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.