Tracking time-varying coefficient-functions

Henrik Aa. Nielsen[a,c] Torben S. Nielsen[a] Alfred K. Joensen[a]

Henrik Madsen[a] Jan Holst[b]

[a] Department of Mathematical Modelling, Technical University of Denmark,
DK-2800 Lyngby, Denmark

[b] Department of Mathematical Statistics, Lund University, Lund Institute of
Technology, S-221 00 Lund, Sweden

[c] Corresponding author: Henrik Aa. Nielsen, Dept. of Math. Modelling,
Technical University of Denmark, DK-2800 Lyngby, Denmark.
Phone: +45 4525 3418, fax: +45 4588 1397, e-mail: han@imm.dtu.dk

Abstract

A method for adaptive and recursive estimation in a class of non-linear autoregressive models with external input is proposed. The model class considered is conditionally parametric ARX-models (CPARX-models), which is conventional ARX-models in which the parameters are replaced by smooth, but otherwise unknown, functions of a low-dimensional input process. These coefficient-functions are estimated adaptively and recursively without specifying a global parametric form, i.e. the method allows for on-line tracking of the coefficient-functions. Essentially, in its most simple form, the method is a combination of recursive least squares with exponential forgetting and local polynomial regression. It is argued, that it is appropriate to let the forgetting factor vary with the value of the external signal which is the argument of the coefficient-functions. Some of the key properties of the modified method are studied by simulation.

Keywords: Adaptive and recursive estimation; Non-linear models; Time-varying functions; Conditional parametric models; Non-parametric method.

1 Introduction

The conditional parametric ARX-model (CPARX-model) is a non-linear model formulated as a linear ARX-model in which the parameters are replaced by smooth, but otherwise unknown, functions of one or more explanatory variables. These functions are called coefficient-functions. In (Nielsen, Nielsen & Madsen 1997) this class of models is used in relation to district heating systems to model the non-linear dynamic response of network temperature on supply temperature and flow at the plant. A particular feature of district heating systems is, that the response on supply temperature depends on the flow. This is modelled by describing the relation between temperatures by an ARX-model in which the coefficients depend on the flow.

For on-line applications it is advantageous to allow the function estimates to be modified as data become available. Furthermore, because the system may change slowly over time, observations should be down-weighted as they become older. For this reason a time-adaptive and recursive estimation method is proposed. Essentially, the estimates at each time step are the solution to a set of weighted least squares regressions and therefore the estimates are unique under quite general conditions. For this reason the proposed method provides a simple way to perform adaptive and recursive estimation in a class of non-linear models. The method is a combination of the recursive least squares with exponential forgetting (Ljung & Söderström 1983) and locally weighted polynomial regression (Cleveland & Devlin 1988). In the paper *adaptive estimation* is used to denote, that old observations are down-weighted, i.e. in the sense of *adaptive in time*. Some of the key properties of the method are discussed and demonstrated by simulation.

Cleveland & Devlin (1988) gives an excellent account for non-adaptive estimation of a regression function by use of local polynomial approximations. Non-adaptive recursive estimation of a regression function is a related problem, which has been studied recently by Thuvesholmen (1997) using kernel methods and by Vilar-Fernández & Vilar-Fernández (1998) using local polynomial regression. Since these methods are non-adaptive one of the aspects considered in these papers is how to decrease the bandwidth as new observations become available. This problem do not arise for adaptive estimation since old observations are down-weighted and eventually disregarded as part of the algorithm. Hastie & Tibshirani (1993) considered varying-coefficient models which are similar in structure to conditional parametric models and have close resemblance to additive models (Hastie & Tibshirani 1990) with respect to estimation. However, varying-coefficient models include additional assumptions on the structure. Some specific time-series counterparts of these models are the functional-coefficient autoregressive models (Chen & Tsay 1993*a*) and the non-linear additive ARX-models (Chen & Tsay 1993*b*).

The paper is organized as follows. In Section 2 the conditional parametric model is introduced and a procedure for estimation is described. Adaptive and recursive estimation in the model are described in Section 3, which also contains a summary of the method. To illustrate the method some simulated examples are included in

Section 4. Further topics, such as optimal bandwidths and optimal forgetting factors are considered in Section 5. Finally, we conclude on the paper in Section 6.

2 Conditional parametric models and local polynomial estimates

When using a conditional parametric model to model the response y_s the explanatory variables are split in two groups. One group of variables $\mathbf{x}_s$ enter globally through coefficients depending on the other group of variables $\mathbf{u}_s$, i.e.

$$y_s = \mathbf{x}_s^T \boldsymbol{\theta}(\mathbf{u}_s) + e_s, \tag{1}$$

where $\boldsymbol{\theta}(\cdot)$ is a vector of coefficient-functions to be estimated and e_s is the noise term. Note that $\mathbf{x}_s$ may contain lagged values of the response. The dimension of $\mathbf{x}_s$ can be quite large, but the dimension of $\mathbf{u}_s$ must be low (1 or 2) for practical purposes (Hastie & Tibshirani 1990, pp. 83-84). In (Nielsen et al. 1997) the dimensions 30 and 1 is used. Estimation in (1), using methods similar to the methods by Cleveland & Devlin (1988), is described for some special cases in (Cleveland 1994) and (Hastie & Tibshirani 1993). A more general description can be found in (Nielsen et al. 1997). To make the paper self-contained the method is outlined below.

The functions $\boldsymbol{\theta}(\cdot)$ in (1) are estimated at a number of distinct points by approximating the functions using polynomials and fitting the resulting linear model locally to each of these *fitting points*. To be more specific let $\mathbf{u}$ denote a particular fitting point. Let $\theta_j(\cdot)$ be the j'th element of $\boldsymbol{\theta}(\cdot)$ and let $\mathbf{p}_{d(j)}(\mathbf{u})$ be a column vector of terms in the corresponding d-order polynomial evaluated at $\mathbf{u}$, if for instance $\mathbf{u} = [u_1 \; u_2]^T$ then $\mathbf{p}_2(\mathbf{u}) = [1 \; u_1 \; u_2 \; u_1^2 \; u_1 u_2 \; u_2^2]^T$. Furthermore, let $\mathbf{x}_s = [x_{1,s} \dots x_{p,s}]^T$. With

$$\mathbf{z}_s^T = \left[x_{1,s}\mathbf{p}_{d(1)}^T(\mathbf{u}_s) \dots x_{j,s}\mathbf{p}_{d(j)}^T(\mathbf{u}_s) \dots x_{p,s}\mathbf{p}_{d(p)}^T(\mathbf{u}_s) \right] \tag{2}$$

and

$$\boldsymbol{\phi}_u^T = [\boldsymbol{\phi}_{u,1}^T \dots \boldsymbol{\phi}_{u,j}^T \dots \boldsymbol{\phi}_{u,p}^T], \tag{3}$$

where $\boldsymbol{\phi}_{u,j}$ is a column vector of local coefficients at $\mathbf{u}$ corresponding to $x_{j,s}\mathbf{p}_{d(j)}(\mathbf{u}_s)$. The linear model

$$y_s = \mathbf{z}_s^T \boldsymbol{\phi}_u + e_s; \quad i = 1, \dots, N, \tag{4}$$

is then fitted locally to $\mathbf{u}$ using weighted least squares (WLS), i.e.

$$\hat{\boldsymbol{\phi}}(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{N} w_u(\mathbf{u}_s)(y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2, \tag{5}$$

for which a unique closed-form solution exists provided the matrix with rows $\mathbf{z}_s^T$ corresponding to non-zero weights has full rank. The weights are assigned as

$$w_u(\mathbf{u}_s) = W\left(\frac{||\mathbf{u}_s - \mathbf{u}||}{\hbar(\mathbf{u})} \right), \tag{6}$$

where $||\cdot||$ denotes the Euclidean norm, $\hbar(\mathbf{u})$ is the bandwidth used for the particular fitting point, and $W(\cdot)$ is a weight function taking non-negative arguments. Here we follow Cleveland & Devlin (1988) and use

$$W(u) = \begin{cases} (1-u^3)^3, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases} \tag{7}$$

i.e. the weights are between 0 and 1. The elements of $\boldsymbol{\theta}(\mathbf{u})$ are estimated by

$$\hat{\theta}_j(\mathbf{u}) = \mathbf{p}_{d(j)}^T(\mathbf{u})\,\hat{\boldsymbol{\phi}}_j(\mathbf{u}); \quad j = 1, \ldots p, \tag{8}$$

where $\hat{\boldsymbol{\phi}}_j(\mathbf{u})$ is the WLS estimate of $\boldsymbol{\phi}_{u,j}$. The estimates of the coefficient-functions obtained as outlined above are called *local polynomial estimates.* For the special case where all coefficient-functions are approximated by constants we use the term local constant estimates.

If $\hbar(\mathbf{u})$ is constant for all values of $\mathbf{u}$ it is denoted a fixed bandwidth. If $\hbar(\mathbf{u})$ is chosen so that a certain fraction α of the observations fulfill $||\mathbf{u}_s - \mathbf{u}|| \leq \hbar(\mathbf{u})$ then α is denoted a nearest neighbour bandwidth. A bandwidth specified according to the nearest neighbour principle is often used as a tool to vary the actual bandwidth with the local density of the data.

Interpolation is used for approximating the estimates of the coefficient-functions for other values of the arguments than the fitting points. This interpolation should only have marginal effect on the estimates. Therefore, it sets requirements on the number and placement of the fitting points. If a nearest neighbour bandwidth is used it is reasonable to select the fitting points according to the density of the data as it is done when using k-d trees (Chambers & Hastie 1991, Section 8.4.2). However, in this paper the approach is to select the fitting points on an equidistant grid and ensure that several fitting points are within the (smallest) bandwidth so that linear interpolation can be applied safely.

3 Adaptive estimation

As pointed out in the previous section local polynomial estimation can be viewed as local constant estimation in a model (4) derived from the original model (1). This observation forms the basis of the method suggested, which is described as a generalization of estimation in (4). For simplicity the adaptive estimation method is described as a generalization of exponential forgetting. However, the more general forgetting methods described by Ljung & Söderström (1983) could also serve as a basis.

3.1 The proposed method

Using exponential forgetting and assuming observations at time $s = 1, \ldots, t$ are available, the adaptive least squares estimate of the parameters $\boldsymbol{\phi}$ relating the explanatory variables $\mathbf{z}_s$ to the response y_s using the linear model $y_s = \mathbf{z}_s^T \boldsymbol{\phi} + e_s$ is found as

$$\hat{\boldsymbol{\phi}}_t = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \sum_{s=1}^{t} \lambda^{t-s} (y_s - \mathbf{z}_s^T \boldsymbol{\phi})^2, \tag{9}$$

where $0 < \lambda < 1$ is called the forgetting factor, see also (Ljung & Söderström 1983). The estimate can be seen as a local constant approximation in the direction of time. This suggests that the estimator may also be defined locally with respect to some other explanatory variables $\mathbf{u}_t$. If the estimates are defined locally to a fitting point $\mathbf{u}$, the adaptive estimate corresponding to this point can be expressed as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s) (y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2, \tag{10}$$

where $w_u(\mathbf{u}_s)$ is a weight on observation s depending on the fitting point $\mathbf{u}$ and $\mathbf{u}_s$, see Section 2.

In Section 3.2 it will be shown how the estimator (10) can be formulated recursively, but here we will briefly comment on the estimator and its relations to non-parametric regression. A special case is obtained if $\mathbf{x}_s$ in (1) is 1 for all s and $d(1)$ is chosen to be 0, then it follows from (2) that $\mathbf{z}_s = 1$ for all s, and simple calculations show that

$$\hat{\phi}_t(\mathbf{u}) = \frac{\sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s) y_s}{\sum_{s=1}^{t} \lambda^{t-s} w_u(\mathbf{u}_s)}, \tag{11}$$

for $\lambda = 1$ this is a kernel estimator of $\phi(\cdot)$ in $y_s = \phi(\mathbf{u}_s) + e_s$, cf. (Härdle 1990, p. 30). For this reason (11) is called an adaptive kernel estimator of $\phi(\cdot)$ and the estimator (10) may be called an adaptive local constant estimator of the coefficient-functions $\boldsymbol{\phi}(\cdot)$ in the conditional parametric model $y_s = \mathbf{z}_s^T \boldsymbol{\phi}(\mathbf{u}_s) + e_s$. Using the same techniques as in Section 2 this can be used to implement adaptive local polynomial estimation in models like (1).

3.2 Recursive formulation

Following the same arguments as in Ljung & Söderström (1983) it is readily shown that the adaptive estimates (10) can be found recursively as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) + w_u(\mathbf{u}_t) \mathbf{R}_{u,t}^{-1} \mathbf{z}_t \left[y_t - \mathbf{z}_t^T \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}) \right] \tag{12}$$

and

$$\mathbf{R}_{u,t} = \lambda \mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t) \mathbf{z}_t \mathbf{z}_t^T \ . \tag{13}$$

It is seen that existing numerical procedures implementing adaptive recursive least squares for linear models can be applied, by replacing $\mathbf{z}_t$ and y_t in the existing procedures with $\mathbf{z}_t\sqrt{w_u(\mathbf{u}_t)}$ and $y_t\sqrt{w_u(\mathbf{u}_t)}$, respectively. Note that $\mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ is a predictor of y_t locally with respect to $\mathbf{u}$ and for this reason it is used in (12). To predict y_t a predictor like $\mathbf{z}_t^T\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}_t)$ is appropriate.

3.3 Modified updating formula

When $\mathbf{u}_t$ is far from the particular fitting point $\mathbf{u}$ it is clear from (12) and (13) that $\hat{\boldsymbol{\phi}}_t(\mathbf{u}) \approx \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u})$ and $\mathbf{R}_{u,t} \approx \lambda\mathbf{R}_{u,t-1}$, i.e. old observations are down-weighted without new information becoming available. This may result in abruptly changing estimates if $\mathbf{u}$ is not visited regularly, since the matrix $\mathbf{R}$ is decreasing exponentially in this case. Hence it is proposed to modify (13) to ensure that the past is weighted down only when new information becomes available, i.e.

$$\mathbf{R}_{u,t} = \lambda v(w_u(\mathbf{u}_t);\lambda)\mathbf{R}_{u,t-1} + w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T, \tag{14}$$

where $v(\cdot\ ;\lambda)$ is a nowhere increasing function on $[0;1]$ fulfilling $v(0;\lambda) = 1/\lambda$ and $v(1;\lambda) = 1$. Note that this requires that the weights span the interval ranging from zero to one. This is fulfilled for weights generated as described in Section 2. In this paper we consider only the linear function $v(w;\lambda) = 1/\lambda - (1/\lambda - 1)w$, for which (14) becomes

$$\mathbf{R}_{u,t} = (1 - (1-\lambda)w_u(\mathbf{u}_t))\mathbf{R}_{u,t-1} + \ w_u(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T. \tag{15}$$

It is reasonable to denote

$$\lambda^u_{eff}(t) = 1 - (1-\lambda) w_u(\mathbf{u}_t) \tag{16}$$

the *effective forgetting factor* for point $\mathbf{u}$ at time t.

When using (14) or (15) it is ensured that $\mathbf{R}_{u,t}$ can not become singular if the process $\{\mathbf{u}_t\}$ moves away from the fitting point for a longer period. However, the process $\{\mathbf{z}_t\}$ should be persistently excited as for linear ARX-models. In this case, given the weights, the estimates define a global minimum corresponding to (10).

3.4 Nearest neighbour bandwidth

Assume that $\mathbf{u}_t$ is a stochastic variable and that the pdf $f(\cdot)$ of $\mathbf{u}_t$ is known and constant over t. Based on a nearest neighbour bandwidth the actual bandwidth can then be calculated for a number of fitting points $\mathbf{u}$ placed within the domain of $f(\cdot)$ and used to generate the weights $w_u(\mathbf{u}_t)$. The actual bandwidth $\hbar(\mathbf{u})$ corresponding to the point $\mathbf{u}$ will be related to the nearest neighbour bandwidth α by

$$\alpha = \int_{\mathbb{D}_u} f(\boldsymbol{\nu}) d\boldsymbol{\nu}, \tag{17}$$

where $\mathbb{D}_u = \{\boldsymbol{\nu} \in \mathbb{R}^d \mid ||\boldsymbol{\nu} - \mathbf{u}|| \leq \hbar(\mathbf{u})\}$ is the neighbour-hood, d is the dimension of $\mathbf{u}$, and $||\cdot||$ is the Euclidean norm. In applications the density $f(\cdot)$ is often unknown. However, $f(\cdot)$ can be estimated from data, e.g. by the empirical pdf.

3.5 Effective number of observations

In order to select an appropriate value for α the effective number of observations used for estimation must be considered. In Appendix A it is shown that under certain conditions, when the modified updating (15) is used,

$$\tilde{\eta}_u = \frac{1}{1 - E[\lambda^u_{eff}(t)]} = \frac{1}{(1-\lambda)E[w_u(\mathbf{u}_t)]} \tag{18}$$

is a lower bound on the effective number of observations (in the direction of time) corresponding to a fitting point $\mathbf{u}$. Generally (18) can be considered an approximation. When selecting α and λ it is then natural to require that the number of observations within the bandwidth, i.e. $\alpha\tilde{\eta}_u$, is sufficiently large to justify the complexity of the model and the order of the local polynomial approximations.

As an example consider $u_t \sim N(0,1)$ and $\lambda = 0.99$ where the effective number of observations within the bandwidth, $\alpha\tilde{\eta}_u$, is displayed in Figure 1. It is seen that $\alpha\tilde{\eta}_u$ depends strongly on the fitting point u but only moderately on α. When investigating the dependence of $\alpha\tilde{\eta}_u$ on λ and α it turns out that $\alpha\tilde{\eta}_u$ is almost solely determined by λ. In conclusion, for the example considered, the effective forgetting factor $\lambda^u_{eff}(t)$ will be affected by the nearest neighbour bandwidth, so that the effective number of observations within the bandwidth will be strongly dependent on λ, but only weakly dependent on the bandwidth (α). The ratio between the rate at which the weights on observations goes to zero in the direction of time and the corresponding rate in the direction of u_t will be determined by α.

As it is illustrated by Figure 1 the effective number of observations behind each

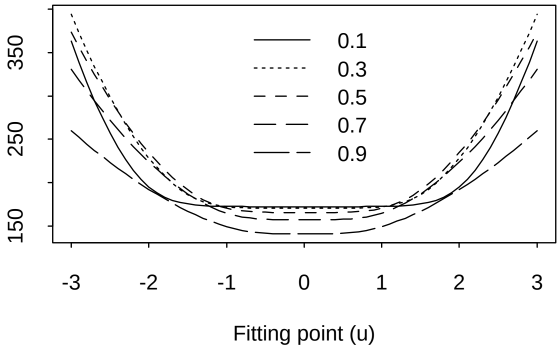

Figure 1: *Effective number of observations within the bandwidth* $(\alpha\tilde{\eta}_u(u))$ *for* $\alpha = 0.1, \ldots, 0.9$ *and* $\lambda = 0.99$.

of the local approximations depends on the fitting point. This is contrary to the non-adaptive nearest neighbour method, cf. Section 2, and may result in a somewhat unexpected behaviour of the estimates. If the system follows a linear ARX-model and if the coefficients of the system are estimated as coefficient-functions then both adaptive and non-adaptive nearest neighbour approaches will be unbiased. However, for this example the variance of local constant estimates will decrease for increasing values of $|u|$. This is verified by simulations, which also show that local linear and quadratic approximations results in increased variance for large $|u|$. Note that, when the true function is not a constant, the local constant approximation may result in excess bias, see e.g. (Nielsen et al. 1997).

If λ is varied with the fitting point as $\lambda(\mathbf{u}) = 1 - 1/(T_0 E[w_u(\mathbf{u}_t)])$ then $\tilde{\eta}_u = T_0$. Thus, the effective number of observations within the bandwidth is constant across fitting points. Furthermore, T_0 can be interpreted as the memory time constant. To avoid highly variable estimates of $E[w_u(\mathbf{u}_t)]$ in the tails of the distribution of $\mathbf{u}_t$ the estimates should be based on a parametric family of distributions. However, in the remaining part of this paper λ is not varied across fitting points.

3.6 Summary of the method

To clarify the method the actual algorithm is briefly described in this section. It is assumed that at each time step t measurements of the output y_t and the two sets of inputs $\mathbf{x}_t$ and $\mathbf{u}_t$ are received. The aim is to obtain adaptive estimates of the coefficient-functions in the non-linear model (1).

Besides λ in (15), prior to the application of the algorithm a number of fitting points $\mathbf{u}^{(i)}$; $i = 1, \ldots, n_{fp}$ in which the coefficient-functions are to be estimated has to be selected. Furthermore the bandwidth associated with each of the fitting points $\hbar^{(i)}$; $i = 1, \ldots, n_{fp}$ and the degrees of the approximating polynomials $d(j)$; $j = 1, \ldots, p$ have to be selected for each of the p coefficient-functions. For simplicity the degree of the approximating polynomial for a particular coefficient-function will be fixed across fitting points. Finally, initial estimates of the coefficient-functions in the model corresponding to local constant estimates, i.e. $\hat{\boldsymbol{\phi}}_0(\mathbf{u}^{(i)})$, must be chosen. Also, the matrices $\mathbf{R}_{u^{(i)},0}$ must be chosen. One possibility is $\mathrm{diag}(\epsilon, \ldots, \epsilon)$, where ϵ is a small positive number.

In the following description of the algorithm it will be assumed that $\mathbf{R}_{u^{(i)},t}$ is nonsingular for all fitting points. In practice we would just stop updating the estimates if the matrix become singular. Under the assumption mentioned the algorithm can be described as:

For each time step t: Loop over the fitting points $\mathbf{u}^{(i)}$; $i = 1, \ldots, n_{fp}$ and for each fitting point:

- Construct the explanatory variables corresponding to local constant estimates using (2):

  $\mathbf{z}_t^T = [x_{1,t}\mathbf{p}_{d(1)}^T(\mathbf{u}_t) \ldots x_{p,t}\mathbf{p}_{d(p)}^T(\mathbf{u}_t)]$.

- Calculate the weight using (6) and (7):

  $w_{u^{(i)}}(\mathbf{u}_t) = (1 - (||\mathbf{u}_t - \mathbf{u}^{(i)}||/\hbar^{(i)})^3)^3$, if $||\mathbf{u}_t - \mathbf{u}^{(i)}|| < \hbar^{(i)}$ and zero otherwise.

- Find the effective forgetting factor using (16):

  $\lambda_{eff}^{(i)}(t) = 1 - (1 - \lambda)w_{u^{(i)}}(\mathbf{u}_t)$.

- Update $\mathbf{R}_{u^{(i)},t-1}$ using (15):

  $\mathbf{R}_{u^{(i)},t} = \lambda_{eff}^{(i)}(t)\mathbf{R}_{u^{(i)},t-1} + w_{u^{(i)}}(\mathbf{u}_t)\mathbf{z}_t\mathbf{z}_t^T$.

- Update $\hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)})$ using (12):

  $\hat{\boldsymbol{\phi}}_t(\mathbf{u}^{(i)}) = \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)}) + w_{u^{(i)}}(\mathbf{u}_t)\mathbf{R}_{u^{(i)},t}^{-1}\mathbf{z}_t \left[y_t - \mathbf{z}_t^T \hat{\boldsymbol{\phi}}_{t-1}(\mathbf{u}^{(i)})\right]$.

- Calculate the updated local polynomial estimates of the coefficient-functions using (8):

  $\hat{\theta}_{jt}(\mathbf{u}^{(i)}) = \mathbf{p}_{d(j)}^T(\mathbf{u}^{(i)})\, \hat{\boldsymbol{\phi}}_{j,t}(\mathbf{u}^{(i)})$; $j = 1, \ldots p$

The algorithm could also be implemented using the matrix inversion lemma as in (Ljung & Söderström 1983).

4 Simulations

Aspects of the proposed method are illustrated in this section. When the modified updating formula (15) is used the general behaviour of the method for different bandwidths is illustrated in Section 4.1. In Section 4.2 results obtained using the two updating formulas (13) and (15) are compared.

The simulations are performed using the non-linear model

$$y_t = a(t, u_{t-1})y_{t-1} + b(t, u_{t-1})x_t + e_t, \tag{19}$$

where $\{x_t\}$ is the input process, $\{u_t\}$ is the process controlling the coefficients, $\{y_t\}$ is the output process, and $\{e_t\}$ is a white noise standard Gaussian process. The coefficient-functions are simulated as

$$a(t, u) = 0.3 + (0.6 - \frac{1.5}{N}t) \exp\left(-\frac{(u - \frac{0.8}{N}t)^2}{2(0.6 - \frac{0.1}{N}t)^2}\right)$$

and

$$b(t, u) = 2 - \exp\left(-\frac{(u + 1 - \frac{2}{N}t)^2}{0.32},\right)$$

where $t = 1, \dots, N$ and $N = 5000$, i.e. $a(t, u)$ ranges from -0.6 to 0.9 and $b(t, u)$ ranges from 1 to 2. The functions are displayed in Figure 2. As indicated by the figure both coefficient-functions are based on a Gaussian density in which the mean and variance varies linearly with time.

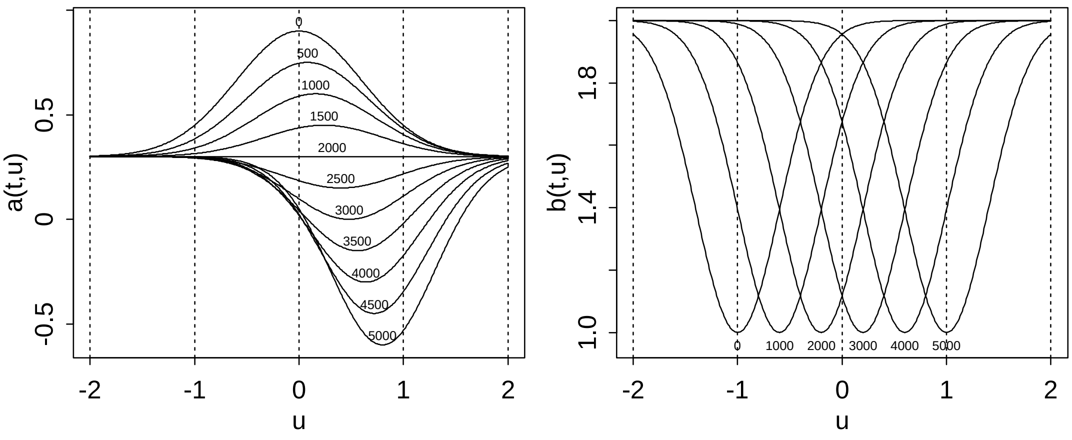

Figure 2: *The time-varying coefficient-functions plotted for equidistant points in time as indicated on the plots.*

Local linear ($d(j) = 1$ for all j) adaptive estimates of the functions $a()$ and $b()$ are then found using the proposed procedure with the model

$$y_t = a(u_{t-1})y_{t-1} + b(u_{t-1})x_t + e_t. \tag{20}$$

In all cases initial estimates of the coefficient-functions are set to zero and during the initialization the estimates are not updated, for the fitting point considered, until ten observations have received a weight of 0.5 or larger.

4.1 Highly correlated input processes

In the simulation presented in this section a strongly correlated $\{\mathbf{u}_t\}$ process is used and also the $\{\mathbf{x}_t\}$ process is quite strongly correlated. This allows us to illustrate various aspects of the method. For less correlated series the performance is much improved.

The data are generated using (19) where $\{x_t\}$ and $\{u_t\}$ are zero mean $AR(1)$-processes with poles in 0.9 and 0.98, respectively. The variance for both series is one and the series are mutually independent. In Figure 3 the data are displayed. Based on these data adaptive estimation in (20) are performed using nearest neighbour bandwidths, calculated assuming a standard Gaussian distribution for u_t.

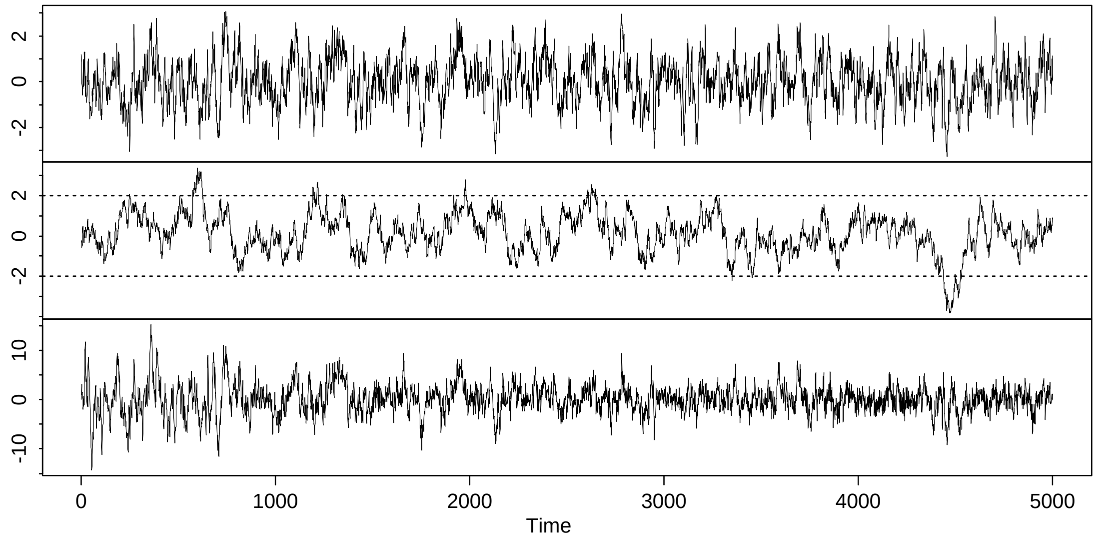

Figure 3: *Simulated output (bottom) when x_t (top) and u_t (middle) are $AR(1)$-processes.*

The results obtained using the modified updating formula (15) are displayed for fitting points $u = -2, -1, 0, 1, 2$ in Figures 4 and 5. For the first 2/3 of the period the estimates at $u = -2$, i.e. $\hat{a}(-2)$ and $\hat{b}(-2)$, only gets updated occasionally. This is due to the correlation structure of $\{u_t\}$ as illustrated by the realization displayed in Figure 3.

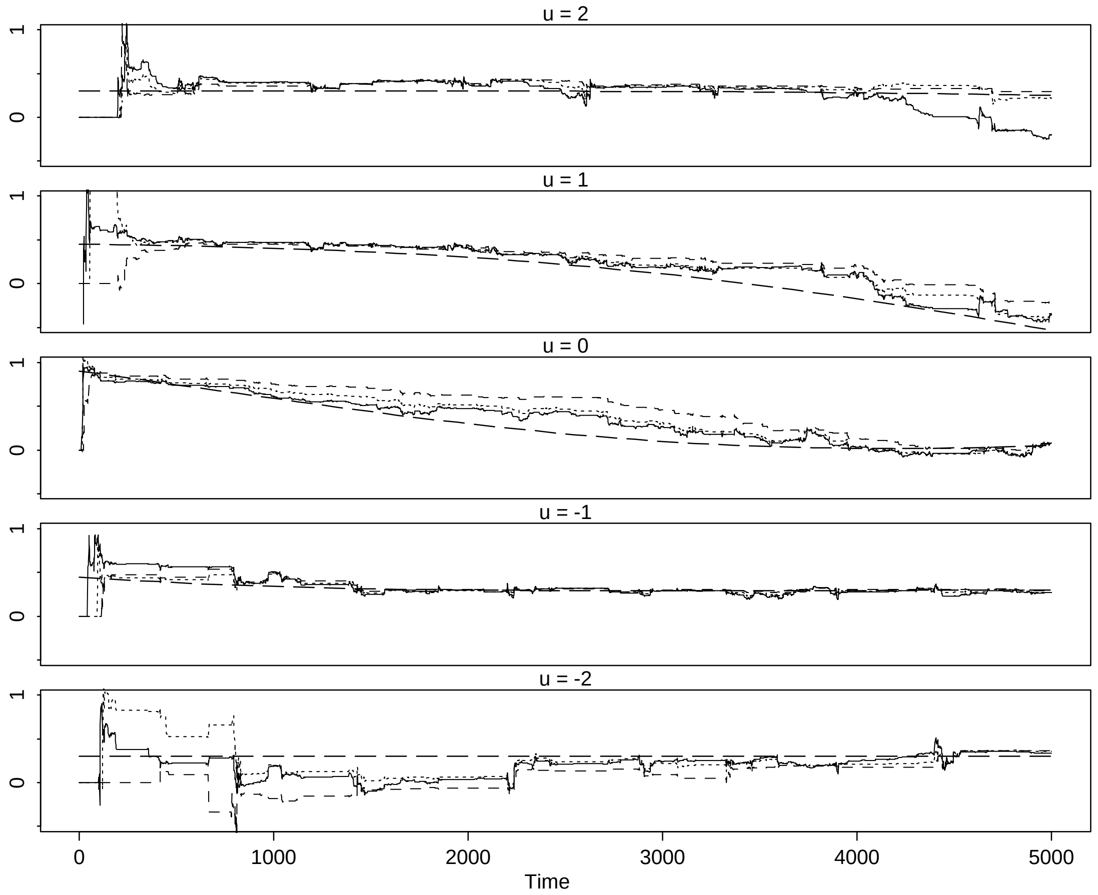

Figure 4: *Adaptive estimates of $a(u)$ using local linear approximations and nearest neighbour bandwidths 0.3 (dashed), 0.5 (dotted), and 0.7 (solid). True values are indicated by smooth dashed lines.*

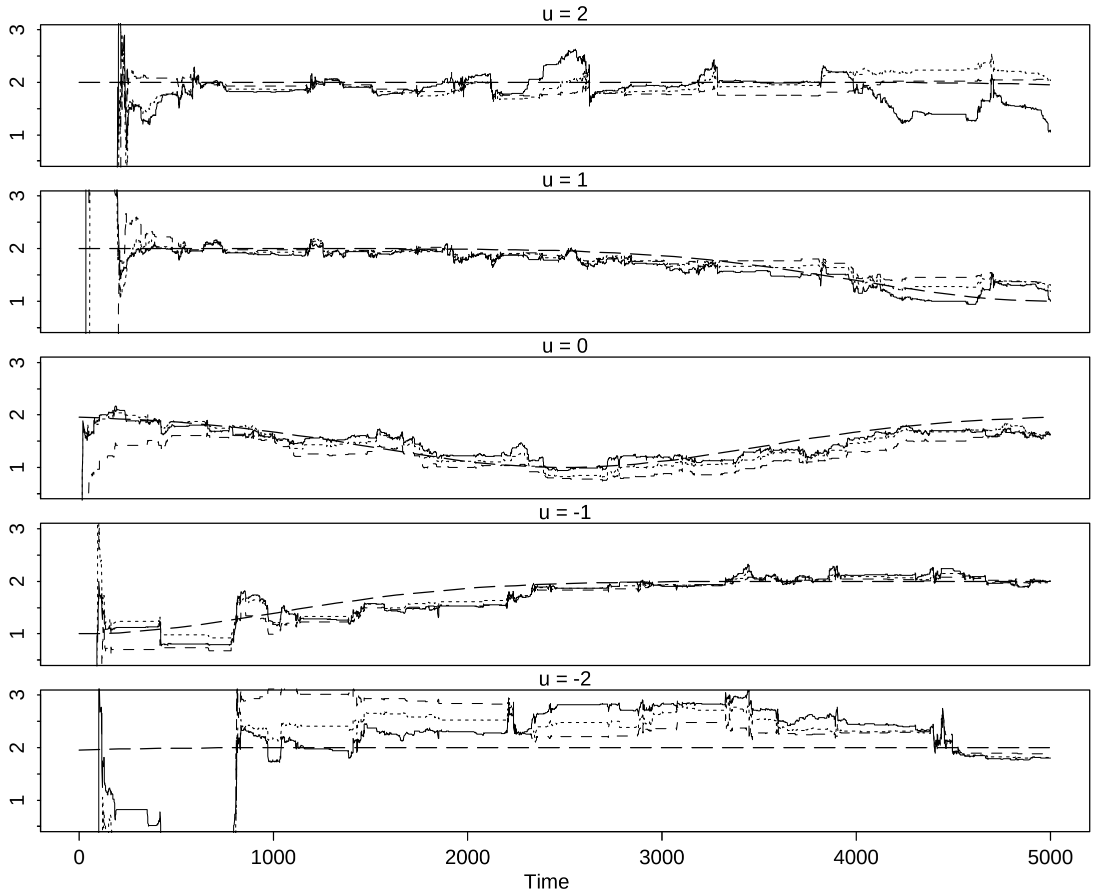

Figure 5: *Adaptive estimates of $b(u)$ using local linear approximations and nearest neighbour bandwidths 0.3 (dashed), 0.5 (dotted), and 0.7 (solid). True values are indicated by smooth dashed lines.*

For both estimates the bias is most pronounced during periods in which the true coefficient-function changes quickly for values of u_t near the fitting point considered. This is further illustrated by the true functions in Figure 2 and it is, for instance clear that adaption to $a(t, 1)$ is difficult for $t > 3000$. Furthermore, $u = 1$ is rarely visited by $\{u_t\}$ for $t > 3000$, see Figure 3. In general, the low bandwidth ($\alpha = 0.3$) seems to result in large bias, presumably because the effective forgetting factor is increased on average, cf. Section 3.5. Similarly, the high bandwidth ($\alpha = 0.7$) result in large bias for $u = 2$ and $t > 4000$. A nearest neighbour bandwidth of 0.7 corresponds to an actual bandwidth of approximately 2.5 at $u = 2$ and since most values of u_t are below one, it is clear that the estimates at $u = 2$ will be highly influenced by the actual function values for u near one. From Figure 2 it is seen that for $t > 4000$ the true values at $u = 1$ is markedly lower that the true values at $u = 2$. Together with the fact that $u = 2$ is not visited by $\{u_t\}$ for $t > 4000$ this explains the observed bias at $u = 2$, see Figure 6.

4.2 Abrupt changes in input signals

One of the main advantages of the modified updating formula (15) over the normal updating formula (13) is that it does not allow fast changes in the estimates at fitting points which has not been visited by the process $\{u_t\}$ for a longer period. If, for instance, we wish to adaptively estimate the stationary relation between the heat consumption of a town and the ambient air temperature then $\{u_t\}$ contains an annual

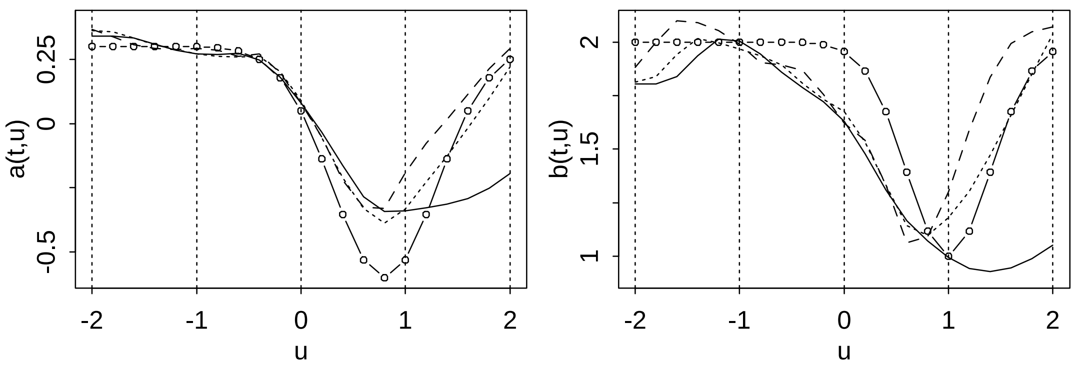

Figure 6: *Adaptive estimates for the example considered in Section 4.1 at* $t = 5000$ *for* $\alpha = 0.3$ *(dashed),* 0.5 *(dotted),* 0.7 *(solid). True values are indicated by circles and fitting points ranging from -2 to 2 in steps of 0.2 are used.*

fluctuation and at some geographical locations the transition from, say, warm to cold periods may be quite fast. In such a situation the normal updating formula (13) will, essentially, forget the preceding winter during the summer, allowing for large changes in the estimate at low temperatures during some initial period of the following winter. Actually, it is possible that, using the normal updating formula will result in a nearly singular $\mathbf{R}_t$.

To illustrate this aspect 5000 observations are simulated using the model (19). The sequence $\{x_t\}$ is simulated as a standard Gaussian $AR(1)$-process with a pole in 0.9.

Furthermore, $\{u_t\}$ is simulated as an iid process where

$$u_t \sim \begin{cases} N(0,1), & t = 1, \ldots, 1000 \\ N(3/2, 1/6^2), & t = 1001, \ldots, 4000 \\ N(-3/2, 1/6^2), & t = 4001, \ldots, 5000 \end{cases}$$

To compare the two methods of updating, i.e. (13) and (15), a fixed λ is used in (15) across the fitting points and the effective forgetting factors are designed to be equal. If $\tilde{\lambda}$ is the forgetting factor corresponding to (13) it can be varied with u as

$$\tilde{\lambda}(u) = E[\lambda^u_{eff}(t)] = 1 - (1 - \lambda)E[w_u(u_t)],$$

where $E[w_u(u_t)]$ is calculated assuming that u_t is standard Gaussian, i.e. corresponding to $1 \leq t \leq 1000$. A nearest neighbour bandwidth of 0.5 and $\lambda = 0.99$ are used, which results in $\tilde{\lambda}(0) = 0.997$ and $\tilde{\lambda}(\pm 2) = 0.9978$.

The corresponding adaptive estimates obtained for the fitting point $u = -1$ are shown in Figure 7. The figure illustrates that for both methods the updating of the estimates stops as $\{u_t\}$ leaves the fitting point $u = -1$. Using the normal updating (13) of $\mathbf{R}_t$ its value is multiplied by $\tilde{\lambda}(-1)^{3000} \approx 0.00015$ as $\{u_t\}$ returns to the vicinity of the fitting point. This results in large fluctuations of the estimates, starting at $t = 4001$. As opposed to this, the modified updating (15) does not lead to such fluctuations after $t = 4000$.

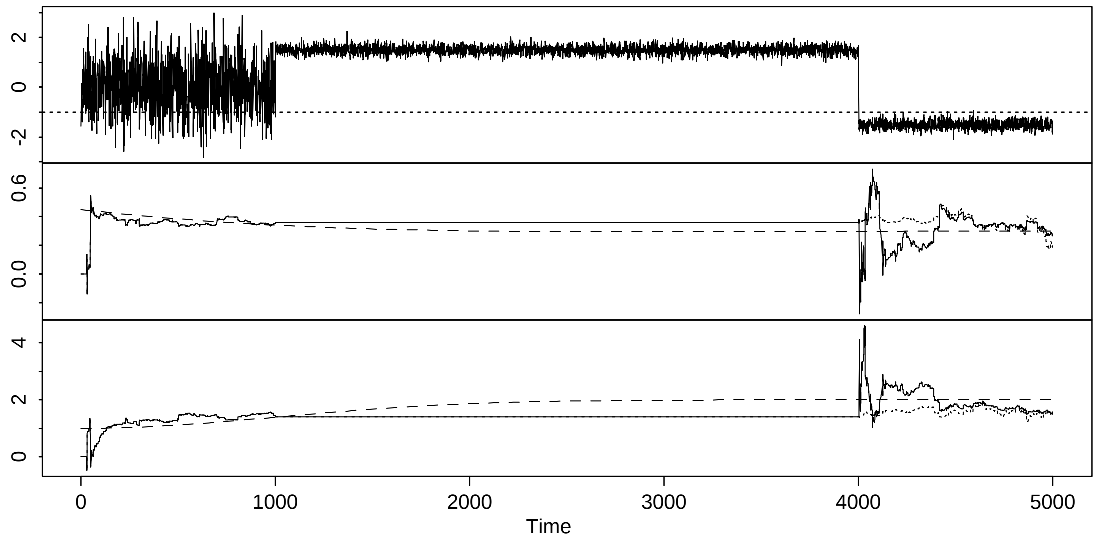

Figure 7: *Realization of* $\{u_t\}$ *(top) and adaptive estimates of* $a(-1)$ *(middle) and* $b(-1)$ *(bottom), using the normal updating formula (solid) and the modified updating formula (dotted). True values are indicated by dashed lines.*

5 Further topics

Optimal bandwidth and forgetting factor: So far in this paper it has been assumed that the bandwidths used over the range of $\mathbf{u}_t$ is derived from the nearest neighbour bandwidth α and it has been indicated how it can be ensured that the average forgetting factor is large enough.

However, the adaptive and recursive method is well suited for forward validation (Hjorth 1994) and hence tuning parameters can be selected by minimizing, e.g. the root mean square of the one-step prediction error (using observed $\mathbf{u}_t$ and $\mathbf{x}_t$ to predict y_t, together with interpolation between fitting points to obtain $\hat{\boldsymbol{\theta}}_{t-1}(\mathbf{u}_t)$).

There are numerous ways to define the tuning parameters. A simple approach is to use (λ, α), cf. (15) and (17). A more ambiguous approach is to use both λ and $\hbar$ for each fitting point $\mathbf{u}$. Furthermore, tuning parameters controlling scaling and rotation of $\mathbf{u}_s$ and the degree of the local polynomial approximations may also be considered.

If n fitting points are used this amounts to $2n$, or more, tuning parameters. To make the dimension of the (global) optimization problem independent of n and to have $\lambda(\mathbf{u})$ and $\hbar(\mathbf{u})$ vary smoothly with $\mathbf{u}$ we may choose to restrict $\lambda(\mathbf{u})$ and $\hbar(\mathbf{u})$, or appropriate transformations of these (logit for λ and log for $\hbar$), to follow a spline basis (de Boor 1978, Lancaster & Salkauskas 1986). This is similar to the smoothing of spans described by Friedman (1984).

Local time-polynomials: In this paper local polynomial approximations in the direction of time is not considered. Such a method is proposed for usual ARX-models by Joensen, Nielsen, Nielsen & Madsen (2000). This method can be combined with the method described here and will result in local polynomial approximations where cross-products between time and the conditioning variables ($\mathbf{u}_t$) are excluded.

6 Conclusion and discussion

In this paper methods for adaptive and recursive estimation in a class of non-linear autoregressive models with external input are proposed. The model class considered is conditionally parametric ARX-models (CPARX-model), which is a conventional ARX-model in which the parameters are replaced by smooth, but otherwise unknown, functions of a low-dimensional input process. These functions are estimated adaptively and recursively without specifying a global parametric form. One possible application of CPARX-models is the modelling of varying time delays, cf. (Nielsen et al. 1997).

The methods can be seen as generalizations or combinations of recursive least squares with exponential forgetting (Ljung & Söderström 1983), local polynomial regression (Cleveland & Devlin 1988), and conditional parametric fits (Cleveland 1994). Hence, the methods constitutes an extension to the notion of local polynomial estimation. The so called modified method is suggested for cases where the process controlling the coefficients are highly correlated or exhibit seasonal behaviour. The estimates at each time step can be seen as solutions to a range of weighted least squares regressions and therefore the solution is unique for well behaved input processes. A particular feature of the modified method is that the effective number of observations behind the estimates will be almost independent of the actual bandwidth. This is accomplished by varying the effective forgetting factor with the bandwidth. The bandwidth mainly controls the rate at which the weights corresponding to exponential forgetting goes to zero relatively to the rate at which the remaining weights goes to zero.

For some applications it may be possible to specify global polynomial approximations to the coefficient-functions of a CPARX-model. In this situation the adaptive recursive least squares method can be applied for tracking the parameters defining the coefficient-functions for all values of the input process. However, if the argument(s) of the coefficient-functions only stays in parts of the space corresponding to the possible values of the argument(s) for longer periods this may seriously affect the estimates of the coefficient-functions for other values of the argument(s), as it corresponds to extrapolation using a fitted polynomial. This problem is effectively solved using the conditional parametric model in combination with the modified updating formula.

A Effective number of observations

Using the modified updating formula, as described in Section 3.3, the estimates at time t can be written as

$$\hat{\boldsymbol{\phi}}_t(\mathbf{u}) = \underset{\boldsymbol{\phi}_u}{\operatorname{argmin}} \sum_{s=1}^{t} \beta(t,s) w_u(\mathbf{u}_s)(y_s - \mathbf{z}_s^T \boldsymbol{\phi}_u)^2,$$

where

$$\beta(t,t) = 1,$$

and, for $s < t$

$$\beta(t,s) = \prod_{j=s+1}^{t} \lambda_{eff}^{u}(j) = \lambda_{eff}^{u}(t)\beta(t-1,s),$$

where $\lambda_{eff}^{u}(t)$ is given by (16). It is then obvious to define the effective number of observations (in the direction of time) as

$$\eta_u(t) = \sum_{i=0}^{\infty} \beta(t, t-i) = 1 + \lambda_{eff}^{u}(t) + \lambda_{eff}^{u}(t)\lambda_{eff}^{u}(t-1) + \ldots \tag{21}$$

Suppose that the fitting point $\mathbf{u}$ is chosen so that $E[\eta_u(t)]$ exists. Consequently, when $\{\lambda_{eff}^{u}(t)\}$ is i.i.d. and when $\bar{\lambda}_u \in [0,1)$ denotes $E[\lambda_{eff}^{u}(t)]$, the average effective number of observations is

$$\bar{\eta}_u = 1 + \bar{\lambda}_u + \bar{\lambda}_u^2 + \ldots = \frac{1}{1-\bar{\lambda}_u}.$$

When $\{\lambda_{eff}^{u}(t)\}$ is not i.i.d., it is noted that since the expectation operator is linear, $E[\eta_u(t)]$ is the sum of the expected values of each summand in (21). Hence, $E[\eta_u(t)]$ is independent of t if $\{\lambda_{eff}^{u}(t)\}$ is strongly stationary, i.e. if $\{\mathbf{u}_t\}$ is strongly stationary. From (21)

$$\eta_u(t) = 1 + \lambda_{eff}^{u}(t)\eta_u(t-1) \tag{22}$$

is obtained, and from the definition of covariance it then follows, that

$$\bar{\eta}_u = \frac{1 + Cov[\lambda_{eff}^{u}(t), \eta_u(t-1)]}{1-\bar{\lambda}_u} \geq \frac{1}{1-\bar{\lambda}_u}, \tag{23}$$

since $0 < \lambda < 1$ and assuming, that the covariance between $\lambda^u_{eff}(t)$ and $\eta_u(t-1)$ is positive. Note that, if the process $\{\mathbf{u}_t\}$ behaves such that if it has been near $\mathbf{u}$ for a longer period up to time $t-1$ it will tend to be near $\mathbf{u}$ at time t also, a positive covariance is obtained. It is the experience of the authors that such a behaviour of a stochastic process is often encountered in practice.

As an alternative to the calculations above $\lambda^u_{eff}(t)\eta_u(t-1)$ may be linearized around $\bar{\lambda}_u$ and $\bar{\eta}_u$. From this it follows, that if the variances of $\lambda^u_{eff}(t)$ and $\eta_u(t-1)$ are small then

$$\bar{\eta}_u \approx \frac{1}{1-\bar{\lambda}_u}.$$

Therefore we may use $1/(1-\bar{\lambda}_u)$ as an approximation to the effective number of observations, and in many practical applications it will be an lower bound, c.f. (23). By assuming a stochastic process for $\{\mathbf{u}_t\}$ the process $\{\eta_u(t)\}$ can be simulated using (22) whereby the validity of the approximation can be addressed.

References

Chambers, J. M. & Hastie, T. J., eds (1991), *Statistical Models in S*, Wadsworth, Belmont, CA.

Chen, R. & Tsay, R. S. (1993*a*), 'Functional-coefficient autoregressive models', *Journal of the American Statistical Association* **88**, 298–308.

Chen, R. & Tsay, R. S. (1993*b*), 'Nonlinear additive ARX models', *Journal of the American Statistical Association* **88**, 955–967.

Cleveland, W. S. (1994), Coplots, nonparametric regression, and conditionally parametric fits, *in* T. W. Anderson, K. T. Fang & I. Olkin, eds, 'Multivariate Analysis and Its Applications', Institute of Mathematical Statistics, Hayward, pp. 21–36.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**, 596–610.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer Verlag, Berlin.

Friedman, J. H. (1984), A variable span smoother, Technical Report 5, Laboratory for Computational Statistics, Dept. of Statistics, Stanford Univ., California.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London/New York.

Hastie, T. & Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 757–796.

Hjorth, J. S. U. (1994), *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap*, Chapman & Hall, London/New York.

Joensen, A. K., Nielsen, H. A., Nielsen, T. S. & Madsen, H. (2000), 'Tracking time-varying parameters with local regression', Accepted for publication in *Automatica*.

Lancaster, P. & Salkauskas, K. (1986), *Curve and Surface Fitting: An Introduction*, Academic Press, New York/London.

Ljung, L. & Söderström, T. (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.

Nielsen, H. A., Nielsen, T. S. & Madsen, H. (1997), ARX-models with parameter variations estimated by local fitting, *in* Y. Sawaragi & S. Sagara, eds, '11th IFAC Symposium on System Identification', Vol. 2, pp. 475–480.

Thuvesholmen, M. (1997), 'An on-line crossvalidation bandwidth selector for recursive kernel regression', Lic thesis, Department of Mathematical Statistics, Lund University, Sweden.

Vilar-Fernández, J. A. & Vilar-Fernández, J. M. (1998), 'Recursive estimation of regression functions by local polynomial fitting', *Annals of the Institute of Statistical Mathematics* **50**, 729–754.