Valery V. Fedorov and Peter Hackl and Werner Müller

Forecasting with Optimized Moving Local Regression

Working Paper

Forecasting with Optimized Moving Local Regression

Valery V. Fedorov, Peter Hackl, Werner Müller

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 30
November 1992

http://statmath.wu-wien.ac.at/

Forecasting with Optimized Moving Local Regression

by

Valery V. Fedorov [1]

Peter Hackl [2]

Werner G. Müller [2]

Abstract:

This paper empirically demonstrates the relative merits of the optimal choice of the weight function in a moving local regression as suggested by Fedorov et al., (1993) over traditional weight functions which ignore the form of the local model. The discussion is based on a task that is imbedded into the smoothing methodology, namely the forecasting of business time series data with the help of a one-sided moving local regression model.

1 Introduction

In the moving local regression approach parameters are estimated by weighting down the observations so that the weights reflect the "distance" of the observations from the forecast point. This gives the flexibility to parametrize the model depending on local conditions. Given that the true model is locally approximated and a certain form of the approximation error (such as the remainder term of a local series expansion) is suspected to be relevant at times, it is possible to choose the weights such that optimal

[1]Department of Applied Statistics, University of Minnesota, St.Paul, Minnesota, USA.

[2]Department of Statistics, University of Economics and Business Administration, Vienna, Austria.

forecasting power is achieved. Such models are particularly useful for describing or forecasting time series that are generated by time-varying processes.

In the literature several suggestions for the choice of the weight function in moving local regression models can be found [e.g. McLain, (1971), Cleveland, (1979)]. A common feature of these weighting schemes is that they are chosen taking no regard of the model specification. The approach presented here aims at maximizing the forecast accuracy and takes a possible model misspecification into account.

In Section 2 the model and the estimation method are introduced. Section 3 presents three weight functions that are to be compared in Section 4. This comparison is based on a time series from bank business that is a typical candidate for nonparametric analysis.

2 The Method

Let $\{x_1, \ldots, x_T\}$ be a given set of supporting points, i.e., points where observations $\{y_1, \ldots, y_T\}$ are available, and let $d_t = x_{T+1} - x_t, \quad t = 1, \ldots, T$, be the "distances" from the point of interest x_{T+1}. Then

$$y_t = \theta^T f(d_t) + \delta\varphi(d_t) + \varepsilon_t\,, \quad t = 1, \ldots, T \tag{2.1}$$

will be called a one-sided regression model. It consists of a main term $\theta^T f(d_t)$ describing the model, that locally approximates the true model, a "nuisance term" $\delta\varphi(d_t)$ describing the approximation error, and an error term ε_t following the usual assumptions $E[\varepsilon_t] = 0$ and $E[\varepsilon_t \varepsilon_{t'}] = \sigma_\varepsilon^2$. The number of components of the parameter vector θ is determined by the structure of the approximating model. For φ, an appropriate function has to be specified; the "nuisance parameter" δ is unknown.

Setting $t = T + 1$ in (2.1) allows us to calculate a forecast for y_{T+1}. If we make the reasonable

Assumption: $f_1(d) \equiv 1$, $f_j(d) \to 0$ for $d \to 0$ and $j \geq 2$, and all components of $\varphi(d)$ also vanish [usually faster than $f_j(d)$] for $d \to 0$, the forecast

$$\hat{y}_{T+1} = \hat{\theta}_1 \tag{2.2}$$

is the first component of the (weighted least squares) estimator

$$\hat{\theta} = M^{-1}Y,$$

with $M = \sum_{t=1}^{T} \lambda(d_t) f(d_t) f^T(d_t)$ and $Y = \sum_{t=1}^{T} \lambda(d_t) f(d_t) y_t$.

The mean squared error matrix of the estimator $\hat{\theta}$ is

$$R = \mathrm{E}\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\} = M^{-1} M_{12} \delta \delta^T M_{12} M^{-1} + \sigma^2 M^{-1} \mathcal{M} M^{-1}, \tag{2.3}$$

where $M_{12} = \sum_{t=1}^{T} \lambda(d_t) f(d_t) \varphi^T(d_t)$ and $\mathcal{M} = \sum_{i=t}^{T} \lambda^2(d_t) f(d_t) f^T(d_t)$. The choice of the weight function $\lambda(d_t)$ which reflects the reliability of the local approximation is discussed in the subsequent section.

$\hat{\theta}$ is generally biased:

$$\mathrm{E}\{\hat{\theta}\} = \theta + M^{-1} M_{12} \delta, \tag{2.4}$$

A detailed treatment of the estimation properties is given in the nonparametric regression literature such as Cleveland & Devlin, (1988) or Buja et al., (1989).

Models of type (2.1) used in local fitting are particularly helpful for time series whose characteristics change over time. For cases where higher order terms reflected by $\delta\varphi(d_t)$ are suspected to have some effect, Fedorov et al., (1993) suggest choosing the weight function $\lambda(d_t)$ so that a suitably chosen scalar function of the m.s.e. matrix R is minimized. Adapted to the forecasting problem, this means direct minimization of the mean square error of the forecast $\hat{\theta}_1$. It is performed under the restriction $\lambda(d_t) \geq 0$ for all d_t and $\sum_t \lambda(d_t) = 1$. The weight function depends on the nuisance parameter δ. Therefore, in its derivation in a particular situation, δ has to be estimated in a preliminary step. The weight function is entirely determined by the model specification and the data.

In a forecasting context this method will be sequentially applied, i.e., forecasts are calculated for time points $T+1, T+2, \ldots$, each estimate being based on the currently available amount of information. This implies that the weight function is derived in each forecast point anew. This generalization of the estimation process is straightforward and so we do not record the corresponding formulae.

Example 1 *As an illustration, the optimal weight function is derived for the model* $y_t = \theta + \delta d_t^2 + \varepsilon_t$, *i.e. the moving average specification with a quadratic "nuisance" term. We consider a collection of* $2n+1$ *equally spaced points in the interval* $[-1,1]$ *and derive the value of the weight function for the central point. For the average quadratic distance* $\bar{d}^2$ *and its variance we obtain* $\frac{n+1}{3n}$ *and* $\frac{(n+1)(4n^2+4n-3)}{45n^3}$ *respectively. The optimal weights are:*

$$\lambda(d_i) = \frac{1}{n} - \frac{\frac{\delta^2}{\sigma_\varepsilon^2}[d_i^2 - \bar{d}^2]\bar{d}^2}{1 + n\frac{\delta^2}{\sigma_\varepsilon^2}\operatorname{var}(d_i^2)},$$

cf Fedorov et al., (1993). Note that they are linear in d_i^2.

The form of the weight function and the number of supporting observations that have nonzero weights (the "window width"), and consequently the degree of smoothing crucially affects the estimate $\hat{\theta}$. A weight function that is too concentrated around the forecast point results in undue variation as it allows reaction to local time series characteristics; a too flat weight function smoothes out local tendencies.

The use of moving averages, i.e., application of the model from Example 1, is suitable for the description of the long wave changes in a time series but smoothes away short term effects. Using a linear moving regression that includes the term θd allows us to identify changes which occur within the period covered by the weight function.

3 Comparison of weight functions

When applying moving regression to a set of time series that differ considerably with respect to its characteristics, the smoothing interval has to be long enough to cover the longest period of changes in these characteristics.

In the literature several recommendations for the choice of the weight function are given. Out of practical considerations McLain, (1971) suggested

$$\lambda(d) = \exp\frac{-\|d\|^2/{d_n}^2}{\|d\|^2+\rho}, \tag{3.1}$$

where d_n is the average distance between neighbouring data points and the constant $\rho = 10^{d_n} - 1$ prevents numerical accuracy problems. A computationally simpler function, the so-called tricube,

$$\lambda(d) = \begin{cases} [1 - (\|d\|/d_q)^3]^3 & 0 \leq \|d\|/d_q \leq 1 \\ 0 & \text{else} \end{cases} \tag{3.2}$$

with d_q being the distance of the $q.n$ nearest point to x, is used by Cleveland, (1979). This function smoothly decreases from 1 to 0 with increasing $\|d\|$. The weight functions (3.1) and (3.2) have in common that they are chosen without regard of the local model, and the possibility of a nuisance term is neglected.

Following the recommendations by Fedorov et al., (1993) the weight function can be chosen such that the mean squared error matrix R [see (2.3)] is minimized in a certain sense. In model (4.1) this approach should be clearly superior to techniques that are based on weight functions such as (3.1) or (3.2). A demonstration of the relative capabilities in applications will be given in the following section by means of an example in which a forecast of bank account data is required.

Example 2 *Let $y_1, \ldots, y_T$ be observations from locations $-1 \leq x_1 < \ldots < 0 < \ldots < x_T \leq 1$ symmetrically arranged around 0. The aim is to get a prediction $\hat{y}$ at the forecast point x. If a linear model with a quadratic nuisance term (cf. next section) is assumed, the optimal weights λ^* for $T = 10$ and $x = 0$, $x = 5$ and $x = 1$ are shown in Figure 1.*

4 Comparison of the Weight Functions: A Case Study

For comparing various weight functions the model

$$y_t = \theta_1 + \theta_2 d_t + \delta d_t^2 + \varepsilon_t\,, \quad t = 1, \ldots, T \tag{4.1}$$

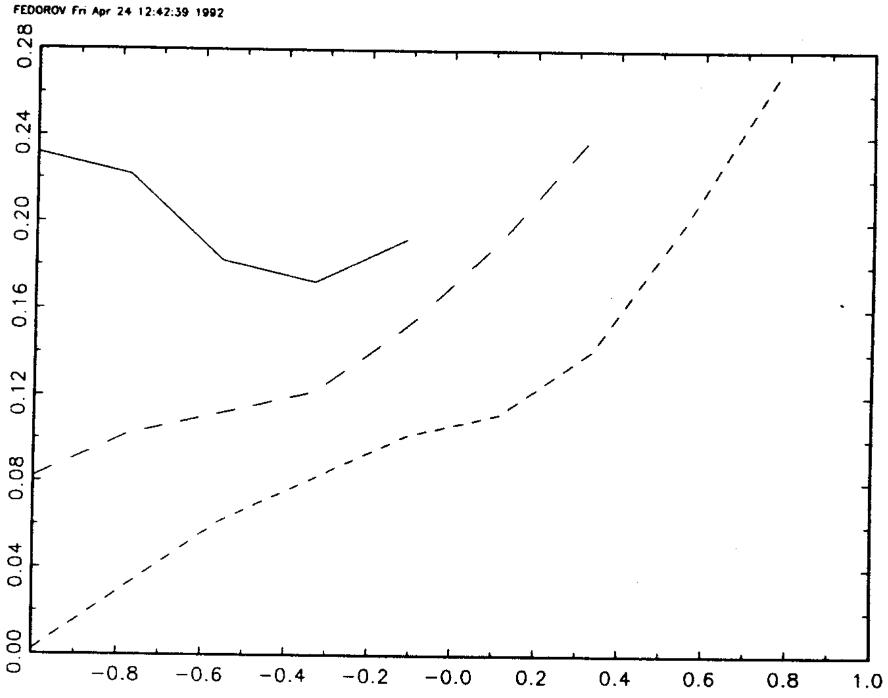

Figure 1: Weight functions for optimal forecasts at $x = 0$ (solid), $x = 0.5$ (dashed) and $x = 1$ (short dashed)

was chosen. It implies that linearity is considered as a suitable description of the local behaviour, and that a possible effect of a quadratic term is allowed to be corrected via the weights of the local regression.

The comparison is based on a time series from the bank business, which is given in Table 1. The data analyzed in the example are the fractions, to which the creditline of a typical small Austrian enterprise is used, observed weekly over a period of 14 months, a 100% exhausted creditline gives a value of 1 in the corresponding series. The bank utilizes these fractions to decide whether the credit should be prolonged or not.

As a first step moving averages were constructed for all possible window lengths (from 5 to 290 days) and all past time points. They can be interpreted as the simplest one step ahead forecasts. The forecasts with the lowest average squared forecast error, corresponding to a window length of 65 days, were used as a reference point for the comparison, as well as for the preestimation of the residual variance $\hat{\sigma}_\epsilon^2$, which gave 0.0172.

Next, minimal average squared forecast errors were found for weight functions

1	1.092860	31	1.131808	61	1.174947	91	1.173454	121	1.210698
2	1.113694	32	1.131808	62	1.174947	92	1.173454	122	1.199281
3	1.092860	33	1.131808	63	1.205920	93	1.173454	123	1.174031
4	1.092860	34	1.131808	64	1.205920	94	1.190121	124	1.325726
5	1.092860	35	1.131808	65	1.205920	95	1.200121	125	1.212815
6	1.092860	36	1.152641	66	1.241076	96	1.200121	126	1.212815
7	0.821910	37	1.055290	67	1.207743	97	1.172621	127	1.214639
8	1.251459	38	1.177804	68	1.207743	98	1.172621	128	1.214639
9	0.914887	39	1.136138	69	1.207743	99	1.171991	129	1.131306
10	0.948220	40	1.136138	70	1.207410	100	1.171991	130	1.131468
11	0.914887	41	1.136138	71	1.207410	101	1.171991	131	1.131468
12	0.914887	42	1.136138	72	1.207410	102	1.297045	132	1.131468
13	0.914887	43	1.136138	73	1.207410	103	1.172045	133	1.131468
14	0.914887	44	1.136138	74	1.207410	104	1.072271	134	1.131522
15	0.914887	45	1.261138	75	1.207410	105	1.172153	135	1.131522
16	0.914887	46	1.136138	76	1.234076	106	1.173977	136	0.881794
17	0.935720	47	1.146138	77	1.234076	107	1.173977	137	1.131522
18	0.935720	48	1.146138	78	1.234076	108	1.173977	138	1.131522
19	1.131808	49	1.146138	79	1.171576	109	1.173977	139	1.131522
20	1.131808	50	1.146192	80	1.171576	110	1.173977	140	1.131522
21	1.131808	51	1.149839	81	1.197622	111	1.173977	141	1.131522
22	1.131808	52	1.149839	82	1.171631	112	1.173977	142	1.131522
23	1.256808	53	1.149839	83	1.171631	113	1.270315	143	1.131522
24	1.131808	54	1.149839	84	1.171631	114	1.174031	144	1.131522
25	1.131808	55	1.174839	85	1.171631	115	1.174031	145	1.168189
26	1.131808	56	1.149839	86	1.173454	116	1.174031	146	1.168189
27	1.131808	57	1.174839	87	1.173454	117	1.174031	147	1.168189
28	1.131808	58	1.312339	88	1.200121	118	1.174031	148	1.131522
29	1.131808	59	1.174947	89	1.173454	119	1.260698	149	1.133346
30	1.131808	60	1.174947	90	1.173454	120	1.210698	150	0.885651

Table 1: Analyzed data-set t and y - part I

151	1.133346	181	0.709575	211	0.789596	241	0.659909	271	0.435733
152	1.133346	182	0.801418	212	0.756419	242	0.659909	272	0.414900
153	1.133346	183	0.859931	213	0.847993	243	0.659909	273	0.414900
154	0.805013	184	0.859931	214	0.768086	244	0.659909	274	0.417909
155	0.805013	185	0.826418	215	0.756419	245	0.659909	275	0.419733
156	0.805013	186	0.801418	216	0.824753	246	0.659909	276	0.419733
157	0.805013	187	0.801418	217	0.824753	247	0.659909	277	0.419733
158	0.843346	188	0.837006	218	0.758086	248	0.660623	278	0.419733
159	0.805013	189	0.921416	219	0.758086	249	0.661669	279	0.466630
160	0.841679	190	0.834855	220	0.658086	250	0.661669	280	0.468936
161	0.841679	191	0.836679	221	0.685583	251	0.686790	281	0.481679
162	0.841679	192	0.836679	222	0.699753	252	0.686790	282	0.481679
163	0.805013	193	0.837929	223	0.711774	253	0.688613	283	0.481679
164	0.905013	194	0.837929	224	0.658086	254	0.688613	284	0.511262
165	0.805013	195	0.837929	225	0.658086	255	0.688613	285	0.444596
166	0.805013	196	0.837929	226	0.658086	256	0.688613	286	0.444596
167	0.805013	197	0.837929	227	0.658086	257	0.688613	287	0.444596
168	0.805013	198	0.895419	228	0.658086	258	0.563613	288	0.444596
169	0.805013	199	0.883762	229	0.658086	259	0.563613	289	0.450180
170	0.805013	200	0.837929	230	0.658086	260	-0.06570	290	0.450180
171	0.806836	201	0.754596	231	0.658086	261	0.416113	291	0.450180
172	0.806836	202	0.754596	232	0.741419	262	0.415400		
173	0.806836	203	0.754596	233	0.658086	263	0.415400		
174	0.806836	204	0.754596	234	0.659909	264	0.415400		
175	0.806836	205	0.775429	235	0.659909	265	0.415400		
176	0.870484	206	0.754596	236	0.659909	266	0.415400		
177	0.806836	207	0.754596	237	0.659909	267	0.415400		
178	0.862669	208	0.754596	238	0.659909	268	0.415983		
179	0.834751	209	0.754596	239	0.659909	269	0.445566		
180	0.834751	210	0.754596	240	0.659909	270	0.445566		

Table 2: Analyzed data-set t and y - part II

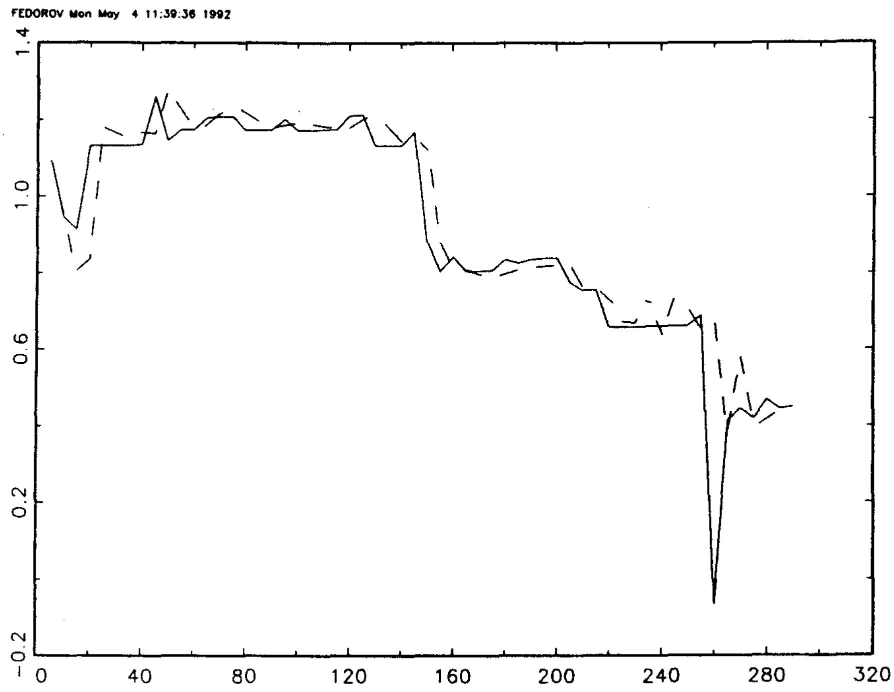

Figure 2: dashed - bank account data, solid - one step ahead forecasts (daily scale)

(3.1) and (3.2). In applying the numerical algorithm for weight optimization from Fedorov et al., (1993), for simplification of the calculation process we firstly assumed that δ is constant over time. For comparison of the results from the three weighting regimes one has to define a common measure of smoothness. Simple to calculate is the sum of squared second differences as an estimate of the local curvature, which is commonly used for penalizing in spline regression.

The data and optimal forecasts are displayed in Figure 2. Figure 3 presents the average squared forecast error for the alternative weighting procedures.

The proposed method with the "optimal" weight function is clearly superior to alternative weighting schemes. The average squared forecast error over all time points lies uniformly below the respective errors for the forecasts using weight functions (3.1) or (3.2) for comparable smoothness levels greater than 0.1. Moreover, its minimum value is 0.0151, which is considerably below (around 6%) the minimum values of 0.0160 and 0.0165 for (3.1) [with $\delta\lambda(d)$] and (3.2), respectively.

Alternatively, to avoid the assumption of constancy in δ,we applied a two step

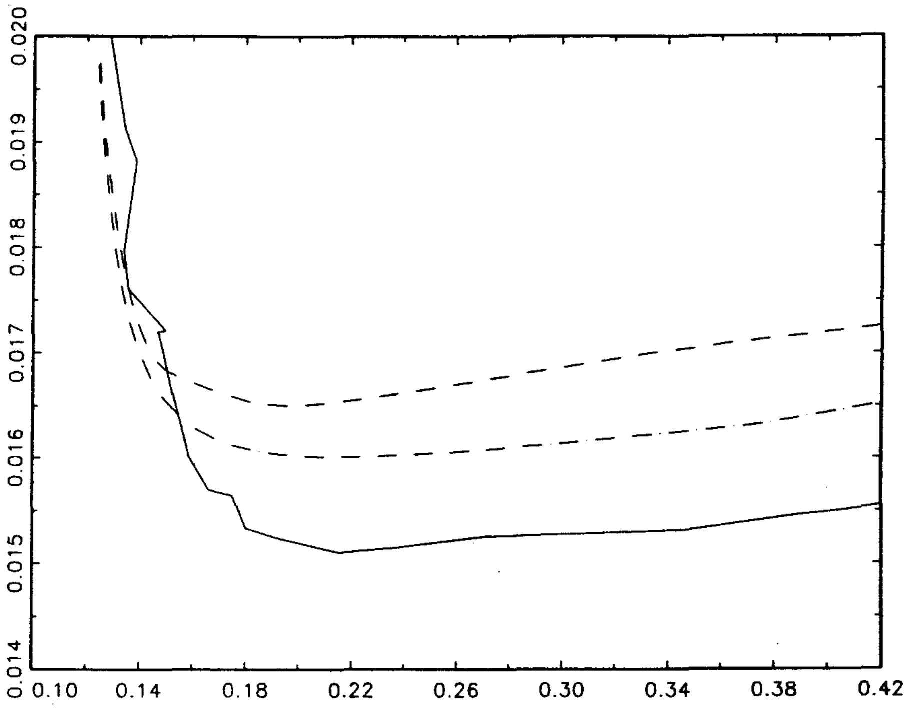

Figure 3: Average forecast error vs. smoothing level: dashed - Clevelands, dot-dashed - McLains, solid - Fedorovs weight function

procedure. In the first step a moving quadratic regression was performed to preestimate δ for each forecast point. Using those estimates in the weight optimizing procedure result in an average squared forecast error of 0.01440, another improvement of around 5%.

5 Conclusions

The comparison of forecast errors obtained by the optimized moving local regression approach and two traditional weighting schemes indicates a clear superiority of the former technique. This superiority strongly supports the choice of this technique in this and similar applications. Of course it has to be noted that for cases where the assumed model does not hold the different weighting schemes compete on the same level and one might then perform accidentally better than another.

In addition to the forecasts a lot of valuable information can be gained from the estimators. A continuously performed discriminant analysis for instance allows various enterprises to be distinguished by their economic status. A related example utilitizing such an approach is presented by Müller, (1992).

References

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models with discussion. *The Annals of Statistics*, 17(2):453–555, 1989.

W.S. Cleveland and S. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

V.V. Fedorov, P.Hackl, and W. Müller. Moving local regression: the weight function. *Journal of Nonparametric Statistics*, in preparation, 1993.

D.H. McLain. Drawing contours from arbitrary data points. *The Computer Journal*, 17(4):318–324, 1971.

W.G. Müller. The evaluation of bank accounts using optimized moving local regressions. In L.Fahrmeir, B.Francis, R.Gilchrist, and G.Tutz, editors, *Advances in GLIM and Statistical Modelling*, Springer-Verlag, 1992.