Nonparametric Forecasting Model
:Application to Forecasts of El Niño/Na niña Events

○Young-Il Moon[1)]

1. Introduction

Short term forecasts of streamflow and lake levels are made routinely by various methods and are used to guide the operation of water resource facilities. Recently (Lall et al., in press, a; Abarbanel et al., 1996; Kember et al., 1993; Smith, 1991; Yakowitz and Karlsson, 1987), nonparametric regression methods have been proposed for forecasting hydrologic time series. Lall et al. (in press, a) were able to forecast the volume of the Great Salt Lake (GSL) for up to 4 years in advance during extreme conditions. They formulated a forecasting model using recent techniques (Abarbanel et al., 1993) for reconstructing the dynamics of a nonlinear system from a single observed state variable. Multivariate Adaptive Regression Splines (MARS) due to Friedman (1991) were used to nonparametrically recover the nonlinear forecasting function from the time series of GSL volume. Such methods for time series analysis are computationally intensive, and can also require long high quality records.

Efficient parameter selection is important for nonparametric function approximation. The strategy provided here is capable of automatically selecting the size of the neighborhood and the order of the polynomial used at each point of estimate. This allows one to represent linear (e.g., classical AR models) or polynomial dynamics, as well as locally approximating more complex dynamics.

In this paper, it is presented that the application of multivariate, locally weighted polynomial regression with locally chosen parameters for nonparametrically approximating the dynamics of the system at each point of prediction. First, Singular Spectrum Analysis(Vautard et al., 1992; Keppenne and Ghil, 1992) is applied to a time series of the Southern Oscillation index(SOI) and then the SOI is filtered out unrelated variability to El Niño events or La Niña events. Blind forecasting El Niño/La Niña up to two years using the filtered SOI are presented.

2. The Nonparametric Forecasting Model

The forecast $f(x_n)$ at time T is obtained through the solution to a general regression model given as

$$y_i = f(x_i) + e_i \qquad i=1,\dots,n \tag{1}$$

1) Department of Civil Engineering, Assistant Professor
University of Seoul, Seoul, Korea

where the function f(.) can be thought of as a regression function.

A nonparametric regression problem results if consider a solution of this problem such that (1) no prior assumption is made about the explicit functional form of f(.), (2) the interest is in approximating f(.) at each desired location, presuming that it belongs to a fairly rich class of functions (e.g., differentiable functions), and (3) the estimate is "local," i.e., the influence of distant points on the regression at a given point diminishes with distance. The target function f(.) may be approximated at a point x_n by retaining the leading terms in its Taylor's series expansion.

This is equivalent to a low order polynomial approximation of the function at that point using k neighboring data points. The idea is illustrated for the univariate case in Figure 1. In the multivariate case one uses k neighbors x_j , j =1...k, of x_n in a vector space of dimension d, to evaluate a low order polynomial regression using the corresponding y_j. The k neighbors are found as the state vectors that are closest in distance to the vector x_n. Thus, in the time series context it is located the k data patterns that are most similar to the state vector $\overline{v_t}$, and evaluated a low-order polynomial regression with these data as an approximation to f($\overline{v_t}$).

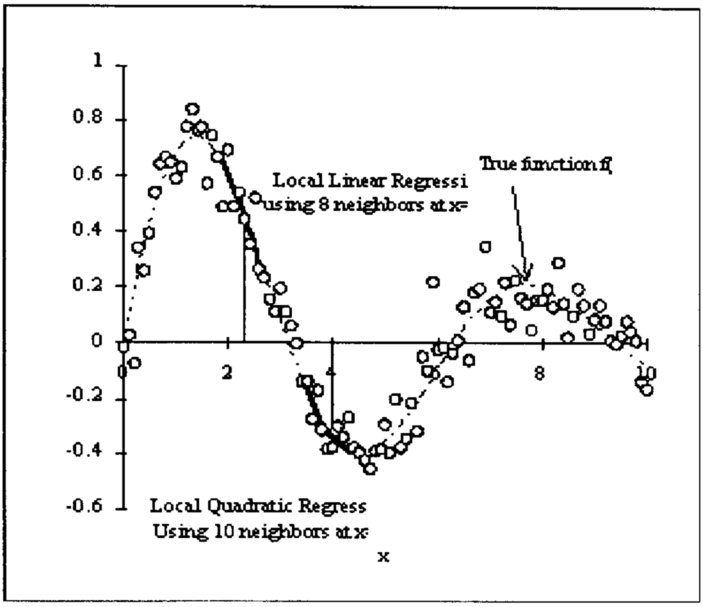

Figure 1. Local linear and local quadratic approximation of f(x) = sin (x)$e^{-0.2x}$ at two points. This "damped" oscillation is representative of the quasi-periodic oscillations seen in the SOI (Southern Oscillation Index) data upon bandpassing it at frequencies that have high spectral power. The data (circles) were generated using equation (1), with $e_i \sim N(0,0.1)$. The function f(x) is shown as the dashed line, and the local regressions are shown as heavy solid lines. For forecasting, x would be a d dimensional vector in state space, the neighbors would be the closest points in IR^d, and a multivariate local regression will be needed.

A detailed exposition of weighted local regression may be found in Cleveland(1979), Cleveland and Devlin (1988), Cleveland et al. (1988), and Lall et al. (in press, a).

Localization of the regression is achieved by using only k neighbors of the prediction point, and also by weighting the data with a monotonic weight function, with weights decreasing as a function of distance of the neighbor from the prediction point.

In this paper, locally linear (p=1) , quadratic (p=2), and quadratic with cross products (p=2') approximations are considered. Say Z is denoted as a data matrix formed by augmenting the matrix $\{x\}_{k,n}$ of k nearest neighbors of x_n to complete a polynomial basis of order p. If p =1, i.e., a linear regression is needed, then Z is formed by augmenting $\{x\}_{k,n}$ by a column with all entries equal to 1, to represent the constant term in the regression. If p=2, one also adds the square of each column of $\{x\}_{k,n}$ If p=2', then the all unique cross products across columns in $\{x\}_{k,n}$, are also added to Z. The number of neighbors k considered ranges from 2*d' to n, where d' is the column dimension of Z. Thus global linear and quadratic models are parts of the set considered. Any data vector x_i is similarly mapped into a data matrix z_i.

The order p weighted local regression using k nearest neighbors is then defined through the model

$$y = Z\beta + e \qquad (2)$$

where y is a k*1 vector, Z is a k*d' matrix, β is a d'*1 vector of regression coefficients and e is a k*1 vector of residuals that are assumed to be independent and locally homogeneous.

The coefficients β are evaluated as the solution to

$$\min_{\beta}(y - Z\beta)^T W (y - Z\beta) \qquad (3)$$

which is given as

$$\beta = (Z^T W Z)^{-1} Z^T W y \qquad (4)$$

which gives us the desired forecast as

$$SOI_{t+1} = z_n \beta \qquad (5)$$

where z_n is the polynomial basis representation of the prediction state vector V_T.

The quality of such a low-order weighted polynomial approximation depends on the size of the neighborhood and the order of the polynomial. For a given order, as the size of the neighborhood increases, the variance of estimate decreases while the bias of estimate may increase. Likewise, increasing the order of the polynomial may reduce the bias or approximation error, while increasing the variance of estimate if the number of points in the neighborhood is kept the same. This bias-variance trade-off suggests the possibility of searching for an optimal model for local estimation by varying the order of the local polynomial, and the size of the neighborhood.

3. Application

The SOI(Southern Oscilliation Index) is formed as the monthly mean difference in sea-level pressure (in mb) (SLP) at Tahiti (approximately 150W, 17.5S) and Darwin (approximately 131E, 12.5S). The El Niño Southern Oscillation (ENSO) refers to an event in the tropical Pacific Ocean that is a significant perturbation of general atmospheric circulation. The ENSO has a typical life cycle of about 3-8 years. When the SOI is a low negative value, a warm event (El Niño) is in progress, i.e., the atmospheric pressure in the eastern Pacific decreases, and the trade winds usually weaken. Then, the warm-water pool extends eastward, piling up off the coast of Peru and southern Ecuador. The opposite phase of the ENSO is called a La Niña event. An important aspect of an ENSO event is the change in precipitation patterns over the world. Western United States precipitation and streamflow are enhanced during an El Nino event and drought occurs over the continental United States with a La Niña event(Kahya and Dracup, 1993; Ropelewski and Halpert, 1987; Keppenne and Ghil, 1992).

The application in this paper is the forecast of Southern Oscillation Index. First, Singular Spectrum Analysis(Vautard and Ghil, 1989; Rasmusson, 1990; Vautard et al., 1992; Keppenne and Ghil, 1992) is applied to a time series of the Southern Oscillation index(SOI). The SSA filters out variability unrelated to the Souther Oscillation and separates the high frequency variability from the lower frequency El Niño cycle(Moon and Lall , 1996). In Figure 2, the filtered SOI(RC1-6) is compared with a 5 month moving average of the original SOI time series (dotted). The white arrows points to El Niño events and the black ones to La Niña events. The correlation between the RC 1-6 and 5-month moving average is 0.95. The 5 month moving average is recommended by U.S National Meteorologic Center (NMC) for noise reduction of the ENSO. The blind forecasts of the filtered SOI(RC1-6) from different states for 2 years into the future from the date of forecast. The forecasted values are then compared with the RC1-6 that were from SSA. They are presented in Figure 3. The lag τ was selected as 5 as in the range of the first minimum of the average mutual information (Moon et al., 1995) and it was based on experimentation to get the best predictions (min predictive squared error). An embedding of m=6 was selected after experimentation with various values in the range 1 to 9. This value corresponded to the one that most commonly minimized LGCV(Locally Generalized Cross Validation). It was searched over k1=30 to k2=150 nearest neighbors. Locally linear and quadratic fits were considered. Typically a linear fit was selected. In Figure 3, Apr. 1995-Mar. 1997 (yp1) and Apr. 1997-Mar. 1999 (yp2) blind forecasts of RC 1-6 (τ=5 and M=6) are presented, using only data from Sep. 1932 to Mar. 1995 and from Sep. 1932 to Mar. 1999. The behaviors in the forecast and RC1-6 filtered data are coincidental.

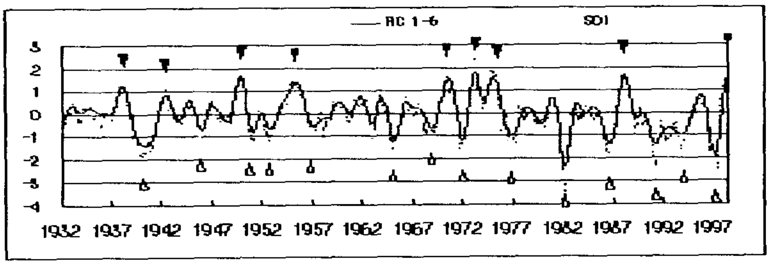

Figure 2. Comparison of the filtered SOI(RC1-6) with the SOI.

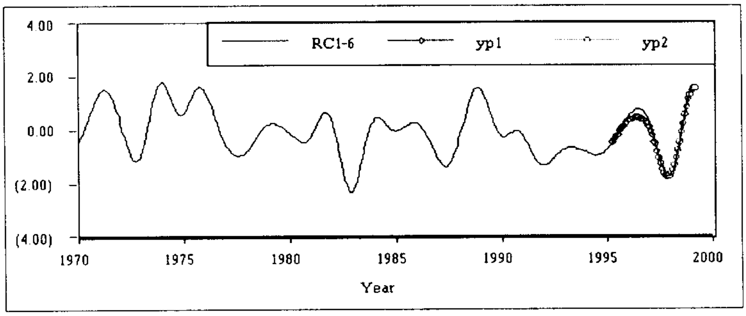

Figure 3. Apr. 1995-Mar. 1997 (yp1) and Apr. 1997-Mar. 1999 (yp2) blind forecasts of RC1-6 (τ =5 and M=6), using only data from Sep 1932 to Mar. 1995 and from Sep 1932 to Mar. 1999 respectively.

4. Conclusions

The utility of a locally weighted polynomial regression approach is demonstrated through an application to nonparametric short term forecasts of the filtered Southern Oscillation Index. Locally weighted polynomials consider the approximation of the target function through a Taylor series expansion of the function in the neighborhood of the point of estimate. This locally weighted polynomial algorithm is an useful tool for the SOI series forecasting. However, the purpose of this paper was in exploring the utility of the local polynomial regression approach for the time series prediction. Applications to various hydrologic time series forecasting and spatial surface reconstruction are also in progress.

Acknowledgements

Partial support of this work by the USGS grant # 1434-92-G-226 and NSF grant # EAR-9205727 is acknowledged.

5. References

Abarbanel, H.D.I., R. Brown, J.J. Sidorowich, and L.S. Tsimring, The analysis of observed chaotic data in physical systems, Rev. of Modern Phys., Vol.65 (N4), 1331-1392, 1993.

Abarbanel, H.D.I., U. Lall, Young-Il Moon, M. Mann, and T. Sangoyomi, Nonlinear dynamics of the Great Salt Lake: A predictable indicator of regional climate, Energy, Vol.21(7/8), 655-665, 1996.

Cleveland, W.S., Robust locally weighted regression and smoothing scatterplots, J. Amer. Stat. Assoc., Vol.74 (368), 829-836, 1979.

Cleveland, W.S., and S.J. Devlin, Locally weighted regression: An approach to regression. analysis by local fitting, J. Amer. Stat. Assn., Vol.83 (403), 596-610,

1988.
Cleveland, W.S., S.J. Devlin, and E. Grosse, Regression by local fitting, J. Econometrics, 37, 87-114, 1988.
Friedman, J.H., Multivariate adaptive regression splines, Ann. Statist., Vol.*19* (1), 1-141, 1991.
Kahya, E., and J.A. Dracup, U.S. streamflow patterns in relation to the El Niño/Southern Oscillation, Water Resources Research, Vol.29(8), 2491-2503, 1993.
Kember, G., A.C. Flower, and J. Holubeshen, Forecasting river flow using nonlinear dynamics, Stoch. Hydrol. Hydraul., Vol.7, 205-212, 1993.
Keppenne, C.L., and M. Ghil, Adaptive filtering and prediction of the southern oscillation index, J. of Geophysical Research, Vol.97, 20449-20454, 1992.
Kiladis, G.N., and H.F. Diaz, Global climatic anomalies associated with extremes in the Southern Oscillation, J. of Climate, Vol.2, 1069-1090, 1989.
Lall, U., Young-Il Moon, and K. Bosworth, Locally weighted polynomial regression: Parameter choice and application to forecasts of the Great Salt Lake, ASCE J. of Hydrologic Engineering, in press, a.
Lall, U., T. Sangoyomi, and H.D.I. Abarbanel, Nonlinear dynamics Of the Great Salt Lake: Nonparametric short term forecasting. Water Res. Res., in press, b.
Moon, Young-Il, B. Rajagopalan, and U. Lall, Estimation of mutual information using kernel density estimators, Physical Review E, Vol.52(3), 2318-2321, 1995.
Moon, Young-Il, and U. Lall, Atmospheric flow indices and interannual Great Salt Lake variability, ASCE J. of Hydrologic Engineering, Vol.1(2), 55-62, 1996.
Rasmusson, E. M., X. Wang, and C. F. Ropelewski, The Biennial Component of ENSO Variability, J. Mar. Sys., Vol.1, 71-96, 1990.
Ropelewski, C.F., and M.S. Halpert, Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation, Monthly Weather Review, Vol.115,1606-1626, 1987.
Smith, J.A., Long-range streamflow forecasting using nonparametric regression, Water Resour. Bull., Vol.27 (1), 39-46, 1991.
Vautard, R., and M. Ghil, Singular Spectrum Analysis in Nonlinear Dynamics, with Applications to Paleoclimatic Time Series, Phys. D, Amsterdam, Vol.35, 395-424, 1989.
Vautard R., S. Yiou, and M. Ghil, Singular Spectrum Analysis: A Toolkit for Short, Noisy Chaotic Signals, Physica D, Vol.58, 95-126, 1992.
Yakowitz, S., and M. Karlsson, Nearest neighbor methods with application to rainfall/runoff prediction, in Stochastic Hydrology. edited by J.B. Macneil, and G.J. Humphries, 149-160, D. Reidel, Hingham, MA, 1987.