# Local Polynomial Smoothing Under Qualitative Constraints

J.S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599–3260
U.S.A.

B.A. Turlach
CMA/Statistics
Australian National University
Canberra ACT 0200
Australia

M.P. Wand
AGSM
University of New South Wales
Sydney NSW 2052
Australia

## Abstract

Some nonparametric regression settings involve auxiliary information, e.g., the support function of a convex set is characterized by the fact that the sum with its second derivative is non negative. Another well known example is the branching curve problem (Silverman and Wood, 1987). Use of local polynomial smoothers in such settings leads to a quadratic programming problem. We shall show that it is feasible to solve this quadratic programming problem and discuss the performance of the resulting smoother using the above examples.

## 1 Introduction

We shall discuss two examples of nonparametric regression settings which involve auxiliary information. This auxiliary information imposes some constraints upon the nonparametric regression fit. In this paper we present a (quite general) methodology to solve such nonparametric regression problems using local polynomial or kernel smoothing.

Much of the work in constrained nonparametric regression has been done in the context of splines, because constraints have been viewed as simpler and easier to incorporate in that context. The perception that constraints are not easily incorporated into a kernel smoother seems to have arisen from the fact that traditional kernel estimators have been written as an explicit estimator rather than as a solution to an optimization problem. However, in recent years, attention has switched to the more appealing local polynomial form of the kernel smoother (Stone, 1977; Cleveland, 1979; Fan, 1992; Hastie and Loader, 1993) which is usually viewed as a solution to an optimization problem. We shall show that constraints can be imposed in a natural way on a kernel smoother from this point of view. If the constraints are linear in the resulting fit, this approach leads to a quadratic programming problem.

The paper is organized in the following way: details of our approach are described in Section 2 and applied to two examples in Section 3. These examples motivate some modification to the original approach which we will discuss in Section 4.

## 2 General Methodology

Nonparametric regression data are of the form $\{(X_i, Y_i) : i = 1, \ldots, n\}$, where

$$Y_i = m(X_i) + \varepsilon_i,$$

for $\varepsilon_1, \ldots, \varepsilon_n$ independent, mean zero random errors. The goal is to estimate the curve $m(x)$ under minimal assumptions such as "smoothness".

Local polynomial smoothing is one approach to nonparametric regression (Wand and Jones, 1995; Fan and Gijbels, 1996). It is based on weighted local fitting of a polynomial. The form of the weights is determined by a kernel function, $K$. The width of the window over which local fitting is done is determined by a bandwidth, $h$, through using weights based on the rescaled kernel function, $K_h(\bullet) = K(\bullet/h)/h$. For a given point $t$ the coefficients of the local polynomial fit of degree $p$ are the minimizers $\beta_0, \ldots, \beta_p$ of

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - t)^j \right\}^2 K_h(X_i - t) \qquad (2.1)$$

The local polynomial estimate of $m(x)$ is then the central value of the polynomial in the $K_h$ window, $\hat{m}(x) = \hat{\beta}_0$.

In matrix notation we can write $\hat{m}(x) = \mathbf{e}_1^T \hat{\beta}_t$, where $\mathbf{e}_1$ is the first unit vector in $\mathbb{R}^{p+1}$, $\mathbf{W}_t$ is a diagonal matrix with $K_h(X_i - t)$ as $i$th diagonal entry and

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^T, \quad \beta = (\beta_0, \ldots, \beta_p)^T,$$

$$\mathbf{X}_t = \begin{pmatrix} 1 & X_1 - t & \cdots & (X_1 - t)^p \\ 1 & X_2 - t & \cdots & (X_2 - t)^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n - t & \cdots & (X_n - t)^p \end{pmatrix}$$

and

$$\hat{\beta}_t = \underset{\beta}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}_t\beta)^T \mathbf{W}_t (\mathbf{Y} - \mathbf{X}_t\beta).$$

An important feature of constraints is that they are "global" in nature (with the potential for values of the estimator at any point to be linked with values at any other point). This motivates a more global view of the estimator. A means of doing this is to extend the parameter vector to a vector of functions, $\beta(t) = (\beta_0(t), \ldots, \beta_p(t))^T$, and to combine the pointwise optimization problems by integrating (since the integral is optimized by combining the pointwise optimizations). This gives

$$\hat{m}(\bullet) = \mathbf{e}_1^T \underset{\beta(\bullet)}{\operatorname{argmin}} \int ((\mathbf{Y} - \mathbf{X}_t\beta(t))^T \mathbf{W}_t (\mathbf{Y} - \mathbf{X}_t\beta(t)))\, dt. \tag{2.2}$$

This global formulation of the optimization problem makes it straightforward to build in constraints. Simply replace the minimization in (2.2) by a minimization over the function space which fulfills the constraints.

In general, this optimization problem no longer allows pointwise optimization, so a numerical approach based on discretization is suggested. Suppose it is desired to estimate $m$ at a grid of values $t_1, \ldots, t_g$. The $g$ optimization problems (2.1) can be combined into a single large optimization problem (which naturally allows global constraints) as follows. Use block diagonal matrix notation to construct

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{t_1} & & \\ & \ddots & \\ & & \mathbf{X}_{t_g} \end{pmatrix},$$

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{t_1} & & \\ & \ddots & \\ & & \mathbf{W}_{t_g} \end{pmatrix},$$

combine parameter vectors across the $t_j$ in a corresponding way:

$$\beta = (\beta_{t_1}^T, \ldots, \beta_{t_g}^T)^T,$$

where

$$\beta_{t_j} = (\beta_{0,j}, \ldots, \beta_{p,j})^T,$$

and extend the data vector and unit vectors to:

$$\mathcal{Y} = (\mathbf{Y}^T, \ldots, \mathbf{Y}^T)^T, \quad \mathbf{e} = (\mathbf{e}_1^T, \ldots, \mathbf{e}_1^T)^T.$$

The vector $\hat{\mathbf{m}} = (\hat{m}(t_1), \ldots, \hat{m}(t_g))^T$ of estimates can now be written in the form

$$\hat{\mathbf{m}} = \mathbf{e}^T \underset{\beta}{\operatorname{argmin}} (\mathcal{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathcal{Y} - \mathbf{X}\beta). \tag{2.3}$$

This is analogous to (2.2), and shares the property of being a rewriting of the local polynomial estimator (2.1) that allows simple, natural inclusion of "global type" constraints. Simply replace the minimization in (2.3) by a minimization over the subspace of $\mathbb{R}^{g(p+1)}$ whose elements fulfill the constraints. If the constraints are linear in $\beta$ the resulting optimization problem belongs to a class known as "quadratic programming problems". In the next section we will discuss two such examples.

A variety of algorithms for solving quadratic programming problems are known (see, among others, Gill et al., 1991). We found that the algorithm proposed by Goldfarb and Idnani (1983) can efficiently solve the quadratic programming problems which we encountered. This algorithm is a "dual algorithm", i.e., it starts from the unconstraint fit and then checks whether any constraint is violated. If a constraint is violated, the estimator is changed such that the constraint will no longer be violated while maintaining the optimality of the solution. This process is iterated until all constraints are satisfied.

# 3 Examples

## 3.1 Support function of a convex set

This example addresses a problem that arises variously in medical imaging and robotic vision, of estimating a convex set from noisy measurements of its support function. The support function $m(\theta)$ of a convex set $\mathcal{C}$, relative to an origin $O$, equals the perpendicular distance from $O$ to that tangent to $\mathcal{C}$ which makes angle $\theta \in (-\pi, \pi]$ to a given direction in the plane, see Figure 3.1. If $m''$ exists and is continuous everywhere, then a necessary and sufficient condition (Santaló, 1976, p. 2) for convexity of $\mathcal{C}$ is

$$m(\theta) + m''(\theta) \geq 0, \qquad \theta \in (-\pi, \pi]. \tag{3.1}$$

A common method of construction an estimate of the set from data is to assume that its boundary is piecewise linear and fit the straight line segments comprising its boundary by using a variant of constrained maximum likelihood under the assumption of either Normally or Uniformly distributed errors (see, among others, Prince and Willsky, 1990).

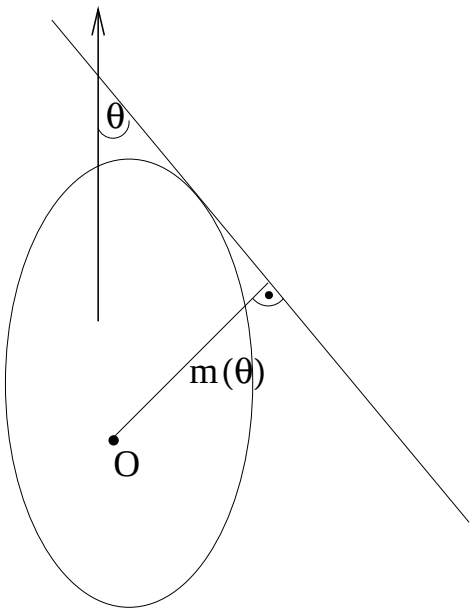


***Figure 3.1:*** *Illustration of the definition of a support function.*

Fisher et al. (1997) propose to estimate $m$ nonparametrically, e.g., via a local polynomial smoother. Essentially, they search for the bandwidth with least bias such that (3.1) is still fulfilled.

With the approach described in Section 2 such careful calibration of the bandwidth is not necessary. Assume that the points $t_1, \ldots, t_g$ are equidistant and in increasing order with $\Delta = t_2 - t_1$ (and $t_1 + 2\pi = t_g + \Delta$ for periodicity reason). Approximating $m''$ by second differences, we see that the discretized version of (3.1) is (with the obvious identification due to periodicity):

$$\beta_{0,i} + \frac{\beta_{0,i-1} - 2\beta_{0,i} + \beta_{0,i+1}}{\Delta^2} \geq 0, \quad i = 1, \ldots, g \quad (3.2)$$

These (inequality) constraints are linear in $\beta$ and hence minimizing (2.3) over all $\beta$ which fulfill (3.2) is a quadratic programming problem.

Figure 3.2 demonstrates this method on a simulated data set. We chose as convex set an ellipse with major axes of one and four units. Figure 3.2(a) shows the support function together with 400 noisy observations. An estimate for the support function using a local linear fit with bandwidth $h = 0.05$ and Gaussian kernel is shown as the dotted line in panel (b). The resulting fit which satisfies the constraints (3.2) is shown as the dashed line. Finally, the boundary obtained from this fit is drawn as the dashed line in panel (c) together with the original set.

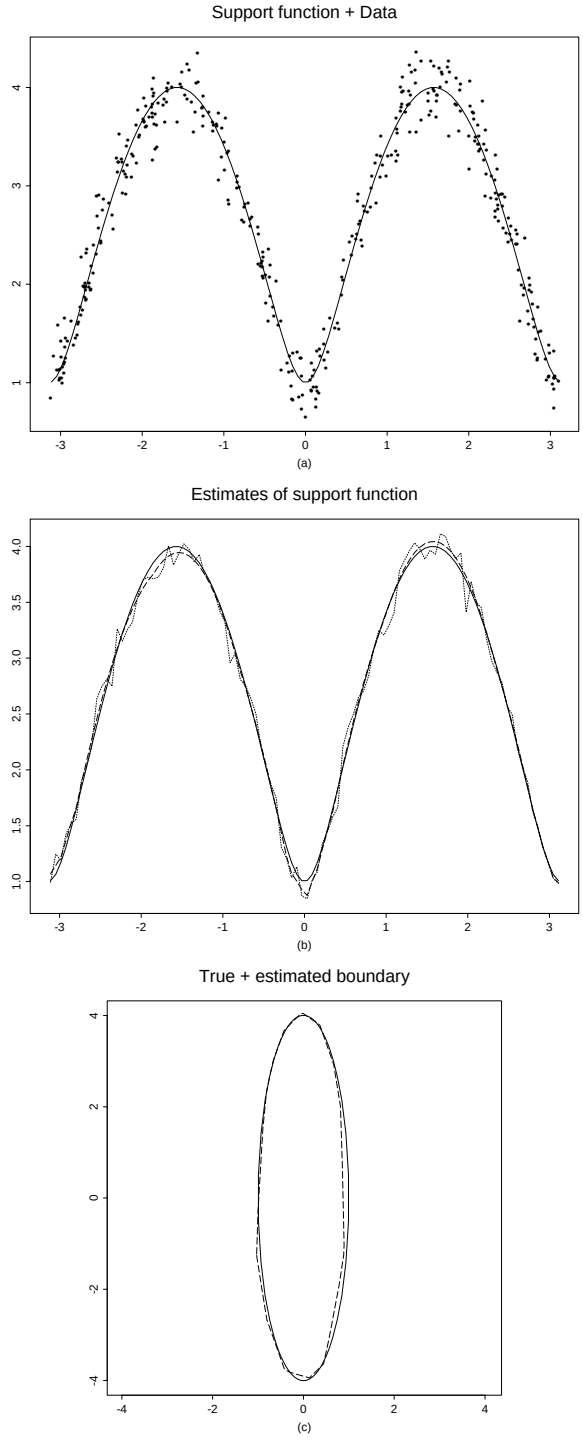


***Figure 3.2:*** *Panel (a) shows the support function of the ellipse shown in panel (c) together with the observed data. Panel (b) shows the support function (solid line) together with the initial fit to the data (dotted line) and the resulting constrained fit (dashed line). The boundary due to the constrained fit is shown as dashed line in panel (c).*

### 3.2 Branching curves

Steer and Hocking (1985) carried out an experiment to test the effect of applying nitrogen to sunflowers at different stages of growth. These data were analyzed by Silverman and Wood (1987) using spline smoothing techniques (see also Green and Silverman, 1994, Section 6.2).

The experiment was done on five groups of sunflowers. In the first group, the control, no nitrogen was applied. To the other four groups a nitrogen compound was applied at a given time after sowing, 38, 56, 63 and 70 days respectively. At various times the nitrogen content of plants taken from the different groups was measured destructively.

Before the time the nitrogen compound was applied there is no difference between the control group and the treatment group. Hence, when fitting regression curve to these data, it is natural to impose that up to the time of treatment the curve of the treatment group will coincide with the curve of the control group.

Although notation is a bit tedious, it is straightforward to write this smoothing problems within the framework of Section 2. Taking the grid $t_1, \ldots, t_g$ we fit five local polynomial smoothers to the data from the five groups. Introducing a further index which denotes the fitted curve, this leads to a minimization problem similar to (2.3) where the minimization is over all $\beta$ such that certain entries in $\beta$ (corresponding to the appropriate fits) are equal.

Obviously this is a quadratic programming problem since we have only (equality) constraints which are linear in $\beta$. Solving this problem using a bandwidth $h = 8$ and the Gaussian kernel leads to the solution shown in Figure 3.3(b). Panel (a) shows the initial, unconstrained fits.

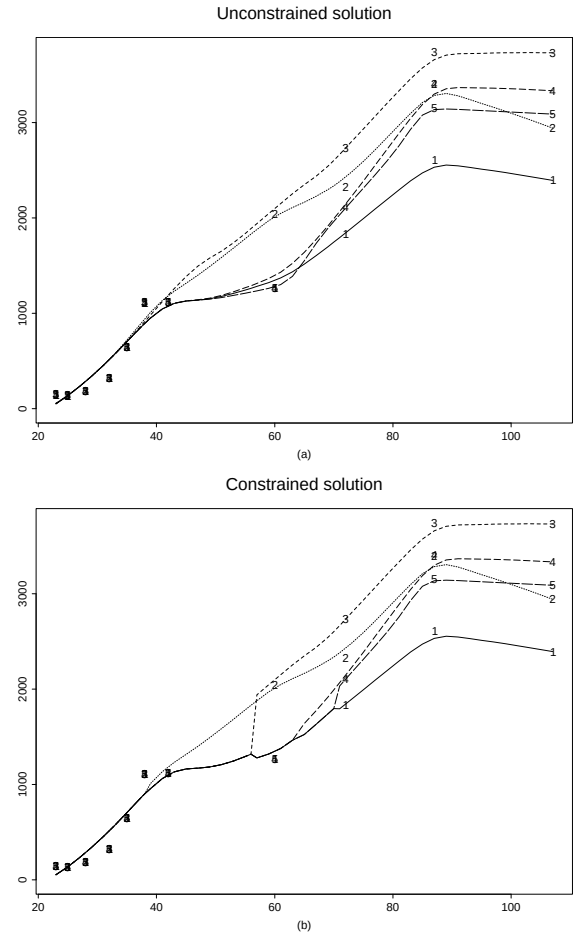


***Figure 3.3:*** *Panel (a) shows separate local linear fits to each of the groups. Panel (b) displays the resulting fits if equality up to the time point of treatment is imposed.*

# 4 Modifications

Using this approach, the final fits (e.g., those displayed in Figure 3.3(b)) have sometimes an "unsmooth" appearance. The reason for this becomes clear if one analyzes the way in which Goldfarb and Idnani's (1983) algorithm finds the (unique) optimal solution to the quadratic programming problem.

As mentioned above, the algorithm starts initially with the optimizer to the unconstrained problem (2.3). After obtaining this optimizer $\beta$ all constraints are checked. If one constraint is violated the minimizer $\beta$ is changed such that the constraint will hold while keeping its optimality property. This process is iterated until all constraints are satisfied. Each time a constraint is enforced the changes done to $\beta$ involve only those components which belong to any of the $t_j$ involved in the violated constraint. Hence, changes done to $\beta$ are somewhat "local" and enforcing of inequality and equality constraints can lead to "kinks" respectively "jumps" in the final solution. The problem of introducing kinks via inequality constraints is not obvious in Figure 3.2 but is noticeable if this approach is used to enforce that the final fit is monotone as we note in upcoming work.

While some smoothing problems suggest ad hoc methods to cope with this problem, a general approach is to introduce a "smoothness penalty" in (2.3). One obvious way of doing this is to change the minimization problem in (2.3) to

$$\hat{\mathbf{m}} = \mathbf{e}^T \operatorname*{argmin}_{\beta} (\mathcal{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathcal{Y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{g} \widehat{m''}(t_j)^2. \tag{4.1}$$

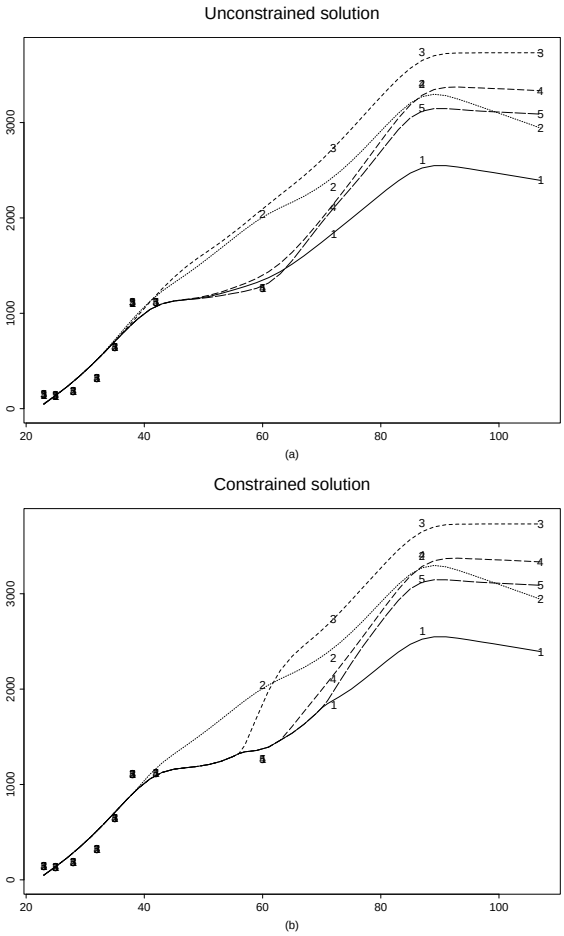


***Figure 4.1:*** *Panel (a) shows separate local linear fits to each of the groups. Panel (b) displays the resulting fits if equality up to the time point of treatment is imposed. Here (2.3) was changed to incorporate a smoothness penalty.*

If $\widehat{m''}(t_j)$ is calculated via second differences using the $\beta_{0,j}$s we can still write the resulting optimization problem as a quadratic programming problem.

Figure 4.1 shows the fits to the branching curve data if such a smoothing penalty is used. The same bandwidth and kernel as in Figure 3.3 are used. We see that the initial fits are not noticeably influenced by adding the smoothness penalty term whereas the constrained fits are markedly smoother.

The properties of the smoother stemming from (4.1) are not obvious and currently under investigation. If we chose $g = n$ and take the observation $X_i$ as grid points $t_j$ then, with $h \to 0$, (4.1) tends to a criterion similar to the one used in spline smoothing (see also Green and Silverman, 1994). Likewise, with $\lambda \to 0$ (4.1) tends to (2.3). Hence, the resulting smoother seems to be some kind of hybrid between local polynomial smoothing and spline smoothing.

Thus, if anything, the approach proposed here gives the user increased flexibility since he can choose the grid $\{t_j\}$ different from the set of observations $\{X_i\}$. Further flexibility can be achieved by using local bandwidths, i.e., different bandwidths at each $t_j$. In general, any modification to local polynomial smoothers used in "standard" nonparametric regression settings would be feasible to use as long as it does not destroy the quadratic programming structure of the problem.

We conclude that the methodology presented here can be used to construct natural local polynomial estimators which satisfy a variety of constraints. As long as these constraints are linear in the local polynomial estimator we are lead to a quadratic programming problem whose (unique) solution can be calculated without (numerical) problems using existing algorithms.

# References


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

Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association* **87**(420): 998–1004.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Application*, Vol. 66 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

Fisher, N.I., Hall, P., Turlach, B.A. and Watson, G.S. (1997). On the estimation of a convex set from noisy data on its support function, *Journal of the American Statistical Association* **92**(437): 84–91.

Gill, P.E., Murray, W., Saunders, M.A. and Wright, M.H. (1991). Inertia-controlling methods for general quadratic programming, *Siam Review* **33**(1): 1–36.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs, *Mathematical Programming* **27**: 1–33.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Hastie, T.J. and Loader, C.R. (1993). Local regression: automatic kernel carpentry (with discussion), *Statistical Science* **8**: 120–143.

Prince, J.L. and Willsky, A.S. (1990). Reconstructing convex sets from support line measurements, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**: 377–389.

Santaló, L.A. (1976). *Integral Geometry and Geometric Probability*, Addison–Wesley, Reading, Massachusetts.

Silverman, B.W. and Wood, J.T. (1987). The nonparametric estimation of branching curves, *Journal of the American Statistical Association* **82**: 551–558.

Steer, B.T. and Hocking, R.A. (1985). The optimum timing of nitrogen application to irrigated sunflowers, *Procceedings of the Eleventh International Sunflower Conference, Mar del Plata, Argentina*, Asociacion Argetina de Girasol, Buenos Aires, pp. 221–226.

Stone, C.J. (1977). Consistent nonparametric regression (with discussion), *Annals of Statistics* **5**: 595–645.

Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing*, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.