Statistical Science
2001, Vol. 16, No. 3, 232–248

A General Projection Framework for Constrained Smoothing

E. Mammen, J. S. Marron, B. A. Turlach and M. P. Wand

Abstract. There are a wide array of smoothing methods available for finding structure in data. A general framework is developed which shows that many of these can be viewed as a projection of the data, with respect to appropriate norms. The underlying vector space is an unusually large product space, which allows inclusion of a wide range of smoothers in our setup (including many methods not typically considered to be projections). We give several applications of this simple geometric interpretation of smoothing. A major payoff is the natural and computationally frugal incorporation of constraints. Our point of view also motivates new estimates and helps understand the finite sample and asymptotic behavior of these estimates.

Key words and phrases: Kernel smoothing, local polynomials, smoothing splines, constrained smoothing, monotone smoothing, additive models.

1. INTRODUCTION

Smoothing as a means of modeling nonlinear structure in data is enjoying increasingly widespread acceptance and use in applications. In many of these it is required that the curve estimates obtained from smoothing satisfy certain constraints; several such examples are discussed in Section 2. However, many of the usual formulations of smoothing are not very amenable to the incorporation of constraints. This is because it is not clear in which sense, if any, they are a *projection*, that is, the solution to a minimization problem with respect to some norm. In this paper we develop a framework in which a number of popular smoothing methods are *exactly* a projection with respect to a particular norm. Our framework is a product vector space that is larger than those usually considered for analyzing smoothing methods. The benefit of this type of geometric view of smoothing is that it reveals a natural way to incorporate constraints, since the modified smoother is defined as the projection onto the constrained set of functions.

The framework that we develop encompasses spline methods which are implicitly defined to be projections. Arguably, this is the reason why spline methods are considered to be the method of choice for constrained smoothing as they seem to incorporate many types of constraints in a natural way. In particular, smoothing splines are defined as minimizers of a penalized sum of squares, so constrained smoothing splines are easily defined as minimizers over the constrained set of functions.

Here we show that the essence of this idea is not restricted to smoothing splines, but applies quite generally, for example, to kernel and local polynomial methods. The key is to work with much larger normed vector spaces than are usually considered in the analysis of smoothers. Our framework, developed in Section 4, is a product structure, that is, we consider "vectors of objects," where the objects are functions, vectors or even sets of functions or vectors. In each case suitable norms are defined for our product space, which correspond to the sums of squares that are usually considered (see Section 3)

E. Mammen is Associate Professor, Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, 69120 Heidelberg, Germany (*e-mail: enno@statlab.uniheidelberg.de*). *J. S. Marron is Professor, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina, 27599-3260* (*e-mail: marron@stat.unc.edu*). *B. A. Turlach is Lecturer, Department of Mathematics and Statistics, University of Western Australia, Nedlands, WA 6907, Australia* (*e-mail: bturlach@maths.uwa.edu.au*). *M. P. Wand is Associate Professor, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115* (*e-mail: mwand@hsph.harvard.edu*).

and thus give a representation of the smoothers as projections. By this device a much broader class of smoothers can be viewed as projections, as shown in Section 4, which allows a natural incorporation of constraints for these methods.

In Section 5 our framework is seen to include smoothing splines and other penalized methods, through the development of Sobolev type norms on our general vector space. A number of asides are given in Section 6, including detailed discussion of the case of monotone smoothing, some remarks about loss functions, decompositions of sums of squares, numerical implementation and asymptotics. Extensions to local polynomials are given in Section 7. Application of our approach to additive models is discussed in Section 8.

For more background on smoothing, see any of a number of monographs: Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1966), Hart (1997), Bowman and Azzalini (1997), Efromovich (1999), Eubank (1999) and Loader (1999).

2. SOME EXAMPLES

2.1 Monotone Smoothing

A constraint that is frequently imposed on a regression curve by some physical or economic theory is monotonicity. As an illustration we use part of the "cars" data from the 1983 ASA Data Exposition. These data are available at the StatLib Internet site (http://lib.stat.cmu.edu/datasets/cars.data) at Carnegie Mellon University. The use of smoothing methods to model regression curves is illustrated in Figure 2.1. Here fuel efficiency, in miles per gallon, is studied as a function of engine output, in horsepower, and data points (X_i, Y_i) are displayed as a scatterplot. The curve in Figure 2.1 is a simple smooth, that is, moving average, as described in (3.1).

This smooth is not monotonically decreasing. But since one expects that more powerful engines consume more fuel, it is sensible to request that the smooth be decreasing. The result of using the sophisticated projection idea of Section 5.1 is shown in Figure 2.2. Starting with the simple smooth in Figure 2.1 that smooth is projected onto the space of monotone functions using a (discretized) Sobolev norm. Note that essentially the increasing parts of the smooth have been "rounded off."

2.2 Parallel Regression Curves

Ratkowsky (1983) provides data from an experiment on the relation between yield of onion plants and the density of planting. The measurements have been taken at two different locations in South Australia, Purnong Landing and Virginia. A scatterplot of the data shows clear differences in yield between the two locations.

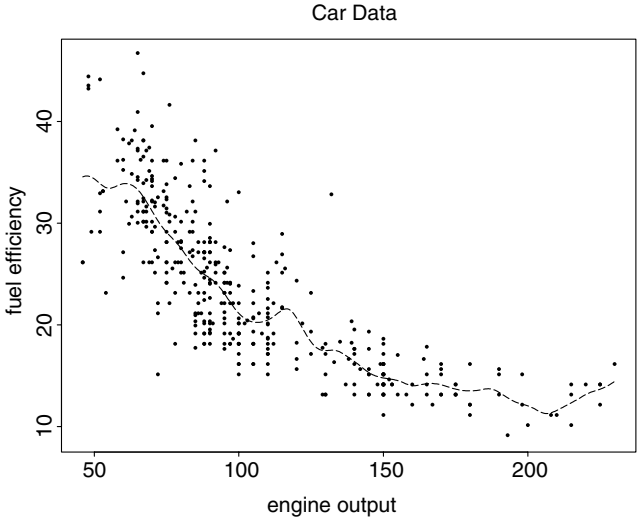

FIG. 2.1. *Raw data and simple smooth for fuel efficiency as a function of engine output. Smooth is Nadaraya–Watson type with Gaussian kernel and bandwidth* $h = 4$.

Bowman and Azzalini (1997, Chapter 6.5) revisit these data and investigate whether one can reasonably assume that regression curves fitted to the data from each location, using the logarithm of yield as response variable and density of planting as regressor, would be parallel. They adapted a two-stage method proposed by Speckman (1988) to obtain two parallel regression smooths. This

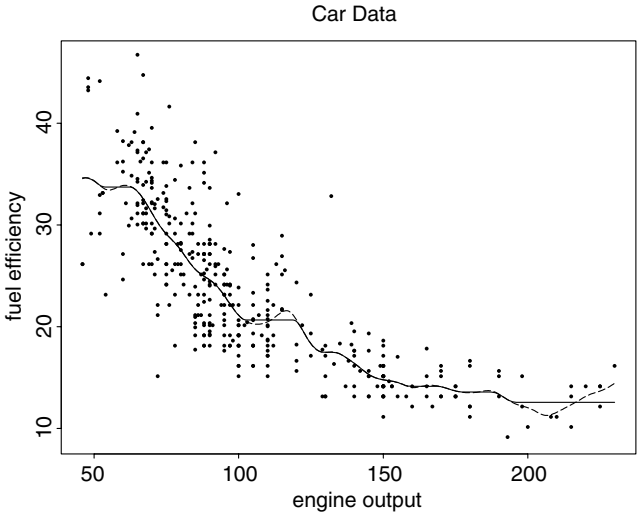

FIG. 2.2. *Raw data and monotonicity constrained smooth for fuel efficiency as a function of engine output. Smooth is Nadaraya–Watson type with Gaussian kernel and bandwidth* $h = 4$.

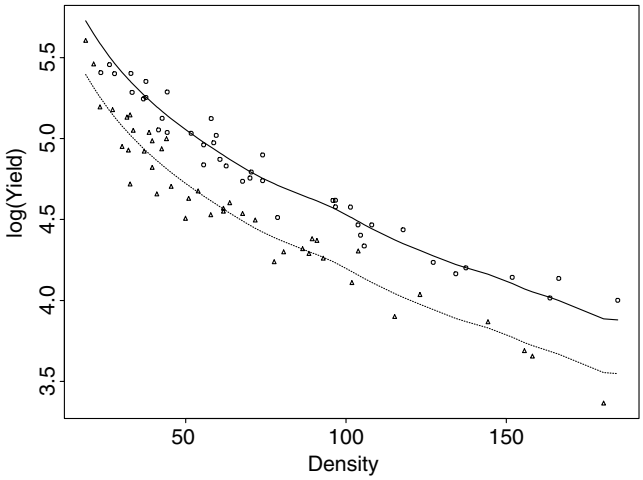

FIG. 2.3. *Raw data for the onions data. Circles and triangles denote the measurements taken at Purnong Landing and Virginia, respectively. The parallel smooths are of local linear type with Gaussian kernel, bandwidth* $h = 11$ *and* $\hat{\alpha} = 0.332$ *apart.*

procedure first calculates a low bias estimate $\hat{\alpha}$ of the distance α between the regression curves using a low bias (i.e., a small bandwidth) regression smoother. During the second step the final curve estimates are obtained by using $\hat{\alpha}$ from the first step and a "more reasonable" smoothing parameter for the regression smoother.

The constraint that two (or more) regression curves should be parallel is easily handled by the framework developed here. First we fit separate simple smooths to each of the groups. Next these smooths are projected into the space of parallel (smooth) curves. The distance between the two curves is chosen such that the residual sum of squares is minimized. The result of this procedure applied to the onions data is shown in Figure 2.3. In this case the distance between the two curves is estimated to be $\hat{\alpha} = 0.332$, a value similar to the one found by Bowman and Azzalini (1997).

2.3 Support Function of a Convex Set

This example addresses a problem that arises variously in medical imaging and robotic vision. Given noisy measurements of the support function of a convex set, the aim is to estimate the boundary of that set. The support function $m(\theta)$, $\theta \in (-\pi, \pi]$, is defined relative to a given origin O and a fixed direction in the plane. More precisely, the support function $m(\theta)$ of a convex set $\mathscr{C}$, is defined as the perpendicular distance from O to that tangent to $\mathscr{C}$ that has angle θ with the given direction. This definition is illustrated in Figure 2.4. If m'' exists and is continuous everywhere, then a necessary and sufficient condition for convexity of $\mathscr{C}$ is (Santaló, 1976, page 2)

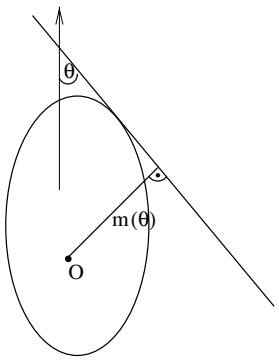

FIG. 2.4. *Illustration of the definition of a support function.*

$$m(\theta) + m''(\theta) \geq 0, \qquad \theta \in (-\pi, \pi]. \tag{2.1}$$

A common method for constructing an estimate of the set from data on the support function 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).

Fisher, Hall, Turlach and Watson (1997) propose to estimate m nonparametrically using a kernel smoother. Essentially, they search for the bandwidth that produces the smooth with least bias for which (2.1) holds. This approach is quite successful if the set does not exhibit strong eccentricity. Such eccentricity corresponds to "peaks" and "troughs" in the support function and the bandwidth chosen by Fisher et al. (1997) could lead to serious undersmoothing of the former and oversmoothing of the latter. An additional complication is that their method only guarantees finding an appropriate bandwidth with probability tending to 1. For any given data set it is possible that their method fails.

With the approach developed in this paper such careful calibration of the bandwidth is not necessary. In fact, we are at liberty to fit an initial fit that violates (2.1) but has low bias. This smooth is then projected into the space of functions that satisfy (2.1).

Figure 2.5 demonstrates our method on a simulated data set. We chose as a convex set an ellipse with major axes of one and four units. [Note that for this example the method of Fisher et al. (1997) has considerable bias problems.] Figure 2.5(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).

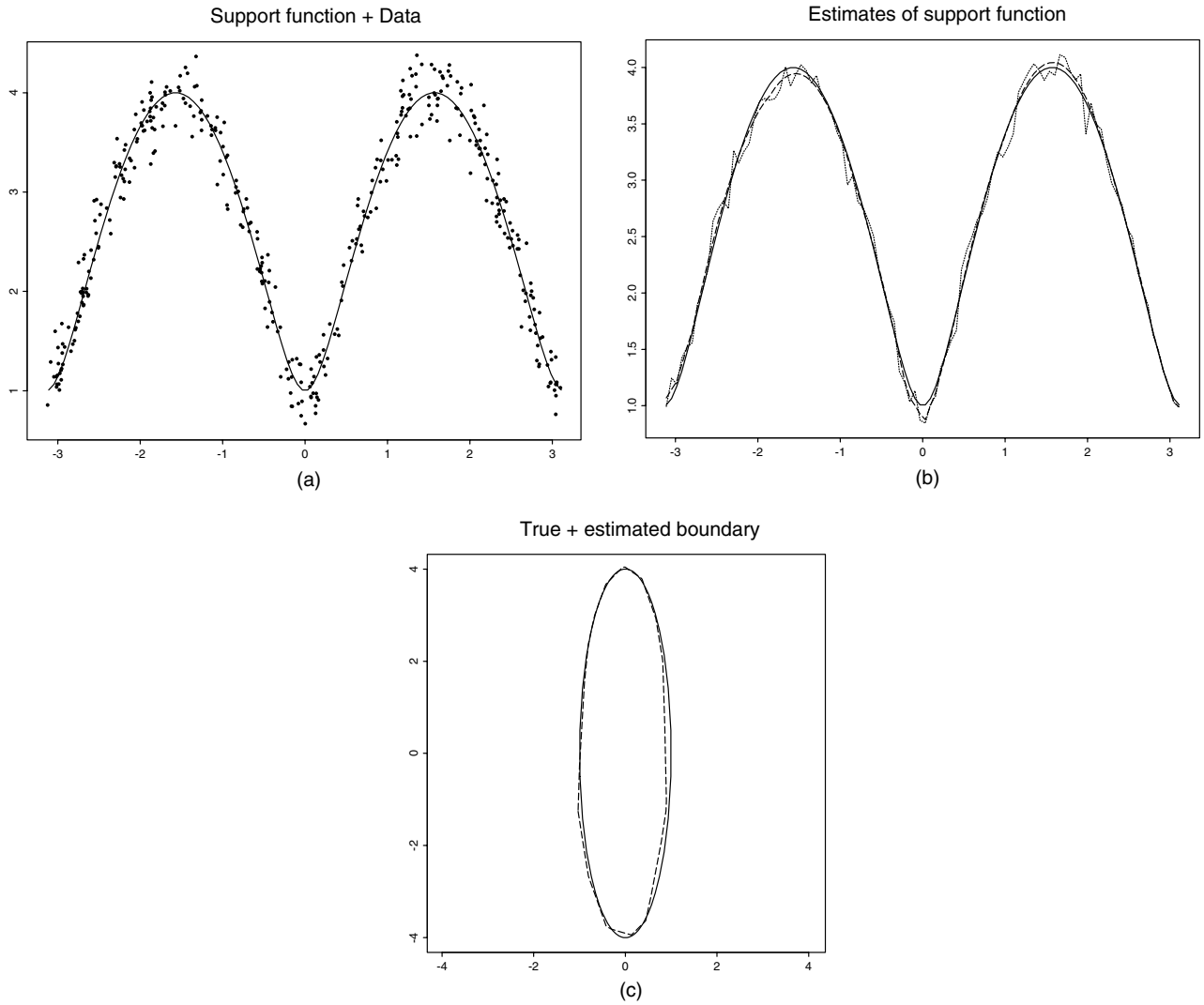

FIG. 2.5. *Panel* (a) *shows the support function of the ellipse shown in panel* (c) *together with the* (*simulated*) *observed data. Panel* (b) *shows the support function* (*solid line*) *together with the initial fit to the data* (*dotted line*) *and the resulting constrained fit* (*dashed line*). *The boundary corresponding to the constrained fit is shown as the dashed line in panel* (c).

The resulting fit which satisfies (2.1) 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.

2.4 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 have been analyzed by Silverman and Wood (1987) using spline smoothing techniques (see also Green and Silverman, 1994, Section 6.2).

The experiment included 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 at which the nitrogen compound was applied there is no difference between the control group and the treatment group(s). Hence, when fitting regression curves to these data, it is natural to impose that for each treatment group the regression curve will coincide with the curve of the control group up to the time of treatment.

It is straightforward to apply our framework to this smoothing problem. The result, using a local linear fit, employing the Gaussian kernel and bandwidth $h = 8$, and using the sophisticated projection idea of Section 5.1 is shown in Figure 2.6(b). Panel (a) shows the initial, unconstrained fits.

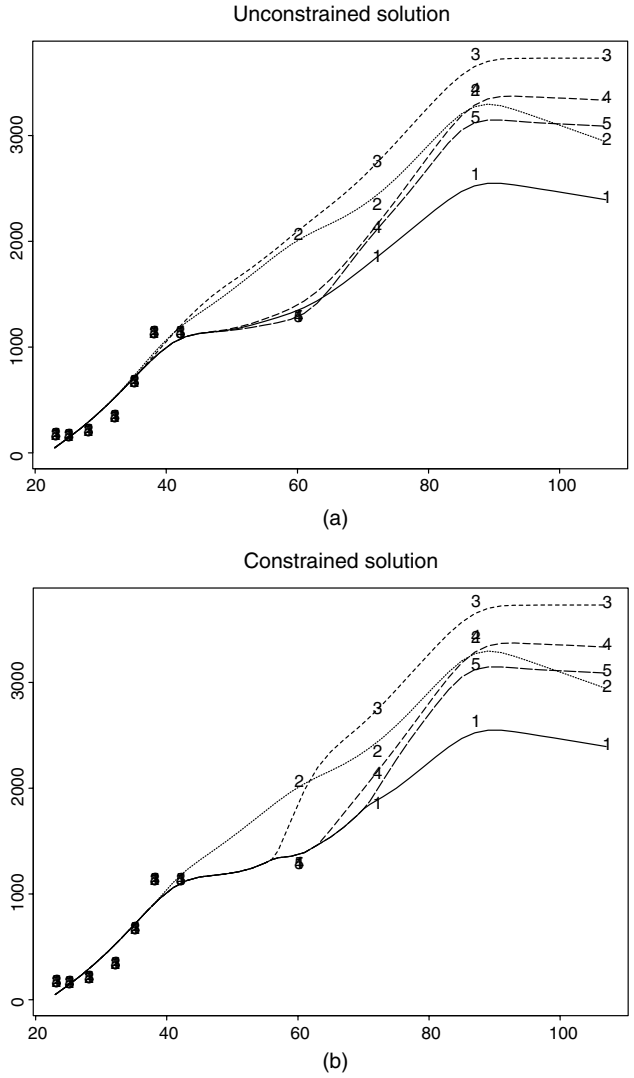

FIG. 2.6. *Panel* (a) *shows separate local linear fits to each of the groups in the sunflower data. Panel* (b) *displays the resulting fits if equality up to time of treatment is imposed.*

3. SIMPLE SMOOTHING AS MINIMIZATION

Before developing our general vector space framework, we first show how simple smoothing, as shown in Figure 2.1, can be written as a minimization problem. Then we show how this viewpoint can be used to do constrained smoothing. A mathematical formulation of smoothing has data $(X_1, Y_1), \ldots, (X_n, Y_n)$, for example, as shown in the scatterplot of Figure 2.1, that are modeled as

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where ε_i, $i = 1, \ldots, n$, are mean 0 error random variables and m is some smooth regression function.

The dashed curve in Figure 2.1 is a "simple smooth" of the form

$$\hat{m}_S(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}, \tag{3.1}$$

that is, a moving (in x) weighted average of the Y_i. The weights $w_i(x)$ used in Figure 2.1 are of Nadaraya–Watson type, as discussed in Section 4.1. See Härdle (1990) and Wand and Jones (1995) for an introduction to the basics of this nonparametric regression estimator.

Note that there are several points where the simple smooths shown in Figures 2.1, 2.5(b) and 2.6(a) do not fulfill the desired constraints. An approach to constraining this type of smooth to satisfy given constraints is to recognize that it can be written as

$$\hat{m}_S = \arg\min_m \int \frac{1}{n} \sum_{i=1}^n \{Y_i - m(x)\}^2 \times w_i(x)\, \nu(dx), \tag{3.2}$$

where $\int$ means definite integration over the real line, and where ν is some measure. A natural choice is $\nu(dx) = dx$, corresponding to Lebesgue integration. However, other measures such as some form of counting measure might also be considered [e.g., $\nu(dx) = dF_n(x)$ where F_n is the empirical distribution]. For the unconstrained estimator the integration with respect to $\nu(dx)$ has no effect because the minimum can be found for each x individually, that is,

$$\hat{m}_S(x) = \arg\min_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (Y_i - m)^2 w_i(x). \tag{3.3}$$

But the integral is included because it reveals that simple smoothing is a projection as developed below. This is the key to our natural formulation of constrained smoothing. If C is a set of functions satisfying some constraint, such as those considered in Section 2, then a constrained version of the simple smooth is:

$$\hat{m}_{S,C} = \arg\min_{m \in C} \int \frac{1}{n} \sum_{i=1}^n \{Y_i - m(x)\}^2 \times w_i(x)\, \nu(dx). \tag{3.4}$$

In contrast to the unconstrained case, the weight measure ν is now no longer negligible because the minimizers at different points x are linked through the constraints. To calculate the constrained smooth in practice, one would typically choose ν to be the counting measure on the points at which one wants to evaluate the (constrained) smooth. For theoretical analysis, a natural choice would be the Lebesgue measure. In Figure 3.1, a discretized version of Lebesgue measure is used, that is, a grid

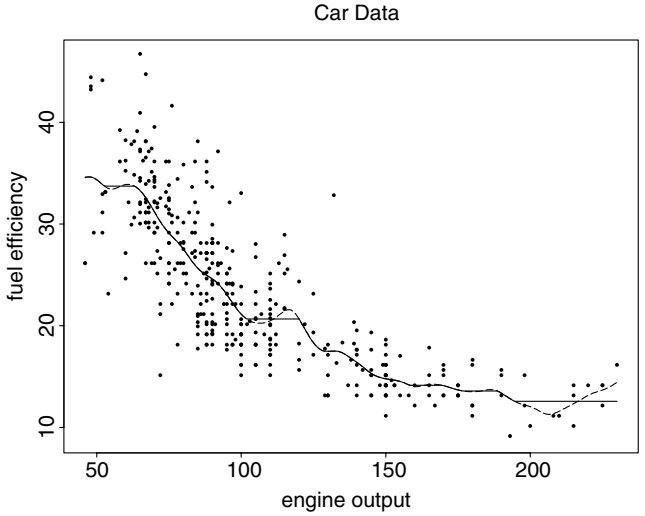

FIG. 3.1. *Unconstrained and constrained (monotone) smooths, for fuel efficiency as a function of engine output, as in Figure* 2.1. *The constrained smooth has "kinks" which have been smoothed out in the more sophisticated constrained smooth of Figure* 2.2.

of equidistant points was selected at which the estimate was evaluated.

While this estimate appears natural, the monotonicity constraint introduces some "kinks" in Figure 3.1, essentially at "break points where $\hat{m}_S$ is not monotone." Insight into these kinks and other aspects of constrained smoothing comes from a particular normed vector space structure that will be introduced in the next section. See Section 6.1 for further discussion, and methods to "round off these corners" as shown in Figure 2.2.

4. SIMPLE SMOOTHERS VIEWED AS PROJECTIONS

In this section we shall introduce a normed vector space, say $\mathcal{Y}_S$, that contains the space of data vectors and the space of candidate regression functions. Within this vector space we shall identify the subspace of data vectors Y, say $\mathcal{Y}_S^Y$, and the subspace of candidate regression functions, say $\mathcal{Y}_S^m$. These two subspaces contain all the information relevant to the smoothing problem at hand and hence reflect the full structure of smoothing.

It will become apparent that in this space simple smoothers appear as a projection of the data vector into the subspace $\mathcal{Y}_S^m$. To capture all of these aspects, it is not enough to simply work with n-dimensional vectors or with functions. A vector space which includes both the data vector Y and the candidate smooths $m(x)$ is a product space containing n-tuples of linear objects,

$$\mathcal{Y}_S = \left\{ \underset{\rightarrow}{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} : v_i \in V,\ i = 1, \ldots, n \right\},$$

where V is some normed vector space. Note that this space $\mathcal{Y}_S$ contains many more elements besides those in $\mathcal{Y}_S^Y$ and $\mathcal{Y}_S^m$ (both defined below). However, only elements in these two subspaces are relevant for the smoothing problem and have direct interpretations. If the vector space V is chosen appropriately then $\mathcal{Y}_S$ will be a vector space that contains these two subspaces of interest and in which the simple smooth can be interpreted as a projection of the data vector.

The choice of V will vary depending on the type of smoother considered. When the result of the smooth is a function, as in the rest of this section, and in Section 5, V will be an appropriate space of functions. But when the result of the smooth is a vector, for example, when the smooth is evaluated only at the design points, V is a set of ordinary vectors.

For the rest of this section, we shall consider V to be a space of functions, so

$$\mathcal{Y}_S = \left\{ \underset{\rightarrow}{f} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} : f_i \colon \mathbb{R}^q \to \mathbb{R},\ i = 1, \ldots, n \right\}.$$

The data vector $Y = [Y_1, \ldots, Y_n]$ can be viewed as an element $\underset{\rightarrow}{Y}$ of $\mathcal{Y}_S$, which is an n-tuple of constant functions, $\overrightarrow{f_i(x)} \equiv Y_i$, $i = 1, \ldots, n$. The subspace of such n-tuples of constant functions will be called $\mathcal{Y}_S^Y$. For a candidate smooth $m \colon \mathbb{R}^q \to \mathbb{R}$, we write $\overrightarrow{m}$ for the n-tuple where each entry is $m(x)$, that is, $f_i(x) \equiv m(x)$, $i = 1, \ldots, n$. The subspace of such n-tuples with identical entries is denoted by $\mathcal{Y}_S^m$. When $w_i(x) \geq 0$, we may define an inner product on $\mathcal{Y}_S$,

$$\langle \underset{\rightarrow}{f}, \underset{\rightarrow}{g} \rangle = \int \frac{1}{n} \sum_{i=1}^{n} f_i(x) g_i(x) w_i(x)\, \nu(dx),$$

and its induced norm on $\mathcal{Y}_S$ is given by

$$\| \underset{\rightarrow}{f} \|^2 = \int \frac{1}{n} \sum_{i=1}^{n} f_i(x)^2 w_i(x)\, \nu(dx). \tag{4.1}$$

Strictly speaking, this defines only a bilinear form and a seminorm if, for some i, $w_i(x) = 0$ on a set of x whose ν-measure is not zero (which happens, e.g., for kernel smoothing with a compactly supported kernel). By identifying functions that are equivalent under this seminorm we can view (4.1) as a norm, that is, implicitly we work on classes of functions. We shall also assume that $\mathcal{Y}_S$ is complete with

respect to this norm (which is possible by specifying an appropriate space for the f_i in the definition of $\mathcal{V}_S$).

This notation shows that both the unconstrained and constrained simple smooths are projections, because (3.2) and (3.4) can be rewritten as

$$\hat{m}_S = \operatorname*{arg\,min}_{m:\ \underrightarrow{m} \in \mathcal{V}_S^m} \left\| \underrightarrow{Y} - \underrightarrow{m} \right\|^2, \tag{4.2}$$

$$\hat{m}_{S,C} = \operatorname*{arg\,min}_{m:\ \underrightarrow{m} \in \mathcal{C}_S^m} \left\| \underrightarrow{Y} - \underrightarrow{m} \right\|^2, \tag{4.3}$$

where $\mathcal{C}_S^m \subset \mathcal{V}_S^m$ is the subset of n-tuples with (identical) entries that are constrained, for example monotone in x.

Using a Pythagorean relationship, the minimization problem (4.3) can be substantially simplified. This yields important computational advantages, and also gives some important insights. In particular, for $\underrightarrow{m} \in \mathcal{V}_S^m$ we have

$$\left\| \underrightarrow{Y} - \underrightarrow{m} \right\|^2 = \left\| \underrightarrow{Y} - \underrightarrow{\hat{m}}_S \right\|^2 + \left\| \underrightarrow{\hat{m}}_S - \underrightarrow{m} \right\|^2, \tag{4.4}$$

because $\underrightarrow{\hat{m}}_S$ is the projection of $\underrightarrow{Y}$ onto the subspace $\mathcal{V}_S^m$, whence $\underrightarrow{Y} - \underrightarrow{\hat{m}}_S$ is orthogonal to $\underrightarrow{\hat{m}}_S - \underrightarrow{m}$ with respect to the inner product; see, for example, Rudin (1987, Theorem 4.11). Furthermore,

$$\begin{aligned} \left\| \underrightarrow{\hat{m}}_S - \underrightarrow{m} \right\|^2 &= \int \frac{1}{n} \sum_{i=1}^{n} [\hat{m}_S(x) - m(x)]^2 w_i(x)\, \nu(dx) \\ &= \int [\hat{m}_S(x) - m(x)]^2 w(x)\, \nu(dx), \end{aligned}$$

where $w(x) = \frac{1}{n} \sum_{i=1}^{n} w_i(x)$. An immediate consequence of this is the following proposition.

PROPOSITION 1. *Assuming that each $w_i(x) \geq 0$, the constrained simple smooth can be represented as a constrained minimization over ordinary functions (i.e., over $m \in C$) as*

$$\begin{aligned} \hat{m}_{S,C}(x) &= \operatorname*{arg\,min}_{m:\ \underrightarrow{m} \in \mathcal{C}_S^m} \left\| \underrightarrow{\hat{m}}_S - \underrightarrow{m} \right\|^2 \\ &= \operatorname*{arg\,min}_{m \in C} \int \{\hat{m}_S(x) - m(x)\}^2 \\ &\qquad \times w(x)\, \nu(dx). \end{aligned} \tag{4.5}$$

The geometric interpretation of Proposition 1 is that the projection of the data vector Y onto $\mathcal{C}_S^m$, (in our enlarged vector space $\mathcal{V}_S$) is the same as the projection (in the space of ordinary functions) of the unconstrained smooth onto C.

The relation (4.4) and similar geometric considerations give other types of insight about constrained smoothing. It is straightforward to check that the orthogonality used in the Pythagorean theorem (4.4) follows from direct calculation of

$$\left\langle \underrightarrow{Y} - \underrightarrow{\hat{m}}_S, \underrightarrow{\hat{m}}_S - \underrightarrow{m} \right\rangle = 0.$$

At first glance, one might suspect that the subspaces $\mathcal{V}_S^Y$ and $\mathcal{V}_S^m$ are orthogonal, but they are not, because they have the intersection $\mathcal{V}_S^C$, the n-tuples of constant functions that are all the same. However, even $\mathcal{V}_S^Y \cap (\mathcal{V}_S^C)^\perp$ (the orthogonal complement of $\mathcal{V}_S^C$ in $\mathcal{V}_S^Y$) and $\mathcal{V}_S^m \cap (\mathcal{V}_S^C)^\perp$ are not orthogonal, as can be seen from direct calculation, or from the fact that this would imply that the projection of Y onto $\mathcal{V}_S^m$ lies in $\mathcal{V}_S^C$ and thus is everywhere constant.

Visual understanding of Proposition 1 is given by Figure 4.1. The horizontal plane represents the subspace $\mathcal{V}_S^m$ of $\mathcal{V}_S$. The diagonal line represents the subspace $\mathcal{V}_S^Y$ (*not* orthogonal to $\mathcal{V}_S^m$). The set $\mathcal{C}_S^m$ is shown as the shaded horizontal region. Proposition 1 states that the point in $\mathcal{C}_S^m$ that is closest to Y is also the point in $\mathcal{C}_S^m$ that is closest to $\hat{m}_S(x)$.

Proposition 1 also suggests which statistical loss functions are associated with choices of the weight measure ν. In particular, if $m_0(x)$ is the "true" function, then the loss (conditional on $X_1, \ldots, X_n$) function

$$L(\hat{m}, m_0) = \int \{\hat{m}(x) - m_0(x)\}^2 w(x)\, \nu(dx) \tag{4.6}$$

is essentially optimized by $\hat{m}_S(x)$ over $\mathcal{V}_S^m$ and by $\hat{m}_{S,C}(x)$ over $\mathcal{C}_S^m$. Specifics of L are discussed in Section 6.2.

Proposition 1 shows that the constrained estimate can be calculated in two relatively straightforward steps:

1. Compute the unconstrained estimate $\hat{m}_S$.

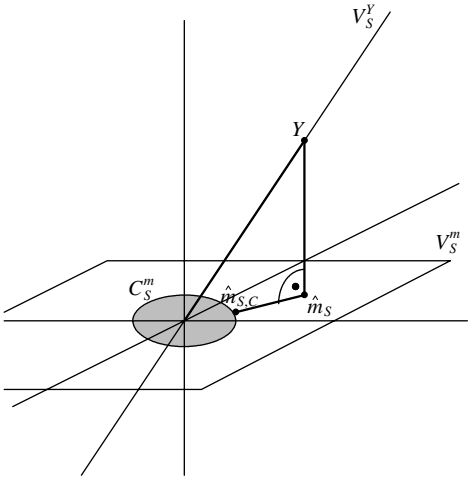

FIG. 4.1. *Diagram representing location of data and unconstrained and constrained smooths, in the vector space $\mathcal{V}_S$.*

2. Project $\hat{m}_S$ onto the constrained set of functions.

Implementation of each of these two steps is relatively straightforward and much simpler than direct computation of (3.4). We shall come back to this point in Section 6.4.

4.1 Some Remarks and Specific Simple Smoothers

Representations of the type (3.2) have been used for many purposes. For example they provide easy understanding of how local polynomial methods, discussed in detail in Section 7, extend conventional kernel smoothers (see Fan and Gijbels, 1996). A different purpose is the motivation of "robust M-smoothing" as introduced in Härdle and Gasser (1984) and Tsybakov (1986), where the square in (3.2) is replaced by a "robust ρ function." Application of our approach to these smoothers will not be discussed here.

It is straightforward to show that Proposition 1 still holds when some of the $w_i(x) < 0$, as long as $w(x) \geq 0$. This is important in the following.

Here are some specifics to show that many types of smoothers can be written in the form (3.1), that is, (3.2). Much of this approach to generality was developed by Földes and Révesz (1974) and Walter and Blum (1979) in the context of density estimation.

1. *Nadaraya–Watson smoother.* Here the weight functions have the form

$$w_i(x) = K_h(x - X_i),$$

where K is a nonnegative, integrable "kernel function" or "window function" (often taken to be a symmetric probability density), and where the "bandwidth" or "smoothing parameter" h controls the amount of smoothing, that is, local averaging, via $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$.

2. *Gasser–Müller smoother.* This is a somewhat different "kernel type" smoother, where

$$w_i(x) = \int_{s_{i-1}}^{s_i} K_h(x - t)\,dt,$$

for "in between points" s_i, where $s_0 < X_1 \leq s_1 < X_2 \leq \cdots \leq s_{n-1} < X_n \leq s_n$. See Müller (1988) for discussion of many properties of this estimator. See Chu and Marron (1991) for comparison of this smoother with the Nadaraya–Watson smoother.

3. *Bandwidth variation.* Our geometric approach extends to the case that the bandwidth h depends on x, for example, $w_i(x) = K_{h(x)}(x - X_i)$ in the case of Nadaraya–Watson smoothing.

4. *Orthogonal Series.* For an orthogonal basis $\{\psi_j\}$, for example, the Fourier basis, or a wavelet basis, a simple class of smoothers is

$$\hat{m}_{OS}(x) = \sum_{j \in S} \hat{\theta}_j \psi_j(x), \tag{4.7}$$

where the "empirical Fourier coefficients" are $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^n Y_i \psi_j(X_i)$, and where S is some set of "coefficients containing most of m_0," for example, low frequency coefficients in the Fourier case or unthresholded coefficients in the wavelet case. Interchanging the order of summation shows that this type of smoother is of the form (3.1) where

$$w_i(x) = \frac{1}{n}\sum_{j \in S} \psi_j(X_i)\psi_j(x).$$

A short description of orthogonal series estimates, including wavelets, can be found in Section 3.2 of Ramsay and Silverman (1997) where additional references are given for particular choices of function bases.

5. *Regression splines.* A class of simple smoothers with a form that is related to (4.7) is the class of regression splines

$$\hat{m}_{RS}(x) = \sum_{j \in S} \hat{\theta}_j B_j(x),$$

but the functions $B_j(x)$ are no longer orthogonal. Now they take the form $B_j(x) = x^j$, for $j = 1, \ldots, p$ and $B_j(x) = (x - k_j)_+^p$ for $j > p$, where the k_j are some given "knot points." The coefficients $\hat{\theta}_j$ are computed by least squares, so they are still linear combinations of Y. Thus this type of smoother can be written in the form (3.1) by interchanging order of summation as above. See Section 7.2 of Eubank (1999) for discussion of many properties of estimators of this form and see Stone, Hansen, Kooperberg and Truong (1997) for related estimators in more complicated models.

6. *Others.* A variation on kernel type smoothers is local polynomials, which are discussed in detail in Section 7. A different type of spline is the smoothing spline discussed in detail in Section 5.

5. EXTENSION TO SMOOTHING SPLINES

Much of the work in constrained nonparametric regression has been done in the context of splines. Smoothing splines are defined as minimizers of a penalized sum of squares; see (5.1). Constraints can be easily incorporated by minimizing over the restricted set. Tantiyaswasdikul and Woodroofe (1994) consider the case where the smoothing order

parameter p [see (5.1)] is equal to 1, but most of the work seems to concentrate on the case $p = 2$. Theoretical properties of constrained smoothing splines are discussed by Utreras (1985), Micchelli and Utreras (1988) and Mammen and Thomas-Agnan (1999).

Algorithms for calculating constrained smoothing splines typically use specially constructed bases of spline functions. If the basis is appropriately chosen then the constraints on the shape of the curve estimate correspond to "simple" constraints on the coefficient of each basis function. These coefficients can then be determined by solving a constrained least-squares problem. Thus, most algorithms are tailored for specific shape constraints, for example, the regression curve is supposed to be monotone (Ramsay, 1988; Gaylord and Ramirez, 1991; Dole, 1999), convex (Dierckx, 1980; Schmidt, 1987; Elfving and Andersson, 1988; Irvine, Marin and Smith, 1986; Dole, 1999), or convex–concave (Schmidt and Scholz, 1990).

A flexible algorithm that uses B-splines and can be easily adapted to a variety of shape constraints is given in Schwetlick and Kunert (1993). An algorithm that starts with the unconstrained solution of (5.1) and adaptively enforces constraints until the smooth fulfills the desired shape constraints is proposed in Turlach (1997). Villalobos and Wahba (1987) consider a bivariate constrained smoothing problem. They use a bivariate spline smoother and enforce the constraints on a fine grid of points. Even though this does not guarantee that the fitted smooth will satisfy the shape constraints everywhere, one can be reasonably sure that it does if the grid on which one imposes the constraints is fine enough.

Applications of constrained splines smoothing are discussed, among others, by Ramsay (1988), Kelly and Rice (1990), and in the books by Wahba (1990) and Green and Silverman (1994). A recent overview on work on shape restricted splines is given in Delecroix and Thomas-Agnan (2000).

Despite this impressive amount of work on constrained smoothing splines no unified framework has emerged. Further insight into how constrained smoothing splines work comes from another type of generalization of the framework of Section 3. The basic smoothing spline of order p is usually written as

$$\hat{m}_{SS}(x) = \arg\min_m \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2 + \lambda \int m^{(p)}(x)^2, \tag{5.1}$$

where λ is the smoothing parameter. See Wahba (1990), Green and Silverman (1994), and Eubank (1999) for discussion of many aspects of this estimator. It can be written in a form which generalizes both (4.1) and (5.1) as

$$\hat{m}_{SS}(x) = \arg\min_{m:\ \underrightarrow{m} \in \mathscr{Y}_S^m} \left\| \underrightarrow{Y} - \underrightarrow{m} \right\|^2$$

where the norm on $\mathscr{Y}_S$ is now generalized to

$$\left\| \underrightarrow{f} \right\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|f_i(x)\|_p^2, \tag{5.2}$$

where $\|\cdot\|_p$ denotes the Sobolev type norm

$$\|f(x)\|_p^2 = \int [f(x)]^2 w_i(x)\, \nu(dx) + \lambda \int \left[f^{(p)}(x)\right]^2 dx.$$

The conventional smoothing spline (5.1) is the special case where $w_i(x) = 1$ and ν is the empirical measure of the design points $X_1, \ldots, X_n$. The norm (4.1) is the special case where $\lambda = 0$.

As above, it is natural to write constrained smoothing splines as

$$\hat{m}_{SS,C}(x) = \arg\min_{m:\ \underrightarrow{m} \in \mathscr{C}_S^m} \left\| \underrightarrow{Y} - \underrightarrow{m} \right\|^2$$

This constrained minimization is simplified, exactly as at (4.4), using a Pythagorean relationship. Following the arguments of Section 4 yields the proposition.

PROPOSITION 2. *The constrained smoothing spline can be represented as a constrained minimization over ordinary functions as*

$$\begin{aligned} \hat{m}_{SS,C}(x) &= \arg\min_{m:\ \underrightarrow{m} \in \mathscr{C}_S^m} \left\| \underrightarrow{\hat{m}}_{SS} - \underrightarrow{m} \right\|^2 \\ &= \arg\min_{m \in C} \int \{\hat{m}_{SS}(x) - m(x)\}^2 w(x)\, \nu(dx) \\ &\qquad + \lambda \int \left\{\hat{m}_{SS}^{(p)}(x) - m^{(p)}(x)\right\}^2 dx. \end{aligned} \tag{5.3}$$

Proposition 2 is proved in Mammen and Thomas-Agnan (1999). There this representation of the smoothing spline was used to study asymptotics and algorithms for shape restricted smoothing splines; see also Section 6.5.

5.1 Sobolev Projection of Smoothers

Motivated by Proposition 2 we propose to mix ideas from spline smoothing and other smoothing approaches. We consider the following class of constrained smoothers. For an arbitrary (unconstrained) smoother $\hat{m}_S$ that is constructed such

that it has p derivatives we define the constrained smoother as

$$\hat{m}_{S,C}(x) = \underset{m:\ \underrightarrow{m} \in \mathscr{C}_S^m}{\arg\min} \left\| \underrightarrow{\hat{m}}_S - \underrightarrow{m} \right\|^2$$

$$= \underset{m \in C}{\arg\min} \int \{\hat{m}_S(x) - m(x)\}^2 w(x)\, \nu(dx) + \lambda \int \{\hat{m}_S^{(p)}(x) - m^{(p)}(x)\}^2 dx.$$

This means that the constrained smoother $\hat{m}_{S,C}$ is the projection of the unconstrained estimator $\hat{m}_S$ onto the constrained set C. Here, the projection is taken with respect to the Sobolev norm

$$\|f\|^2 = \int f(x)^2 w(x)\, \nu(dx) + \lambda \int f^{(p)}(x)^2\, dx. \tag{5.4}$$

This estimate has two advantages:

1. The unconstrained estimate $\hat{m}_S$ will only be changed if it violates any of the constraints and then only in the neighborhood of this violation. In particular, for monotone smoothing $\hat{m}_S$ will only be changed in neighborhoods of sets where the monotonicity was violated by $\hat{m}_S$. Hence, away from such neighborhoods the constrained estimate has the same (theoretical) properties as the unconstrained estimator since it is identical to the latter. More important, the good interpretability of the unconstrained estimator carries over to the constrained estimator away from such neighborhoods.
2. The constrained estimate $\hat{m}_{S,C}$ is a smooth function. The reason is that the penalty term $\lambda \int [m^{(p)}(x)]^2\, dx$ of the Sobolev norm forces $\hat{m}_{S,C}$ to be smooth. In particular, for monotone smoothing with a choice $p \geq 1$ we get an estimate that is differentiable. This means that this estimate does not have the kinks observed in Figure 3.1 for monotone local linear fits. This is shown in Figure 2.2 where the constrained smoother of Figure 2.1 is shown. That projection is calculated with respect to (5.4) where the penalty term has been replaced by a discretized version. This has been done for computational reasons. For a more detailed discussion of algorithms using local polynomial smoothers see Mammen, Marron, Turlach and Wand (2001). Delecroix, Simioni and Thomas-Agnan (1995, 1996) consider a related two-step procedure for Priestley–Chao type kernel smoothers.

6. ASIDES

6.1 The Monotone Case

For monotone smoothing, $\hat{m}_{S,C}(x)$ is a version of the older idea of "smooth, then monotonize" discussed, for example, in Barlow and van Zwet (1970), Wright (1982), Friedman and Tibshirani (1984), Mukerjee (1988), Kelly and Rice (1990) and Mammen (1991a) (see also Cheng and Lin, 1981; Ramsay, 1998; Mammen et al., 2001). Moreover, to our knowledge, the fact that $\hat{m}_{S,C}$ is the projection onto a constrained set has not been recognized before.

It can be shown that for monotone (increasing) smoothing, (4.5) implies that

$$\hat{m}_{S,C}(x) = \max_{u \leq x} \min_{v \geq x} \frac{\int_u^v \hat{m}_S(s) w(s)\, \nu(ds)}{\int_u^v w(s)\, \nu(ds)}. \tag{6.1}$$

A proof of (6.1) for discrete measures ν can be found in the books by Barlow, Bartholomew, Bremner and Brunk (1972) or Robertson, Wright and Dykstra (1988). The case of general ν is discussed in Mammen et al. (2001). A careful inspection of (6.1) shows that one obtains the monotone function $\hat{m}_{S,C}$ from $\hat{m}_S$ by replacing parts of $\hat{m}_S$ by constant pieces. In an interval where $\hat{m}_{S,C}$ is constant it is equal to a weighted average of $\hat{m}_S$ over this interval. At the boundary of such intervals $\hat{m}_{S,C}$ may not be differentiable. This explains the kinks that were observed for the monotone smoother of the data in Section 3. See Figure 3.1.

Mammen (1991a) also considers other proposals for monotone smoothing that are of the form "monotonize then smooth," denoted by $\hat{m}_{C,S}$, which is a smooth of the monotonized data denoted by Y_C. Insight into how this type of smoother compares with $\hat{m}_{S,C}(x)$ comes from Figure 6.1. In both Figures 6.1(a) and 6.1(b), the subspace $\mathscr{V}_S^m$ (of ordinary functions) is shown as a horizontal line, and the subset $\mathscr{C}_S^m$ (of constrained functions) is the heavily shaded portion. The subspace $\mathscr{V}_S^Y$ (of ordinary vectors) is shown as a diagonal line, and the subset $\mathscr{C}_S^Y$ (of vectors satisfying the constraint) is the heavily shaded portion. Figure 6.1(a) corresponds to the case that the smoother $\hat{m}_{C,S}$ is "monotonicity preserving" (i.e., when applied to monotone data, the result is monotone), and Figure 6.1(b) is the case where the smoother is not monotonicity preserving, which can happen for example for local polynomial smoothers, as shown in Figure 7.1.

When the smoother is monotonicity preserving, the set $\mathscr{C}_S^m$ "covers all the area directly underneath $\mathscr{C}_S^Y$," since smooths of monotone data are again monotone. So when the data Y are first monotonized (i.e., projected onto $\mathscr{C}_S^Y$) to get Y_C, the resulting smooth $\hat{m}_{C,S}$ (which comes from projecting Y_C onto $\mathscr{V}_S^m$), will typically be "inside $\mathscr{C}_S^m$." This means that this approach will tend to "round out the sharp corners in $\hat{m}_{S,C}(x)$."

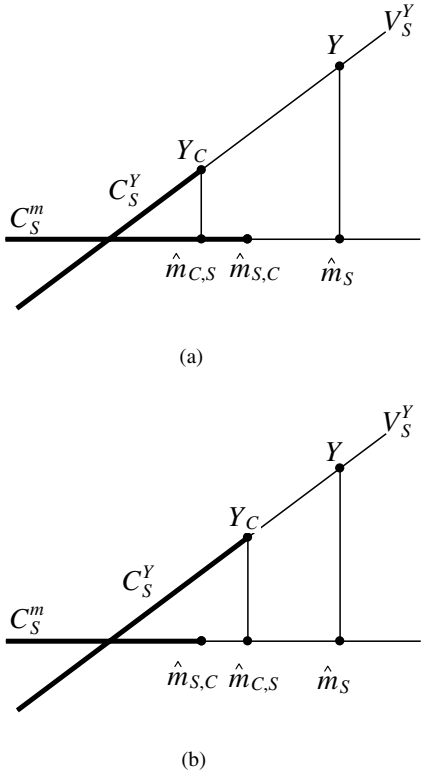

FIG. 6.1. *Diagram showing relation of "monotonicity preserving smoothers." Panel* (a) *and "non-monotonicity preserving smoothers." Panel* (b) *in the vector space* $\mathcal{V}_S$.

When the smoother is not monotonicity preserving, the smooth $\hat{m}_{C,S}$ of the monotonized data Y_C, that is, the projection of Y_C onto $\mathcal{V}_S^m$, need not be monotone, as shown in Figure 6.1(b). Another illustrative example for the situation in Figure 6.1(b) are functions that are constrained to go through the origin. A projection of a function f onto the constrained set is achieved by replacing the single value $f(0)$ by 0. This example highlights the fact that the resulting estimate of the approach "smooth then constrain" may not be smooth. Furthermore the idea "constrain then smooth" may not lead to a constrained estimate. The Sobolev projection method described in Section 5.1 is a way of addressing this problem.

6.2 Remarks on Implied Loss Functions

The constrained estimate minimizes a weighted L_2 distance from the smoothed estimate. Different choices of the weight measure ν lead to different L_2 norms. For different forms of the simple smoother (3.1), this entails different versions of the implied loss (4.6) .

For Nadaraya–Watson weights, $w(x) = \frac{1}{n} \times \sum_{i=1}^n K_h(x - X_i)$ is a kernel density estimator, so under reasonable assumptions (see, e.g., Silverman, 1986; Wand and Jones, 1995) $w(x)$ is approximately $f(x)$ the density of $X_1, \ldots, X_n$, so this estimator is approximately optimizing

$$\int \{\hat{m}(x) - m_0(x)\}^2 f(x)\, \nu(dx).$$

For situations where "f weighting" is desirable in Nadaraya–Watson smoothing, $\nu(dx) = dx$ is appropriate. When "no weighting" is desired, then the choice $\nu(dx) = w(x)^{-1}\, dx$ is natural.

For Gasser–Müller weights, $w(x) = \frac{1}{n} \sum_{i=1}^n \times \int_{s_{i-1}}^{s_i} K_h(x-t)\, dt = \frac{1}{n} \int_{s_0}^{s_n} K_h(x-t)\, dt$. Under reasonable assumptions (either x is away from boundary regions, or $s_0 = -\infty$, $s_n = \infty$), $w(x)$ is approximately constant, so this estimator is essentially optimizing

$$\int \{\hat{m}(x) - m_0(x)\}^2\, \nu(dx).$$

Thus $\nu(dx) = dx$ gives "no weighting" and "f weighting" can be obtained from $\nu(dx) = \frac{1}{n} \times \sum_{i=1}^n K_h(x - X_i)\, dx$.

Next we study the effect of the weight function w under constraints. For some constraints, the projection of the smoother onto the constraint set leads only to "local" changes of the smoother. Consider, for example, the case of monotone smoothing and assume that the smoother is nearly monotone with the exception of some local wiggles. As noted at (6.1) one achieves the monotone smoother by replacing the local wiggles by constant local pieces where the estimate is taken as a local weighted average. Such local averages do not depend strongly on the weight function w or on the measure ν, unless the sample size is small (careful investigation of this is done in Mammen et al., 2001). So usually the choice of the weight measure ν is of relatively minor importance.

6.3 ANOVA Decompositions and Model Choice

Our projection framework provides further insight into model choice and comparison between models. For example assume that we have a class of nested submodels $\mathcal{C}_{S,1}^m \subset \cdots \subset \mathcal{C}_{S,k}^m \subset \mathcal{V}_S^m$ given. Our approach allows us to compare the corresponding estimates using the norm (4.1) or its generalization (5.2). Define for $j = 1, \ldots, k$ the constrained estimates analogous to (4.3),

$$\underrightarrow{\hat{m}}_{S,C,j} = \arg\min_{\underrightarrow{m} \in \mathcal{C}_{S,j}^m} \left\| \underrightarrow{Y} - \underrightarrow{m} \right\|^2.$$

If the submodels $\mathcal{C}_{S,1}^m, \ldots, \mathcal{C}_{S,k}^m$ are vector spaces, repeated application of the Pythagorean theorem

yields

$$\begin{aligned}
&\left\|\underset{\rightarrow}{Y} - \underset{\rightarrow}{\hat{m}}_{S,C,1}\right\|^2 \\
&\quad = \left\|\underset{\rightarrow}{Y} - \underset{\rightarrow}{\hat{m}}_{S,C,k}\right\|^2 + \left\|\underset{\rightarrow}{\hat{m}}_{S,C,k} - \underset{\rightarrow}{\hat{m}}_{S,C,k-1}\right\|^2 \\
&\qquad + \cdots + \left\|\underset{\rightarrow}{\hat{m}}_{S,C,2} - \underset{\rightarrow}{\hat{m}}_{S,C,1}\right\|^2 \\
&\quad = \left\|\underset{\rightarrow}{\hat{m}}_{S} - \underset{\rightarrow}{\hat{m}}_{S,C,k}\right\|^2 + \left\|\underset{\rightarrow}{\hat{m}}_{S,C,k} - \underset{\rightarrow}{\hat{m}}_{S,C,k-1}\right\|^2 \\
&\qquad + \cdots + \left\|\underset{\rightarrow}{\hat{m}}_{S,C,2} - \underset{\rightarrow}{\hat{m}}_{S,C,1}\right\|^2.
\end{aligned}$$

Unfortunately, the summands in this decomposition are, as opposed to "traditional" ANOVA decompositions, typically not independent.

This observation holds for finite samples as well as asymptotically. To appreciate why, suppose that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with standard normal $N(0,1)$ distribution and consider $\mathcal{V}_S$ endowed with the norm (4.1). It follows that $\underset{\rightarrow}{Y}$ has a standard normal multivariate distribution on the vector subspace $\mathcal{V}_S^Y$.

Consider, the next two projections, say $\Pi_1\underset{\rightarrow}{Y}$ and $\Pi_2\underset{\rightarrow}{Y}$, of $\underset{\rightarrow}{Y}$ onto orthogonal subspaces L_1 and L_2 of $\mathcal{V}_S^m$ as illustrated by Figure 6.2. Specifically, take L_1 and L_2 as the orthogonal complements of $\mathcal{C}_{S,j}^m$ in $\mathcal{C}_{S,j+1}^m$ for two different values of j, that is, $L_1 = \mathcal{C}_{S,j}^{m,\perp} \cap \mathcal{C}_{S,j+1}^m$ and $L_2 = \mathcal{C}_{S,j'}^{m,\perp} \cap \mathcal{C}_{S,j'+1}^m$ for $j \neq j'$. Hence, $\Pi_1\underset{\rightarrow}{Y}$ is $\underset{\rightarrow}{\hat{m}}_{S,C,j+1} - \underset{\rightarrow}{\hat{m}}_{S,C,j}$ and $\Pi_2\underset{\rightarrow}{Y}$ is $\underset{\rightarrow}{\hat{m}}_{S,C,j'+1} - \underset{\rightarrow}{\hat{m}}_{S,C,j'}$.

With this choice of L_1 and L_2, neither of the two subspaces is contained in $\mathcal{V}_S^Y$ nor are they orthogonal to $\mathcal{V}_S^Y$ (see the discussion in Section 4). Therefore we cannot conclude in general that $\Pi_1\underset{\rightarrow}{Y}$ and $\Pi_2\underset{\rightarrow}{Y}$ are independent. As an extreme case consider the simple two-dimensional plot of Figure 6.2. Here, $\underset{\rightarrow}{Y}$ has a one(!)-dimensional normal distribution on the line $\mathcal{V}_S^Y$ and $\Pi_1\underset{\rightarrow}{Y}$ depends deterministically on $\Pi_2\underset{\rightarrow}{Y}$. This implies, in particular, that they are not independent.

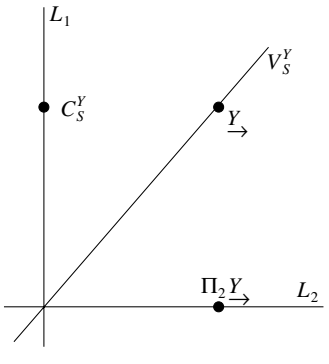

FIG. 6.2. *Diagram showing the data vector $\underset{\rightarrow}{Y}$ lying in a one-dimensional space $\mathcal{V}_S^Y$, and that the projections $\Pi_1\underset{\rightarrow}{Y}$ and $\Pi_2\underset{\rightarrow}{Y}$ onto the orthogonal spaces $L_1 = \mathcal{C}_{S,j}^{m,\perp} \cap \mathcal{C}_{S,j+1}^m$ and $L_2 = \mathcal{C}_{S,j'}^{m,\perp} \cap \mathcal{C}_{S,j'+1}^m$ need not be independent.*

Furthermore, in general the summands $\|\underset{\rightarrow}{\hat{m}}_{S,C,k} - \underset{\rightarrow}{\hat{m}}_{S,C,k-1}\|^2$ do not have an (asymptotic) χ^2 distribution; see for example, Härdle and Mammen (1993) who propose using bootstrap methods to avoid these problems. The situation is a little bit simpler for orthogonal series estimates. For a general discussion of lack-of-fit tests in nonparametric regression see Hart (1997).

6.4 Numerical Implementation

According to Proposition 1 for the calculation of constrained estimates we have only to calculate the unconstrained smoother and to calculate the projection of the smoother onto the constrained set. This yields a big computational gain. For example, if ν is counting measure on an equally spaced grid of g values of x, then instead of minimizing over vectors of dimension $n \cdot g$, as required for (4.3), only vectors of dimension g need to be considered for (4.5). In addition, established algorithms may be used on the reduced problem. The reduced problem (in its discretized form) is a constrained (weighted) least squares problem. Algorithms for such problems are well studied in the numerical literature. Solutions can be iteratively calculated by active set methods (see, e.g., McCormick, 1983), by the method of iterative projections (see, e.g., Dykstra, 1983; Robertson et al., 1988), or primal-dual methods (see, e.g., Goldfarb and Idnani, 1983). For monotone smoothing the pool adjacent violators algorithm, which calculates effectively projections onto monotone vectors, can be used in the second step. For a discussion of this algorithm and other constrained least squares algorithms see the books by Barlow et al. (1972) and Robertson, Wright and Dykstra (1988). General optimization algorithms are discussed, among others, in Fletcher (1987), den Hertog (1994) and Nash and Sofer (1996).

6.5 Asymptotics for Constrained Estimates

Asymptotics for unconstrained kernel-type estimates is quite well developed. For some examples the asymptotic results of the unconstrained estimates carry over to the constrained estimates. Trivially, this is the case if the unconstrained estimate fulfils the constraint with probability tending to 1. This implies that, with probability tending to 1, the constrained estimate coincides with the unconstrained. An important example for this case is monotone smoothing: under appropriate conditions, the derivative m' of the regression function is

consistently estimated by the derivative of kernel smoothers. Then, if m' is bounded away from 0, the constrained estimate is monotone with probability tending to 1. So asymptotics of the constrained estimate is reduced to the unconstrained case (see, e.g., Mukerjee, 1988 and Mammen, 1991a). This does not hold for monotonicity constraints of higher order derivatives. Under such conditions the constrained estimate can achieve faster rates of convergence than the unconstrained estimate. This has been shown in Mammen and Thomas-Agnan (1999) for smoothing splines; see also the results in Mammen (1991b) on constrained least squares estimates. An essential mathematical tool for showing rates of convergence of restricted smoothers is given by empirical process theory; see van de Geer (1990).

7. EXTENSION TO LOCAL POLYNOMIALS

Now we extend our projection framework for smoothing to local polynomial smoothers. For simplicity of notation, we assume now that the covariables X_i are one-dimensional and that the regression function m goes from $\mathbb{R}$ to $\mathbb{R}$. Given a set of weights $w_i(x)$, such as those of Section 4.1, a local polynomial smoother of order p, can be written as

$$\hat{m}_{LP}(x) = \hat{\beta}_0(x),$$

where

$$\hat{\beta}(x) = \begin{bmatrix} \hat{\beta}_0(x) \\ \vdots \\ \hat{\beta}_p(x) \end{bmatrix} = \arg\min_{\beta} \int \frac{1}{n} \times \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j(x)(x - X_i)^j \right\}^2 \times w_i(x)\, \nu(dx). \tag{7.1}$$

As for $\hat{m}_S$, the integral and the weight measure ν play no role, because the minimization can be done individually for each x.

To write this smoother as a projection we use an expanded version of the normed vector space $\mathcal{V}_S$ which is the set of $n(p+1)$ tuples of functions,

$$\mathcal{V}_{LP} = \left\{ \underset{\rightarrow}{f} = \begin{bmatrix} f_{1,0}(x) \\ \vdots \\ f_{1,p}(x) \\ \vdots \\ f_{n,0}(x) \\ \vdots \\ f_{n,p}(x) \end{bmatrix} : f_{i,j}: \mathbb{R} \to \mathbb{R},\ i = 1, \ldots, n,\ j = 0, \ldots, p \right\}.$$

Now the data vector $Y^T = [Y_1, \ldots, Y_n]$ is viewed as an element $\underset{\rightarrow}{Y}$ of $\mathcal{V}_{LP}$, which is an $n(p+1)$-tuple of the form $\underset{\rightarrow}{Y}^T = [Y_1, 0, \ldots, 0, Y_2, 0, \ldots, 0, Y_n, 0, \ldots, 0]$, that is, within blocks of $p+1$, only the first entries may be nonzero,

$$f_{i,j}(x) \equiv \begin{cases} Y_i, & j = 0, \\ 0, & j = 1, \ldots, p, \end{cases} \qquad i = 1, \ldots, n.$$

The subspace of such $n(p+1)$-tuples is called $\mathcal{V}_{LP}^{Y}$. A candidate smooth now involves several functions $\beta_j: \mathbb{R} \to \mathbb{R}$, which are elements of $\mathcal{V}_{LP}$ of the form $\underset{\rightarrow}{\beta}$, that are $n(p+1)$-tuples where entries are common across i, and for each j are $\beta_j(x)$, that is, $f_{i,j}(x) = \beta_j(x)$, $i = 1, \ldots, n$, $j = 0, \ldots, p$. The subspace of $n(p+1)$-tuples with entries that are identical across i is denoted by $\mathcal{V}_{LP}^{m}$. The appropriate analog of the norm (4.1) on $\mathcal{V}_{LP}$ is

$$\left\| \underset{\rightarrow}{f} \right\|^2 = \int \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=0}^{p} f_{i,j}(x)(x - X_i)^j \right\}^2 \times w_i(x)\, \nu(dx). \tag{7.2}$$

This notation represents local polynomial smooths as a projection, because $\hat{m}_{LP}(x) = \hat{\beta}_0(x)$, where (7.1) can be rewritten as

$$\hat{\beta}(x) = \arg\min_{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{V}_{LP}^{m}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \right\|^2. \tag{7.3}$$

Now given a set of constrained $n \cdot (p+1)$ tuples $\mathcal{C}_{LP}^{m} \subset \mathcal{V}_{LP}^{m}$, for example $\beta_0(x)$ monotone, a natural constrained local polynomial smoother is $\hat{m}_{LP,C}(x) = \hat{\beta}_{0,C}(x)$, where

$$\hat{\beta}_C(x) = \arg\min_{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}_{LP}^{m}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \right\|^2. \tag{7.4}$$

This constrained minimization is simplified, exactly as at (4.4), using a Pythagorean relationship. Following the same arguments (with nearly the same notation) as in Section 3 yields the proposition.

PROPOSITION 3. *The constrained local polynomial smooth can be represented as a constrained minimization over ordinary functions as* $\hat{m}_{LP,C}(x) = \hat{\beta}_{0,C}(x)$ *where*

$$\begin{aligned} \hat{\beta}_C(x) &= \arg\min_{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}^{m}{}_{LP}} \left\| \underset{\rightarrow}{\hat{\beta}} - \underset{\rightarrow}{\beta} \right\|^2 \\ &= \arg\min_{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}^{m}{}_{LP}} \int \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=0}^{p} \left(\hat{\beta}_j(x) - \beta_j(x) \right) \times (x - X_i)^j \right]^2 w_i(x)\, \nu(dx) \\ &= \arg\min_{\beta \in C_{LP}} \int \sum_{j=0}^{p} \sum_{j'=0}^{p} \left(\hat{\beta}_j(x) - \beta_j(x) \right) \times \left(\hat{\beta}_{j'}(x) - \beta_{j'}(x) \right) U_{j+j'}(x)\, \nu(dx), \end{aligned} \tag{7.5}$$

where

$$U_j(x) = \frac{1}{n}\sum_{i=1}^{n}(x - X_i)^j\, w_i(x) \quad \text{for } j = 0, \ldots, 2p.$$

As does Proposition 1 in Section 4 for kernel smoothing, Proposition 3 gives geometric insights, as well as computational gains. Again, the computational problem is reduced to a constrained least squares problem. So the remarks of Section 6.4 apply. In many cases the set of constrained functions $\beta \in C_{LP}$ will involve constraints only on some of the β_j. For example, in monotone regression, a simple constraint is that only $\beta_0(x)$ is increasing, but it could also be desirable to assume in addition that $\beta_1(x) \geq 0$.

Proposition 3 shows that, as for kernel smoothing, constrained smoothing leads to estimates of the form: "smooth then constrain." Again, one could try estimates based on the idea "first constrain then smooth." For local polynomials this idea does not work: smoothing by local polynomials is not monotonicity preserving. This can be seen from Figure 7.1 that shows some artificial monotone data with a local linear fit that is not monotone. This is in contrast to the Nadaraya–Watson smoother that always preserves monotonicity (see, Mukerjee, 1988; Mammen and Marron, 1997). Sufficient conditions for a smoother to be monotonicity preserving are given in Mammen and Marron (1997). They also discuss a modification of the local linear smoother which is monotonicity preserving. A detailed discussion of monotone local polynomials can be found in Mammen et al. (2001).

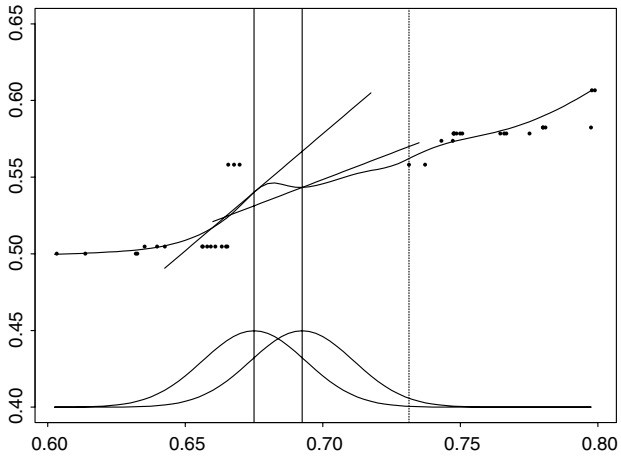

FIG. 7.1. *Monotone artificial data with nonmonotone local linear fit.*

8. ADDITIVE MODELS

We now consider smoothing estimates for additive models. For simplicity this will be done for Nadaraya–Watson smoothing. Using the ideas from the last section the approach can be easily generalized to local polynomial smoothing. For details see Mammen, Linton and Nielsen (1999). In this model the additive Nadaraya–Watson smoother can be calculated by the backfitting algorithm. Our geometric point of view can be used to show that this algorithm converges under weak conditions. Furthermore, our geometric representations can be used as essential tools to give asymptotic distributions of additive Nadaraya–Watson smoothers and additive local polynomial smoothers. We now describe how our projection framework carries over to this model. Our constraint on the regression function m: $\mathbb{R}^q \to \mathbb{R}$ is that

$$m(x) = m_0 + m_1(x_1) + \cdots + m_q(x_q) \quad \text{for } x = (x_1, \ldots, x_q), \tag{8.1}$$

where m_0 is a constant and $m_1, \ldots, m_q$ are functions from $\mathbb{R}$ to $\mathbb{R}$. For identifiability, it is assumed that $E\, m_l(X_{i,l}) = 0$, $i = 1, \ldots, n$; $l = 1, \ldots, q$. Discussion of the additive model can be found in Hastie and Tibshirani (1990).

The constrained and unconstrained Nadaraya–Watson smoother (or more generally simple smoother) is defined as in (4.2) and (4.3). The space $\mathscr{C}_S^m \subset \mathscr{V}_S^m$ is now the subset of n-tuples with (identical) entries that are additive, that is,

$$\mathscr{C}_S^m = \left\{ \vec{f} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} : \begin{array}{c} f_i(x) = g_1(x_1) + \cdots + g_q(x_q) \\ \text{for some functions } g_1, \ldots, \\ g_q : \mathbb{R}^q \to \mathbb{R} \text{ for } i = 1, \ldots, n \end{array} \right\}.$$

In this model we do not recommend first calculating the unrestricted estimate (and then projecting this estimate on the subspace $\mathscr{C}_S^m$). The reason is that the calculation of the unrestricted estimate involves many unknown parameters. If the data are too sparse this calculation would be unstable or the estimate may not even be defined at many locations. A standard method to calculate the constrained (i.e., additive) estimate is the backfitting algorithm (see Hastie and Tibshirani, 1990). It is based on iterative minimization of $\|\vec{Y} - \vec{m}\|^2$. In each minimization step the norm is minimized over one additive component while letting the other components be fixed. In each cycle of the algorithm this is done for each component k. It can be easily seen that each step in a cycle of the algorithm is a projection onto an appropriate subspace of the space $\mathscr{C}_S^m$. That means that, in our geometry, backfitting is based on iterative application of projections. This is much easier

to understand as iterative application of smoothing operators. In particular, it can be used to show that under weak conditions, backfitting converges to the minimizer with exponential speed (see Mammen, Linton and Nielsen, 1999). This implies not only consistency of the backfitting algorithm, it shows also that for getting the asymptotic distribution of the estimate it suffices to consider the result of the backfitting algorithm after $O(\log n)$ cycles. Using this approach Mammen, Linton and Nielsen (1999) show that the local linear estimate for one additive component achieves the same asymptotic normal limit as the oracle estimate based on knowing the other components. For an asymptotic result for another additive local polynomial backfitting estimate that does not achieve the asymptotic oracle limit see Opsomer (2000) and Opsomer and Ruppert (1997).

9. EXTENSIONS

In this paper we have only discussed constrained smoothing of regression functions. Similar problems arise in other settings like density estimation, generalized regression, white noise models and nonparametric time series models. Another field of possible applications are semiparametric models where constraints are put on the nonparametric components.

Here, we mention other variations from nonparametric regression.

• *Boundary conditions.* A regression function m, that is defined on $[0, 1]$, say, is assumed to be zero at the boundary point 0. Or more generally, m is supposed to take fixed known values in certain regions. He and Ng (1999) note that U.S. Army Construction Engineers use the flashing condition index (FCI) as a measurement for roof condition on buildings. Naturally, without interference the condition cannot improve and at the time of construction a roof is assumed to have an index of 100. Hence, He and Ng (1999) consider fitting a decreasing regression function m with $m(0) = 100$ and $0 \le m(x) \le 100$.

• *Additive models with monotone components.* The regression function m: $\mathbb{R}^q \to \mathbb{R}$ is supposed to be of additive form $m(x_1, \ldots, x_q) = m_1 \times (x_1) + \cdots + m_q(x_q)$ where the additive components (or a subset of them) are monotone.

• *Observed derivatives.* One observes r samples corresponding to r regression functions (as in the last point) with now $r = 2$. Now it is assumed that m_2 coincides with the derivative of m_1; see Cox (1988).

• *Bivariate extreme-value distributions.* Hall and Tajvidi (2000) study methods to estimate the dependence function of a bivariate extreme-value distribution. Their methods requires estimating a convex function m such that $m(0) = m(1) = 1$ and $m(x) \ge \max(x, 1 - x)$ for $x \in [0, 1]$.

• *Positivity constraints.* Imposing positivity constraints on wavelet estimators, especially if used for density estimation is discussed in Dechevsky, MacGibbon and Penev (1998) and Dechevsky and MacGibbon (1999).

REFERENCES

BARLOW, R. E. and VAN ZWET, W. R. (1970). Asymptotic properties of isotonic estimators for generalized failure rate function I. Strong consistency. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 159–173. Cambridge Univ. Press.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York

BOWMAN, A. W. and AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford Univ. Press.

CHENG, K. F. and LIN, P. E. (1981). Nonparametric estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **57** 223–233.

CHU, C. K. and MARRON, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.* **6** 404–436.

COX, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist.* **16** 694–712.

DECHEVSKY, L. and MACGIBBON (1999). Asymptotically minimax nonparametric function estimation with positivity constraints i. Unpublished manuscript.

DECHEVSKY, L., MACGIBBON, B. and PENEV, S. (2001). Numerical methods for asymptotically minimax nonparametric function estimation with positivity constraints i. *Sankhyā*. To appear.

DELECROIX, M., SIMIONI, M. and THOMAS-AGNAN, C. (1995). A shape constrained smoother: simulation study. *Comput. Statist.* **10** 155–175.

DELECROIX, M., SIMIONI, M. and THOMAS-AGNAN, C. (1996). Functional estimation under shape constraints. *J. Nonparametr. Statist.* **6** 69–89.

DELECROIX, M. and THOMAS-AGNAN, C. (2000). Spline and kernel smoothing under shape restrictions. In *Smoothing and Regression: Approaches, Computation and Application* (M. Schimek, ed.) 109–134. Wiley, New York.

DEN HERTOG, D. (1994). *Interior Point Approach to Linear, Quadratic and Convex Programming*. Kluwer, Dordrecht.

DIERCKX, P. (1980). An algorithm for cubic spline fitting with convexity constraints. *Computing* **24** 349–371.

DOLE, D. (1999). C_0S_{m0}: A constrained scatterplot smoother for estimating convex monotone transformations. *J. Bus. Econom. Statist.* **17** 444.

DYKSTRA, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **77** 621–628.

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer, New York.

ELFING, T. and ANDERSSON, L. E. (1988). An algorithm for computing constrained smoothing spline functions. *Numer. Math.* **52** 583–595.

EUBANK, R. L. (1999). *Smoothing Splines and Nonparametric Regression*, 2nd ed. Dekker, New York.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Application*. Chapman and Hall/CRC, 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. *J. Amer. Statist. Assoc.* **92** 84–91.

FLETCHER, R. (1987). *Practical Methods of Optimization*, 2nd ed. Wiley, New York.

FÖLDES, A. and RÉVÉSZ, P. (1974). A general method for density estimation, *Studia Sci. Math. Hungar.* **9** 81–92.

FRIEDMAN, J. H. and TIBSHIRANI, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26** 243–250.

GAYLORD, C. K. and RAMIREZ, D. E. (1991). Monotone regression splines for smoothed bootstrapping. *Comput. Statist. Quarterly* **6** 85–97.

GOLDFARB, D. and IDNANI, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Programming* **27** 1–33.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, London.

HALL, P. and TAJVIDI, N. (2000). Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli* **6** 835–844.

HÄRDLE, W. (1990). *Applied Non-parametric Regression*. Cambridge Univ. Press.

HÄRDLE W. and GASSER, T. (1984). Robust nonparametric function fitting, *J. Roy. Statist. Soc. Ser. B* **46** 42–51.

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947.

HART, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall/CRC, London.

HE, X. and NG, P. (1999). COBS: qualitatively constrained smoothing via linear programming. *Comput. Statist.* **14** 315–337.

IRVINE, L. D., MARIN, S. P. and SMITH, P. W. (1986). Constrained interpolation and smoothing. *Constr. Approx.* **2** 129–151.

KELLY, C. and RICE, J. (1990). Monotone smoothing with application to dose–response curves and the assessment of synergism. *Biometrics* **46** 1071–1085.

LOADER, C. (1999). *Local Regression and Likelihood, Statistics and Computing*. Springer, New York.

MAMMEN, E. (1991a). Estimating a smooth monotone regression function. *Ann. Statist.* **19** 724–740.

MAMMEN, E. (1991b). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490.

MAMMEN, E. and MARRON, J. S. (1997). Mass recentered kernel smoothers. *Biometrika* **84** 765–778.

MAMMEN, E., MARRON, J. S., TURLACH, B. A. and WAND, M. P. (2001). Monotone local polynomial smoothers. Unpublished manuscript.

MAMMEN, E. and THOMAS-AGNAN, C. (1999). Smoothing splines and shape restrictions. *Scand. J. Statist.* **26** 239–252.

MCCORMICK, G. P. (1983). *Nonlinear Programming: Theory, Algorithms and Applications*. Wiley, New York.

MICCHELLI, C. A. and UTRERAS, F. I. (1988). Smoothing and interpolation in a convex subset of a Hilbert space. *SIAM J. Sci. Statist. Comput.* **9** 728–746.

MUKERJEE, H. (1988). Monotone nonparametric regression. *Ann. Statist.* **16** 741–750.

MÜLLER, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York.

NASH, S. G. and SOFER, A. (1996). *Linear and Nonlinear Programming*. McGraw-Hill, New York.

OPSOMER, J. D. (2000). Aysmptotic properties of backfitting estimators. *J. Multivariate Anal.* **73** 166–179.

OPSOMER, J. D. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

PRINCE, J. L. and WILLSKY, A. S. (1990). Reconstructing convex sets from support line measurements. *IEEE Trans. Pattern Anal. Machine Intelligence* **12** 377–389.

RAMSAY, J. O. (1988). Monotone regression splines in action (with discussion). *Statist. Sci.* **3** 425–461.

RAMSAY, J. O. (1998). Estimating smooth monotone functions. *J. Roy. Statist. Soc. Ser. B* **60** 365–375.

RAMSAY, J. O. and SILVERMAN, B. W. (1997). *Functional Data Analysis*. Springer, New York.

RATKOWSKY, D. A. (1983). *Nonlinear Regression Modeling*. Dekker, New York.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

RUDIN, W. (1987). *Real and Complex Analysis*. McGraw-Hill, New York.

SANTALÓ, L. A. (1976). *Integral Geometry and Geometric Probability*. Addison-Wesley, Reading, MA.

SCHMIDT, J. W. (1987). An unconstrained dual program for computing convex C^1-spline approximants. *Computing* **39** 133–140.

SCHMIDT, J. W. and SCHOLZ, I. (1990). A dual algorithm for convex-concave data smoothing by cubic C^2-splines. *Numer. Math.* **57** 333–350.

SCHWETLICK, H. and KUNERT, V. (1993). Spline smoothing under constraints on derivatives. *BIT* **33** 512–528.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, London.

SILVERMAN, B. W. and WOOD, J. T. (1987). The nonparametric estimation of branching curves. *J. Amer. Statist. Assoc.* **82** 551–558.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

SPECKMAN, P. L. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

STEER, B. T. and HOCKING, R. A. (1985). The optimum timing of nitrogen application to irrigated sunflowers. In *Proceedings of the Eleventh International Sunflower Conference* 221–226. Asociación Argetina de Girasol, Buenos Aires.

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. and TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1424.

TANTIYASWASDIKUL, C. and WOODROOFE, M. B. (1994). Isotonic smoothing splines under sequential designs. *J. Statist. Plann. Inference* **38** 75–88.

TSYBAKOV, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems Inform. Transmission* **22** 133–146.

TURLACH, B. A. (1997). Constrained smoothing splines revisited. Statistics Research Report SRR 008-97, Center for Math. and Its Applications, Australian National Univ. Canberra.

UTRERAS, F. I. (1985). Smoothing noisy data under monotonicity constraints: Existence, characterization and convergence rates, *Numer. Math.* **47** 611–625.

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

VILLALOBOS, M. and WAHBA, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.* **82** 239–248.

WAHBA, G. (1990). *Spline Functions for Observational Data*. SIAM, Philadelphia.

WALTER, G. G. and BLUM, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328–340.

WAND, M. P. and JONES, M. C. (1995). *Kernel smoothing*. Chapman and Hall/CRC, London.

WRIGHT, F. T. (1982). Monotone regression estimates for grouped observations. *Ann. Statist.* **10** 278–286.