Moving Local Regression: The Weight Function

Fedorov, Valery V.; Hackl, Peter; Müller, Werner

Published: 01/01/1992

Document Version
Publisher's PDF, also known as Version of record

Moving Local Regression: The Weight Function

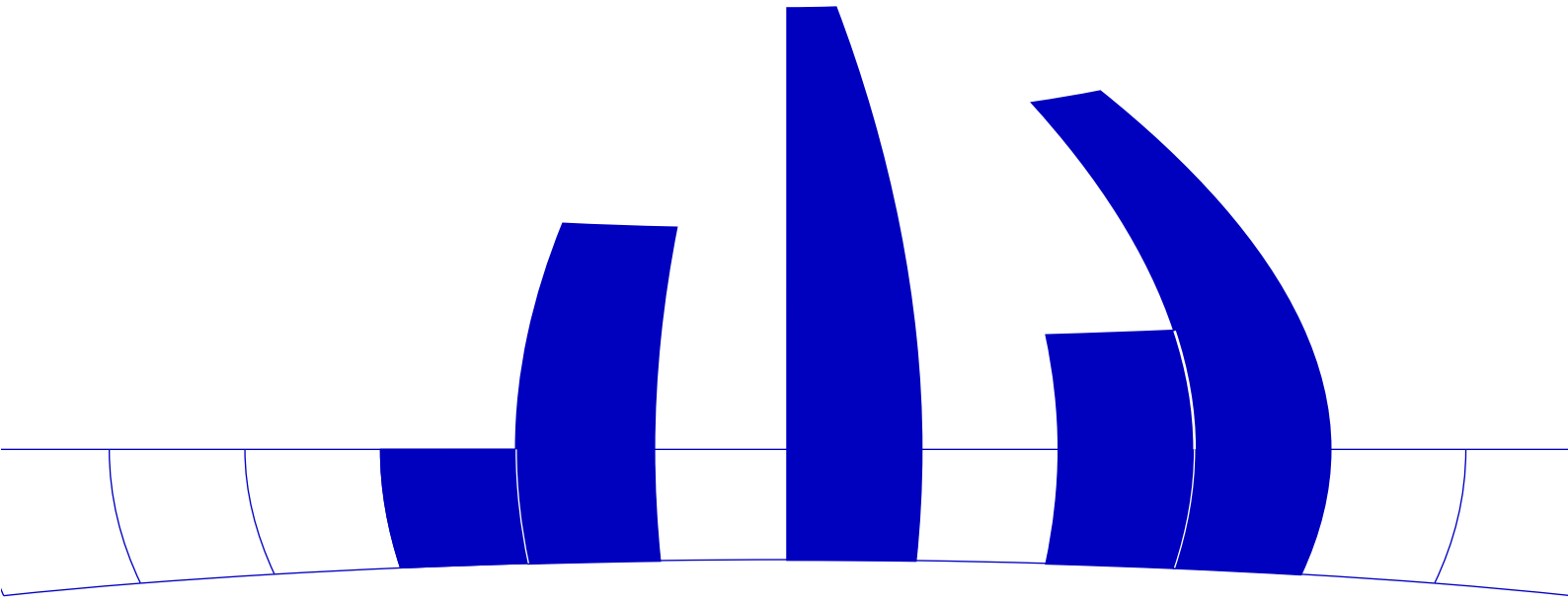
Valery V. Fedorov, Peter Hackl, Werner Müller

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 25
January 1992

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



Moving Local Regression: The Weight Function

by

Valery V. Fedorov [1]

Peter Hackl [2]

Werner Müller [3]

Abstract: Moving local regression is a nonparametric technique for smoothing, interpolating and forecasting by means of locally fitted regression models. The paper explores the "optimal" structure of the weight function, taking into account the location of supporting points and the suspected behaviour of the remainder term, and surveys results on choice of weight functions in traditional moving local regression approaches.

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

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

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

1 Introduction

Nonparametric regression methods can effectively be used for various purposes, either to build up ideas about a possible parametric model during the explorative phase of a study or due to its rich properties itself they could be applied for inference about a (presumably nonlinear) process. Pelto *et al.* (1968) probably were the first who utilized the concept of local regression schemes to interpolate a given surface with a sparse amount of irregulary spaced data points. They called this variant of a nonparametric regression algorithm "moving weighted least squares" estimation and derived some important properties of it. Cleveland (1979) used the moving regression algorithm for extracting information from fuzzy scatterplots and for smoothing of strongly scattering time-series. He called the method "locally weighted (or *loess*-) regression", worked out computational algorithms, and considered its statistical properties.

Moving local regression analysis is an intuitively simple statistical method for smoothing, interpolating, and forecasting, similar to kernel estimators or related techniques. The main idea is to calculate estimates by weighting down the observations so that the weights reflect the "distance" of the observations from the point of interest. This gives the flexibility to parametrize the model depending on local conditions. Of course, the choice of the weights is crucial for the method. In this paper we suggest a new method of choosing the weights or more generally the weight function: According to our method an *optimal weight function* is chosen such that an appropriate scalar function of the covariance matrix of the estimated coefficients is minimized.

The paper is organized as follows. In Section 2 we introduce the necessary notation and state the problem. Section 3 surveys the desired properties of weight functions in moving local regression. Optimal weight functions are discussed in Section 4. Special emphasis is given to time series applications where one-sided weight functions are of interest for smoothing and forecasting. After discussing optimization of the weight function the case of one-sided weight functions is treated in detail in Section 5. We illustrate our results in a final Section.

2 Statement of the Problem

The basic idea of the moving local regression approach is as follows. Over a sufficiently small region of the predictor space a considerably smooth function can be approximated by a "simpler", such as a polynomial, function. This new function should be flexible enough to approximate the response function of the regression model on a given set of points of interest, i.e., points where an interpolated or smoothed or forecasted value is desired. We assume that the points of interest are surrounded by or mixed with points where observations are available. For interpolation, smoothing or forecasting a weighted least squares method on the approximate response function is performed with a weight function that decrease for increasing distance between point of interest and observation.

Let X_p be the set of points of interest, and $X_n = \{x_1, \ldots, x_n\}$ the "supporting set", i.e., the set of points where observations are available. In the following, we formulate the model just for a single point $x \in X_p$; the more general formulation where more than one point of interest is considered is a straightforward extension. We shall consider the class of models that are obtained by expanding the model for the response y_i that is obtained at $x_i \in X_n$ in a Taylor series at point x

$$y_i = \theta^T f(d_i) + \delta^T \varphi(d_i) + \varepsilon_i \tag{2.1}$$

for $i = 1, \ldots, n$. Here, the m-component function $f(z) = [f_1(z), \ldots, f_m(z)]^T$ has the argument $d_i = x_i - x$, the distance between the point of interest x and the support point x_i, the m-vector θ contains unknown parameters, $\delta^T \varphi(d_i)$ describes the remainder term, and ε_i is the error of the observation taken at x_i. For ε_i we assume $\mathrm{E}\{\varepsilon_i\} \equiv 0$, $\mathrm{E}\{\varepsilon_i^2\} \equiv \sigma^2$, and $\mathrm{E}\{\varepsilon_i \varepsilon_{i'}\} \equiv 0$ for $i = 1, \ldots, n$ and $i \neq i'$. The functions f are known.

In general, the remainder term contains several parameters which can be included in the analysis of the quality of approximation. Here, for sake of simplicity we assume that the components of δ are small and bounded by means of "common sense" considerations.

In the following 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 $d \to 0$.

Under this assumption, the first component θ_1 of vector θ coincides with the value of the response function at the point of interest x.

Example 1 *Let the model*

$$y_i = \theta + \delta\varphi(d_i) + \varepsilon_i$$

describe the response y_i obtained at the support point x_i as a function of the distance d_i. Under the above-stated assumptions the response at x is estimated to be $\hat{\theta}$.

In the following we will also use the abbreviated notation f_i and φ_i for $f(d_i)$ and $\varphi(d_i)$, respectively.

The problem of interest is to estimate θ_1 or - more generally - the parameter vector θ. Given that the function $\varphi(d)$ is unknown (or neglected in the analysis) we can hardly propose a better approach for estimating θ than the weighted least squares method. For any point $x \in X_p$ we define

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \lambda_i (y_i - \theta^T f_i)^2 , \tag{2.2}$$

where the use of weights $\lambda_i = \lambda(d_i)$ gives us the opportunity to make the contribution of each of the terms in the sum depending on the distance $\|d_i\|$. We assume that the weights λ_i are nonnegative. The estimator for θ is

$$\hat{\theta} = M^{-1}Y \tag{2.3}$$

with $M = \sum_{i=1}^{n} \lambda_i f_i f_i^T$ and $Y = \sum_{i=1}^{n} \lambda_i f_i y_i$. The estimator is biased:

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

and for the mean squared error matrix R we obtain

$$R = \mathrm{E}\{(\hat{\theta} - \theta_t)(\hat{\theta} - \theta_t)^T\} = M^{-1}M_{12}\delta_t\delta_t^T M_{12}M^{-1} + \sigma^2 M^{-1}\mathcal{M}M^{-1} ; \tag{2.5}$$

the subscript "t" indicates true values (i.e., values that provide the best appoximation of the response function with no observation errors), $M_{12} = \sum_{i=1}^{n} \lambda_i f_i \varphi_i^T$, and $\mathcal{M} = \sum_{i=1}^{n} \lambda_i^2 f_i f_i^T$.

3 The Weight Function

It is obvious that the estimator (2.3) is determined by the form of the weight function $\lambda(d)$. Two properties are essential to allow for a sensible interpretation:

(a) $\lambda(d)$ is a nonincreasing function of the norm $\|d\|$ of d $[\lambda(0) = \max \lambda]$.

(b) $\lim_{\|d\| \to \infty} \lambda(d) = 0$.

Choosing the weight function within this framework gives us a wide range of possibilities to adjust the method to specific problems.

If the given approach is used for interpolation of a response function, an additional condition on the weight function is needed to ensure that the surface fits exactly to the data. Pelto et al., (1968) provide the proof that

(c) $\lim_{\|d\| \to 0} \lambda(d) = \infty, \qquad$ s.t. $\int_{-\infty}^{\infty} \lambda(d) = 1$

ensures interpolation, if the support points are among the points of interest. Through this condition the smoothness of the estimated response function can be controlled, as Cleveland, (1979), Cleveland et al., (1988), and Cleveland & Devlin, (1988) do in their applications. Buja et al., (1989) mention the problem of choosing smoothness parameters as a function of the data which destroys the linearity of the estimator $\hat{\theta}_1$. This is generally avoided by the assumption that the parameters are fixed a priori.

Another important notion is the concept of locality of the method. As Ripley, (1981) indicates it is not guaranteed that $\hat{\theta}_1$ is a weighted average of mainly local values if the region of interest is expanded unless

(d_1) $\lim_{\|d\| \to \infty} \|d\|^p \lambda(d) = 0,$

where p can, for instance, be defined as the order of the Taylor's expansion of the response function. The proof is given by Ripley, (1981) and by Müller, (1991). This condition assures independence of the results from the choice of the region of interest. Cleveland, (1979), Cleveland et al., (1988), and Cleveland & Devlin, (1988) avoid such difficulties by assigning weight zero to a certain percentage q of data points due to their remoteness, i.e.,

(d_2) $\lambda(d) = 0$ if $\|d\| > d_q$,

where d_q is the distance of the $q.n$ nearest point to x. Thus only points in the neighbourhood of x contribute to the estimate. This approach provides another tool for handling locality problems that, however, can cause computational problems in multidimensional cases. This trimming also reduces the undesirable biasing effect of "influential" data points. On the other hand, the non-continuous form of the weight function may cause a loss of smoothness of the estimated response function as will be shown below.

The weight function λ should be chosen so that a considerably smooth and as many times as possible differentiable estimated response function $\hat{\theta}_1$ is guaranteed. Silvey, (1980) states that, in general, response functions generated by obeying the conditions (a), (b), (c), and (d_1) are as many times differentiable as $1/\lambda(\|d\|)$. This can be checked by getting the derivatives $\partial\hat{\theta}_1/\partial x$ of (2.3).

An example of such a weight function is

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

suggested out of practical considerations by McLain, (1971) where d_n is the average distance between neighbouring data points and the constant $\delta = 10^{d_n} - 1$ prevents from arithmetic overflow. In the example of Section 5 the normalized function $\delta\lambda(d)$ will be used.

A computationally simpler function that fulfills (a), (b), and (d_2), the so-called tricube

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

is used by Cleveland, (1979), Cleveland et al., (1988), and Cleveland & Devlin, (1988). This function smoothly decreases from 1 to 0 in the interval $[0, d_q]$ as can be seen from *Figure 1.*

insert Figure 1

As shown by Müller, (1987), the asymptotic equivalence between certain moving

local regression techniques and kernel smoothers for the univariate case implies the restriction

$$(d_3)\ \lambda(d) = 0 \text{ if } \|d\| > d_t,$$

where d_t is a real constant; this condition is equivalent to (d_2) for equally-spaced data points. Among the weight functions that fulfill (d_3), the symmetric polynomial weight function

$$\lambda_i^* = \sum_{k=0,\text{even}}^{2\mu} \frac{(-1)^{k/2}(k+2)(2\mu+2)!}{(1+k/2)!(\mu-k/2)!(\mu+1)!2^{2\mu+3}} x_i^k 1_{[-1,1]}$$

not only minimizes the asymptotic variance of the estimator but also the asymptotic variances of the μ-th derivative [see Müller, (1987)].

4 Optimal Weight Function

For a given set X_p of points of interest and given observations, the variance or, more general, the mean squared error matrix R of $\hat{\theta}$ is a function $R(\lambda)$ of the weights $\lambda_1, \ldots, \lambda_n$. Hence, improving the quality of the estimated response function $\hat{\theta}$, i.e., reducing $R(\lambda)$, implies an appropriate choice of the weights $\lambda_1, \ldots, \lambda_n$, or, more general, of the weight function $\lambda(d)$.

Similarly to optimization in optimal experimental design theory we can minimize only a scalar function of the dispersion matrix R. Consequently, choosing optimal weights is equivalent to the optimization problem

$$\lambda^* = \arg\min_{\lambda} \Psi[R(\lambda)]\,. \tag{4.1}$$

where Ψ can be any convex objective function.

In this section we will discuss the choice of an optimal weight function for (a) the case where $\theta^T f(d_i)$ in (2.1) degenerates to a constant θ and for (b) the general case (2.1). The former case is known as moving average model.

4.1 Moving Average

The model

$$y_i = \theta + \delta\varphi(d_i) + \varepsilon_i\,, \tag{4.2}$$

already shortly discussed in Section 2, is a frequently applied version of (2.1). To calculate an estimator $\hat{\theta}$ of the scalar parameter θ we need the quantities $M = \sum_{i=1}^{n} \lambda_i$, $\mathcal{M} = \sum_{i=1}^{n} \lambda_i^2$, and $M_{12} = \sum_{i=1}^{n} \lambda_i\varphi_i$. The dispersion matrix $R(\lambda)$ degenerates to the scalar function

$$R(\lambda) = [\delta^2(\lambda^T\varphi)^2 + \sigma^2\lambda^T\lambda](l^T\lambda)^{-2}\,, \tag{4.3}$$

where we make use of the vectors $\lambda = (\lambda_1, \ldots, \lambda_n)^T$, $\varphi = (\varphi_1, \ldots, \varphi_n)^T$, and $l = (1, \ldots, 1)^T$. Multiplication of λ by a real constant does change neither $\hat{\theta}$ nor $R(\lambda)$. Therefore, we assume in the following that $\sum_{i=1}^{n} \lambda_i = 1$.

Our aim is to find the optimal weight function

$$\lambda^* = \arg\min_{\lambda} R(\lambda)\,.$$

This problem can be solved on the basis of

Assertion 1 *The necessary and sufficient condition for λ^* to be optimal is the equality*

$$\min_i(\Omega\lambda^*)_i = \sigma^{-2}R(\lambda^*) \tag{4.4}$$

where $\Omega = (I + \gamma^2\varphi\varphi^T)$ and $\gamma^2 = \delta^2/\sigma^2$.

Proof:

The function $R(\lambda)$ and the set λ are convex. Therefore (see the results from convex experimental design theory), the necessary and sufficient condition for the optimality of $R(\lambda)$ is that the directional derivative at point λ^* fulfills

$$\min_{\lambda} \lim_{\alpha\to 0} \frac{\partial}{\partial\alpha} R[(1-\alpha)\lambda^* + \alpha\lambda] \geq 0\,. \tag{4.5}$$

From (4.3) follows that

$$\lim_{\alpha\to 0} \sigma^{-2} \frac{\partial}{\partial\alpha} R[(1-\alpha)\lambda^* + \alpha\lambda] = \lambda^T\Omega\lambda^* - \sigma^{-2}R(\lambda^*)\,.$$

The minimum

$$\min_{\lambda} \lambda^T \Omega \lambda^* = \min_{\lambda} \sum_{i,j=1}^{n} \lambda_i \Omega_{ij} \lambda_j^* = \min_{i} \sum_{j=1}^{n} \Omega_{ij} \lambda_j^*$$

is reached if we put all weight ($\sum_{i=1}^{n} \lambda_i = 1$) to the minimal component. It follows that (4.5) is equivalent to the inequality

$$\min_{i} (\Omega \lambda^*)_i \geq \sigma^{-2} R(\lambda^*) . \tag{4.6}$$

Taking into account that $\lambda^{*T} \Omega \lambda^* = \sigma^{-2} R(\lambda^*)$ and that $\lambda^{*T} \Omega \lambda^* \geq \min_i (\Omega \lambda^*)_i$, we find that (4.6) implies (4.4). □

Assertion 1 provides some recommendations on the structure of the optimal weight function. Moreover, similarly to the well-developed experimental design theory we can use [see Fedorov (1972), Silvey (1980)] some simple algorithms for the construction of λ^* given that matrix Ω (or φ) is known. Unfortunately, in practical situations information on φ, σ^2, and δ^2 is rather uncertain. Usually, it is accumulated in experiments or it stems from intuition.

Remark In some situations the form

$$\min_{i} [\lambda_i^* + \gamma^2 \varphi_i M_{12}(\lambda^*)] = \sigma^{-2} R(\lambda^*)$$

of (4.4) is more convenient.

To construct an optimal weight function λ^* for model (4.2) we first look for a candidate for λ^* on the basis of common sense ideas and subsequently prove its optimality. Assume that a weight function λ^* satisfies

$$\Omega \lambda^* = \sigma^{-2} R(\lambda^*) l . \tag{4.7}$$

As λ^* fulfills (4.4) it is optimal. Equality (4.7) holds if

$$\lambda^* = \sigma^{-2} R(\lambda^*) \Omega^{-1} l \text{ and } \lambda^* \geq 0 .$$

From matrix algebra we know that

$$(A + BCB^T)^{-1} = A^{-1} - A^{-1} B (C^{-1} + B^T A^{-1} B)^{-1} B^T A^{-1} .$$

Therefore,

$$(I + \gamma^2 \varphi\varphi^T)^{-1} = I - \frac{\gamma^2 \varphi\varphi^T}{1 + \gamma^2 \varphi^T \varphi}$$

or

$$\lambda^* = \sigma^{-2} R(\lambda^*) \left[l - \frac{\gamma^2 \varphi(\varphi^T l)}{1 + \gamma^2 \varphi^T \varphi} \right]$$

and, consequently,

$$\lambda_i^* = \sigma^{-2} R(\lambda^*) \left[1 - \frac{\gamma^2 \varphi_i(\varphi^T l)}{1 + \gamma^2 \varphi^T \varphi} \right] \tag{4.8}$$

for $i = 1, \ldots, n$. If γ^2 is sufficiently small (e.g., δ^2 is small or σ^2 is large), then $\lambda_i > 0$ for all i and (4.8) defines the optimal λ^*. Dividing (4.8) by $l^T \lambda^* = 1$ gives

$$\lambda_i^* = n^{-1} \left[1 - \frac{\gamma^2 (\varphi_i - \overline{\varphi})\overline{\varphi}}{n^{-1} + \gamma^2 v(\varphi)} \right], \tag{4.9}$$

where $\overline{\varphi} = n^{-1} \sum_{i=1}^n \varphi_i$ and $v(\varphi) = n^{-1} \sum_{i=1}^n (\varphi_i - \overline{\varphi})^2$. Direct calculation gives

$$R(\lambda^*) = \sigma^2 n^{-1} \left[1 + \frac{\gamma^2 \overline{\varphi}^2}{n^{-1} + \gamma^2 v(\varphi)} \right]. \tag{4.10}$$

The analysis is more complicated if some of the weights λ_i^* defined by (4.9) are negative. This can happen if either γ^2 or φ_i (or both) are large. It is typical in time series analysis when we face a long series of x_i. In such cases the following iterative procedure can be recommended:

> Start with a set of n_0 points x_i that are neighbouring the given point of interest, and construct $\lambda^*(n_0)$. At stage s of the procedure,
>
> (a) remove all points with $\lambda^*(n_s) < 0$; construct for the set of n'_{s+1} remaining points the weights $\lambda^*(n'_{s+1})$;
>
> (b) if the new weights fulfill $\lambda^*(n'_{s+1}) > 0$, add Δn nearest points and repeat (a) with $n_{s+1} = n'_{s+1} + \Delta n$.
>
> The procedure is continued until $R[\lambda^*(n_s)]$ cannot be further decreased.

When x_i is one-dimensional we have no problem to choose points to be added or removed. In multi-dimensional cases, it is not a simple question to define what a "nearest point" is. In this case, (4.6) can guide the process. For instance, to add a point we have to find

$$i^*(s) = \arg\min_i \varphi_i [\varphi^T \lambda^*(n_s)],$$

where the minimization is taken over the set of points that are candidates for being added.

It should be emphasized that for practical application we need a good prior information about γ^2.

Example 2 *Estimation of a model with curvature.*

Let $y_1, \ldots, y_n$ be observations from locations $-1 \leq x_1 < \ldots \leq 0 \leq \ldots < x_n \leq 1$ symmetrically arranged around 0. We want to get a prediction $\hat{y}$ for x. If we assume that only a linear trend must be accounted for, optimal weights λ^* are to be chosen using $\varphi_i = x_i - x$. For $n = 10$ and $x = 0$, $x = 0.5$ and $x = 1$, respectively, we get λ^* as shown in Figure 2.

If x is in the vicinity of a maximum (minimum) of the response function, it is suitable to take $\varphi_i = (x_i - x)^2$. For this case, the typical structure of the corresponding weight function is shown in Figure 2. *In practice, when the actual curvature is unknown, one has to compromise these two weight functions. For most remote points, x^2 becomes dominating.*

In general, the structure of λ^* is given by

$$\lambda_i^* = a - b\varphi_i \,,$$

where a and b are defined by (4.9) and all $x_i, i = 1, \ldots, n$, are assumed to be fixed.

insert Figure 2

A general recommendation is to investigate the structure of λ^* for a given set $X_n = \{x_1, \ldots, x_n\}$. As the corresponding values $\varphi_1, \ldots, \varphi_n$ are usually unknown, prior knowledge from theory or past experience must be used to make assumptions on the form of the weight function. Then, for a special application, the so obtained general structure can be normalized and adapted to the features of the situation in mind. The consideration of some simple cases help to clarify the dependence of the weight function on the location of the supporting points.

Example 3 *We assume that* $\varphi_i = x_i^2$ *and that the supporting set* $\mathcal{X} = \{-x, 0, x\}$ *consists of three points. For the point of interest* $x = 0$, *we find* $\overline{\varphi} = 2x^2/3$ *and* $v(\varphi) = 2x^4/9$. *For the weight function we obtain*

$$\lambda(\pm x) = \frac{1}{3 + 2\gamma^2 x^4},$$

$$\lambda(0) = \lambda_2 = \frac{1 + 2\gamma^2 x^4}{3 + 2\gamma^2 x^4}.$$

Similar results can also be found for nonsymmetrical designs. If we take $\varphi_i = x_i$ *and* $\mathcal{X} = \{-x, 0\}$, *then* $\overline{\varphi} = -x/2$, $v(\varphi) = x^2/4$ *and*

$$\lambda(-x) = \frac{1}{2 + \gamma^2 x^2}, \quad \lambda(0) = \frac{1 + \gamma^2 x^2}{2 + \gamma^2 x^2}.$$

For fixed x *the results turn out to be special cases of concluding result in Example 2.*

4.2 Moving regression: The general case

The analysis of the general case is based on the ideas proposed in the last section. For convenience, we assume that the objective function is of the form

$$\Psi[R(\lambda)] = \operatorname{tr} AR(\lambda)$$

where A is nonnegative definite and $R(\lambda)$ is defined in (2.5). According to the assumption of Section 2, in moving regression analysis only the upper left element of $R(\lambda)$ is of interest, i.e., we set $A_{11} = 1$ and all other elements of A to zero. However, more general situations can be considered. Generalizations of the results to other convex objective functions for Ψ are possible.

Similar to the moving average case, the following theorem can be stated.

Theorem 1 *The necessary and sufficient condition for* λ^* *to be optimal is that the* λ^* *fulfill the equality*

$$\min_i \{f_i^T [\lambda_i^* M^{-1}(\lambda^*) - \sigma^{-2} R(\lambda^*)] A M^{-1}(\lambda^*) f_i \tag{4.11}$$

$$+ \sigma^{-2} f_i^T A M^{-1}(\lambda^*) M_{12}(\lambda^*) \delta \delta^T \varphi_i\} = 0$$

or

$$\min_i \{\lambda_i^* f_i^T M^{-1}(\lambda^*) A M^{-1}(\lambda^*) f_i - \sigma^{-2} f_i^T R(\lambda^*) A M^{-1}(\lambda^*) f_i$$
$$+ \sigma^{-2} f_i^T A M^{-1}(\lambda^*) M_{12}(\lambda^*) \delta\delta^T \varphi_i\} = 0\,.$$

Proof:

The proof is based on the relations

$$\lim_{\alpha\to 0} \frac{\partial M^{-1}(\overline{\lambda})}{\partial \alpha} = M^{-1}(\lambda^*) - M^{-1}(\lambda^*) M(\lambda) M^{-1}(\lambda^*),$$

$$\lim_{\alpha\to 0} \frac{\partial}{\partial \alpha} M_{12}(\overline{\lambda}) \delta\delta^T M_{12}(\overline{\lambda}) = 2M_{12}(\lambda^*)\delta\delta^T M_{12}(\lambda) - 2M_{12}(\lambda^*)\delta\delta^T M_{12}(\lambda^*),$$

$$\lim_{\alpha\to 0} \frac{\partial}{\partial \alpha} \mathcal{M}(\overline{\lambda}) = 2\mathcal{M}(\lambda^*, \lambda) - 2\mathcal{M}(\lambda^*),$$

where $\overline{\lambda} = (1-\alpha)\lambda^* + \alpha\lambda$ with $0 \le \alpha \le 1$, and $\mathcal{M}(\lambda^*, \lambda) = \sum_{i=1}^n \lambda_i^* \lambda_i f_i f_i^T$. We find that

$$\frac{\sigma^{-2}}{2} \lim_{\alpha\to 0} \frac{\partial R(\overline{\lambda})}{\partial \alpha}$$
$$= M^{-1}(\lambda^*) \left[M_{12}(\lambda^*)\delta\delta^T M_{12}(\lambda^*) - M(\lambda) M^{-1}(\lambda^*) M_{12}(\lambda^*)\delta\delta^T M_{12}(\lambda^*) \right.$$
$$\left. + \mathcal{M}(\lambda^*, \lambda) - M(\lambda) M^{-1}(\lambda^*) \mathcal{M}(\lambda^*) \right] M^{-1}(\lambda^*).$$

Similarly to Section 4.1 the necessary and sufficient condition for the optimality of λ^* is that the inequality

$$\lim_{\alpha\to 0} \frac{\partial \operatorname{tr} A R(\overline{\lambda})}{\partial \alpha} \ge 0 \qquad (4.12)$$

is fulfilled for any α. Application of the one-point weight function λ to (4.12) leads to the final result (4.11). □

Numerical algorithms for constructing an optimal weight function can be designed similarly to the algorithms used in the experimental design theory. For instance, we can use the following iterative algorithm:

Start with any choice $\lambda(0)$. At stage s of the procedure proceed as follows:

(a) Given $\lambda(s)$, find

$$i_s^* = \arg\min_i \Phi_i[\lambda(s)]\,,$$

where

$$\begin{aligned}\Phi_i(\lambda) &= \sigma^2\lambda_i f_i^T M^{-1}(\lambda)AM^{-1}(\lambda)f_i - f_i^T R(\lambda)AM^{-1}(\lambda)f_i \\ &\quad + f_i^T AM^{-1}(\lambda)M_{12}(\lambda)\delta\delta^T\varphi_i\,.\end{aligned}$$

(b) Construct the new set of weights

$$\lambda(s+1) = (1-\alpha_s)\lambda(s) + \alpha_s\lambda(i_s^*)\,,$$

where $\lambda(i_s^*) = 1$ at point $x_{i_s^*}$ and zero otherwise.

A suitable choice for the series $\{\alpha_s\}$ is given by $\alpha_s \sim (s+m)^{-1}$, where m is the dimension of θ. This algorithm is robust with respect to the initial choice $\lambda(0)$ and converges under general conditions.

Of course, the iterative procedure is mainly of theoretical interest as it crucially depends upon σ^2, δ, and φ. Nevertheless, it can be helpful to clarify the general structure of λ^*.

5 An Illustration: Moving Regression in Forecasting

In this section we will show how the procedure discussed so far can be used to calculate forecasts for time series data that are optimal in the above-stated sense.

Let $\{x_1, \ldots, x_n\}$ be a given set of supporting points where observations $\{y_1, \ldots, y_n\}$ are available, and let $d_i = x_i - x_{n+1}$. Then

$$y_i = \theta^T f(d_i) + \varepsilon_i + \delta\varphi(d_i)\,, \quad i = 1, \ldots, n \tag{5.1}$$

will be called a one-sided regression model. In a typical situation of this kind, the

subscript represents time.

Setting $i = n+1$ allows to calculate a forecast for y_{n+1} from (5.1). If we again make the assumption stated in Section 2, the forecast for y_{n+1} is

$$\hat{y}_{n+1} = \hat{\theta}_{n+1,1}$$

the first component of the (least squares) estimator

$$\hat{\theta}_{n+1} = M^{-1}Y$$

witch is attained along the lines given in Section 2. In the one-sided situation, $\hat{\theta}_{n+1}$ again is generally biased [see (2.4)]. However, the mean squared error matrix R of $\hat{\theta}_{n+1}$, given by (2.5), is usually "worse" than the one we find in the two-sided moving regression case. Of course, one-sided moving regression is not effected by right-side boundary conditions.

One-sided moving regression is a suitable method of analysis if we are interested in the current condition of an object that develops in time. This is of particular value if we cope with time series that are short relative to the smoothness interval.

5.1 Application to bank account data

For applying moving regression to a set of time series that differ considerably with respect to its characteristics such as bank account data for different firms, the smoothing interval has to be long enough to cover the longest period of changes in these characteristics. The weight function itself can be chosen in accordance with the recommendations of the previous section.

For the linear moving regression (of the first order) the weight function is chosen to be $\lambda(x) \sim a - bx^2$, with the possible normalization $\lambda(0) = 1$ and $\lambda(-T) = 0$, where T is the length of the smoothing interval. Moving average can describe the long wave changes in the activities of enterprises, but smoothes away short term effects. Adding the term θx, i.e., using a linear moving regression, allows to identify at least those changes which are significant within the smoothing interval.

Typically, enterprises are different in all sorts of characteristics such as size and kind of activities. Classification according to probability of failing can be based on one of the following two approaches.

(1) We can take lessons from the history of the particular enterprise. This means that we construct moving regression model for all past time points, i.e., time points that are not affected by the left side boundary conditions. Heavy "negative" tails in the plot of the frequency distribution of each parameter of the moving regression model must be considered as suspicious. An alarming situation at recent cases is indicated by parameter estimates at the tails of the parameter frequency function. Similarly, scatter plots can be used in the two parameter case; e.g., one axis represents the intercept (θ_1) and the other one the slope (θ_2).

(2) The second approach consists in clustering the time series, i.e., the enterprises, in groups with respect to their failing probability. For each group some typical patterns in the frequency functions or the scattering plots will to be found that can be used to classify a new case. Application of moving regression of second kind with random parameters might provide even better results.

An important point of the analysis concerns the length of interval between points of interest for moving regression. The calculation of the moving regression on a daily basis allows to recognize the dynamics (trajectories) of estimated values by visual analysis of frequency functions or scatter plots. However, the estimates of the moving regression parameters must be expected to be strongly correlated for neighboring points and, therefore, do not contain too much information relative to each other. Moreover, the volume of numerical calculations is increasing with decreasing length.

As an illustration, the correlation function is derived for the moving average

$$y_i = \theta + \varepsilon_i ,$$

i.e., the model contains no "bias" term. Using weights $\{\lambda_i\}$, we have for the weighted moving average at time T and $T - \Delta$

$$\hat{\theta}(T) \;=\; \sum_{i=T-n}^{T} \lambda_{i-(T-n)} y_i ,$$

$$\hat{\theta}(T-\Delta) \;=\; \sum_{i=T-\Delta-n}^{T-\Delta} \lambda_{i-(T-\Delta-n)} y_i \,,$$

respectively. For the correlation function $\rho(\Delta)$ we get

$$\begin{aligned} \rho(\Delta) &= \mathrm{E}\{[\hat{\theta}(T)-\theta][\hat{\theta}(T-\Delta)-\theta]\} \\ &= \mathrm{E}\{\sum_{i=T-n}^{T} \lambda_{i-(T-n)}\varepsilon_i \sum_{j=T-\Delta-n}^{T-\Delta} \lambda_{j-(T-\Delta-n)}\varepsilon_j \\ &= \sum_{i=T-n}^{T}\sum_{j=T-\Delta-n}^{T-\Delta} \lambda_{i-(T-n)}\lambda_{j-(T-\Delta-n)}\delta_{ij} \\ &= \sum_{i=T-\Delta-n}^{T-\Delta} \lambda_{i-(T-n)}\lambda_{j-(T-\Delta-n)} \cdot \end{aligned} \tag{5.2}$$

Obviously, $\rho(\Delta) = 0$ when $\Delta > n$. For weights

$$\lambda_l \sim a - b(n-l)$$

that are chosen in accordance with the previous section normalization (n has to depend upon γ^2) it yields

$$\lambda_l = 2\left(1 - \frac{n-l}{n}\right) .$$

Using thess values in (5.2) gives numerical values for $\rho(\Delta)$ as a function of n and T.

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.

W.S. Cleveland, S.J. Devlin, and E. Grosse. Regression by local fitting. *Journal of Econometrics*, 37:87–114, 1988.

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

H.G. Müller. Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82(397):231–238, 1987.

W.G. Müller. *On moving local regression, with special reference to experimental design in economics.* Technical Report, unpublished doctoral thesis at Vienna University, Department of Statistics and Informatics, Vienna, 1991.

C.R. Pelto, T.A. Elkins, and H.A. Boyd. Automatic contouring of irregulary spaced data. *Geophysics*, 33(3):424–430, 1968.

B.D. Ripley. *Spatial Statistics.* Wiley, New York, 1981.

S.D. Silvey. *Optimal Design.* Chapman and Hall, London, 1980.

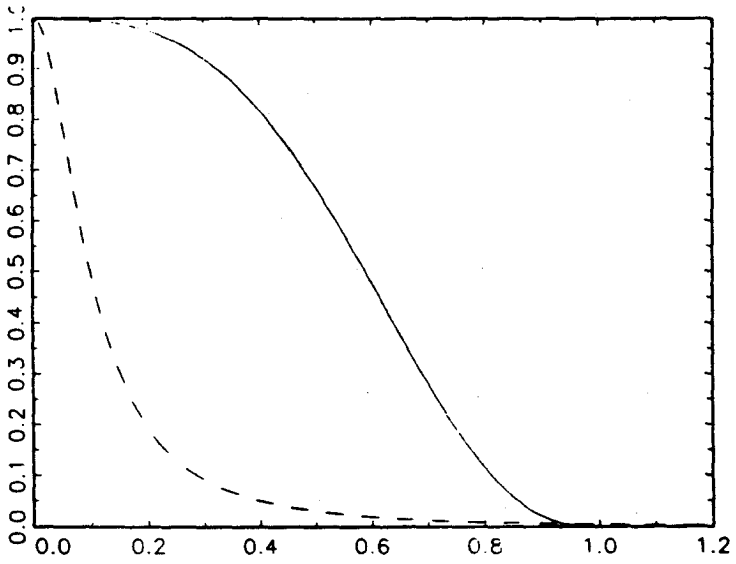


Figure 1: dotted - normalized McLain's (3.1), solid - Cleveland's (3.2) weight function

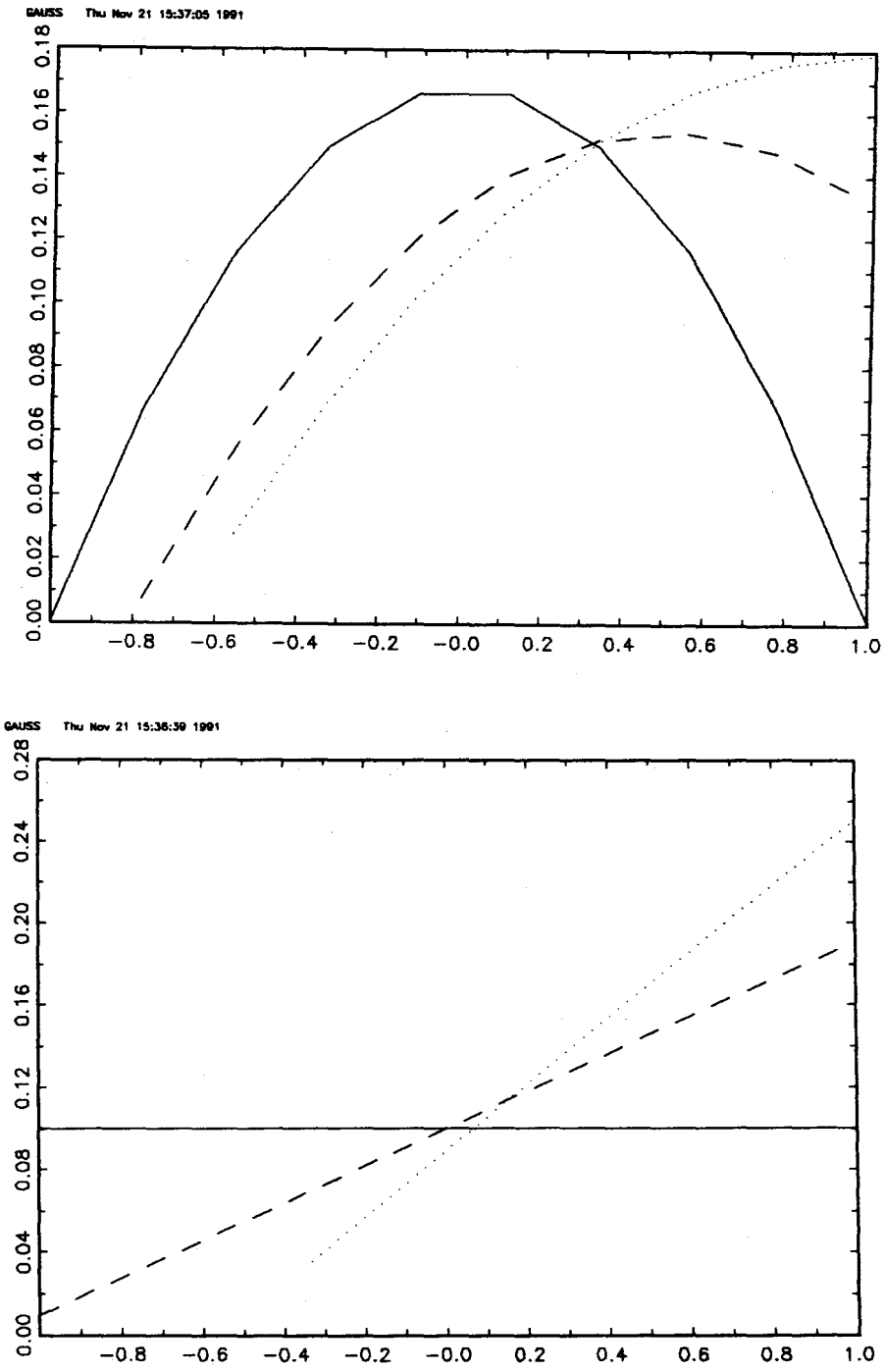


Figure 2: weight functions in Example 2