

# Semiparametric regression smoothing and feature detection in time series

Michael G. Schimek

*Institute for Medical Informatics, Statistics and Documentation,  
Karl-Franzens-University Graz, Austria, Europe; michael.schimek@uni-graz.at*

---

## Abstract

For time series data a semiparametric partial spline model is considered which facilitates the detection of features resulting from external interventions. It is a generalization of Schimek's (2000) unbiased risk approach to semiparametric regression under the usual i.i.d. error assumption. Non-iterative procedures for exact model estimation and smoothing parameter choice under autoregressive errors are introduced. Dummy input is modeled via intervention functions in the parametric part of the model and means of inference are outlined. An exploratory approach to feature detection is described and illustrated on a well-known time series data set from the literature.

*Key words:* AR errors, exact estimation, intervention function, partial spline, semiparametric regression, smoothing parameter choice, time series, unbiased risk

---

## 1 Introduction

Let us assume a regularly spaced time series with white noise or weak stationary autoregressive (abb. AR) errors of known order. Our interest is the detection (i.e. identification and estimation) of such features in the time series that cannot be explained by the long-term trend or the error structure. Typically such features are associated with irregular impacts on the process monitored.

Assuming nonparametric smooth trends and white noise errors, semiparametric regression models are a reasonable choice because they allow screening for impact effects in a parametric manner as known from classical time domain intervention analysis (Box and Tiao, 1975). Unbiased partial spline fitting (Schimek, 2000) is a recent and computationally efficient approach to evaluate such a model. However, correct nonparametric estimation of a smooth trend

in a series of dependent observations asks for specific regression techniques that take care of the error structure to avoid under- respectively over-smoothing (Kohn, Schimek and Smith, 2000). A purely nonparametric approach was proposed by Schimek and Schmaranz (1994) for AR and moving average errors. Here we extend it to the semiparametric setting for AR errors (moving average errors in combination with smoothing splines are of limited interest; see later). Finally it is feasible to perform feature detection as in the white noise case.

Under both error options we introduce the same methodology for feature detection. It is based on artificial dummy input series which are linked to the time series output by transfer functions. In that way we can experimentally analyze certain features of the time series which are of interest beyond long-term trend. These features for instance can be characterized by abrupt directional changes or drifts (due to external known or unknown impacts or interventions) in the time series other than smooth, longer lasting trends.

Whenever we have to fit a smooth curve in a semiparametric regression model the choice of the smoothing parameter or the bandwidth is crucial. There is some evidence concerning independent identically distributed (abb. i.i.d.) errors (Eubank et al., 1998). A still open problem raised and tackled in Schimek and Schmaranz (1994) by means of exploratory tools is the choice of the appropriate degree of smoothing under dependent errors. The severity of this problem becomes even more pronounced in the semiparametric setting when statistical testing of dummy variable input is necessary. Our ability of obtaining significant  $p$ -values is regrettably not independent of the degree of smoothing (for details see Azzalini and Bowman, 1993). Here Eubank's et al. (1998) unbiased risk (UBR) criterion is adapted in such a way that AR errors can be accommodated.

In this paper trend is fitted nonparametrically by cubic smoothing splines and estimation techniques are outlined with special emphasis on dependent observations. It is shown that the computational burden increases when the white noise error assumption is replaced by the AR error assumption. As a matter of fact the UBR estimates are obtained in  $O(n^2)$  steps instead of  $O(n)$  which is reasonable given the complexity of the estimation problem. In both instances the parametric coefficients can still be tested via a simple approximately  $N(0, 1)$  test statistic.

Finally we illustrate the approach on a time series data set from Pankratz (1991).

## 2 The semiparametric regression model

The predictor function of the semiparametric regression model consists of a parametric linear component and an arbitrary nonparametric component depending on a design variable  $t$  denoting time in this paper. The responses  $y_1, \dots, y_n$  are obtained at arbitrary non-stochastic ordered values  $t_1, \dots, t_n$  of  $t$ . Let us then have

$$y_i = u_i^T \gamma + g(t_i) + e_i \quad (1)$$

for  $i = 1, \dots, n$ , where  $u_1, \dots, u_n$  are known  $k$ -dimensional covariate vectors,  $\gamma$  is a  $k$ -dimensional unknown coefficient vector,  $g$  is an unknown smooth function, and the  $e_1, \dots, e_n$  are independent, zero mean random variables with a common variance  $\sigma^2$ . The i.i.d. (white noise) error assumption is crucial for the evaluation of such a model. In matrix notation (1) takes the form

$$\mathbf{y} = \mathbf{U}\gamma + \mathbf{g} + \epsilon$$

where  $\mathbf{y} = (y_1, \dots, y_n)^T$ ,  $\mathbf{U}^T = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ ,  $\mathbf{g} = (g(x_1), \dots, g(x_n))^T$ , and  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  a vector of random errors as specified above.

The corresponding sum of squares equation is

$$SS(\gamma, g) = \sum_{i=1}^n (y_i - u_i^T \gamma - g(t_i))^2 + \lambda RP,$$

where

$$RP = \int_a^b [g''(t)]^2 \quad (2)$$

denotes the roughness penalty of a cubic smoothing spline with knots at  $t_1, \dots, t_n$  for a fixed smoothing parameter  $\lambda > 0$ . For each value of  $\lambda$  let us assume that there is an  $n \times n$  cubic spline smoother matrix  $\mathbf{S}_\lambda = (f_\lambda(x_1), \dots, f_\lambda(x_n))^T$  which is positive semidefinite.

Under the assumptions made, not to forget the i.i.d. errors (white noise) equation (2) describes what is commonly known as partial spline (Wahba, 1990, chap. 10).

## 2.1 Estimation, smoothing parameter choice and inference

Green and Silverman (1994, chap. 4) suggest an estimation concept related to iterative backfitting which yields asymptotically biased coefficients  $\gamma$ , however the bias is not severe for reasonable  $n$  (Schimek, 2000, p. 533ff). Based on results due to Speckman (1988), Eubank et al. (1998) and Schimek (2000) could derive an UBR estimator and cheap  $O(n)$  algorithms. There, different from B-spline approaches, the full design information is relevant for knot placement. This extra effort - essential when extending the regression model to accommodate AR errors (see next section) - is more than compensated by an efficient direct (non-iterative) algorithm, yielding unbiased curve and coefficient estimates.

The estimators satisfying the sum of squares equation (2) are (Schimek, 2000)

$$\gamma = (\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - \mathbf{S}_\lambda) \mathbf{y}, \quad (3)$$

$$\mathbf{g} = \mathbf{S}_\lambda (\mathbf{y} - \mathbf{U} \gamma),$$

and

$$\mu = \mathbf{g} + \mathbf{U} \gamma = \mathbf{H}_\lambda \mathbf{y}$$

where  $\mathbf{H}_\lambda$  is a new smoother matrix, also called hat matrix, depending on  $\lambda$ , with

$$\mathbf{H}_\lambda = \mathbf{S}_\lambda + \tilde{\mathbf{U}} (\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - \mathbf{S}_\lambda),$$

and

$$\tilde{\mathbf{U}} = (I - \mathbf{S}_\lambda) \mathbf{U}.$$

The so-called partial residual vectors as introduced in Speckman (1988, p.414f) are  $\tilde{\mathbf{y}} = (I - \mathbf{S}_\lambda) \mathbf{y}$  and  $\tilde{\mathbf{U}} = (I - \mathbf{S}_\lambda) \mathbf{U}$ . The parameter  $\gamma$  is then computed by regression on partial residuals. This removes the influence of the design variable  $t$  from both  $\mathbf{U}$  and  $\mathbf{y}$  which substantially reduces the asymptotic bias.

We cannot emphasize enough how critical the choice of  $\lambda$  is in the semiparametric context: Apart from estimation aspects our task is to formally test covariate effects via the  $\gamma$  coefficients. Generalized cross-validation could be used here but UBR estimation performs better according to our experience. However the latter requires the consistent estimation of the error variance  $\sigma^2$ . Eubank et al. (1998) provide such an estimator based on pseudo-residuals which can be calculated in only  $O(n)$  steps (for details see there).

The smoothing parameter  $\hat{\lambda}$  in the semiparametric setting can finally be obtained by minimizing the UBR criterion

$$UBR(\lambda) = n^{-1} (\|\mathbf{y} - \mu\|^2 + 2\sigma^2 \text{tr}(\mathbf{H}_\lambda)) \quad (4)$$

where  $\mathbf{H}_\lambda = \mathbf{S}_\lambda + \tilde{\mathbf{U}}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - \mathbf{S}_\lambda)$ .  $\|\cdot\|$  denotes the Euclidean norm and  $\|\mathbf{y} - \mu\| = \|(I - \mathbf{H}_\lambda)\mathbf{y}\|$ .

As far as testing of the parametric component is concerned, Speckman (1988) could prove that the coefficient vector  $\gamma$  is asymptotically NID without distributional assumption on the dependent variable respectively errors. As a result of this a z-distributed test statistic (Eubank, 1999, p.206),

$$z = \hat{\gamma}/SE(\hat{\gamma}). \tag{5}$$

is efficient at least in an asymptotic sense. For the standard errors  $SE$  we need to calculate

$$Var(\gamma) = \sigma^2(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - \mathbf{S}_\lambda)^2 \tilde{\mathbf{U}} (\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1}.$$

### 3 A semiparameric regression model for autoregressive errors

Our intention is to overcome the strict white noise error assumption which characterizes the well-known semiparametric regression model described in section 2. This assumption is rare in time series. Instead of arbitrary ordered values  $t_1, \dots, t_n$  of  $t$  we assume equidistant time points as known from standard time domain analysis.

First of all let us define a  $n$ -dimensional column vector  $e$  representing the errors. We assume  $E(e) = 0$  for the error mean and  $E(ee^T) = D$  for the error dispersion. We consider  $AR(p)$  processes (in operator notation; see Tiao, 2001a, p.54f)

$$\Phi_p(B)x_t = \epsilon_t$$

where  $x_t$  is the input of a linear system at time  $t$ ,  $B$  denoting the backshift operator, and  $\Phi_p(B)$  the AR operator of order  $p$ ,

$$\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p).$$

The random shocks  $\epsilon_t$  are considered to be independently  $N(0, \sigma_\epsilon^2)$  distributed.

We imply restrictions on the AR parameters  $\phi$  for the purpose of stationarity of the  $AR(p)$  process in that we require the roots (regarding  $B$  as a complex variable) of  $\Phi_p(B) = 0$  to be outside the unit circle (Tiao, 2001a, p.58f). As a matter of fact the assumption of weak (covariance) stationarity is sufficient for our purposes. With these prerequisites we can specify the error term  $e_i$  in equation (1) over the design space.



### 3.1 Estimation, smoothing parameter choice and inference

Here our primary goal is to develop a non-iterative algorithm which works for dependent errors and at the same time is not too computationally demanding. Schimek (1988) developed a discrete roughness penalty approach (spline-like curve fitting) for known AR and moving average errors. Here we adapt some of these ideas for the AR case to work with partial (smoothing) splines. This can be achieved by extending the methodology developed in Schimek (2000).

Equation (6) for the partial spline in matrix notation is

$$SS(\gamma, \mathbf{g}) = (\mathbf{y} - \mathbf{U}\gamma - \mathbf{g})^T D^{-1}(\mathbf{y} - \mathbf{U}\gamma - \mathbf{g}) + \lambda \mathbf{g}^T K \mathbf{g} \quad (7)$$

where  $K = QR^{-1}Q^T$  and  $\mathbf{y} = (y_1, \dots, y_n)^T$ ,  $\mathbf{U}^T = [\mathbf{u}_1, \dots, \mathbf{u}_n]$  and  $\mathbf{g} = (g(t_1), \dots, g(t_n))^T$ .  $Q$  and  $R$  are known band matrices from the value-second derivative representation of the cubic spline (Green and Silverman, 1994, p. 12f).

Let  $h_i = t_{i+1} - t_i$  for  $i = 1, \dots, n-1$ . Let  $Q$  be the  $n \times (n-2)$  matrix with elements  $q_{ij}$  for  $i = 1, \dots, n$  and  $j = 2, \dots, n-1$ , given by  $q_{j-1,j} = h_{j-1}^{-1}$ ,  $q_{j,j} = -h_{j-1}^{-1} - h_j^{-1}$ , and  $q_{j+1,j} = h_j^{-1}$  for  $j = 2, \dots, n-1$ , and  $q_{ij} = 0$  for  $|i-j| \geq 2$ . The columns of  $Q$  start with  $j = 2$ .

The symmetric matrix  $R$  is  $(n-2) \times (n-2)$  with elements  $r_{ij}$  for  $i$  and  $j$ , running from 2 to  $(n-1)$ . They are defined as follows:  $r_{ii} = \frac{1}{3}(h_{i-1} + h_i)$  for  $i = 2, \dots, n-1$ ,  $r_{i,i+1} = \frac{1}{6}h_i$  for  $i = 2, \dots, n-2$ , and  $r_{ij} = 0$  for  $|i-j| \geq 2$ .  $R$  is positive definite.

Finally the  $(n-2) \times (n-2)$  matrix  $K$  is symmetric and positive definite. For a weak stationary AR( $p$ ) process the inverted covariance matrix  $D^{-1}$  is symmetric, banded with bandwidth  $2p+1$ , and positive definite. Its elements can be calculated directly.

Our goal is to efficiently estimate the parameter vector  $\gamma$ , the function  $\mathbf{g}$ , and the mean vector  $\mu$ . Minimizing (7) yields a penalized least squares solution. The estimators are

$$\gamma = (\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - (D^{-1} + \lambda K)^{-1} D^{-1}) \mathbf{y}$$

and

$$\mathbf{g} = (D^{-1} + \lambda K)^{-1} D^{-1} (\mathbf{y} - \mathbf{U}\gamma),$$

where

$$\tilde{\mathbf{U}} = (I - (D^{-1} + \lambda K)^{-1} D^{-1}) \mathbf{U}.$$

Further we have

$$\mu = \mathbf{U}\gamma + \mathbf{g} = \mathbf{H}_{\lambda}^* \mathbf{y}$$

with the hat matrix

$$\mathbf{H}_\lambda^* = (D^{-1} + \lambda K)^{-1} D^{-1} + \tilde{\mathbf{U}}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - (D^{-1} + \lambda K)^{-1} D^{-1}).$$

Compared to independent errors for which we have linear algorithms, the algorithm for AR errors is of quadratic nature as long as  $n \gg p$ . Cheaper numerical approaches exploiting the band-limited nature of some matrices are subject to ongoing research. A similar algorithm cannot be designed for moving average errors (would imply full matrices throughout). However the combination of smoothing splines and moving average errors is of limited practical value (see Diggle, 1996, and Schimek, 1992). The estimators for the extended semiparametric regression model are asymptotically unbiased under the assumptions made.

For the UBR criterion (4) the variance estimator of Eubank et al. (1998) cannot be applied any more. Dependent observations result in a loss of consistency (Herrmann, 2000, p. 88). However we might adapt an estimator due to Speckman (1988, p. 426).

The reason is that this variance estimator can be corrected for AR errors via a modified hat matrix  $\mathbf{H}_\lambda^*$ . This modification causes a positive but asymptotically negligible bias,

$$\sigma_{AR}^2 = \frac{RSS}{tr(I - \mathbf{H}_\lambda^*)^T (I - \mathbf{H}_\lambda^*)}, \quad (8)$$

where  $RSS = \frac{1}{n} \|(I - \mathbf{H}_\lambda^*)\mathbf{y}\|^2$ . The computational demand for the calculation of the residual variance is of  $O(n^2)$  and not  $O(n)$  any more as in the i.i.d. setting.

Finally for testing  $\gamma$  coefficients individually we need

$$Var(\gamma) = \sigma_{AR}^2 (\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1} \tilde{\mathbf{U}}^T (I - (D^{-1} + \lambda K)^{-1} D^{-1})^2 \tilde{\mathbf{U}} (\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})^{-1}.$$

It is still possible to apply an approximate z-distributed test statistic

$$z = \gamma / SE(\gamma)$$

without specific distributional assumptions (the results from Speckman, 1988, still hold).

## 4 Dummy input and feature detection

For the special purpose of feature detection the covariate vectors  $\mathbf{u}_i = \mathbf{u}_j = (u_1, \dots, u_k)^T$  are identified by lagged dummy input series. Transfer functions, known from parametric time series analysis (see e.g. Tiao, 2001b, p.366f.), are introduced here in this semiparametric setting. What we aim at is the analysis of intervention effects on the time series which can be separated from long-term trend.

Let us first introduce transfer functions with dummy input (zero/one variable). They are then called intervention functions and convert the intervention input into level effects which might be present in the time series of interest. The features they describe can be studied analytically (see Schimek, 1988b).

What kind of embedded episodes (features) can be analyzed by means of intervention functions? Let us make the general assumption that we have time-dependent pairs  $(X_t, Y_t)$  of variables,  $X_t$  being the input and  $Y_t$  being the output of a linear system. The general form of an intervention function of order  $s$  is given by

$$Y_t = \omega_s(B)X_{t-b} \tag{9}$$

with a polynomial in the backshift operator  $B$

$$\omega_s(B) = (\omega_0 - \omega_1 B - \omega_2 B^2 - \dots - \omega_s B^s),$$

where the coefficients  $\omega$  are called intervention parameters and the parameter  $b$  is denoting the discrete system lag ( $b \geq 0$ ).

We substitute  $y_i$  for  $Y_t$  and scalar  $u_i$  for  $X_t$  and introduce an intervention function in equation (1). The backshift operator  $B$  applied to the  $u_i$  series gives the dummy series for the onset of the intervention effect and those for the lags up to order  $s$ .

We obtain

$$y_i = \omega_s(B)u_{i-b} + g(t_i) + e_i.$$

Such an intervention function allows us to describe intervention effects of step-wise (exponential) behaviour. A single abrupt change from one mean level to another forms the simplest case (zero-order intervention). A number of examples of zero- and first-order interventions are displayed in Schimek (1988b, p.56ff).

When analyzing impacts on an empirical time series a theory of change would be useful, yet such a theory does not exist. There is no purely parametric way to explore data with respect to intervention functions. However, in the semi-parametric context we are in the position to combine nonparametric function fitting under known AR error with exploratory feature detection. The properties of the change agent represented by the time-dependent intervention variable, especially the onset of the influencing event must be known a priori (in most applied research this is the case). The  $\omega$  parameters forming the intervention function are estimated and tested parametrically in the extended model.

All the information in the data which can be attributed to long-term trend and error structure is used for the fitting of the regression curve. Then we try to model the remaining signal, i.e. the short-term features, by an intervention function. In most cases this situation is characterized by insufficient knowledge concerning the features to be analyzed. Hence model selection has to be performed in an exploratory context.

## 5 An application

For the purpose of illustration let us analyze a well known business data set from Pankratz (1991, p.279, Case 6), known as the *Year-End Loading Data*. It is an equidistant monthly time series which starts January 1976 and ends May 1986, comprising  $n = 125$  observations (no missings) of shipments of a consumer product from the manufacturer to the distributors.

There is the following external information: In 1983 top-level managers began pressuring the marketing department to meet strict calendar year shipment targets. We know of such management interventions for December of the years 1983, 1984, and 1985. In Figure 1 the *Year-End Loading Data*-time series is displayed. In addition we see vertical lines indicating the intervention onsets due to the policy of year-end extra shipments to the distributors. One can ask the following question: What is the impact of this strategy (intervention) on the shipments in later months?

Answers to this question based on classical parametric Box-Tiao time domain intervention analysis can be found in Pankratz (1991, p.280ff). Different models have been suggested and results as well as conclusions can differ substantially.

We believe that purely parametric approaches are not flexible enough to allow for feature detection and at the same time trend and error modeling. Box-Jenkins time series analysis for instance requires differencing of non-stationary

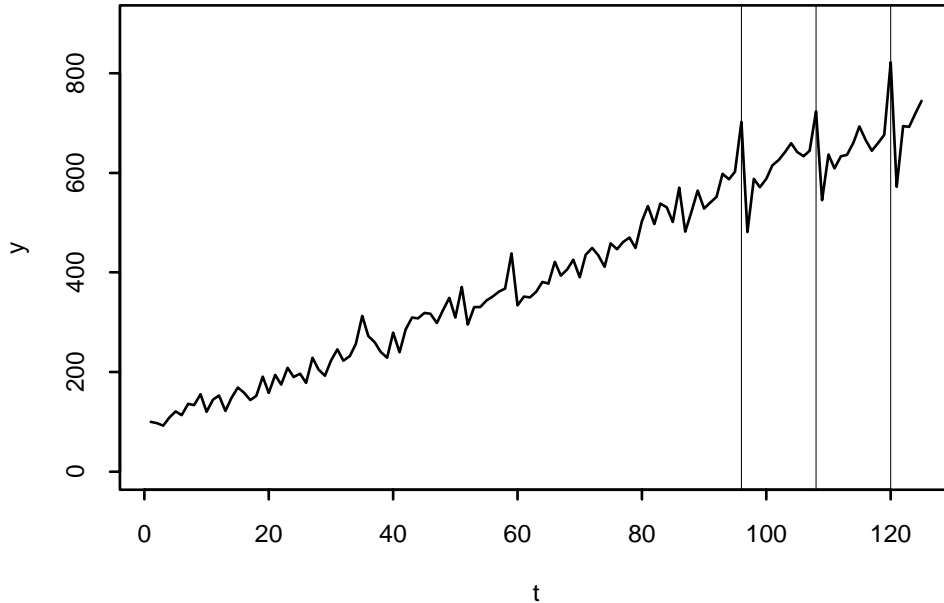


Fig. 1. Monthly time series of shipments and intervention onset for the years 1983, 1984, and 1985

observations. We cannot identify or estimate linear stochastic error processes without mean stationarity. Because of that trend and seasonal components have to be eliminated beforehand. Data transformations such as differencing, assuming an adequate difference order, do not change the error structure, however they can influence the features we want to detect.

The advantage of the semiparametric partial spline approach for impact assessment is its ability to estimate trend while correcting for pre-specified errors. In the semiparametric context we do not have an artificial separation between long-term trend and seasonality as long as these components are smooth. This makes feature detection easier.

We take the following partial spline fitting strategy for the example data set:

- (i) Identify and estimate an AR error model from the pre-intervention series
- (ii) Select  $\hat{\lambda}$  under the specified error model and a preliminary zero-order intervention function via UBR
- (iii) Fit second-order intervention function in three parameters:
  - $\gamma_1$  (i.e.  $\omega_0$ ) for lag=0 effects
  - $\gamma_2$  (i.e.  $\omega_1$ ) for lag=1 effects
  - $\gamma_3$  (i.e.  $\omega_2$ ) for lag=2 effects
- (iv) If (iii) is inadequate fit first-order intervention function in two param-

eters

- (v) If (iv) is inadequate fit zero-order intervention function in one parameter

A significance level of  $\alpha = 0.05$  is assumed throughout this analysis.

### 5.1 Results

As can be seen in Figure 1 the time series is not mean stationary. Because we have to identify an appropriate error process we eliminate the approximately linear trend which characterizes the time series over the whole sample period. Mean stationarity can be obtained by first-order differencing.

The empirical autocorrelation function and the empirical partial autocorrelation function (not displayed here) suggest an AR(2) model with coefficients  $\phi_1 = -0.83$  and  $\phi_2 = -0.36$ . Exploratory prefiltering of the original series with this error model yields some negative observations. Because negative shipments (i.e. goods coming back) are not plausible we decide for a simple AR(1) error process instead (no such effects seen). The stationary AR(1) error model has a coefficient of  $\phi = -0.61$ .

Next, based on the AR(1) error model, a smoothing parameter of  $\hat{\lambda} = 0.50$  is chosen via the UBR criterion. At this point it is worth mentioning that a negative first-order autocorrelation can cause substantial oversmoothing when not accounted for (see Schimek, 1992 p.315ff). In the extended semiparametric regression model we control for this with the benefit of not disguising potential features of interest.

Further exploration of the series suggests a system lag of  $b = 0$  (i.e. an immediate response) for the policy of year-end shipments. The highest possible intervention function order is two. Hence we fit a partial spline model under the assumption of  $b = 0$  and AR(1) errors.

The estimation results for the three  $\omega$  parameters of the second-order intervention function (i.e. the parametric part of the semiparametric model) are summarized in Table 1.

	$\hat{\gamma}$	$SE(\hat{\gamma})$	$z$ -value	significant
Intervention parameter $\omega_0$	106.49	19.65	5.42	yes
Intervention parameter $\omega_1$	-177.68	19.64	-9.04	yes
Intervention parameter $\omega_2$	59.12	19.70	3.00	yes

Table 1

UBR estimation results for the intervention part of the semiparametric model under the correctly specified AR(1) error term with  $\phi = -0.61$

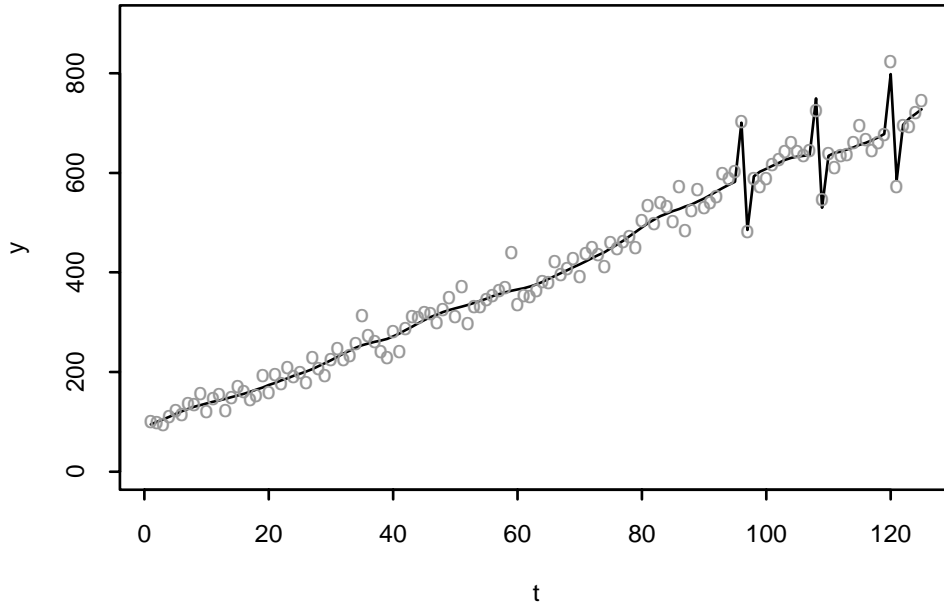


Fig. 2. Semiparametric fit for  $\phi = -0.61$  and  $\lambda = 0.5$ : function  $\mu$  and observations (denoted by circles)

We see that all the  $\omega$ 's are significant. Hence the second-order intervention function is adequate. Plugging in these estimates into the intervention function leads to the final effect features. In Figure 2 the mean function  $\hat{\mu}$  is plotted. It perfectly mimics the reactions to the management interventions (non-smooth feature part of the estimated function). The effect can be characterized in the following way: in each of the years 1983, 1984 and 1985 the artificial increase (positive  $\hat{\omega}_0$ ) in shipments is followed by a sharp decrease (negative  $\hat{\omega}_1$ ), finally returning to the pre-intervention level (positive  $\hat{\omega}_2$ ). From the plot one might conclude that also the long-term trend responds to the interventions with a slightly decreasing gradient, at least for 1983 and 1984.

In Figure 3 the ordinary residuals  $r$  resulting from the above fitted semiparametric model are plotted. The series is not different from white noise, however there are a few remaining peaks of limited size we have no intervention information for. The impacts of interest considered in Pangratz (1991) do not show up again in the residual series. Hence we have a perfect fit under the AR error process.

Let us fit the partial spline model for another time, but now under the simple yet inadequate white noise error assumption. We apply the second-order intervention function from above and the same  $\hat{\lambda} = 0.50$  for comparison (the UBR  $\lambda$ -value is larger, i.e. more smoothing required). The obtained results are

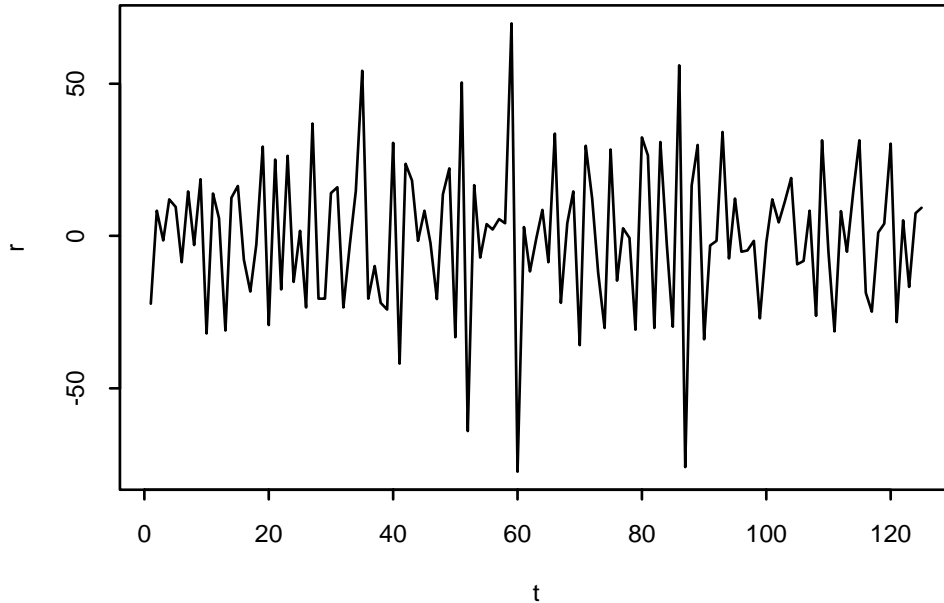


Fig. 3. Semiparametric fit for  $\phi = -0.61$  and  $\lambda = 0.5$ : ordinary residuals

summarized in Table 2.

	$\hat{\gamma}$	$SE(\hat{\gamma})$	$z$ -value	significant
Intervention parameter $\omega_0$	113.29	13.21	8.57	yes
Intervention parameter $\omega_1$	-106.83	13.21	-8.09	yes
Intervention parameter $\omega_2$	-4.18	13.25	-0.32	no

Table 2

UBR estimation results for the intervention part of the semiparametric model under the inadequate white noise error assumption

When comparing with Table 1 we immediately see that the  $\gamma$  estimates for the three intervention parameters have changed and the one for  $\omega_2$  is not significant any more. Why is that? Fluctuations caused by the AR error structure are now part of the trend signal and as a consequence smoothed out. The same happens to feature information which is not much distinct from the rest of the observations and this is certainly true for  $\omega_2$  marking the end of each intervention. With the UBR choice of the  $\lambda$ -value under the wrong white noise assumption we would have seen even more smoothing because of the negative sign of the AR parameter  $\phi$ , masking the intervention effect extra.

## 6 Conclusions

Semiparametric partial spline regression opens new possibilities for intervention effect modeling and feature detection. The here presented approach allows us to extend the scope of semiparametric regression based on an i.i.d. error assumption to time-dependent observations. It has been demonstrated that unbiased estimates can not solely be obtained for white noise errors but also for AR errors. Further a strategy for feature detection similar to time domain intervention analysis based on transfer (intervention) functions has been introduced and its use demonstrated for a well-known time series data set. In summary we can say that the semiparametric approach out-performs classical parametric time series intervention analysis as in our example.

## References

- Azzalini, A. and Bowman, A. W. (1993) On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc., B*, 55, 549-557.
- Box, G. E. P. and Tiao, G. C. (1975) Intervention analysis with applications to economic and environmental problems. *JASA*, 70, 70-79.
- Diggle, P. (1990) *Time series. A biostatistical introduction*. Clarendon Press, Oxford.
- Eubank, R. L. (1999, 2nd ed.). *Nonparametric regression and spline smoothing*. Dekker, New York.
- Eubank, R. L., Kambour, E. L., Kim, J. T., Klipple, K., Reese, C. S. and Schimek, M. (1998). Estimation in partially linear models. *CSDA*, 29, 27-34.
- Galbraith, R. F. and Galbraith, J. I. (1974). On the inverse of some patterned matrices arising in the theory of stationary time series. *J. Appl. Prob.*, 11, 63-71.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach*. Chapman & Hall, London.
- Herrmann, E. (2000). Variance estimation and bandwidth selection for kernel regression. In Schimek, M. G. (ed.) *Smoothing and regression. Approaches, computation and application*. John Wiley, New York, 71-107.
- Kohn, R., Schimek, M. G. and Smith, M. (2000) Spline and kernel smoothing for dependent data. In Schimek, M. G. (ed.) *Smoothing and regression. Approaches, computation and application*. John Wiley, New York, 135-158.
- Pankratz, A. (1991) *Forecasting with dynamic regression models*. John Wiley, New York.
- Pollock, D. S. G. (1979). *The algebra of econometrics*. John Wiley, Chichester.

- Schimek, M. G. (1988a). A roughness penalty approach for statistical graphics. In Edwards, D. and Raun, N. E. (ed.). Proceedings in Computational Statistics 1988. Physica, Heidelberg, 37-43.
- Schimek, M. G. (1988b). How to differentiate between impact and effect using BOX and TIAO intervention functions: some zero and first order cases. *EDV Med. Biol.*, 19, 49-57.
- Schimek, M. G. (1992) Serial correlation in spline smoothing: A simulation study. *Computational Statistics*, 7, 309-327.
- Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *JSPI*, 91, 525-540.
- Schimek, M. G. (2002) Unbiased partial spline fitting under autoregressive errors. In Härdle, W. and Rönz, B. (ed.) *COMPSTAT 2002. Proceedings in Computational Statistics*. Heidelberg: Physica, 605-610.
- Schimek, M. G. and Schmaranz, K. G. (1994) Dependent Error Regression Smoothing: A new method and PC program. *CSDA*, 17, 457-464.
- Speckman, P. (1988). Kernel smoothing in partial linear models. *JRSS, B*, 50, 413-436.
- Tiao, G. C. (2001a). Univariate autoregressive moving-average models. In Peña, D., Tiao, G. C. and Tsay, R. S. *A course in time series analysis*. John Wiley, New York, 53-85.
- Tiao, G. C. (2001b). Vector ARMA models. In Peña, D., Tiao, G. C. and Tsay, R. S. *A course in time series analysis*. John Wiley, New York, 365-407.
- Wahba, G., 1990. *Spline models for observational data*. SIAM, Philadelphia.