amount of information as would just 1 month of independent daily values. Meteorology is one area where long-memory dependence is widespread; it would seem important, for example, to take account of its possible existence when estimating and testing for global warming effects.

Beran did not give much attention to the practical aspects of estimation for models where the high- and medium-frequency components of the spectrum are specified separately from the low-frequency/long-memory component. This is important in practice because the short-range dependence structure may well differ from what would be predicted by a model of the long-memory component alone. The fractionally differenced ARIMA model given by Beran's equation (4) is a flexible framework for modeling the entire spectrum.

Because this is a linear stationary process, one can compute the likelihood exactly using the partial linear regression coefficients as calculated from the Durbin-Levinson recursion (Ramsey, 1974; Hosking, 1982). One can then obtain maximum likelihood estimates by maximizing this with a numerical optimization algorithm that does not use derivatives. However, the required computer time is asymptotically $O(n^2)$ and is in practice large for the long series typical of the areas where time series have been found most often to possess long memory. This would preclude interactive model comparison and exploration in many typical applications. The wind speed time series in Haslett and Raftery (1989) were of length $n = 6{,}574$; meteorological time series are typically this long and often much longer. By contrast, Beran's longest series is about one-tenth as long as this.

In Haslett and Raftery (1989), we developed an approximation to the likelihood that is accurate, reduces asymptotic computer time from $O(n^2)$ to $O(n)$ and in practice reduced computer time by about two orders of magnitude for the wind series. This is given by equations (4.3) through (4.8) in Haslett and Raftery (1989) and consists of approximating the higher-lag partial linear regressions coefficients (lags above M), but using the lower-lag ones exactly. In numerical experiments with $n = 1{,}000$ and $M = 100$, for example, the difference between the exact and approximate likelihoods was typically less than the contribution of a single observation. This opens the way to routine exploratory fitting of such models, both frequentist and Bayesian, even for long series. By comparison, the approximation in Beran's equation (12) requires $O(n^2)$ computer time and may be less exact because it does not use the exact values of the important lower-lag partial linear regression coefficients.

Software to calculate maximum likelihood estimators for fractionally differenced ARIMA models using the approximation of Haslett and Raftery (1989) may be obtained from StatLib. There are two versions: a Fortran version and an S version. The Fortran version may be obtained by sending an e-mail message to *statlib@stat.cmu.edu* containing the single line "send fracdiff from general." The S version may be obtained by sending the message "send fracdiff from S" to the same address. This software yields exact maximum likelihood estimates by setting $M = n$. The S version also includes an S function for simulating the models. It is planned to include these S functions in version 3.1 of S-PLUS.

Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195.

ACKNOWLEDGMENTS

This research was supported by ONR Contract N00014-91-J-1074. I am grateful to Chris Fraley for writing the Fracdiff software and for helpful comments, and to my long-time collaborator John Haslett for many useful discussions.

Comment

Richard L. Smith

Jan Beran has written an excellent and timely review of a topic that is gaining increasing attention in a whole variety of fields. As his review makes clear, the origins of the subject go back a long way and were rooted in practical problems in several fields. However, it is only in recent years, stimulated by the development of the fractional ARIMA model, that the subject has started to receive widespread attention among statisticians. Beran does a superb job of bringing together the extensive results that now exist on the effects of long-range dependence on a whole range of statistical inferences. Nevertheless, I suspect it is in the identification and estimation of long-range models themselves that readers will take the greatest interest, and it is here that I concentrate my comments.

A common feature of long-range models is that the spectral density $f(x)$ satisfies the relation

$$f(\omega) \sim b\omega^{1-2H}, \quad \omega \to 0. \tag{1}$$

Beran's equation (6) is a slight generalization of this, replacing the constant b by a slowly varying function, but for most purposes (1) suffices. One feature of many of the results about the effect of long-range dependence, such as Beran's equation (8), is that they depend on the spectral density only through the constants b and H. In fact, (8) itself depends only on H, but many related results depend also on the scaling constant b. For this reason, it is of interest to look for direct estimators of b and H, rather than assume some parametric model such as fractional ARIMA. I have been particularly interested in estimators based on the periodogram, which are among those reviewed in Section 2.4. If $I_n(\omega)$ denotes the periodogram at frequency ω based on n observations, then it is "well-known" that the sampling distribution of $I_n(\omega_j)$ at the Fourier frequencies $\omega_j = 2\pi j/n$ for $0 \le j < n/2$ is approximately that of independent exponential random variables with means $f(\omega_j)$. If we assume $f(\omega) = b\omega^{1-2H}$ then this suggests that $1 - 2H$ could be estimated as the slope of a linear regression of $\log I_n(\omega_j)$ on $\log \omega_j$. This idea has been suggested by a number of authors, in particular Geweke and Porter-Hudak (1983). Two refinements of Geweke and Porter-Hudak seem desirable:

a. Geweke and Porter-Hudak used least squares regression of log periodogram ordinate on log frequency. In contrast, since the asymptotic distribution of $I_n(\omega_j)$ is exponential, a regression of $\log I_n(\omega_j)$ based on errors from the Gumbel distribution function $1 - \exp(-e^x)$ would seem preferable. I call this the maximum likelihood (ML) approach, in contrast to Geweke and Porter-Hudak's least squares (LS) approach.

b. In addition, it is becoming increasingly clear that it is necessary to restrict the range of frequencies used in the regression, say to $n_0 \le j \le n_1$ where $1 < n_0 < n_1 << n/2$. At the lower end, the difficulty is that the above-mentioned "well-known" properties of the periodogram apparently break down for very low frequencies in the case of a long-range model (see, e.g., Künsch, 1987; Haslett and Raftery, 1989). At the upper end, the problem arises from the fact that (1) is only an asymptotic relation, not an identity, so attention must be restricted to small ω. A more formal argument along these lines was presented in my discussion of Haslett and Raftery (1989).

It seems to me that Graf's HUB00 and HUBINC estimators deal with problem (a), albeit in a quite different way from the ML approach being suggested here, but do not contain anything that corresponds directly to the selection of n_0 and n_1. In view of this, I am somewhat doubtful about the theoretical justification of these estimators.

The rest of this discussion concerns three examples, two of them taken from Beran's paper, which illustrate the importance of appropriate selection of n_0 and n_1 in this approach.

The first of these is the Nile data. Beran's Figure 3 plots the periodogram in log-log coordinates. It can be seen that the plot is decreasing at an approximately

Richard L. Smith is Professor, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27599-3260.