ON ADAPTIVE LOCAL POLYNOMIAL APPROXIMATION WITH VARYING BANDWIDTH

Vladimir Katkovnik

Department of Statistics
University of South Africa (UNISA)
P.O. Box 392, Pretoria 0001, RSA
e-mail : katkov@risc6.unisa.ac.za

ABSTRACT

The local polynomial approximation (LPA) of noisy data is considered with the new adaptive procedure for varying bandwidth selection. The algorithm is simple to implement and nearly optimal within $\ln N$ factor in the point-wise risk for estimating the function and its derivatives. The adaptive varying bandwidth enables the algorithm to be spatial adaptive over a wide range of the classes of functions in the sense that its quality is close to that which one could achieve if smoothness of the estimated function was known in advance. It is shown that the cross-validation adjustment of the threshold parameter of the algorithm significantly improves its accuracy. In particular, simulation demonstrates that the adaptive algorithm with the adjusted threshold parameter performs better than the wavelet estimators.

1. INTRODUCTION

Suppose that we are given by noisy samples of a signal $y(x)$,

$$z_s = y(x_s) + \varepsilon_s, \ s = 1, 2, ..., N, \tag{1}$$

where ε_s i.i.d., $E(\varepsilon_s) = 0$, $E(\varepsilon_s^2) = \sigma^2$. It is assumed that $y(x)$ belongs to the nonparametric class of piece-wise continuous $m-$differentiable functions

$$\mathcal{F} = \{\left|y^{(m)}(x)\right| \leq L_m\}.$$

Our goal is to estimate $y_s = y(x_s)$ depending on observations $(z_s)_{s=1}^N$ with a point-wise mean squared error (MSE) risk which is as small as possible. The following loss function is used in the standard linear LPA (e.g. [3][6][8]):

$$J_h(x) = \sum_{s=1}^{N} \rho_h(x_s - x)(z_s - C^T \phi(x_s - x))^2 \tag{2}$$

$$\begin{aligned} \phi(x) &= (1, x, x^2/2, ..., x^{m-1}/(m-1)!)', \\ C &= (C_0, C_1, ..., C_{m-1})', \end{aligned}$$

where x is the centre and m is the order of the LPA. The window $\rho_h(x) = \rho(x/h)/h$ is a function satisfying the convention properties of the "kernel" estimates, in particular, $\rho(x) \geq 0$, $\rho(0) = \max_x \rho(x)$, $\rho(x) \to 0$ as $|x| \to \infty$ and $\int_{-\infty}^{\infty} \rho(u) du = 1$. Here h is a window 'size' or a bandwidth.

Then minimization of $J_h(x)$ with respect to C

$$\hat{C}(x, h) = \arg \min_{C \in R^m} J_h. \tag{3}$$

gives $\hat{y}(x) \triangleq \hat{C}_0(x, h)$ as an estimate of $y(x)$, and $\hat{y}_k(x) \triangleq \hat{C}_k(x, h)$, $k = 1, ..., m-1$, as estimates of the derivatives $y^{(k)}(x)$. These estimates can be represented in the form of linear filters

$$\hat{y}_k(x, h) = \sum_s g_k(x, x_s, h) y_s, \tag{4}$$

where

$$\sum_s g_k(x, x_s, h) x^l = k! \cdot \delta_{l,k}, \ k, l = 0, 1, ..., m-1, \tag{5}$$

shows that the linear transforms (4) have accurate reproductive properties with respect to polynomial components of $y(x)$ up to degree $m-1$.

The linear estimators (4)-(5) are a very popular tool in signal processing and statistics with application to a wide variety of the fields for smoothing, filtering, interpolation and extrapolation (e.g. [1][3][6][8]) It is well known that bandwidth selection is a crucial point of the efficiency of the LPA estimators. In particular, the essentially varying in x curvature of $y(x)$ requires a varying spatially adaptive bandwidth $h = h(x)$.

This paper is inspired by the novel approach developed in [4] (see also more general results in [5]). It is shown that the LPA equipped with the special

This work was supported by the Foundation of Research Development of South Africa.

adaptive bandwidth possesses simultaneously many attractive asymptotic properties, namely, 1) it is nearly optimal within $\ln N$ factor in the point-wise risk for estimating the function and its derivatives; 2) it is spatial adaptive over a wide range of the classes of $y(x)$ in the sense that its quality is close to that which one could achieve if smoothness of $y(x)$ was known in advance. The intersection of the confidence intervals (ICI) of the estimates with different bandwidths is proposed in [4] for bandwidth selection.

The multi-window LPA with the data-driven adjustment of the threshold parameter of the ICI bandwidth selection is developed in this paper. It is shown that this adjustment results in a very valuable accuracy improvement of the adaptive estimates of $y(x)$. The fast MATLAB implementation of the algorithm is done.

The paper is organized as follows. The ICI bandwidth selection is described in section II. The adjusted ICI bandwidth selection with the data-driven threshold parameter as well as simulation is presented in section III. It is shown in particular that the multiwindow LPA with the adjusted ICI bandwidth selection achieves a better accuracy than the wavelet filters, while it is not the case for the adaptive algorithm with a fixed value of this threshold parameter studied in [4].

2. ADAPTIVE BANDWIDTH SELECTION

Let us present the basic idea of the ICI.

Assume that

$$H = \{h_1 < h_2 < .... < h_{N_h}\} \tag{6}$$

is an increasing sequence of bandwidths. Then the confidence intervals of the estimates $\hat{y}_N(x, h_i)$, $i = 1, 2, ...$, form sequences

$$\begin{aligned} \mathcal{D}_i &= [L_i,\ U_i], \\ U_i &= \hat{y}_k(x, h_i) + 2\chi(\alpha) std_k(x, h_i), \\ L_i &= \hat{y}_k(x, h_i) - 2\chi(\alpha) std_k(x, h_i), \end{aligned} \tag{7}$$

where $std_k(x, h) = \sigma\sqrt{\sum_s g_k^2(x, x_s, h)}$ is the standard deviation of the estimate and $y_k(x) \in \mathcal{D}_i$ with probability $p = 1 - \alpha$. Let us call $\chi(\alpha)$ in (7) the threshold parameter. As it follows from the asymptotic accuracy analysis (e.g. [3]) the segments $\mathcal{D}_i$ narrow and the bias of the estimates $\hat{y}_k(x, h_i)$ grows as i increases.

Consider intersections of the intervals $\mathcal{D}_i$ as $i = 1, 2, ...$ and let i^+ be the largest of those i for which the segments $\mathcal{D}_j$, $j \leq i$, have a point in common. This i^+ determines a data-driven ICI bandwidth as

$$h^+ = h_{i^+} \tag{8}$$

and the adaptive LPA estimator as

$$\hat{y}_k^+(x) \triangleq \hat{y}_k(x, h^+).$$

The adaptive algorithm equipped with the ICI bandwidth selection demonstrates the following convergence rate.

Let $x_s = s/N$, $s = 0, 1, ..., N$ and $\chi = \beta \ln N$, with a constant β then

$$r_{k,m}(x, h^+) = O\left(\left(\frac{\ln N}{N}\right)^{\frac{m-k}{2m+1}}\right). \tag{9}$$

This convergence rate is different only in the factor $\ln N$ from the optimal MSE obtained provided a known smoothness of $y(x)$. It is emphasized that for estimating a function of unknown smoothness this factor cannot be eliminated [11]. Thus the adaptive estimate is adaptive in an optimal way to the unknown smoothness of the function to be estimated. The optimal bandwidths are different for the estimates of the function and its derivatives.

We wish to note that this bandwidth selection procedure requires a knowledge of the estimate and its variance only and is much simpler to implement than for instance "plug-in" methods which require a knowledge of the estimation bias (e.g. [3]).

3. ICI WITH ADJUSTED THRESHOLD PARAMETER

The threshold parameter χ in (7) plays a crucial role in the performance of the algorithm (6)-(8). Too large or too small values of χ result in oversmoothing and undersmoothing data. The standard choices $\chi(\alpha) = 2$ or 3 corresponding to the probabilities $p = 0.95$ and 0.99 are far beyond universal acceptability. The MSE has a minimum with respect to χ depending in particular on the signal-noise ratio, the sampling length N and the bandwidth set H. Minimizing χ significantly improves the accuracy.

We found that the cross-validation method proves to be efficient for finding the optimal χ. For the linear filter (4) the cross-validation loss function can be represented as a weighted sum of squared residuals

$$I_{CV} = \sum_s \left(\frac{z_s - \hat{y}(x_s, h^+(x_s))}{1 - g_k(x_s, x_s, h^+(x_s))}\right)^2. \tag{10}$$

Thus the procedure (6)-(8) is assumed to be repeated for every $\chi \in K$, $K = \{\chi_1, \chi_2, .. \chi_{N_\chi}\}$, and

$$\hat{\chi} = \arg\min_{\chi \in K} I_{CV} \tag{11}$$

gives the adjusted threshold parameter.

The cross-validation in the form (10) presents a very reasonable and effective selector for χ. Our attempts to use instead of the cross-validation another quality-of-fit statistics, in particular the C_P, Akaike criteria and its modifications (see e.g. [7]), which are different from I_{CV} only by the used weights of the residuals, have not shown an improvement in accuracy.

Thus the adaptive LPA estimation consists of the following basic steps:

1. Set $\chi = \chi_r$, $r = 1, 2, ..., N_\chi$ and $x = x_s$, $s = 1, 2, .., N$.

2. For $h = h_i$, $i = 1, ..., N_h$, calculate the estimates $\hat{y}(x_s, h_i)$ and

$$
\begin{aligned}
\bar{U}_i &= \min(\bar{U}_{i-1}, U_i),\ \underline{L}_i = \max(\underline{L}_{i-1}, U_i), \quad (12) \\
\bar{U}_0 &= 0,\ \underline{L}_0 = 0,
\end{aligned}
$$

while $\bar{U}_i \geq \underline{L}_i$.

The largest of those i for which $\bar{U}_i \geq \underline{L}_i$ gives i^+, the bandwidth (8) and estimate $\hat{y}(x_s, h^+(x_s))$.

4. Repeat Step 3 for all x_s, $s = 1, 2, .., N$, and χ_r, $r = 1, 2, ..., N_\chi$.

5. Find $\hat{\chi}$ from (11).

The used in (7) standard deviation σ is estimated by

$$\hat{\sigma} = \{\text{median}(|z_s - z_{s-1}| : s = 2, .., N)\}/0.6745. \quad (13)$$

The average $\frac{1}{N-1}\sum_{n=2}^{N}(z_s - z_{s-1})^2$ could also be applied as an estimate of σ^2. However, we prefer a median (13) as a robust estimate.

For the data given on the regular grid the fast implementation of the algorithm is done in MATLAB. The items of the matrices of estimates $Y = (\hat{y}(x_s, h_i))_{N\times N_h}$ and corresponding confidence intervals are calculated simultaneously for all (x_s, h_i) and the search for $h^+(x_s)$ is produced as a matrix operation. The algorithm is quite efficient even in searching for the adjusted threshold parameter $\hat{\chi}$.

4. SIMULATION

Let ρ_L, ρ_R and ρ_S be the left, right, and symmetric window functions, i.e. $\rho_L(u) = 0$ for $u > 0$, $\rho_R(u) = 0$ for $u < 0$ and $\rho_S(u) = \rho_S(-u)$, and $\hat{y}_L$, $\hat{y}_R$, and $\hat{y}_S$ be the corresponding estimates of $y(x)$.

Then the combined LPA estimate $\hat{y}$ can be produced in the form

$$
\begin{aligned}
\hat{y} &= \lambda_L \hat{y}_L + \lambda_R \hat{y}_R + \lambda_S \hat{y}_S, \quad (14) \\
\lambda_L &= \frac{std_L^{-2}}{std^{-2}},\ \lambda_R = \frac{std_R^{-2}}{std^{-2}},\ \lambda_S = \frac{std_S^{-2}}{std^{-2}}, \\
std^{-2} &= std_R^{-2} + std_L^{-2} + std_S^{-2},
\end{aligned}
$$

with the inverse standard deviations used as weights. The important feature of the linear combination (14) is that the combined estimate $\hat{y}$ again can be considered as a linear estimate with the combined weight

$$g_{k(comb)} = \lambda_L g_{k(L)} + \lambda_R g_{k(R)} + \lambda_S g_{k(S)}, \quad (15)$$

where $g_{k(L)}$, $g_{k(R)}$ and $g_{k(S)}$ are weights of the corresponding left (forward), right (backward) and symmetric filters (4)-(5). In this case the cross-validation loss function in the form (10) can be used for evaluation of the combined estimate.

Thus we arrive at the concept of the LPA filter bank which consists of the elementary filters with weights g_k obtained for the left, right and symmetric window functions with different LPA degrees m, reasonably restricted say to $m = 0, 1, 2$. These combined estimates in the form (15) produce sets of possible estimates.

However, as the combinations in the form (15) does not always result in improving of the estimate, the selection of the best estimate or the stopping rule for the combining is a problem.

We found that the cross-validation in the form (10) once more presents a very reasonable and effective selector.

The following simulation results are given for the filters obtained for the left, right and symmetric rectangular windows and the linear LPA, i.e. $m = 1$. The considered combined estimates are obtained by combining two left and right filters or all three filters, including the symmetric filter. We compare the adaptive LPA estimates with the results achieved by the wavelet filters on the test functions Blocks and HeavySine. These functions, noise, and the conditions of the Monte-Carlo statistical modelling are exactly as given in Table 2 in ([2], p. 1218). The root mean squared errors of estimation are presented in Table 1:

$$SRMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\frac{1}{N}\sum_{s=1}^{N}(y(x_s) - \hat{y}^{[j]}(x_s, h^+(x_s)))^2},$$

where the average over $M = 20$ simulation runs is calculated and the superscript $[j]$ indicates the randomness of the estimate for every $j-th$ run. Figures in the first column are the numbers N of observations. The second and third columns give the $SRMSE$ for the LPA estimator respectively with the adjusted threshold parameter $\chi = var$ and fixed $\chi = 2.2$, as it is used in the simulation given in [4]. The fourth column present the interval of the $SRMSE$ values obtained in [2] for the different wavelet filters.

It is evident from the table that the developed algorithm with the adjusted threshold parameter in all cases achieves a better accuracy than the wavelet estimators, while it is not true for the algorithm with a fixed value of the threshold parameter χ.

The algorithm with the adaptive threshold parameter demonstrates the accuracy improvement around 1.5-2 times in comparison with the algorithm with the fixed value of χ. We wish to note that the estimates corresponding to $\chi = 2.2$ have a better appearance than the ones corresponding to $\chi = var$, but oversmooth data. The accuracy comparison is definitely in favor of the adjusted $\chi = var$.

It is emphasized that in the Monte-Carlo simulation the variance of the adjusted $\chi = var$ between different runs is very small, thus the value of χ found in one simulation run can be used as a fixed value for all other runs. It makes a difference with the adaptive bandwidths h_{i+} which are different for every run. Thus while the adaptive bandwidths given by (8) are random and smooth noise effects for every sampling of observations, the threshold parameter χ is much more conservative and depends only on some total parameters of the problem such as the variance of $y(x)$, signal-to-noise ratio, and N.

5. CONCLUSION

In conclusion we wish to emphasize that the idea of the LPA nonparametric estimation equipped with the ICI statistic used for the bandwidth selection is quite general and very fruitful. It can be modified to problems with nonlinear observations. In particular, for $y(x) = A\exp(j\varphi(x))$ the approach was used in [9] and [10] in order to develop the adaptive LPA estimator of the instantaneous frequency $\varphi^{(1)}(x)$.

6. REFERENCES

[1] Brown R. Smoothing, forecasting and prediction of discrete time series. Prentice-Hall, Englewood Cliffs, N.Y., 1963.

[2] Donoho D.L. and Johnstone I. M. "Adapting to unknown smoothness via wavelet shrinkage". JASA, vol. 90(432), pp. 1200-1224, 1995.

[3] Fan J. and Gijbels I. Local polynomial modelling and its application. London: Chapman and Hall, 1996.

[4] Goldenshluger A. and Nemirovski A. "On spatial adaptive estimation of nonparametric regression". Research report, 5/94, Technion, Israel, November, 1994.

[5] Goldenshluger A. and Nemirovski A. "Adaptive de-noising of signals satisfying differential inequalities". IEEE Trans. Inf. Theory, vol. 43, 3, pp.873-889, 1997.

[6] Hardle W. Applied nonparametric regression. Cambridge, University Press, Cambridge, 1990.

[7] Hurvich C. M. and Simonoff J. S. "Smoothing parameter selection in nonparametric regression using an improved AIC criterion". Journal of the Royal Statistical Society, Ser. B, 1998. To appear.

[8] Katkovnik V. Nonparametric identification and smoothing of data (Local approximation methods). Nauka, Moscow, 1985 (in Russian).

[9] Katkovnik V. "Adaptive local polynomial periodogram for time-varying frequency estimation". Proceedings of the IEEE-SP Int Symposium on Time-Frequency and Time-Scale Analysis (TFTS-96), Paris, pp. 329-332, 1996.

[10] Katkovnik V. and Stanković LJ. "Instantaneous frequency estimation using the Wigner distribution with varying and data-driven window length". IEEE Trans. on Signal Processing. Submitted, 1997.

[11] Lepskii O. "On a problem of adaptive estimation in gaussian white noise". Theory of Probability and Its Applications, vol. 35(3), pp.454-466, 1990.

Table 1. Root mean squared errors of estimation using the LPA and various wavelet methods

N	$\chi = var$	$\chi = 4$	*Wavelets*
Blocks			
256	0.61	1.56	(0.68-1.20)
512	0.46	0.92	(0.59-1.12)
1024	0.33	0.74	(0.47-1.03)
2048	0.25	0.49	(0.41-0.85)
Heavy			
256	0.47	0.79	(0.49-0.62)
512	0.37	0.67	(0.40-0.55)
1024	0.31	0.49	(0.32-0.46)
2048	0.24	0.34	(0.38-0.61)