Optimal Design for Moving Local Regressions

Müller, Werner

Published: 01/01/1992

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

Optimal Design for Moving Local Regressions

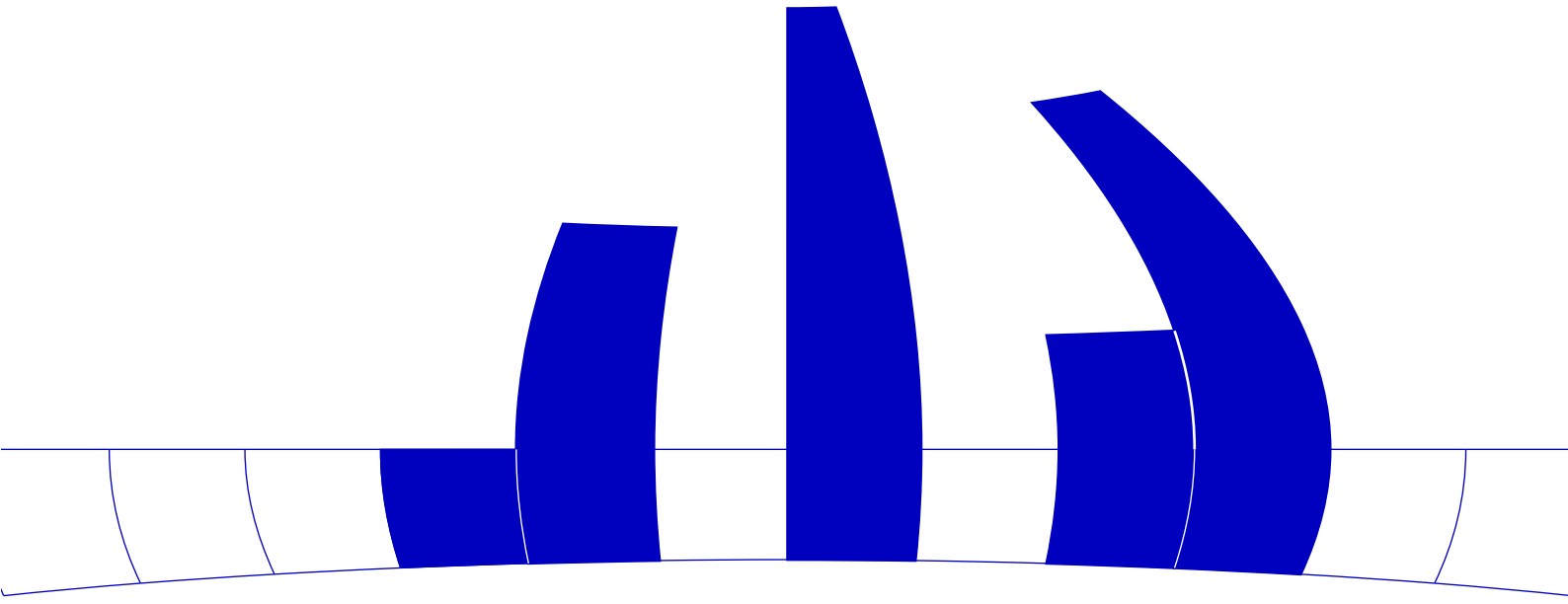
Werner Müller

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 29
November 1992

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



OPTIMAL DESIGN FOR MOVING LOCAL REGRESSIONS

WERNER G. MÜLLER

Abstract

This paper describes the so-called *moving local regression*, a special nonparametric statistical tool, which is firstly discussed thoroughly from analysis point of view. It avoids some drawbacks of its most serious alternatives: spline and kernel methods. The former lack computational simplicity and calculation speed, the latter may introduce high bias due to local constancy.

The methods incorporation into the design framework is given, including the derivation of the necessary formulae. A Kiefer-Wolfowitz type equivalence theorem is formulated. Some geometrical examples illuminate the interrelations of the basic ingredients of the method and establish empirical relations to parametric techniques.

1 Introduction

In any experimental situation one might pose the question how to select certain inputs in order to gain most accurate information about the studied process. This question is referred to as the design problem in the statistical literature, which in the regression context origins in the work of Kiefer, (1959). To be specific, one looks for a set:

$$\begin{pmatrix} p_1, & p_2, & \ldots, & p_n \\ x_1, & x_2, & \ldots, & x_n \end{pmatrix}$$

where $p_i = N_i/N$ and $\sum_{i=1}^{n} p_i = 1$ and which is called the (normalized) design ξ as input to a (nonlinear) regression problem:

$$y_i = \eta(x_i, \theta) + u_i \qquad \text{with} \quad E[u_i] = 0 \tag{1.1}$$

where $y = (y_1, ..., y_i, ..., y_n)$ is a vector of observations and $\eta(x, \theta)$ is the prior given structure called response-function. The weights p_i can be regarded as precision or duration of the measurements.

The main principle behind a well designed experimental plan is obvious: Asserting hypothesis-testing or parameter estimation as the aim of the experiment, the task

remains to take a (possibly given) number of measurements, that is to choose the set ξ, in a way, that either the power of the test or the precision of the estimation of θ is maximized. Additionally restrictions on the experimental conditions (i.e. to a certain experimental region $\mathcal{X}$) might have to be considered.

Note that in (1.1) η, the so-called response function, is assumed to be valid over the whole design space $\mathcal{X}$. For those type of problems and several extensions there exists a very well developed theory of design, see Silvey, (1980) and Ford et al., (1989).

Nevertheless if the region $\mathcal{X}$ is expanded, it is sometimes hard to find simple, say linear models η due to local disturbances. Though the experimenter still wants to be with them, either because of good statistical properties or good interpretability.

Therefore it is reasonable to use the simple models locally, which keeps some advantages but also allows for complex structures on the global level. In principle two approaches apply in this context. The first is to subdivide $\mathcal{X}$ into several intervals or subregions, wherein the simple models are assumed to hold 'locally'. The design problem for this class of models is to some extent equivalent to the standard methods for the global approach, for segmented polynomial regression see Park, (1978). Slightly more complicated but still related considerations hold for spline smoothing, see Micchelli & Wahba, (1981).

The second approach is to approximate the global behaviour locally in a moving fashion, similar to standard moving averages. The most popular method within this class are the so-called kernel estimators, the corresponding design problem is tackled by Müller, (1988).

A different moving approach is intuitively simple and shows up most desirable properties. It was independently developed for either smoothing or interpolation purposes. Though Cleveland, (1979) used it for extracting information from fuzzy scatterplots and somehow claimed the invention of the referred method, it was probably Pelto et al., (1968), who firstly used a variant implemented in an automatic contouring algorithm (the moving average, which is a particular case, is already known for a much longer time). Their aim was to interpolate a given surface η with a sparse amount of irregulary spaced data points by local regression surfaces. They called the method 'moving weighted least squares' estimation and derived some important properties of which, like the conditions for interpolation of data points.

It was Cleveland, (1979) who firstly used the method in a univariate context: the smoothing of strongly scattering time-series. He considered computational algorithms and statistical properties of the method, which he called 'locally weighted regression' (or *loess*-regression).

To avoid confusion the discussed approach will be referred to as 'moving local regression' throughout this paper, though the differences of the Pelto versus the Cleveland approach will be remarked.

Mathematical development in this area is not yet completed and a lot of questions still remained unanswered. The problems of optimal design or optimal weighting function are only examples. The former is thoroughly discussed here as well as in

the univariate context in Müller, (1987), who shows asymptotic equivalence between certain moving local regression and kernel smoothers, whilst the latter is referred to extensively in Fedorov et al., (1990).

2 Linear Models

Most of the cited nonparametric approaches to regression (like kernel estimators, spline functions, moving local linear regressions, running means, bin smoothers) and also least squares regression belong to the class of linear estimators, which Buja et al., (1989) give a detailed comparison of. They have the common linear form:

$$\hat{\eta}(x) = \sum_{i=1}^{n} l_{i,n}(x) y_i \tag{2.1}$$

i.e. they are locally weighted averages of the data, with the weights $l_{i,n}(x)$ independent from y_i.

The weights of the kernel method for instance have general form, with:

$$l_{i,n}(x) = \frac{1}{b} \int_{s_{i-1}}^{s_i} K(\frac{x-u}{b}) du \tag{2.2}$$

Here s_i are interpolating the x_i's, b is the so-called band-width and K the kernel function. Details and examples of kernel estimators can be found in Müller, (1988).

Smoothing splines on the other hand can be thought of as kernel estimators with variable bandwith (for a recent survey see Silverman, (1985)). They loose their applicability in higher dimensions, since this affords tesselation of the region and handling of a growing number of (mixed) second (or higher) derivatives.

For high-dimensional generalizations of non-parametric methods some notes can be found in Ripley, (1981). For a simple polynomial problem see Spruill, (1988).

2.1 Moving local regression

The principle of the approach is that a considerably smooth function η can always be approximated by a 'simpler', say polynomial function over a small region of the regressors space. At any point ${}_jx$ where an interpolated or smoothed value is desired, a weighted least squares regression is performed, with a weight function decreasing with increasing distances from ${}_jx$ to the data points (leading indices indicate fixed points, contrary to design points). A good choice of this function, fulfilling certain requirements, introduces locality, the construction of which is a major point to the problem.

Suppose that at any point x_i the response can be fitted by a local expansion from some arbitrary point ${}_jx$ in the vicinity of x_i, so that:

$$y_i = \gamma(\theta_j, d_{ij}) + \epsilon_{ij}, \qquad i = 1, \ldots, n, \tag{2.3}$$

where $d_{ij} = x_i - {}_jx$ ($\bar{d}_{ij}$ is any suitable distance-measure between x_i and ${}_jx$) and ϵ can be regarded as the local stochastic disturbance (or approximation error), vanishing as $o(d_{ij})$ when $x_i \rightarrow {}_jx$.

Obviously the choice of the approximation function γ is arbitrary, not only the linear case, where

$$\gamma(\theta_j, d_{ij}) = \theta_{0j} + \theta_j^T d_{ij},$$

which will be entirely referred to in what follows, but also polynomial terms of higher order (from a taylor expansion), or even nonlinear functions supported by physical considerations, could be used.

A sensible estimator for θ_0 and θ is now defined as

$$\{\hat{\theta}_{0j}, \hat{\theta}_j\} = \arg \min_{\theta_{0j}, \theta_j} \sum_{i=1}^{n} \lambda(\bar{d}_{ij})[y_i - \theta_{0j} - \theta_j^T d_{ij}]^2, \tag{2.4}$$

where the so-called weight function λ reflects the influence of local stochastics (or the reliability of the given Taylor-expansion). Obviously norm L_q generalizations could equivalently be used. As already mentioned moving local regression, applied rigorously, affords a lot of computations, since at every point ${}_jx$ to be evaluated a standard weighted least squares regression has to be performed. Assuming that $var(\epsilon_{ij}) \propto \lambda(\bar{d}_{ij})^{-1}$ Fedorov, (1989) gives the necessary formulae derived from the well-known technique.

From Equations (2.3) to (2.4) it clearly follows that $\hat{\eta}({}_jx) = \hat{\theta}_{0j}$. Since $\hat{\theta}_{0j}$ is a linear estimator, it hence can be considered a moving locally weighted average of the data (compare Cleveland, (1979) and (2.1)):

$$\hat{\eta}({}_jx) = L_j y, \tag{2.5}$$

where $y^T = [y_1, \ldots, y_n]$ and $L_j = [l(\bar{d}_{1j}), \ldots, l(\bar{d}_{nj})]$. L_j depends upon the x_i's and λ but not on the y_i's.

Though Cleveland, (1979) spends some efforts in investigating the form of the local fit, any complication at that stage does not add to the richness of the method enough to justify the loss in simplicity.

As the remark to (2.3) indicates there exists some freedom in choosing alternative local fitting schemes versus the discussed linear approach. The point is that insufficient local fitting may introduce a bias to the estimator, especially if the curvature of the original surface is high. Cleveland, (1979) argues that increasing the neighbourhood (via weight function λ, see below) tends to decrease the variance term of the given estimators. On the other hand the bias keeps growing along with the neighbourhood, which of course is a fundamental problem to deriving statistical properties. An empirical evidence from a simulation experiment gives Müller, (1991).

It is clear, that through introducing more sophisticated local fitting schemes it would be possible to absorb the bias effect, though of course the computational burden may increase tremendously (as for local quadratic fitting $2p+p(p-1)/2$ parameters are to be estimated). Cleveland, (1979),Cleveland et al., (1988) and Cleveland & Devlin, (1988)

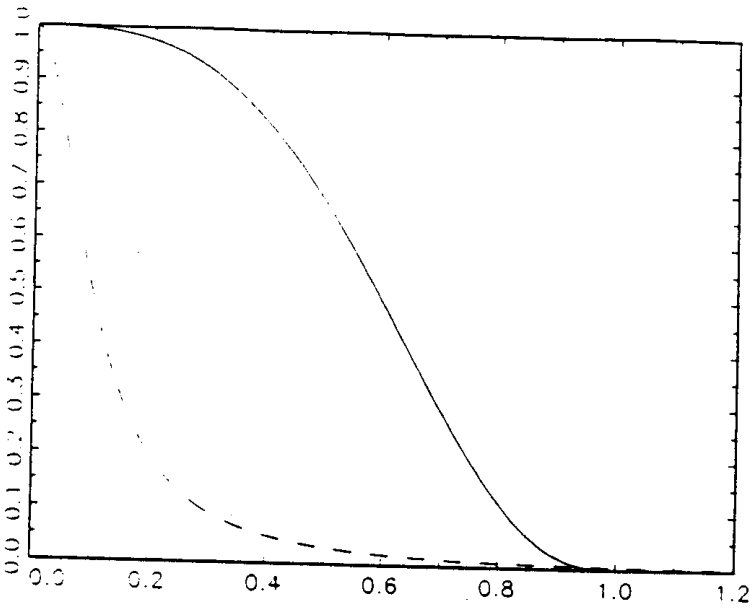


Figure 1: dotted - normalized McLain's (4.8), solid - Cleveland's (4.9) weight function

use the so-called M-plot as a tool for deciding between local polynomial approximation methods, while Fedorov, (1989) demands support for a specific scheme by prior knowledge.

The usefulness of the given method enormously depends upon the used weight function λ. Two properties are essential to allow a sensible interpretation:

(*a*) $\lambda(\bar{d})$ is a nonincreasing function for $\bar{d} \geq 0$ ($\lambda(0) = \max \lambda$).

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

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

Out of practical considerations McLain, (1971) suggested the weight function

$$\lambda(\bar{d}) = \exp(-\bar{d}^2/\bar{d}_n^2)/(\bar{d}^2 + \delta), \tag{2.6}$$

where $\bar{d}_n$ is the average distance between neighbouring data points and constant δ is used to avoid arithmetic overflow. A normalized ($\delta\lambda$) and reparametrized ($\delta = 10^{\bar{d}_n} - 1$) version will be used in the examples.

A computationally simpler function, the so-called tricube, is used by Cleveland, (1979):

$$\lambda(\bar{d}) = \begin{cases} (1 - (\bar{d}/\bar{d}_f)^3)^3 & 0 \leq \bar{d}/\bar{d}_f \leq 1 \\ 0 & \text{else} \end{cases}, \tag{2.7}$$

where $\bar{d}_f$ is the distance of the $f.n$ nearest point to ${}_jx$. This function smoothly decreases from 1 to 0 in the interval [0,1].

Buja et al., (1989) compare different linear estimators by plotting L_j against x_i. These plots, referred to as equivalent kernels, reflect the influence of the weight function and the form of the neighbourhood.

2.2 How to design

In practise the following (design) problem could frequently occur: At some points of interest $({}_1x, \ldots, {}_jx, \ldots, {}_qx)$ of a prescribed region $\mathcal{X}$ the unknown response η has to be interpolated (smoothed). The question is where to take observations (at which x_i's) within $\mathcal{X}$ (which in principle may also be infinitly expanded) in order to estimate $\eta({}_jx)$ in a most efficient way using the scheme given above. Such a set ξ_n^* of x_i's minimizing a so called optimality criterion $\phi(\xi_n)$ is referred to as an optimal design, following traditional experimental design theory (see for instance Silvey, (1980)).

The total (deterministic as well as stochastic) deviation of an estimate should be reflected by such a criterion. But because this makes many problems untreatable only the minimization of the stochastic deviation (the variance) will be regarded here, under the assumption, that $\hat{\eta}$ is an unbiased estimator, which of course holds only in special cases. Nevertheless it seems justified to derive further properties still keeping this assumption, since the amount of the bias can to some extent be governed by the weight function.

If one looks at the presentation $\hat{y} = Ly$ of the estimator, one could guess that it is reasonable to use a scalar function of $\mathrm{Cov}(\hat{\eta}) = L\,\mathrm{Cov}(y)L^T$ as the criterion of the design problem (with $L^T = (L_1, \ldots, L_q)$). Since the D-criterion in this context would be cumbersome to handle Fedorov, (1989) indicates that we can choose a weighted sum of the variance of the estimates $\hat{\eta}({}_jx)$ as a sensible objective function ϕ (e.g. the A-criterion).

The optimal design is then given by:

$$\xi_n^* = \arg\min_{\xi_n} \sum_{j=1}^{q} a_j \ \mathrm{var}(\hat{\eta}({}_jx)), \tag{2.8}$$

where the scalars a_j reflect the importance of ${}_jx$, the variances coming from the diagonal of $\mathrm{Cov}(\hat{\eta})$.

Applying standard weighted least squares techniques the variance of an estimator can be easily expressed as:

$$\begin{aligned} \mathrm{var}(\hat{\eta}({}_jx)) &= \mathrm{tr}\, AM_j^{-1}, & (2.9)\\ &= \mathrm{tr}\, A(F_j^T \,\mathrm{diag}(\lambda_j) F_j)^{-1}, & (2.10) \end{aligned}$$

where A is a null matrix except $A_{11} = 1$ and $F_j^T = \begin{pmatrix} 1 & \cdots & 1 \\ d_{1j} & \cdots & d_{nj} \end{pmatrix}$.

Using (2.9) the optimization criterion can directly be formulated:

$$\phi(\xi_n) = \mathrm{tr} \sum_{j=1}^{m} A_j M_j(\xi_n)^{-1}, \tag{2.11}$$

where $A_j = a_j A$.

Theorem 1 *Following the general theory, the minimizing optimality criterion (2.11) is equivalent to obeying the condition:*

$$\min_i \varphi(x_i, \xi) \geq 0, \tag{2.12}$$

with

$$\varphi(x_i, \xi) = tr \sum_{j=1}^{m} A_j M_j(\xi)^{-1} - \sum_{j=1}^{m} \frac{\lambda(\bar{d}_{ij})}{\sum_{i=1}^{n} \lambda(\bar{d}_{ij})} F_{ij}^T M_j(\xi)^{-1} A_j M_j(\xi)^{-1} F_{ij}, \tag{2.13}$$

where F_{ij} is a particular row from F_j.

Proof:

For the sake of simplicity we now turn from exact designs ξ_n to approximate designs ξ, which impose a measure on the space spanned by the regressors.

It is clear that if ξ^* is the optimal design and $\bar{\xi}$ is any other, that the following inequality holds at point $\xi = \xi^*$:

$$\frac{\partial \phi(\xi)}{\partial \alpha} \geq 0, \tag{2.14}$$

if $\xi = (1-\alpha)\xi^* + \alpha\bar{\xi}$, with $M_j(\xi) = (1-\alpha)M_j(\xi^*) + \alpha M_j(\bar{\xi})$ (see for instance Silvey, (1980)).

Evaluating (2.11) yields:

$$\begin{aligned} \frac{\partial \phi(\xi)}{\partial \alpha} &= \frac{\partial \operatorname{tr} \sum_{j=1}^{m} A_j M_j(\xi)^{-1}}{\partial \alpha} \\ &= \operatorname{tr} \sum_{j=1}^{m} A_j \frac{\partial M_j(\xi)^{-1}}{\partial \alpha} \geq 0, \end{aligned}$$

and following techniques of martix differentiation

$$= -\operatorname{tr} \sum_{j=1}^{m} A_j M_j(\xi)^{-1} \frac{\partial M_j(\xi)}{\partial \alpha} M_j(\xi)^{-1} \geq 0,$$

after inserting $M_j(\xi) = (1-\alpha)M_j(\xi^*) + \alpha M_j(\bar{\xi})$:

$$\operatorname{tr} \sum_{j=1}^{m} A_j M_j(\xi)^{-1} \frac{\partial (1-\alpha)M_j(\xi^*) + \alpha M_j(\bar{\xi})}{\partial \alpha} M_j(\xi)^{-1} \leq 0$$

$$= \operatorname{tr} \sum_{j=1}^{m} A_j M_j(\xi)^{-1} [M_j(\bar{\xi}) - M_j(\xi^*)] M_j(\xi)^{-1}.$$

If one now lets $\alpha \to 0$ then:

$$\operatorname{tr} \sum_{j=1}^{m} A_j M_j(\xi)^{-1} M_j(\bar{\xi}) M_j(\xi)^{-1} \leq \operatorname{tr} \sum_{j=1}^{m} A_j M_j(\xi^*)^{-1}$$

Now it is assumed that $\bar{\xi}$ only consists of one point x_i, then $M_j(\bar{\xi}) = F_{ij}^T \frac{\lambda(\bar{d}_{ij})}{\sum_{i=1}^n \lambda(\bar{d}_{ij})} F_{ij}$ which yields:

$$\varphi(x_i, \xi) = \mathrm{tr} \sum_{j=1}^m A_j M_j(\xi)^{-1} - \sum_{j=1}^m \frac{\lambda(\bar{d}_{ij})}{\sum_{i=1}^n \lambda(\bar{d}_{ij})} F_{ij}^T M_j(\xi)^{-1} A_j M_j(\xi)^{-1} F_{ij}.$$

qed

For his specially restricted moving local regression estimate Müller, (1984) and Müller, (1988) also addresses the problem of optimal design. He uses the integrated mean squared error (no special points of interest) as criterion function and finds the optimal design:

$$\xi^*(x) = \frac{\sigma(x)}{\int_0^1 \sigma(u) du} \tag{2.15}$$

where σ denotes a smooth variance function.

3 Examples

To illustrate the described techniques the following section gives some examples of theoretical, geometrical kind. These examples shall submit an idea of how the optimal design for local moving regression behaves, when various conditions, like the weight function, the local fitting scheme, etc. are changed.

It has to be emphasized that all examples have been calculated thoroughly numerically and the corresponding algorithm was stopped after about 100 iterations, which may lead to small differences to the 'true' optimal designs.

The simplest framework for investigation of the numerical examples to the given approach is provided by the case of constant local fitting and a constant weight function with a locality parameter d_δ:

$$\lambda = \begin{cases} 1 & 0 \leq \bar{d} \leq d_\delta \\ 0 & \text{else} \end{cases} . \tag{3.1}$$

That is of course the special case of moving averages. The points of interest for the following examples is a uniform collection of nine points with equal weights on the interval, that is with intermediate distances of 0.25. For the sake of simplicity the design region consists of the same nine points, which is a fairly good approximation of the interval $[-1, \ldots, 1]$.

It is obvious that if one chooses d_δ such that observations at a design point can influence the estimation at only one point of interest, that is in our case $d_\delta = 0.1$, less than half as long as the interdistances, then the optimal rule for the described situation is of course to take observations only exactly at the points of interest, since any intermediate points have to deliver worse information. The corresponding graph is displayed in Figure 2. The abscissa describes the design region, the height of bars at

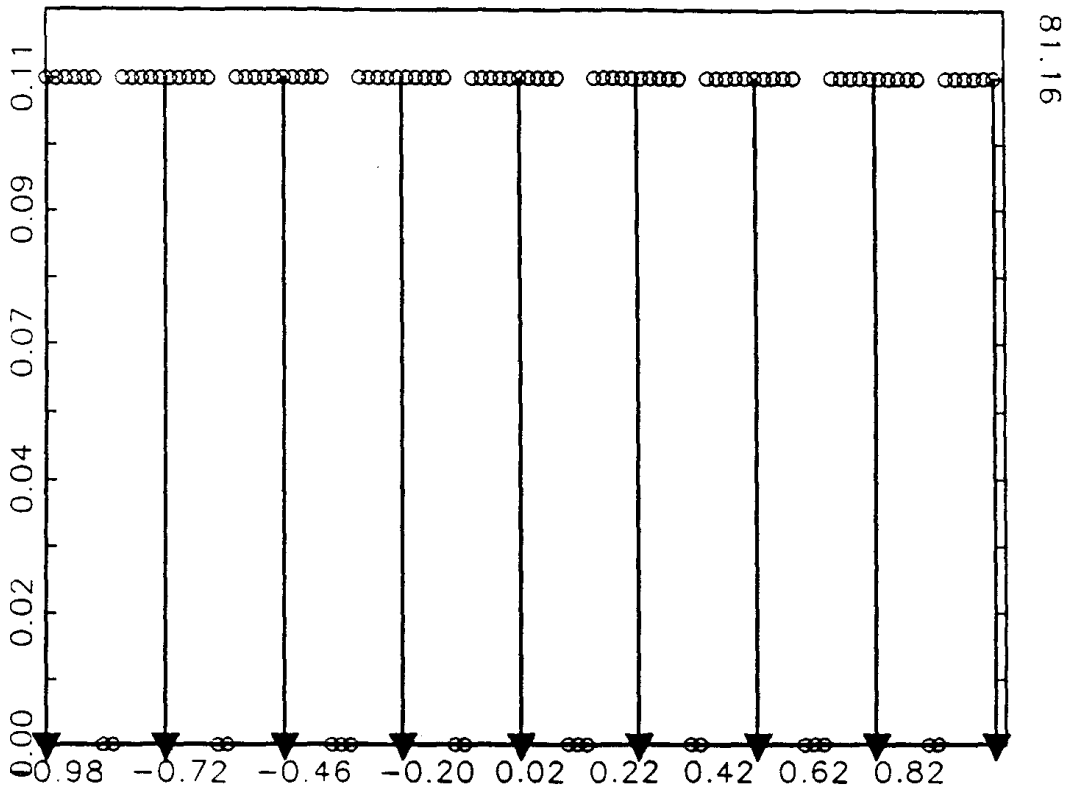


Figure 2: Optimal design for moving averaging with $d_\delta = 0.1$

design points denote their corresponding weights, the black triangles take the positions of the points of interest. The circles represent the value of the dual criterion function, its maximum value and hereby its scaling can be found at the upper right vertex of the graph.

In Figure 2 one can see that the dual function is equally maximum around the points of interest and zero at intermediate points. If one remembers that the optimal design has to take its values where the this function is maximum, the given optimal design is justified.

If we investigate the same design for a situation, where the locality of the weight function is altered such that now observations at a design point may influence estimates for three consecutive points of interest, that is simply $d_\delta = 0.3$, it clearly turns out to be inappropriate. The design does not achieve the maximum of the dual function at every of its supporting points equally, the dual function indicates the adjustment necessary for the unsymmetricity occuring at ends of intervals, when using moving averages. Note that the implicite estimation procedure are moving averages of length 3. The optimal design for the described case is given in Figure 3. It only consists of the three equally weighted points $(-0.75, 0, 0.75)$. Note that the value of the criterion is reduced by almost one third (from 38.33 down to 27.65). That is an equivalent increase in precision of the estimates.

Some more insight to the described methods can be gained from examples, where the local constancy of the weight function is substituted by a monotonically decreasing scheme (as for instance one of those described in the previous sections). For this reason the regarded weight functions have to be calibrated in order to ensure some equivalent degree of smoothing. Buja et al., (1989) introduced the notion degrees of freedom or equivalent number of parameters in this context. The most straightforward approach is to borrow the definition from the traditional linear model ($d.f. = \mathrm{tr}(LL^T)$) Following

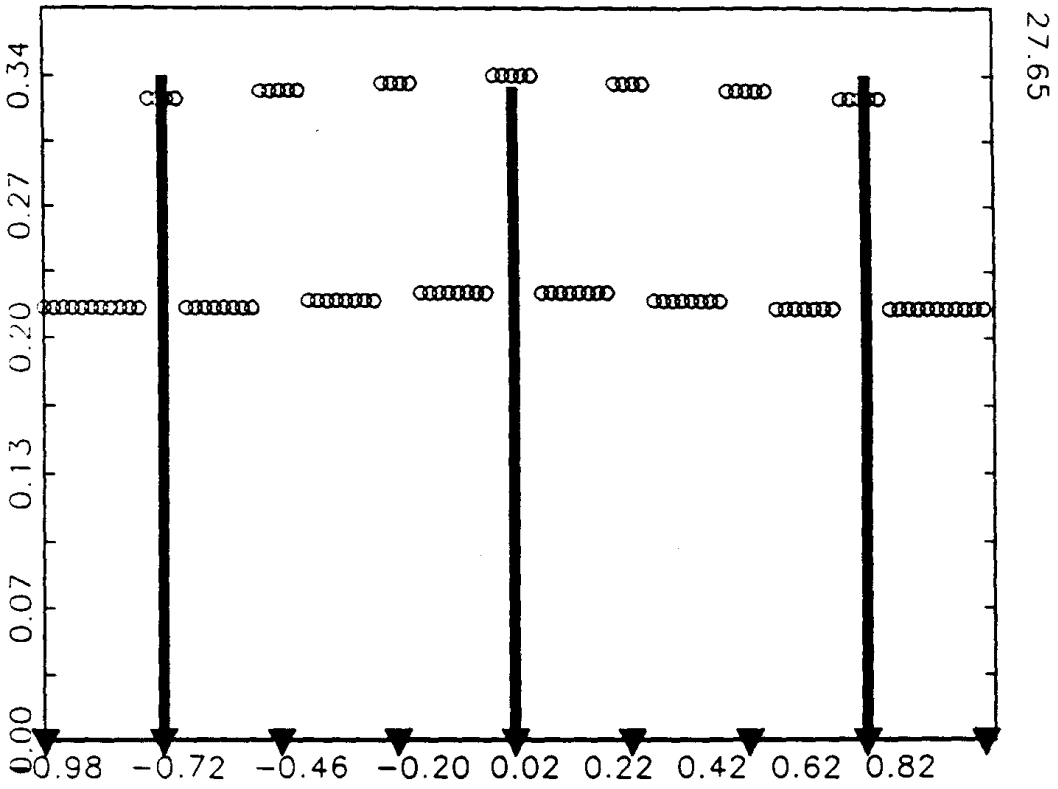


Figure 3: Optimal design for moving averaging with $d_\delta = 0.3$

these definitions the smoothing parameters of the respective weight functions were set to values that gave correspondence of $d.f.$'s.

Following the arguments that local constant fit (moving averaging) should be replaced by a more sophisticated fitting scheme (because of the bias problem), the following examples in this chapter will be devoted to local linear fit. This is the scheme, that is widely used in practise and therefore the illustrations from the examples may be informative also from an application point of view.

Two different weight functions, the same as previously discussed, will be utilized in what follows: Cleveland's (2.7) and McLain's (2.6). All results were carried out for the symmetrical set of nine uniformly distributed points of interest, as before, leaving the investigation of different schemes for situations arising from applications. For an applied example in two dimensions see Müller, (1991).

Firstly the weighting parameters are chosen to give approximately d.f.=3, that is a polynomial of second order model as a corresponding parametric scheme, the optimal design for which (three equally weighted points at $\{-1, 0, 1\}$) is well known from the literature. It can be seen from Figure 4, that the optimal design for moving local regression with both considered weight functions is pretty similar to this classical design, which is in accordance with expectations.

A quite different pattern, displayed in Figure 5, gives the application of the weight functions with $d.f. = 6$. Since now the approximation is more local than the approaches described above and the points of interest are rather remote, the optimal design points will coincide with them. However the weights tend to show a characteristic pattern, which differs remarkably from the results for the corresponding parametric polynomial of 5th order (equally weighted points at $\{-1, -0.8, -0.4, 0.4, 0.8, 1\}$).

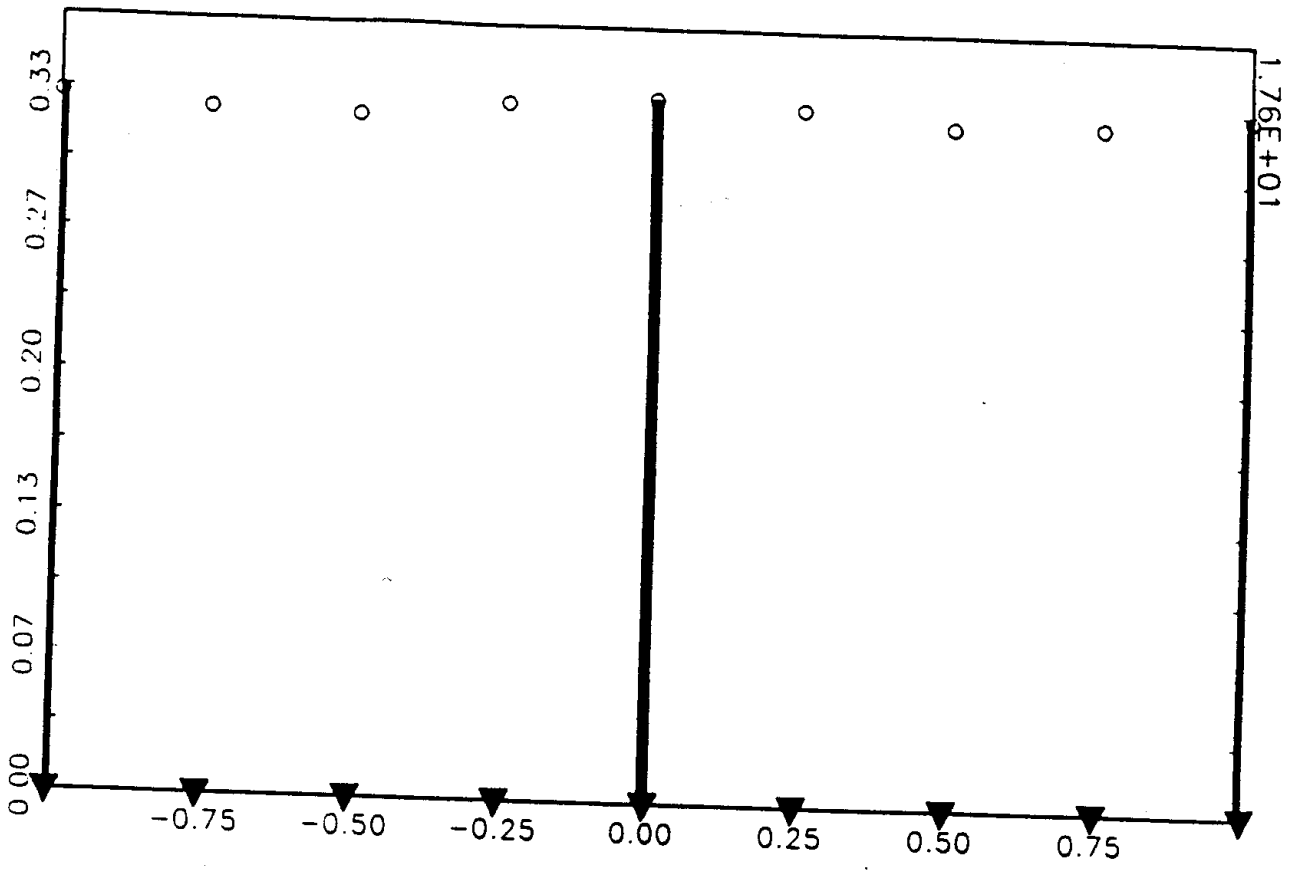


Figure 4: Optimal design for moving local linear regression with Cleveland's and McLain's weight function for $d.f. = 3$

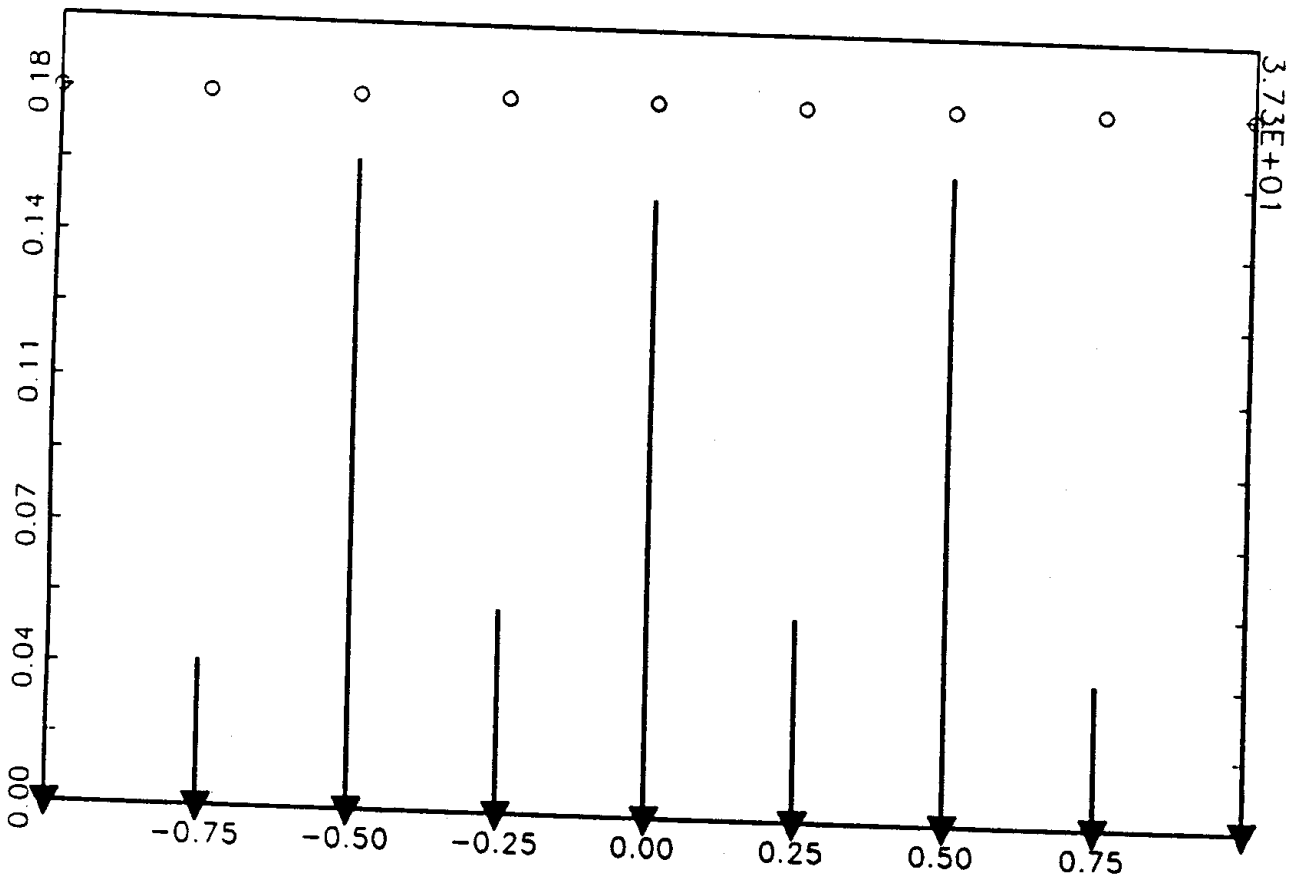


Figure 5: Optimal design for moving local linear regression with Cleveland's and McLain's weight function for $d.f. = 6$

4 Conclusions

Note that if little is known about the structure of the process generating model, one has to confine oneself to the discussed approaches (frequently called nonparametric regression techniques), which afford much less assumptions (just about smoothness and differentiablity) than the standard parametric methods.

Nonparametric methods are frequently applied in the first stage of a study, as exploratory analysis to find the general shape of the underlying function. Nevertheless if it is impossible to justify strong assumptions, those methods remain the only to be reasonably used.

From these few but characteristic and illuminative examples, one can get a feeling for the general pattern of the design in moving local regression experiments: weights are often not directly concentrated at the points of interest but at some distance, obviously dependend on the locality of the weight function and the density of points of interest. Nevertheless sometimes the results for a corresponding parametric problem seem to be a good landmark for the construction of those designs and may for instance serve as a starting design for numerical algorithms.

Acknowledgement

I am most grateful to Prof. V. Fedorov for numerous valuable suggestions and steady encouragement to the topic.

Some contents of this paper are part of the authors phd-thesis at the University Vienna.

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.

V.V. Fedorov. Kriging and other estimators of spatial field characteristics. *Atmospheric Environment*, 23(1):175–184, 1989.

V.V. Fedorov, P. Hackl, and W. Müller. *Estimation and experimental design for second kind regression models.* Technical Report, Economic University, Vienna, A-1090 Wien, Augasse 2-6, 1990.

I. Ford, D.M. Titterington, and C.P. Kitsos. Recent advances in nonlinear experimental design. *Technometrics*, 31(1):49–60, 1989.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, 1990.

J. Kiefer. Optimal experimental designs (with discussion). *Journal of the Royal Statistical Society*, B:272–319, 1959.

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

C.A. Micchelli and G. Wahba. Design problems for optimal surface interpolation. In *Approximation Theory and Applications*, pages 329–349, Academic Press, 1981.

H.G. Müller. Optimal designs for nonparametric kernel regression. *Statistics & Probability Letters*, 2:285–290, 1984.

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

H.G. Müller. *Nonparametric Regression Analysis of Longitudinal Data.* Volume 46 of *Lecture Notes in Statistics*, Springer-Verlag, 1988.

W.G. Müller. *Optimal designs for moving local regression.* Technical Report 281, Institute for Advanced Studies, Research Memorandum, Vienna, 1991.

S.H. Park. Experimental designs for fitting segmented polynomial regression models. *Technometrics*, 20(2):151–154, 1978.

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.

B.W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, B, 47(1):1–52, 1985.

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

M.C. Spruill. *Optimal Designs for Multivariate Interpolation.* Technical Report, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A., 1988.