

Partially Adaptive Bandwidth Used in Prediction with Local Regression

Janis Grabis
Department of Operations Research
Riga Technical University
Kalku 1, Riga LV-1658, Latvia
grabis@itl.rtu.lv

A bandwidth parameter of the local regression model can be set either globally or locally. This paper considers partially adaptive bandwidth selection. The partially adaptive bandwidth is used to adjust the global bandwidth for particular data points. The adjustment takes place if a specified quality criterion of the given local model fails. The partially adaptive bandwidth attempts to encompass both the robustness of the global bandwidth and the flexibility of the local bandwidth. It primary is aimed to reduce large prediction errors. Performance of the partially adaptive bandwidth is evaluated by prediction of several empirical times series coming from the area of hydrology. The results show that more accurate predictions can be obtained.

1. Introduction

A bandwidth parameter determines the performance of local regression. Atkeson et al. (1997) distinguish global and local bandwidth selection procedures. The main strength of the global procedure is its robustness while the local procedure allows adapting to the data density and distribution, variation of noise level and local behavior of an underlying function. The main drawbacks of the local or adaptive bandwidth are potential increase of the prediction errors and risk of overfitting the data. These difficulties can be overcome in smoothing while predictive power of the locally selected bandwidth is questionable. In prediction an effect of adaptation should be assessed prior to start making predictions.

This paper considers a partially adaptive bandwidth of local regression applied in prediction of nonlinear time series. The partially adaptive bandwidth is used to adjust the global bandwidth for data points somehow differing from majority of observations. The adjustment takes place if a specified quality criterion of the given local model fails. Rejection of the quality criterion indicates a risk of large prediction errors and bandwidth adaptation is aimed to reduce this risk. The approach is motivated by believe that the global bandwidth is appropriate for the majority of data points and the adjustment is necessary only for several rarer cases. Therefore the partially adaptive bandwidth is sufficiently

robust and it also encompasses some flexibility of the adaptive bandwidth.

A local linear model (Cleveland, 1979; Fan & Gijbels, 1996) is used for local modeling. Application of local regression in time series analysis is surveyed among others in Heiler (1999) (Tong (1993) and Kantz & Schreiber (1997) provide the general accounts on nonlinear time series analysis). I have a limited information on adaptive prediction with local models and Stenman (1999) is one to refer.

The following section describes the modeling problem and the local linear model. Section three introduces the partially adaptive bandwidth, several alternatives of partial adaptation are considered. Experimental evaluation of the partially adaptive bandwidth is in the forth section. This section compares prediction accuracy of local modeling using both the partially adaptive bandwidth and the global bandwidth with parametric models. Selection between the partially adaptive bandwidth and its global counterpart is also discussed. Empirical time series used in the experimental studies come from the area of hydrology. Section five concludes.

2. Framework

The partially adaptive bandwidth is investigated in the framework of time series prediction.

Given the observed time series containing N points

$$y_1, \dots, y_N,$$

the aim of prediction is to estimate a future value of the time series y_{N+1} (only one-step-ahead prediction will be considered in this paper).

The prediction model is a free form autoregression

$$y_i = f(\mathbf{x}_i) + \mathbf{e}_i, \quad i = 1, \dots, N, \dots$$

where \mathbf{x}_i is the vector of p lagged values of y_t as y_{i-1}, \dots, y_{i-p} and \mathbf{e}_i is an independent random variable with zero mean and finite variance.

The functional relationship $f(\cdot)$ is unknown and suspected to be nonlinear. Local linear regression is used to approximate

the unknown function. Given the point \mathbf{x}_i , the local regression estimate \hat{y}_i of y_i is obtained as

$$\hat{y}_i = \mathbf{x}_i' (\mathbf{X}' \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}_i \mathbf{y}, \quad (1)$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$, $\mathbf{y} = (y_1, \dots, y_N)'$ and \mathbf{W}_i is the diagonal weight matrix with diagonal elements defined as

$$w_{ii} = \begin{cases} w(d_i/h_i), & \text{if } d_i < h_i \\ 0, & \text{if } d_i > h_i \end{cases}$$

A variable d , defined as the Euclidean norm,

$$d_i = \left[\sum_{j=1}^p (x_{ij} - x_{ij})^2 \right]^{1/2}, \quad i = 1, \dots, N \text{ and } i \neq t,$$

measures the distance between \mathbf{x}_i and \mathbf{x}_t . h_i is a bandwidth parameter controlling the size of local area. I use the nearest-neighbor (NN) bandwidth (Cleveland & Loader, 1994) expressed using the integer number k . k accounts for a number of points with non-zero weights. This number has more sensible interpretation in the case of the NN bandwidth than h_i (h_i actually depends upon k). The parameter k is constant for all points in the case of the global bandwidth. The parameter k is determined using standard cross-validation (Hastie & Tibshirani, 1990), that is, the optimal k produces the smallest in-sample prediction error

$$\text{MSPE}(k) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2,$$

$$k = \arg \min(\text{MSPE}(k))$$

Two different weight functions will be considered. The first is the smooth tricube weight function (TW) and the second alternative is the discontinuous uniform weight function (equal to one for all nearest-neighbors) (UW).

The prediction accuracy throughout the paper is measured by MSPE, abbreviation MSFE is used when the true out-of-sample prediction accuracy is considered.

3. Partial adaptive bandwidth

The fixed bandwidth is used to make prediction. The adaptive bandwidth is being employed only if a specified quality criterion of the particular local model fails. The Mallows's C_p criterion is used by Cleveland et al. (1988) and Cleveland & Loader (1994) to evaluate quality of the local model in the case of the fully adaptive bandwidth. Here the localized Aikake's Information criterion (AIC)

$$\text{AIC}_t(k) = \frac{1}{\text{tr}(\mathbf{W}_t)} \sum_{i=1}^k (y_i - \tilde{y}_i)^2 w_{ii} * \exp\left(\frac{2\text{tr}(\mathbf{M}_t)}{\text{tr}(\mathbf{W}_t)}\right)$$

is used. It shows how good the neighbors selected relatively to point t predict each other with penalty on small

bandwidths, \tilde{y} is the local estimate without recalculation of the local model and $\mathbf{M}_t = (\mathbf{X}' \mathbf{W}_t \mathbf{X})^{-1} (\mathbf{X}' \mathbf{W}_t^2 \mathbf{X})$.

The threshold value of the quality criterion is selected relatively to the average value of AIC

$$\overline{\text{AIC}}(k) = c \frac{1}{N} \sum_{t=1}^N \text{AIC}_t(k),$$

where c is the adjustable parameter. The choice of c will be discussed in the following section.

The quality criterion is being calculated using an initial global bandwidth. There are three ways to specify the initial bandwidth.

The first way is to use an arbitrary specified bandwidth k_a (usually small). In this case cross-validation is skipped and large computational savings may be achieved.

The second option considers bandwidth selection according to the results of cross-validation.

The final option involves two consecutive applications of cross-validation. The following steps should be executed to obtain the two-stage bandwidth k_2 :

- i. perform cross-validation using all available data and find optimal k_1 ;
- ii. exclude from the data sets points having AIC larger than the specified threshold;
- iii. perform cross-validation using the refined data set and find k_2 .

The following actions should be carried out to obtain out-of-sample prediction:

- i. compute AIC for the given \mathbf{x} using k_1 and compare it with the threshold value;
- ii. if AIC is smaller than the threshold value then compute the prediction using k_2 otherwise go to the next step;
- iii. if AIC is larger than the threshold value then locally adjust the parameter k and compute the prediction using the locally adjusted bandwidth.

The local adjustment is very simple. AIC is minimized relatively to k by discrete search over selected grid points.

The one stage bandwidth estimate may be too much affected by prediction errors corresponding to points with large AIC. The two-stage procedure is aimed to find the most appropriate bandwidth for the points with small corresponding values of AIC.

4. Applications

Experimental studies are designed to evaluate a predictive power of the partially adaptive bandwidth. Additionally, they have to answer a question when to use the partially adaptive bandwidth and to set the threshold value of the quality criterion.

The studies are based on analysis of several daily river flow time series. These series are available from web: <http://water.usgs.gov>. The data summary is given in Table 1. All series are standardized to have zero mean and unit variance. The first difference of the original series is used in modeling. Graphs (Figure 1) show the changing variation of the series.

The bandwidths are determined using the initial N observations (the initial sample). Successive M observations (the test sample) are used for true out-of-sample forecasting. The vector \mathbf{x}_t is built using up to date information while only observations from the initial sample are used to solve the weighted least squares problem (1) and to perform adaptation.

Table 1. Description of the considered time series. t_s is the starting date of a sample.

Data set	N	t_s	Lags
Little Missouri River Nr Boughton, Ark. (LM1)	365	10.26.73	{1,2,3}
Little Missouri River Nr Boughton, Ark. (LM2)	365	12.25.73	{1,2,3}
Little Missouri River Nr Boughton, Ark. (LM3)	365	02.03.74	{1,2,3}
Patoka River Near Princeton, Ind. (P)	365	10.13.93	{1,2,3}
Wolf River At Langlade, Wi (W)	365	11.4.91	{1,2,3}
Niobrara River Nr. Verdel, Nebr. (N)	365	12.17.90	{1,2}

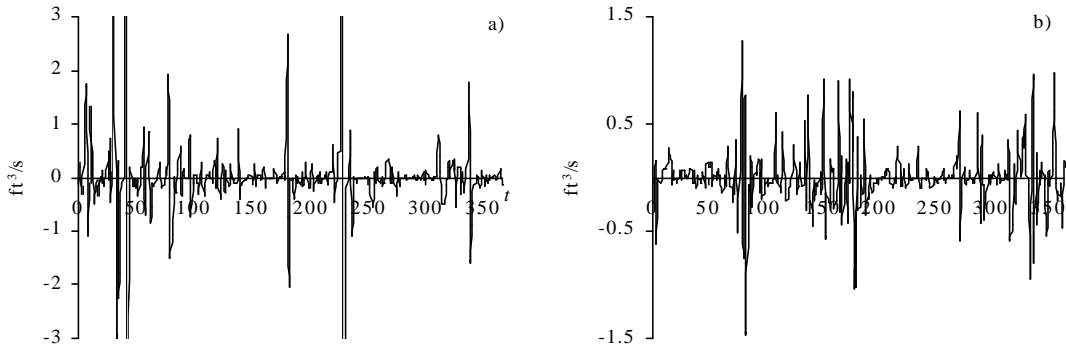


Figure 1. First difference series of a) LM and b) N time series.

Table 2 compares the prediction results depending on choice of the initial bandwidth. The prediction results are reported both for the uniform weights and the tricube weights. The coefficient c is set equal to one.

The arbitrary bandwidth is set as $k_a = N/5 = 73$.

The reported prediction errors are similar for all choices of the initial bandwidth. Only the arbitrary selected bandwidth occasionally gives weaker results as it was expected. The initial bandwidths k_1 and k_2 give the same prediction results for points having AIC larger than the average because in both cases the same partially adapted bandwidth k^* is employed. The accuracy differences may only occur for points with AIC smaller than the average. Table 3 shows the

prediction accuracy according to values of AIC. These results show that the overall prediction accuracy is determined by the points with large AIC. Therefore difference between application of k_1 and k_2 is minor. This also provides additional motivation for the partially adaptive bandwidth. The prediction errors generally are small if AIC is small. Adaptation is computationally costly and large accuracy gains can be mainly achieved for points with large AIC. Therefore adaptation is being applied only in the cases of large AIC.

Additionally, instability of the cross-validation results (Park & Turlach, 1992) is a factor hampering comparison of the initial bandwidths.

Table 2. Prediction accuracy according to different choices of the initial bandwidth. MSFE (AR) is the accuracy of linear autoregressive model, k_1 is the cross-validated bandwidth, k_a is the arbitrary bandwidth, k_2 is the two-stage bandwidth and k^* is the local bandwidth. N_1 is a number of observations with AIC smaller than the threshold value.

Data set	MSFE (AR)	k_1	MSFE (k_s)	AIC (k_1)	MSFE (k_a, k^*)	MSFE (k_1, k^*)	k_2	N_1	MSFE (k_2, k^*)	MSFE (k^*)
UW										
LM1	1.117	312	0.922	0.29	0.866	0.888	140	191	0.877	0.865
LM2	0.757	52	0.353	0.16	0.392	0.396	193	349	0.406	0.435
LM3	0.501	47	0.244	0.17	0.213	0.213	306	357	0.213	0.236
P	0.0052	224	0.0041	0.0113	0.0035	0.0027	10	135	0.0027	0.0027
W	0.0770	351	0.0776	0.0295	0.0840	0.0768	56	107	0.0772	0.0842
N	0.0536	96	0.0554	0.0365	0.0483	0.0475	201	297	0.0473	0.0470
TW										
LM1	1.117	356	0.855	0.325	1.017	0.845	283	339	0.867	1.154
LM2	0.757	343	0.373	0.220	0.568	0.543	81	335	0.559	0.654
LM3	0.501	344	0.151	0.242	1.615	1.533	47	335	1.551	1.698
P	0.0052	357	0.0041	0.0145	0.0034	0.0042	195	336	0.0031	0.0040
W	0.0770	362	0.0739	0.0281	0.145	0.109	117	173	0.114	0.179
N	0.0536	348	0.0529	0.0477	0.0632	0.0636	284	335	0.0649	0.0695

The results are similar for both UW and TW when the global bandwidth k_1 is used (except for the data sets LM1 and LM3). TW gives substantially weaker results than UW with the adaptive bandwidth. Figure 2 and Figure 3 show the results of partial adaptation obtained using UW and TW respectively. The difference between accuracy of both forecasts is due to two large errors at points 391 and 400. These errors are obtained when the adaptive bandwidth equal to the smallest possible value is used. Weight functions like TW require larger bandwidths than UW to achieve the same noise reduction. Although local bandwidths with small values are capable to improve forecasting results occasionally they cause extremely large prediction errors (occasional failure of adaptation partly is caused by a relatively small number of observations and high noise intensity). This is also illustrated by Figure 4 comparing the global cross-validation functions obtained with UW and TW. The tricube weight function gives extremely high MSPE for small bandwidths while the cross-validation curve is relatively smooth for UW.

The results brought by partial adaptation and full adaptation are similar although partial adaptation performs marginally better in three cases (Table 2). Important that partial

adaptation achieves the same (or even better) accuracy using the adaptive bandwidth just to make 28 % (average for all series) of all predictions (Table 3).

The results brought by the adaptive bandwidth qualitatively differ from those of the global bandwidth (Figure 2 and Figure 3). They exhibit better ability to trace variability of the original series.

Table 3. Prediction accuracy according to a value of AIC. Predictions are made using bandwidths k_1 and k^* . $MSFE_s$ is the prediction accuracy calculated for points having AIC smaller than the average, $MSFE_l$ is the prediction accuracy for points with AIC larger than the average, u is the proportion of points with AIC larger than the average.

Data set	UW			TW		
	$MSFE_s$	$MSFE_l$	u	$MSFE_s$	$MSFE_l$	u
LM1	0.428	1.600	0.40	0.346	1.844	0.34
LM2	0.060	1.473	0.24	0.156	1.453	0.3
LM3	0.152	0.388	0.26	0.546	3.692	0.32
P	0.0005	0.0234	0.10	0.0022	0.0165	0.14
W	0.0857	0.0620	0.40	0.0481	0.1878	0.44
N	0.0173	0.1264	0.28	0.0388	0.1359	0.26

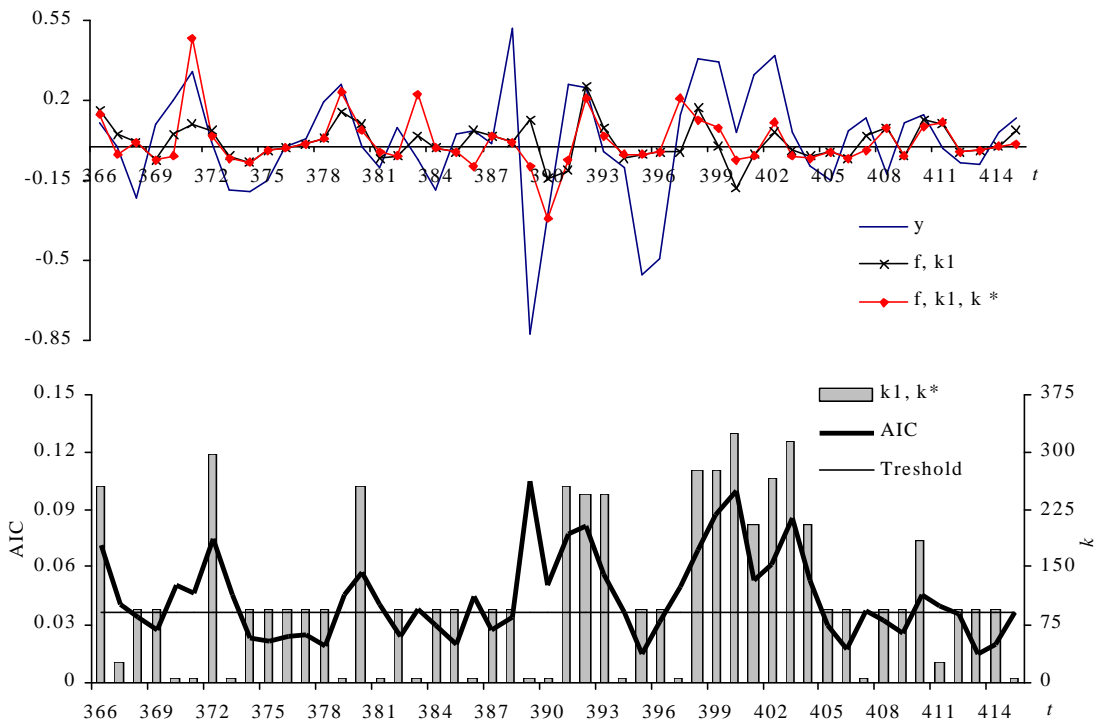


Figure 2. Forecasts and bandwidth values obtained using UW (N data set). y is the actual series and f are the forecasted series.

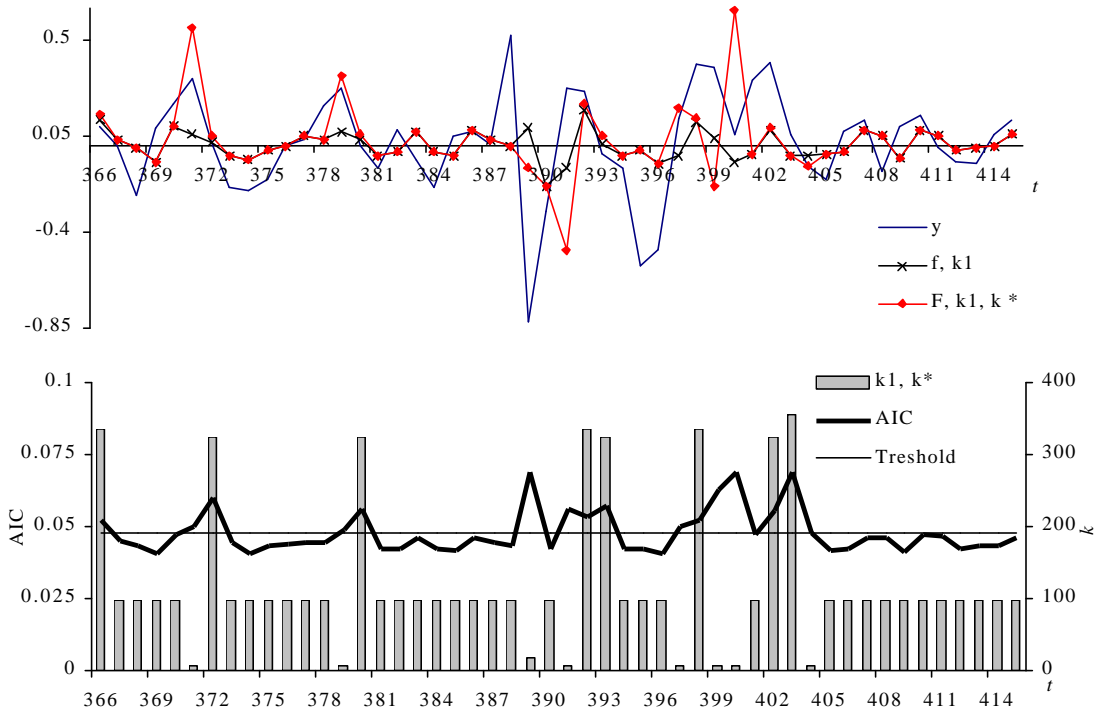


Figure 3. Forecasts and bandwidth values obtained using TW (N data set).

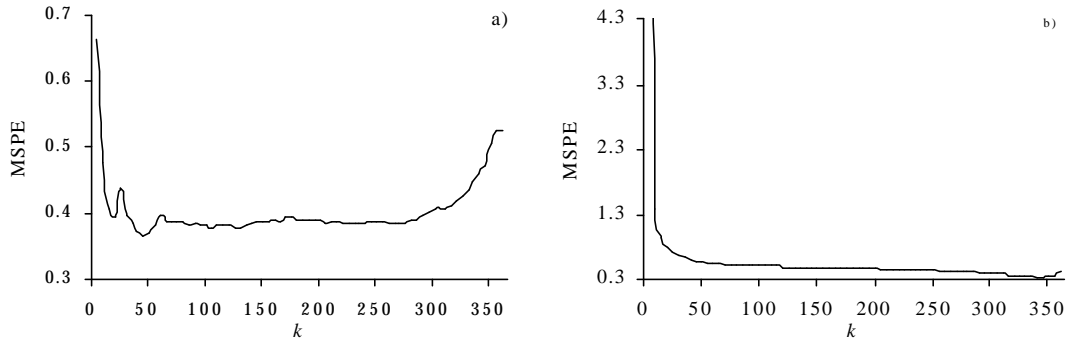


Figure 4. The global cross-validation functions for the LM3 data using a) the uniform weights and b) the tricube weights.

The partially adaptive bandwidth gives better prediction results than the global bandwidth in three out of six case (results are similar in two cases and the global bandwidth performs better in one case). Distributional graphs of the in-sample estimation residuals and AIC (Figure 5) show that partial adaptation as expected deals with long-tailed error distributions. The distributions are highly peaked and several extreme values are present. The fully adaptive bandwidth can be expected to perform well in a case of less peaked distributions.

Figure 6 shows ascending ordered AIC values and corresponding in-sample estimation residuals. It confirms correspondence between large values of AIC and large prediction errors although large errors occur also at the low end of AIC values. That suggests using of adaptation also in cases of small AIC. It is particularly apparent for the W data set. Therefore partial adaptation has not improved the prediction results relatively to the global bandwidth (Table 2).

The uniform character of the AIC curves (Figure 6 a) and c)) is due to a large value of the initial bandwidth k_1 . Large

values of the initial bandwidth leaves a little space for adaptation.

There is a sharp border between small values of AIC and large values of AIC for the LM3 data set. The transition from small values of AIC to large values of AIC is smooth for the N data set. Perhaps therefore partial adaptation gives the best prediction results for the LM3 data while partial adaptation and full adaptation give almost the same results for the N data.

Graphs of Figure 6 type provide a way to select the coefficient c . The coefficient c should be selected to separate points having large AIC from points having small AIC. This border is quite obvious for the LM data sets. The border is not sharp for the N data and the prediction results show that any value could be appropriate from the accuracy point of view. Choice of c significantly influences amount of computations required for execution of partial adaptation. The value $c=1$ occurred to be quite reasonable for the considered time series.

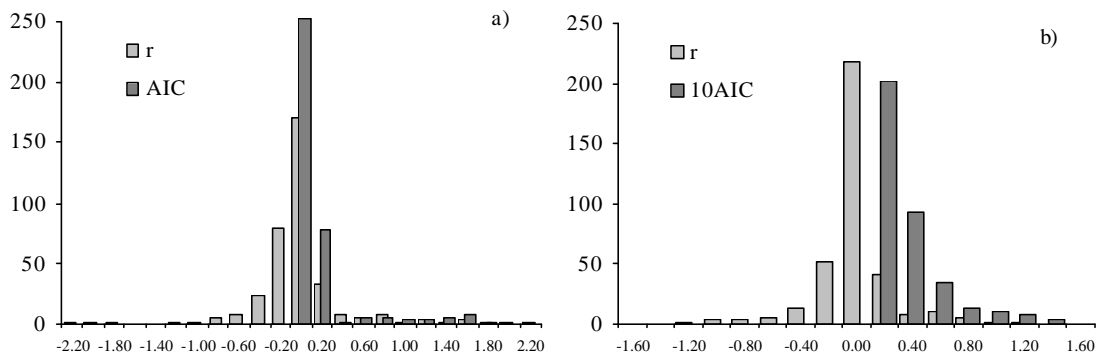


Figure 5. The distributions of AIC and in-sample estimation residuals for a) LM3 and b) N data.

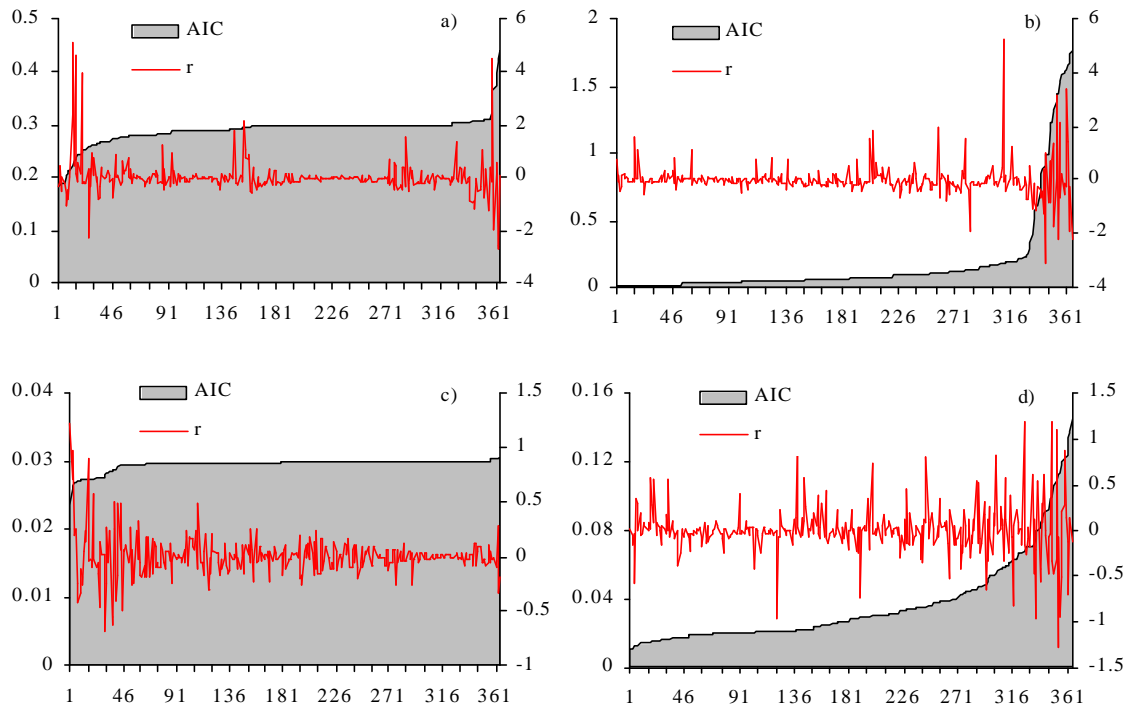


Figure 6. Ascending ordered AIC and corresponding estimation residuals a) LM1, b) LM3, c) W and d) N.

5. Conclusion

The partially adaptive bandwidth of local regression and its application in prediction of nonlinear time series was discussed in this paper. The quality criterion classified prediction situations and according to this classification the global or the local bandwidth is used to obtain predictions. The partially adaptive bandwidth primary is aimed to reduce large prediction errors.

The prediction results showed that in certain situations the partially adaptive bandwidth outperforms the global bandwidth. The partially adaptive bandwidth gave just marginally better results than the local bandwidth. Comparison of partial adaptation and full adaptation in light of computational costs depends from a particular prediction task.

Adaptation failed when the tricube weight function was used. That probably was caused by an insufficient number of observations and high noise intensity. The uniform weight function appears to be more reliable in such circumstances.

The two-stage procedure used to find the initial bandwidth did not give better results than the one stage procedure. The two-stage procedure with the arbitrary initial bandwidth could be a reasonable alternative.

The analysis of the AIC distributions and the ascending ordered AIC graphs allows choosing between application of the global and the partially adaptive bandwidth and to specify the threshold value.

References

- Atkeson, C. G., Moore, A. W., Schaal, S. (1997). Locally Weighted learning, *Artificial Intelligence Review*, 11, 11-73.
- Cleveland, W. S. (1979). Robust locally weighted regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74, 829-836.
- Cleveland, W. S., Devlin, S. J., Grosse, E. (1988). Regression by Local fitting. Methods, Properties, and Computational Algorithms, *Journal of Econometrics*, 37, 87-114.
- Cleveland, W. S., Loader, C. (1994). Smoothing by Local regression: Principles and Methods, AT&T Bell Laboratories Research Report.
- Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.
- Hastie, T. J., Tibshirani, R. J. (1990). *Generalized Additive Models*, London: Chapman & Hall.
- Heiler, S. (1999). A Survey on Nonparametric Time Series Analysis, Working paper, available from <http://ideas.uqam.ca/ideas/data/Papers/knzcofedp9905.html>.

Kantz, H., Schreiber, T. (1997). *Nonlinear Time Series Analysis*, Cambridge: Cambridge University Press.

Park, B. U., Turlach, B. A. (1992). Practical Performance of Several Data-driven Bandwidth Selectors, *Computational Statistics*, 7, 251-271.

Stenman, A. (1999). *Model-on-demand: Algorithms, Analysis and Applications*, Ph. D. thesis, Linköping University.

Tong, H. (1993). *Non-linear Time Series: a Dynamical System Approach*, New York: Oxford University Press.