韓國水資源學會論文集
第33卷 第1號・2000年 2月
pp. 31～38

지역가중다항식을 이용한 예측모형

Locally Weighted Polynomial Forecasting Model

문 영 일*

Moon, Young-Il

Abstract

Relationships between hydrologic variables are often nonlinear. Usually the functional form of such a relationship is not known a priori. A multivariate, nonparametric regression methodology is provided here for approximating the underlying regression function using locally weighted polynomials. 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. The utility of this nonparametric regression approach is demonstrated through an application to nonparametric short term forecasts of the biweekly Great Salt Lake volume.

keywords: locally weighted polynomial, nonparametric regression, Great Salt Lake

요 지

수문변량 사이의 관계는 대부분 비선형 관계를 보이고 있다. 일반적으로 이런 비선형 관계는 어떤 선행하는 명백한 하나의 함수적인 형태로 표현할 수 없는 것이 일반적이다. 본 논문에서는, 비매개변수적 다변량 회귀분석 방법을 지역적으로 가중된 다항식을 이용하여 비선형 예상 함수를 근사 추정하였다. 지역적으로 가중된 다항식은 추정치 각 점에서의 인접한 이웃자료를 가지고 목적 함수를 테일러 급수 확장을 통하여 고려하였다. 이런 비매개변수적 회귀분석 방법의 실용성을 Great Salt Lake의 격주 체적자료에 대한 단기간 예측을 통하여 보여주었다.

핵심용어 : 지역가중다항식, 비매개변수적 회귀식, Great Salt Lake

* 서울시립대학교 토목공학과 조교수
Assistant Professor, Dept. of Civil Engineering, University of Seoul, Seoul 130-143, Korea

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, nonparametric regression methods (Lall et al., in press; Abarbanel et al., 1996; Kember et al., 1993; Smith, 1991; Yakowitz and Karlsson, 1987) have been proposed for forecasting hydrologic time series. Lall et al. (in press) were able to forecast the volume of the Great Salt Lake (GSL) 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. Locally Weighted Polynomial Regression (LWPR) was 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. Blind forecasting the Great Salt Lake volume up to four years using the 1847-1999 time series are presented.

2. Locally Weighted Polynomial Forecasting Model

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

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

where the function $f(\cdot)$ 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 $f(\cdot)$, (2) the interest is in approximating $f(\cdot)$ 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 $f(\cdot)$ 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 Fig. 1. The "damped" oscillation in Fig. 1 is representative of the quasi-periodic oscillations seen in the GSL data upon bandpassing it at frequencies that have high spectral power. The data (circles) were generated using $f(x) = \sin(x)\, e^{-0.2x}$, 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. In the multivariate case one uses k neighbors x_j, $j = 1 \cdots k$, of x_n in a

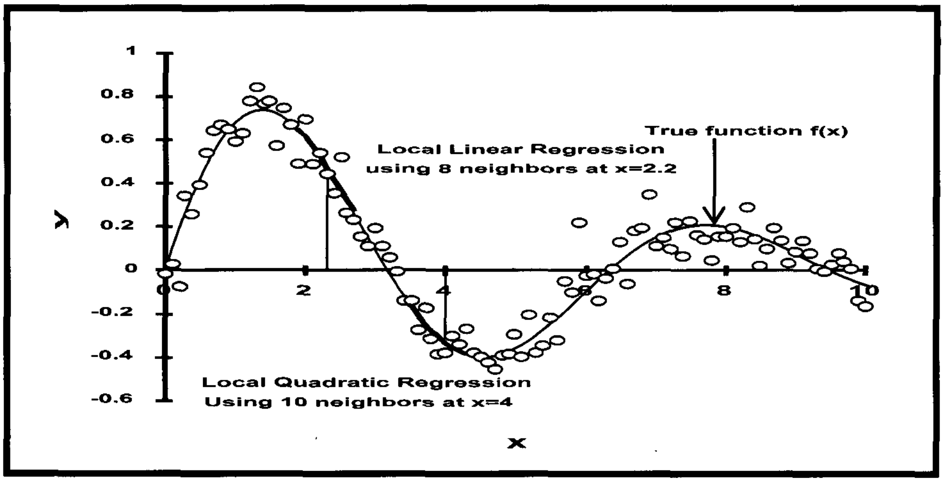

Fig. 1. Local Linear and Local Quadratic Approximation of $f(x) = \sin(x)e^{-0.2x}$ at Two Points

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 $f(\overline{v_t})$. The state space coordinate $\overline{v_t}$ are defined as

$$\overline{V_t} = \{x_{t,}\ x_{t-\tau,}\ x_{t-2\tau,}\ x_{t-3\tau,}\ldots\ldots,\ x_{t-\tau}(m-1)\} \tag{2}$$

where τ is a time delay and m is an embedding dimension.

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). 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 $2\times d'$ 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

$$y = Z\beta + e \tag{3}$$

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

The coefficients β are evaluated through the solution of the weighted least squares problem:

$$\underset{\beta}{\mathrm{Min}}(y - Z\beta)^T W (y - Z\beta) \tag{4}$$

which is given as

$$\beta = (Z^T W Z)^{-1} Z^T W y \tag{5}$$

where W is the weight matrix for estimation at z_i. The resulting forecast is then:

$$\hat{f}(x_i) = z_n \beta \tag{6}$$

where z_n is the polynomial basis representation of the prediction state vector $\overline{V_t}$.

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. Here, the parameter selection method is based on Locally Generalized Cross Validation (LGCV) (Moon, 1997; Lall et al., in press). The LGCV score is then given as:

$$\mathrm{LGCV}(\hat{f}) = \frac{e^T W e}{\left(\frac{k - d'}{k}\right)^2} \tag{7}$$

where the errors e are the residues of the model fitted over the k nearest neighbors; W is the corresponding weight matrix; d' is the number of coefficients fitted.

The appropriate values of k and p can then be obtained as the ones that minimize the LGCV score for the local regression.

3. Application

The primary application considered in this paper is the forecast of the volume of the Great Salt Lake at key points in time from its 1847-1999, biweekly time series. However, it will begin by forecasts of data from two known models to assure itself that the forecasting scheme can work.

3.1 Synthetic Series

The first scenario is a time series of length 200 from an AR(2), or autoregressive model of lag 2. The model is defined by:

$$y_t = y_{t-1} - 0.5 y_{t-2} + e_t \tag{8}$$

where e_t is a normally distributed random variable with mean 0 and variance 1.

In this case, $\tau = 1$ and varied m from 1 to 5 were selected. The number of nearest neighbors to use was varied from 50 to the sample length 200, and linear and quadratic (without cross product terms) fits were considered from 1 to 20 various points in the series. In all cases LGCV was used to select the parameters of interest. The number of nearest neighbors was consistently picked as the full sample size, m was typically picked

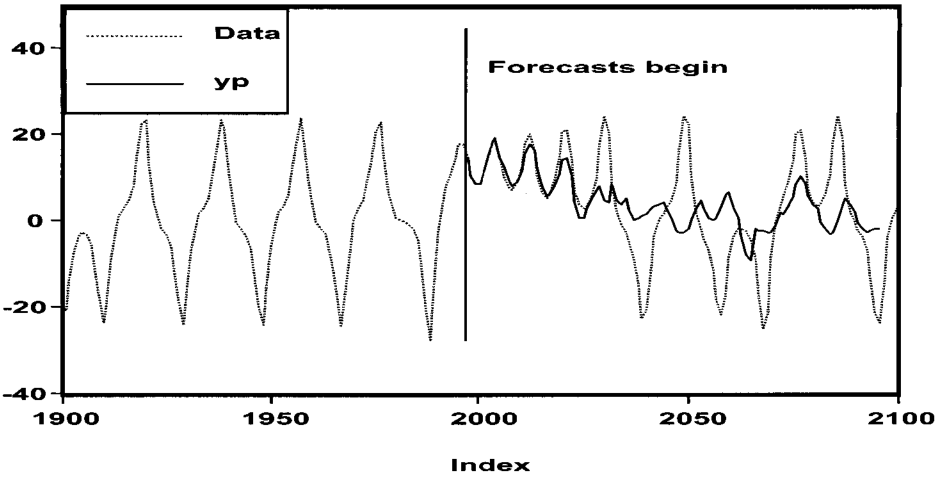

Fig. 2. The Forecasts of Lorenz Data Starting from Index 1997

to be 2, and linear fits were always selected. Since, the models selected were essentially global, linear, autoregressive models each time, the resulting statistical properties were satisfactory.

3.2 Lorenz Equations

The Lorenz equations are given as :

$$\dot{x} = -\sigma(x+y)$$
$$\dot{y} = -xz + rx - y \qquad (9)$$
$$\dot{z} = xy - bz$$

Here $\sigma = 16$, $r = 45.92$, $b = 4$, and $\partial t = 0.05$ were selected, and the x state variable was sampled. This is a chaotic system, that has been well studied by many investigators. From prior work (Moon and Lall, 1996; Moon et al., 1995) it was known that one should expect $\tau = 2$ to 4, and $m = 4$ to 6. Consequently these values and k1=50 to k2=150, and p1=1, p2=2 were investigated. Forecasts from index 1996 of the x time series are presented in Fig. 2. No data after index 1996 were used for the forecasts. The dotted line represents the actual values while the solid lines are the forecasted values ($m=5$, $\tau=3$, $k=50$ at the first 21 points, 90 or 150 at the rest of the points, and p=1 at the first 9 points, p=2 at the rest points). The divergence of the forecasted and the observed trajectories near index 2030, is characteristic of the loss of predictability in the Lorenz system as trajectories pass near the unstable point ($x=y=z=0$). The increase in k and p after the first 20 points may reflect increasing derivatives of $f(\overline{V_t})$ as one approaches the origin.

The Lorenz system has an instability near $x = y = z = 0$. Trajectories that approach this state tend to diverge rapidly. In Fig. 2, the forecasts of the Lorenz x variable are quite good until the trajectory passes near the unstable point. A small uncertainty in the value at index 1996 leads to the trajectories from the numerical simulation of the Lorenz equations diverge similarly. Thus this divergence is intrinsic to this model. Subsequent similarity in the forecast and actual trajectories is coincidental.

3.3 Great Salt Lake Forecasts

The Great Salt Lake (GSL) (latitudes 40° 20′ and 41° 40′ N, and longitudes 111° 52′ and

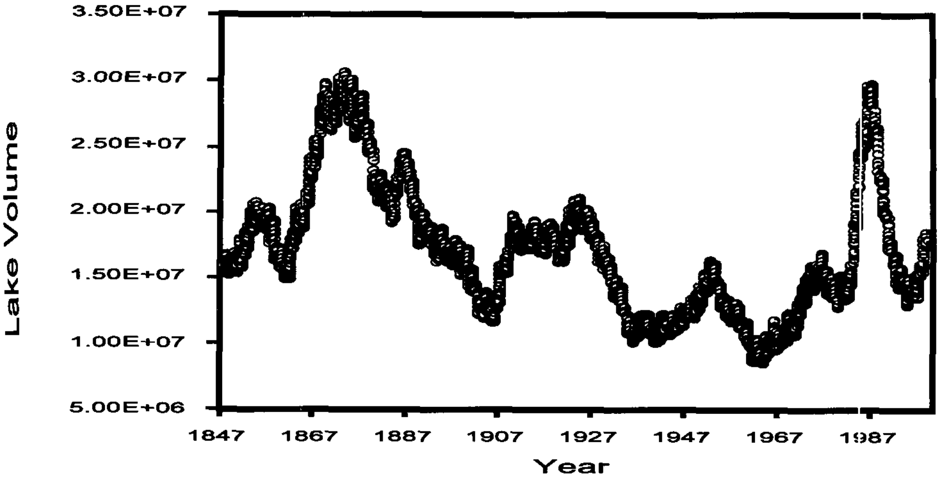

Fig. 3. Biweekly Time Series of the Great Salt Lake, 1847-1999

113° 06′ W) of Utah is a closed lake in the lowest part (elevation 1280 m above Mean Sea Level) of the Great Basin, in the arid Western U.S.A. The GSL is approximately 113 ㎞ long and 48 ㎞ wide, with a maximum depth of 13.1 m and an average depth of 5.0 m. The large surface area and shallow depth make the lake very sensitive to fluctuations in long term climatic variability. Fluctuations of the GSL's level are of direct concern to mineral industries along the shore, the Salt Lake City airport, the Union Pacific Railroad, and Interstate 80. They are also well correlated with regional water supply conditions. As shown in Fig. 3, the lake volume has varied considerably over decadal time scales during the last 150 years. The low frequency character makes this an interesting time series to forecast.

It was considered blind forecasts of the GSL volume from different states for 1 year into the future from the date of forecast. The forecasted values are then compared with the volumes that were actually recorded subsequently. In Fig. 4, the forecasted values for a sequence of 1 year blind forecasts of the GSL, from August 1977 to July 1987, are presented. The dots represent the observed GSL time series. The solid lines represent 12 forecasts, one for each month of the next year. Only data available up to the beginning of the forecast period is used for fitting and forecasting. Given the extreme nature of the 1983-87 period the predictions appear to be quite good. Of particular interest is the forecast starting in August 1983. The predictability is quite poor for this forecast. The lag τ was selected as 10 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 $m = 5$ was selected after experimentation with various values in the range 1 to 9. Usually, this value corresponded to the one that minimized LGCV.

It was searched over k1=50 to k2=150 nearest neighbors and typically selected 120 to 150. Locally linear and quadratic (without x-products) fits were considered. Typically linear was selected. The results are discussed in the figures.

Finally, a forecast of the Great Salt Lake volume for 4 years beginning Feb. 1999 is presented in Fig. 5. The solid line is the

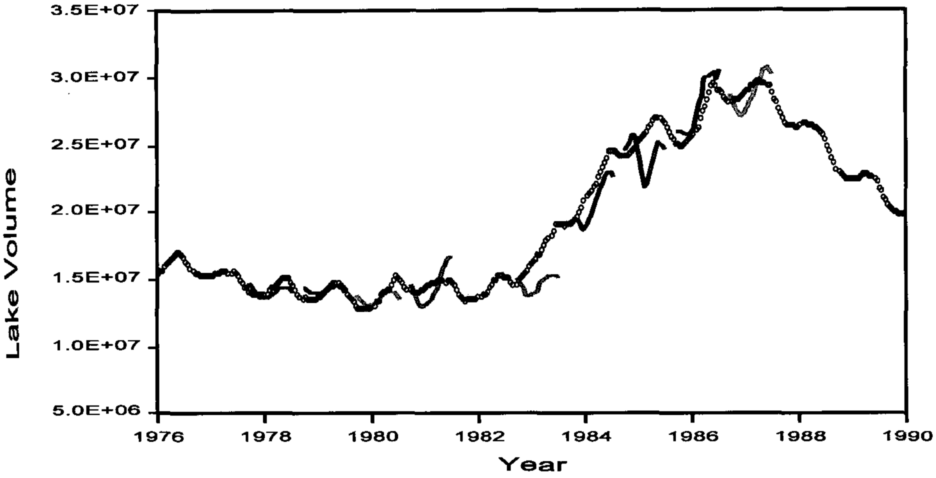

Fig. 4. A Sequence of 1 year Blind Forecasts of the GSL, from August 1977 to July 1987

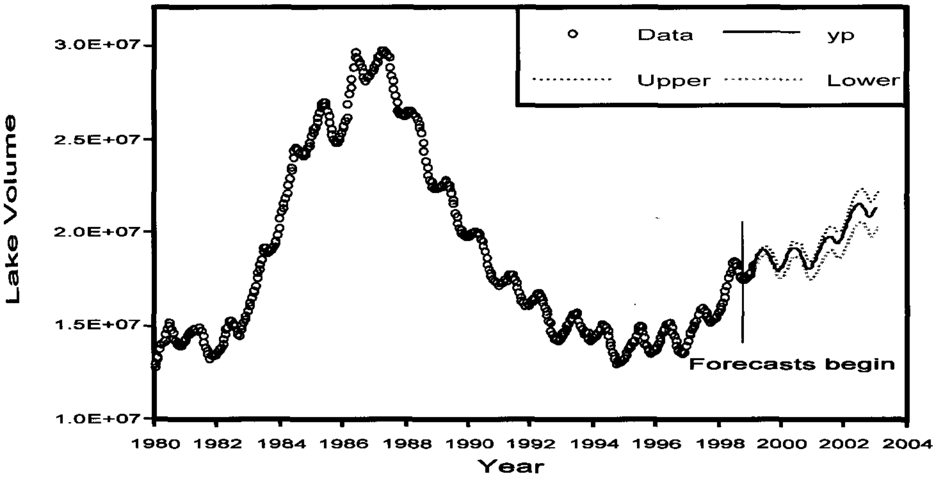

Fig. 5. Forecasts for 4 Years Starting Feb. 1999

forecasted sequence while circles are the actual series. Prediction intervals using LGCV for the forecast are also shown.

4. Conclusions

The utility of a locally weighted polynomial regression approach is demonstrated through an application to nonparametric short term forecasts of the volume of the GSL. 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 model is an useful tool for the GSL 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.

References

Abarbanel, H.D.I., Brown, R. Sidorowich, J.J., and Tsimring, L.S. (1993). "The analysis of observed chaotic data in physical systems." *Rev. of Modern Phys.*, Vol. 65, No. 4, 1331-1392.

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

Cleveland, W.S. (1979). "Robust locally weighted regression and smoothing scatterplots." *J. Amer. Stat. Assoc.*, Vol. 74, No. 368, pp. 829-836.

Cleveland, W.S., and Devlin, S.J. (1988). "Locally weighted regression: An approach to regression. analysis by local fitting." *J. Amer. Stat. Assn.*, Vol. 83, No. 403, pp. 596-610.

Cleveland, W.S., Devlin, S.J., and Grosse, E. (1988). "Regression by local fitting." *J. Econometrics*, Vol. 37, pp. 87-114.

Kember, G., Flower, A.C., and Holubeshen, J. (1993). "Forecasting river flow using nonlinear dynamics." *Stoch. Hydrol. Hydraul.*, Vol. 7, pp. 205-212,

Lall, U., Moon, Young-Il, and Bosworth, K. (in press). "Locally weighted polynomial regression: Parameter choice and application to forecasts of the Great Salt Lake." *J. of Hydrologic Engineering*, ASCE.

Moon, Young-Il, Rajagopalan, B., and Lall, U. (1995). "Estimation of mutual information using kernel density estimators." *Physical Review E*, Vol. 52 No. 3, pp. 2318-2321.

Moon, Young-Il, and Lall, U. (1996). "Atmospheric flow indices and interannual Great Salt Lake variability." *J. of Hydrologic Engineering*, ASCE, Vol. 1, No. 2, pp. 55-62.

Moon, Young-Il. (1997). "A nonparametric nonlinear time series forecasting model application to selected hydrologic variables in Korea." *American Geophysical Union, Fall Meeting*, Vol. 78, pp. 46.

Smith, J.A. (1991). "Long-range streamflow forecasting using nonparametric regression." *Water Resour. Bull.*, Vol. 27, No. 1, pp. 39-46.

Yakowitz, S., and Karlsson, M. (1987). "Nearest neighbor methods with application to rainfall/runoff prediction." *Stochastic hydrology*, Edited by Macneil, J.B., and Humphries, G.J., D. Reidel, Hingham, MA, pp. 149-160.

(논문번호:99-055/접수:1999.07.29/심사완료:1999.11.03)