A general framework for constrained smoothing

E. Mammen* J.S. Marron† B.A. Turlach‡ M.P. Wand§

June 16, 1998

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 it helps to understand the finite sample and asymptotic behaviour of these estimates.

*Institut für Angewandte Mathematik, Ruprecht–Karls–Universität Heidelberg, Germany.

†Department of Statistics, University of North Carolina, U.S.A.

‡Department of Statistics, The University of Adelaide, Australia (This manuscript was mainly written while the author was working at the Centre for Mathematics and its Applications and the Cooperative Research Centre for Advanced Computational Systems, The Australian National University, Australia).

§Department of Biostatistcs, Harvard School of Public Health, U.S.A.

1. Introduction

Smoothing as a means of modelling non-linear 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, such as monotonicity. 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*, i.e. 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.

Smoothing is illustrated in Figure 1.1 we show part of the "cars" data used in 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. Here fuel consumption, 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 1.1 is a simple smooth, i.e. moving average, as described in (2.1).

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. This, and other types of constraints are not natural to incorporate into many types of smoothing, including the simple smooth used in Figure 1.1. Green and Silverman (1994) have pointed out that smoothing splines, where many types of constraints are incorporated in a natural way, are an exception to this rule. 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 3, is a product structure, i.e. we consider "vectors of objects", where the objects are functions, vectors, or even sets of functions or vectors. When the result of the smoothing process is a curve, the

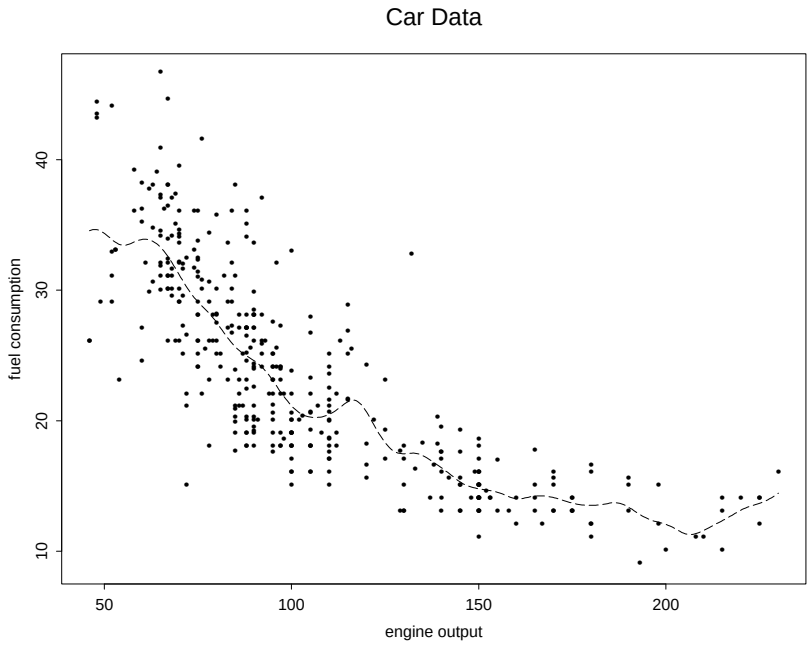


Figure 1.1: Raw data and simple smooth for Fuel consumption as a function of engine output. Smooth is Nadaraya-Watson type with Gaussian kernel and bandwidth $h = 4$.

objects are taken to be functions. When the result is a vector, e.g. the smooth evaluated at the design points, the objects are taken to be vectors. For local polynomial smoothing, projection follows from letting the objects be groups 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 2, and thus give representation of the smoothers as projections. By this device a much broader class of smoothers can be viewed as projections, as shown in Section 3, which allows natural incorporation of constraints for these methods.

In Section 4 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 5, including detailed discussion of the case of monotone smoothing, some remarks about loss functions, decompositions of sums of squares, numerical implementation, and sums of squares. Extensions to local polynomials are given in Section 6. Application of our approach to additive models is discussed in Section 7.

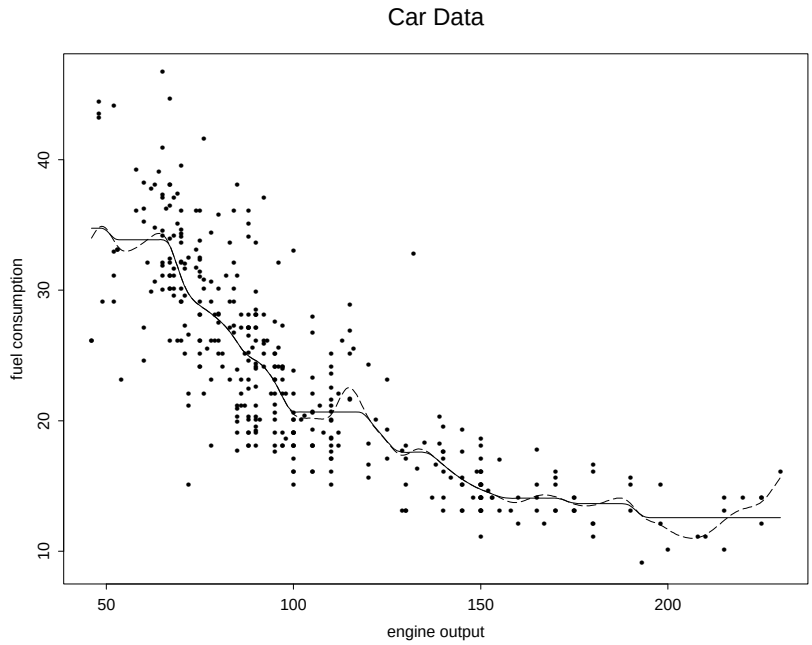


Figure 1.2: Raw data and monotonicity constrained smooth for Fuel consumption as a function of engine output. Smooth is Nadaraya-Watson type with Gaussian kernel and bandwidth $h = 4$.

Figure 1.2 shows the result of the sophisticated projection ideas of Sections 4.1 and 5, starting with the simple smooth in Figure 1.1. Note that essentially the increasing parts of the smooth have been "rounded off".

For more background on smoothing, see any of a number of monographs, e.g. in the last five years, Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Hart (1997) and Bowman and Azzalini (1997).

2. Simple smoothing as minimization

Before developing our general vector space framework, we first show how simple smoothing, as shown in Figure 1.1, 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$, e.g. as

shown in the scatterplot of Figure 1.1, that are modeled as

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

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

The dashed curve in Figure 1.1 is a "simple smooth" of the form

$$\widehat{m}_S(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}, \tag{2.1}$$

i.e. a moving (in x) weighted average of the Y_i. The weights $w_i(x)$ used in Figure 1.1 are of Nadaraya-Watson type, as discussed in Section 3.1. 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 this curve, shown in Figure 1.1 is not monotone decreasing. An approach to constraining this type of smooth to be monotone is to recognize that it can be written as

$$\widehat{m}_S = \operatorname*{argmin}_m \int \frac{1}{n} \sum_{i=1}^n \{Y_i - m(x)\}^2 w_i(x)\, \nu(dx), \tag{2.2}$$

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). The integral is not necessary for this unconstrained estimator, because the minimum can be found for each x individually, i.e.

$$\widehat{m}_S(x) = \operatorname*{argmin}_{m \in I\!R} \frac{1}{n} \sum_{i=1}^n (Y_i - m)^2 w_i(x). \tag{2.3}$$

For the same reason the weight measure ν also has no effect on $\widehat{m}_S(x)$. 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 monotonicity, then a constrained version of the simple smooth is:

$$\widehat{m}_{S,C} = \operatorname*{argmin}_{m \in C} \int \frac{1}{n} \sum_{i=1}^n \{Y_i - m(x)\}^2 w_i(x)\, \nu(dx). \tag{2.4}$$

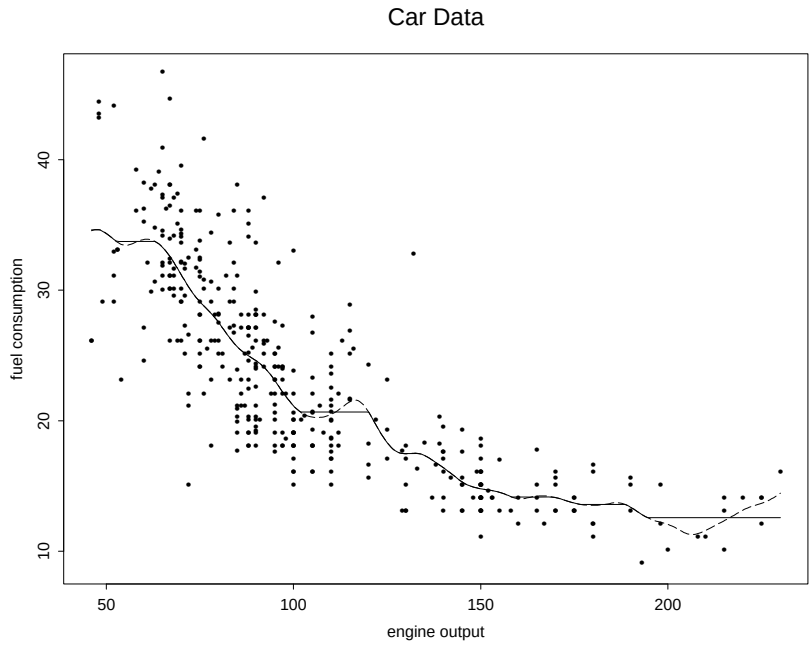


Figure 2.1: Unconstrained and constrained (monotone) smooths, for Fuel consumption as a function of engine output, as in Figure 1.1. The constrained smooth has "kinks" which have been smoothed out in the more sophiticated constrained smooth of Figure 1.2

The weight measure ν now plays an important role, because the minimizers at different points x are linked through the constraints. In Figure 2.1, a discretized version of Lebesgue measure is used.

While this estimate appears natural, the monotonicity constraint introduces some "kinks" in Figure 2.1, essentially at "break points where $\widehat{m}_S$ 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 5.1 for further discussion, and methods to "round off these corners" as shown in Figure 1.2.

3. Simple smoothers viewed as projections

In this section we shall introduce a normed vector space that contains the data vector and the regression functions. We shall show that in this space kernel

smoothers appear as a projection of the data vector onto an appropriate vector subspace. To capture all of these aspects, it is not enough to simply work with n-dimensional vectors, or with functions. A vector space which reflects the full structure of smoothing, i.e. includes both the data vector Y, and the candidate smooths $m(x)$, is a product space containing n-tuples of linear objects

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

where V is some normed vector space. The vector space 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 4, V will be an appropriate space of functions. But when the result of the smooth is a vector, e.g. when the smooth is evaluated only at the design points, V is a set of ordinary vectors. For local polynomial smoothing, V is taken to be vectors of functions (or vectors), as described in Section 6.

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

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

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

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

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

$$\left\| \underset{\rightarrow}{f} \right\|^2 = \int \frac{1}{n} \sum_{i=1}^{n} f_i(x)^2 w_i(x)\, \nu(dx). \tag{3.1}$$

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 (3.1) as a norm, i.e. implicitly we work on classes of functions. We shall also assume that $\mathcal{V}_S$ 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 (2.2) and (2.4) can be rewritten as

$$\widehat{m}_S = \operatorname*{argmin}_{m:\, \underset{\rightarrow}{m} \in \mathcal{V}_S^m} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|^2, \tag{3.2}$$

$$\widehat{m}_{S,C} = \operatorname*{argmin}_{m:\, \underset{\rightarrow}{m} \in \mathcal{C}_S^m} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|^2, \tag{3.3}$$

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

Using a Pythagorean relationship, the minimization problem (3.3) can be substantially simplified. This yields important computational advantages, and also gives some important insights. In particular, for $\underset{\rightarrow}{m} \in \mathcal{V}_S^m$ we have

$$\left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|^2 = \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\widehat{m}}_S \right\|^2 + \left\| \underset{\rightarrow}{\widehat{m}}_S - \underset{\rightarrow}{m} \right\|^2, \tag{3.4}$$

because $\underset{\rightarrow}{\widehat{m}}_S$ is the projection of $\underset{\rightarrow}{Y}$ onto the subspace $\mathcal{V}_S^m$, whence $\underset{\rightarrow}{Y} - \underset{\rightarrow}{\widehat{m}}_S$ is orthogonal to $\underset{\rightarrow}{\widehat{m}}_S - \underset{\rightarrow}{m}$ with respect to the inner product, see e.g. Rudin (1987, Theorem 4.11). Furthermore,

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

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:*

$$\widehat{m}_{S,C}(x) = \underset{m:\underset{\rightarrow}{m} \in \mathcal{C}_S^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{\widehat{m}}_S - \underset{\rightarrow}{m} \right\|^2 = \underset{m \in C}{\operatorname{argmin}} \int \{\widehat{m}_S(x) - m(x)\}^2 w(x)\, \nu(dx). \quad (3.5)$$

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 (3.4), 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 (3.4) follows from direct calculation of

$$\left\langle \underset{\rightarrow}{Y} - \underset{\rightarrow}{\widehat{m}}_S, \underset{\rightarrow}{\widehat{m}}_S - \underset{\rightarrow}{m} \right\rangle = 0.$$

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. But even $\mathcal{V}_S^Y \cap \left(\mathcal{V}_S^C\right)^\perp$ (the orthogonal complement of $\mathcal{V}_S^C$ in $\mathcal{V}_S^Y$) and $\mathcal{V}_S^m \cap \left(\mathcal{V}_S^C\right)^\perp$ 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 3.1. 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 $\widehat{m}_S(x)$.

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

$$L(\widehat{m}, m_0) = \int \{\widehat{m}(x) - m_0(x)\}^2 w(x)\, \nu(dx) \quad (3.6)$$

is essentially optimized by $\widehat{m}_S(x)$ over $\mathcal{V}_S^m$ and by $\widehat{m}_{S,C}(x)$ over $\mathcal{C}_S^m$. Specifics of L are discussed in Section 5.2.

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

(1) Compute the unconstrained estimate $\widehat{m}_S$.

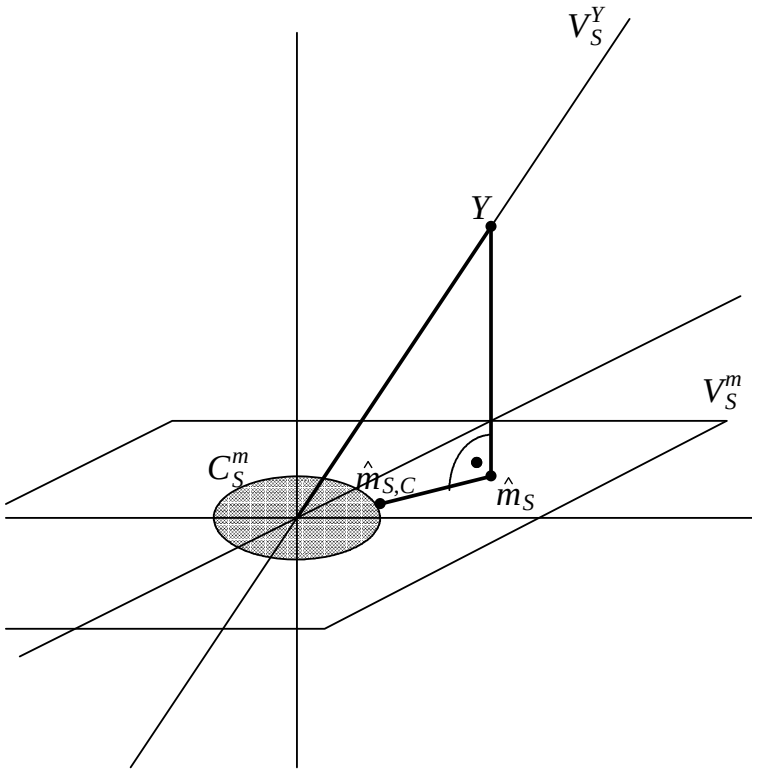


Figure 3.1: Diagram representing location of data and unconstrained and constrained smooths, in the vector space $\mathcal{V}_S$.

(2) Project $\widehat{m}_S$ onto the constrained set of functions.

Implementation of each of these two steps is relatively straightforward and much simpler than direct computation of (2.4). We shall come back to this point in Section 5.4.

3.1. Some remarks and specific simple smoothers

Representations of the type (2.2) have been used for many purposes. For example they provide easy understanding of how local polynomial methods, discussed in detail in Section 6, 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 (2.2) 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 the 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 (2.1), i.e. (2.2). Much of this approach to generality was developed by Földes and Revesz (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, i.e. local averaging, via $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$.

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.

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

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

$$\widehat{m}_{OS}(x) = \sum_{j \in S} \widehat{\theta}_j \psi_j(x), \tag{3.7}$$

where the "empirical Fourier coefficients" are $\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^n Y_i \psi_j(X_i)$, and where S is some set of "coefficients containing most of m_0", e.g. 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 (2.1) 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 (3.7) is the class of regression splines,

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

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 $\widehat{\theta}_j$ are computed by least squares, so they are still linear combinations of Y. Thus this type of smoother can be written in the form (2.1) by interchanging order of summation as above. See Section 7.2 of Eubank (1988) for discussion of many properties of estimators of this form and see Stone *et al.* (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 6. A different type of spline is the smoothing spline discussed in detail in Section 4.

4. 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 (4.1). Constraints can be easily incorporated by minimizing over the restricted set. For work on constrained smoothing splines see Dierckx (1980), Utreras (1985), Irvine *et al.* (1986), Schmidt (1987), Villalobos and Wahba (1987), Elfving and Andersson (1988), Micchelli and Utreras (1988), Ramsay (1988), Fritsch (1990), Kelly and Rice (1990), Schmidt and Scholz (1990), Gaylord and Ramirez (1991), Schwetlick and Kunert (1993), Tantiyaswasdikul

and Woodroofe (1994), Dole (1996), and Mammen and Thomas-Agnan (1998). Some applications are discussed in the books by Wahba (1990) and Green and Silverman (1994). Overviews on work on shape restricted splines are given in Delecroix and Thomas-Agnan (1997). Insight into how constrained smoothing splines work comes from another type of generalization of the framework of Section 2. The basic smoothing spline of order p is usually written as

$$\widehat{m}_{SS}(x) = \operatorname*{argmin}_{m} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 + \lambda \int m^{(p)}(x)^2, \tag{4.1}$$

where λ is the smoothing parameter. See Eubank (1988), Wahba (1990) and Green and Silverman (1994) for discussion of many aspects of this estimator. It can be written in a form which generalizes both (3.1) and (4.1) as

$$\widehat{m}_{SS}(x) = \operatorname*{argmin}_{m: \underset{\rightarrow}{m} \in \mathcal{V}_S^m} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|^2$$

where the norm on $\mathcal{V}_S$ is now generalized to

$$\left\| \underset{\rightarrow}{f} \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|f_i(x)\|_p^2 , \tag{4.2}$$

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 (4.1) 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 (3.1) is the special case where $\lambda = 0$.

As above it is natural to write constrained smoothing splines as

$$\widehat{m}_{SS,C}(x) = \operatorname*{argmin}_{m: \underset{\rightarrow}{m} \in \mathcal{C}_S^m} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|^2$$

This constrained minimization is simplified, exactly as at (3.4), using a Pythagorean relationship. Following the arguments of Section 3 yields:

Proposition 2: *The constrained smoothing spline can be represented as a constrained minimization over ordinary functions as:*

$$\begin{aligned} \widehat{m}_{SS,C}(x) &= \operatorname*{argmin}_{m: \underset{\rightarrow}{m} \in \mathcal{C}_S^m} \left\| \underset{\rightarrow}{\widehat{m}}_{SS} - \underset{\rightarrow}{m} \right\|^2 \\ = \operatorname*{argmin}_{m \in C} &\int \{\widehat{m}_{SS}(x) - m(x)\}^2 w(x) \, \nu(dx) + \lambda \int \{\widehat{m}_{SS}^{(p)}(x) - m^{(p)}(x)\}^2 dx. \end{aligned} \tag{4.3}$$

Proposition 2 is proved in Mammen and Thomas-Agnan (1998). There this representation of the smoothing spline was used to study asymptotics and algorithms for shape restricted smoothing splines, see also Section 5.5.

4.1. 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 $\widehat{m}_S$ that is constructed such that it has p derivatives we define the constrained smoother as:

$$\begin{aligned}\widehat{m}_{S,C}(x) &= \operatorname*{argmin}_{m:\underset{\rightarrow}{m}\in\mathcal{C}_S^m} \left\| \underset{\rightarrow}{\widehat{m}}_S - \underset{\rightarrow}{m} \right\|^2 \\ &= \operatorname*{argmin}_{m\in C} \int \{\widehat{m}_S(x) - m(x)\}^2 w(x)\,\nu(dx) \\ &\qquad +\lambda \int \{\widehat{m}_S^{(p)}(x) - m^{(p)}(x)\}^2 dx.\end{aligned}$$

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

$$\|f\|^2 = \int f(x)^2 w(x)\,\nu(dx) + \lambda \int f^{(p)}(x)^2 dx. \tag{4.4}$$

This estimate has two advantages:

(1) The unconstrained estimate $\widehat{m}_S$ 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 $\widehat{m}_S$ will only be changed in neighborhoods of sets where the monotonicity was violated by $\widehat{m}_S$. 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 importantly, the good interpretability of the unconstrained estimator carries over to the constrained estimator away from such neighborhoods.

(2) The constrained estimate $\widehat{m}_{S,C}$ is a smooth function. The reason is that the penalty term $\lambda \int \left[m^{(p)}(x)\right]^2 dx$ of the Sobolev norm forces $\widehat{m}_{S,C}$ 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 2.1 for monotone local linear fits. This is shown in Figure 1.2 where the constrained smoother of Figure 1.1 is shown. That projection is calculated with respect to (4.4) 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 *et al.* (1998). Delecroix *et al.* (1995, 1996) consider a related two step procedure for Priestley-Chao type kernel smoothers.

5. Asides

5.1. The monotone case

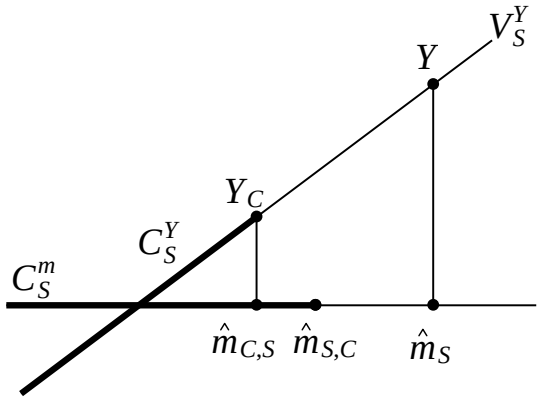
For monotone smoothing, $\widehat{m}_{S,C}(x)$ is a version of the older idea of "smooth, then monotonize" discussed e.g. in Barlow and van Zwet (1970), Wright (1982), Friedman and Tibshirani (1984), Mukerjee (1988), Kelly and Rice (1990) and Mammen (1991a) (see also Cheng and Lin, 1981; Ramsay, 1998;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

It can be shown that for monotone (increasing) smoothing (3.5) implies that

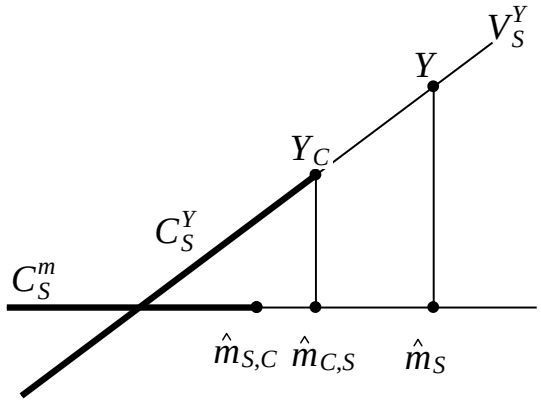
$$\widehat{m}_{S,C}(x) = \max_{u \le x} \min_{v \ge x} \frac{\int_u^v \widehat{m}_S(s) w(s)\, \nu(ds)}{\int_u^v w(s)\, \nu(ds)}. \tag{5.1}$$

A proof of (5.1) for discrete measures ν can be found in the books by Barlow *et al.* (1972) or Robertson *et al.* (1988). The case of general ν is discussed in Mammen *et al.* (1998). A careful inspection of (5.1) shows that one obtains the monotone function $\widehat{m}_{S,C}$ from $\widehat{m}_S$ by replacing parts of $\widehat{m}_S$ by constant pieces. In an interval where $\widehat{m}_{S,C}$ is constant it is equal to a weighted average of $\widehat{m}_S$ over this interval. At the boundary of such intervals $\widehat{m}_{S,C}$ may not be differentiable. This explains the kinks that were observed for the monotone smoother of the data in Section 2, see Figure 2.1.

Mammen (1991a) also considers other proposals for monotone smoothing that are of the form "monotonize then smooth", denoted by $\widehat{m}_{C,S}$, which is a smooth of the monotonized data denoted by Y_C. Insight into how this type of smoother compares with $\widehat{m}_{S,C}(x)$ comes from Figure 5.1. In both Figures 5.1(a) and 5.1(b), the subspace $\mathcal{V}_S^m$ (of ordinary functions) is shown as a horizontal line, and the



(a)



(b)

Figure 5.1: Diagram showing relation of "monotonicity preserving smoothers", Panel (a), and "non-monotonicity preserving smoothers", Panel (b), in the vector space $\mathcal{V}_S$.

subset $\mathcal{C}_S^m$ (of constrained functions) is the heavily shaded portion. The subspace $\mathcal{V}_S^Y$ (of ordinary vectors) is shown as a diagonal line, and the subset $\mathcal{C}_S^Y$ (of vectors satisfying the constraint) is the heavily shaded portion. Figure 5.1(a) corresponds to the case that the smoother $\widehat{m}_{C,S}$ is "monotonicity preserving" (i.e. when applied to monotone data, the result is monotone), and Figure 5.1(b) is the case where

the smoother is not monotonicity preserving, which can happen for example for local polynomial smoothers, as shown in Figure 6.1.

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

When the smoother is not monotonicity preserving, the smooth $\widehat{m}_{C,S}$ of the monotonized data Y_C, i.e. the projection of Y_C onto $\mathcal{V}_S^m$, need not be monotone, as shown in Figure 5.1(b). Another illustrative example for the situation in Figure 5.1(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 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 4.1 is a way of addressing this problem.

5.2. 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 (2.1), this entails different versions of the implied loss (3.6) .

For Nadaraya-Watson weights, $w(x) = \frac{1}{n}\sum_{i=1}^n K_h(x - X_i)$ 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

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

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, $w(x) = \frac{1}{n}\sum_{i=1}^n \int_{s_{i-1}}^{s_i} K_h(x-t)dt = \frac{1}{n}\int_{s_0}^{s_n} K_h(x-t)dt$. 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

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

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

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 e.g. the case of monotone smoothing and assume that the smoother is nearly monotone with the exception of some local wiggles. As noted at (5.1) 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.*, 1998). So usually the choice of the weight measure ν is of relatively minor importance.

5.3. ANOVA decompositions and model choice

Our projection framework can also be used for comparison of models and model choice. For example assume that we have a class of nested submodels $\mathcal{C}_{S,1}^m \subset \ldots \subset \mathcal{C}_{S,k}^m \subset \mathcal{V}_S^m$ given. Our approach allows us to compare the corresponding estimates using the norm (3.1) or its generalisation (4.2). Define for $j = 1, \ldots, k$ the constrained estimates analogous to (3.3):

$$\underset{\rightarrow}{\widehat{m}}_{S,C,j} = \underset{\underset{\rightarrow}{m} \in \mathcal{C}_{S,j}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{m} \right\|^2 .$$

If the submodels $\mathcal{C}_{S,2}^m$, ..., $\mathcal{C}_{S,k}^m$ are vector spaces, repeated application of the Pythagorean Theorem yields:

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

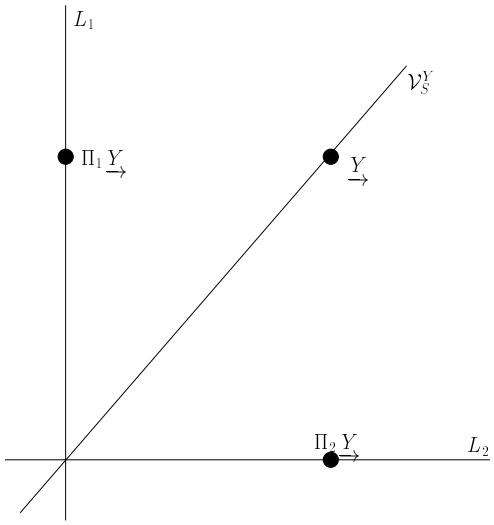


Figure 5.2: Diagram showing the data vector $\underset{\rightarrow}{Y}$ and projections $\Pi_1 \underset{\rightarrow}{Y}$ and $\Pi_2 \underset{\rightarrow}{Y}$ onto the orthogonal spaces $L_1 = \mathcal{C}_{S,j}^{m,\perp} \cap \mathcal{C}_{S,j+1}^m$ and $L_2 = \mathcal{C}_{S,j'}^{m,\perp} \cap \mathcal{C}_{S,j'+1}^m$.

As opposed to "traditional" ANOVA decompositions, the summands in this decomposition are usually 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 (3.1). It follows that $\underset{\rightarrow}{Y}$ has a standard normal multivariate distribution on the vector subspace $\mathcal{V}_S^Y$.

Consider 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 5.2. 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, i.e. $L_1 = \mathcal{C}_{S,j}^{m,\perp} \cap \mathcal{C}_{S,j+1}^m$ 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 $\underset{\rightarrow}{\widehat{m}}_{S,C,j+1} - \underset{\rightarrow}{\widehat{m}}_{S,C,j}$ and $\Pi_2 \underset{\rightarrow}{Y}$ is $\underset{\rightarrow}{\widehat{m}}_{S,C,j'+1} - \underset{\rightarrow}{\widehat{m}}_{S,C,j'}$.

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 3). 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 5.2. Here, $\underset{\rightarrow}{Y}$ has a one (!)-dimensional normal distribution on the line $\mathcal{V}_S^Y$ and $\Pi_1 \underset{\rightarrow}{Y}$ depends determin-

istically on $\Pi_2 \underset{\rightarrow}{Y}$. This implies, in particular, that they are not independent.

Furthermore, in general the summands $\left\| \widehat{\underset{\rightarrow}{m}}_{S,C,k} - \widehat{\underset{\rightarrow}{m}}_{S,C,k-1} \right\|^2$ do not have an (asymptotic) χ^2 distribution, see e.g. Härdle and Mammen (1993) who propose using bootstrap methods to avoid these problems. The situation is a little bit simpler for orthogonal series estimates, see Section 3.1. For a general discussion of lack-of-fit tests in nonparametric regression see Hart (1997).

5.4. 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 (3.3), only vectors of dimension g need to be considered for (3.5). 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 studied well 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 *et al.* (1988). General optimization algorithms are discussed, among others, in Fletcher (1987), den Hertog (1994) and Nash and Sofer (1996).

5.5. 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 one. This implies that, with probability tending to one, 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 one. 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 (1998) 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).

6. 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 3.1, a local polynomial smoother of order p, can be written as

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

where

$$\widehat{\beta}(x) = \begin{bmatrix} \widehat{\beta}_0(x) \\ \vdots \\ \widehat{\beta}_p(x) \end{bmatrix} = \operatorname*{argmin}_{\beta} \int \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j(x)(x - X_i)^j \right\}^2 w_i(x)\, \nu(dx). \tag{6.1}$$

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

It is possible to represent $\widehat{m}_{LP}(x)$ in the form (2.1), as

$$\widehat{\beta}(x) = \operatorname*{argmin}_{\beta} \{Y - X(x)\beta(x)\}^T W(x) \{Y - X(x)\beta(x)\}$$

where

$$X(x) = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ 1 & X_2 - x & \cdots & (X_2 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix},$$

$$W(x) = \begin{bmatrix} w_1(x) & 0 & \cdots & 0 \\ 0 & w_2(x) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_n(x) \end{bmatrix}.$$

Standard linear algebra yields

$$\widehat{\beta}(x) = \left\{X(x)^T W(x) X(x)\right\}^{-1} X(x)^T W(x) Y. \tag{6.2}$$

Hence, $\widehat{m}_{LP}(x)$ can also be written as a "simple smoother" in the form (2.1). Note, that the weights, say $\tilde{w}_i(x)$, used when writing $\widehat{m}_{LP}(x)$ in the form (2.1) differ from $w_i(x)$ used in (6.1) to define the local polynomial smoother. Moreover, it is possible that $\tilde{w}_i(x)$ becomes negative but, since $\widehat{m}_{LP}(x)$ reproduces constant functions, we are assured that $\sum_{i=1}^n w_i(x) = 1$ and Proposition 1 holds as noted in Section 3.1. The calculations in Section 2 could now be used for constrained smoothing, but there are some limitations to this setup. In particular only constraints on $\widehat{\beta}_0(x)$ would be allowed.

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

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

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 $\underset{\rightarrow}{Y}^T = [Y_1, 0, \ldots, 0, Y_2, 0, \ldots, 0, Y_n, 0, \ldots, 0]$, i.e. within blocks of $p+1$, only the first entries may be nonzero, i.e.

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

The subspace of such $n(p+1)$-tuples is called $\mathcal{V}_{LP}^Y$. A candidate smooth now involves several functions $\beta_j : I\!R \to I\!R$, 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 $\beta_j(x)$, i.e. $f_{i,j}(x) = \beta_j(x)$, $i = 1, \ldots, n$, $j = 0, \ldots, p$. 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 (3.1) on $\mathcal{V}_{LP}$ is

$$\left\| \underset{\rightarrow}{f} \right\|^2 = \int \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=0}^{p} f_{i,j}(x) \, (x - X_i)^j \right\}^2 w_i(x) \, \nu(dx). \tag{6.3}$$

This notation represents local polynomial smooths as a projection, because $\widehat{m}_{LP}(x) = \widehat{\beta}_0(x)$,where (6.1) can be rewritten as

$$\widehat{\beta}(x) = \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{V}_{LP}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \right\|^2. \tag{6.4}$$

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 $\widehat{m}_{LP,C}(x) = \widehat{\beta}_{0,C}(x)$, where

$$\widehat{\beta}_C(x) = \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}_{LP}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \right\|^2, \tag{6.5}$$

This constrained minimization is simplified, exactly as at (3.4), using a Pythagorean relationship. Following the same arguments (with nearly the same notation) as in Section 2 yields:

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

$$\begin{aligned} \widehat{\beta}(x) &= \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{\widehat{\beta}} - \underset{\rightarrow}{\beta} \right\|^2 \\ &= \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}^m}{\operatorname{argmin}} \int \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=0}^{p} \left(\widehat{\beta}_j(x) - \beta_j(x) \right) (x - X_i)^j \right]^2 w_i(x) \, \nu(dx) \\ &= \underset{\beta \in C_{LP}}{\operatorname{argmin}} \int \sum_{j=0}^{p} \sum_{j'=0}^{p} \left(\widehat{\beta}_j(x) - \beta_j(x) \right) \left(\widehat{\beta}_{j'}(x) - \beta_{j'}(x) \right) U_{j+j'}(x) \, \nu(dx) \end{aligned} \tag{6.6}$$

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 Proposition 1 in Section 3 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 5.4 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 $\beta_1(x) \geq 0$, see below for the latter case.

Suppose that the restricted $\beta_j(x)$ are grouped into a vector as $\beta_-(x)^T = (\beta_0(x), \cdots, \beta_{q-1}(x))$, and that $\beta_+(x)^T = (\beta_q(x), \cdots, \beta_p(x))$ is a grouping of the unrestricted ones. Then the minimization problem (6.6) can be further simplified, by explicitly minimizing in $\beta_+(x)$ for fixed $\beta_-(x)$. Useful notation is

$$\begin{aligned} \delta_-(x)^T &= \left(\widehat{\beta}_0(x) - \beta_0(x), \cdots, \widehat{\beta}_{q-1}(x) - \beta_{q-1}(x)\right), \\ \delta_+(x)^T &= \left(\widehat{\beta}_q(x) - \beta_q(x), \cdots, \widehat{\beta}_p(x) - \beta_p(x)\right), \end{aligned}$$

$$U(x) = \begin{bmatrix} U_0(x) & \cdots & U_p(x) \\ \vdots & & \vdots \\ U_p(x) & \cdots & U_{2p}(x) \end{bmatrix} = X(x)^T W(x) X(x).$$

Also let $U_{--}(x)$, $U_{+-}(x)$, and $U_{++}(x)$ denote respectively the upper left $q \times q$, lower left $(p-q) \times q$, and lower right $(p-q) \times (p-q)$ submatrices of $U(x)$. Calculations as done for (6.2) show that for given $\beta_-(x)$, the minimizer of (6.6), i.e.

$$\int [\delta_-(x)^T, \delta_+(x)^T] U(x) \begin{bmatrix} \delta_-(x) \\ \delta_+(x) \end{bmatrix} \nu(dx)$$

over $\beta_+(x)$ is given by the $\beta_+(x)$ component of

$$\delta_+(x) = -U_{++}(x)^{-1} U_{+-}(x) \delta_-(x)$$

Hence, the minimization problem (6.6) can be reduced to minimizing

$$\int \delta_-(x)^T \left\{U_{--}(x) - U_{++}(x)^{-1} U_{+-}(x)\right\} \delta_-(x) \, \nu(dx)$$

over $\beta_-(x)$.

In the case $q = 1$, this reduces to

$$\widehat{m}_{LP,C}(x) = \operatorname*{argmin}_m \int \{\widehat{m}_{LP}(x) - m(x)\}^2 \rho(x) \, \nu(dx)$$

where $\rho(x) = U_{--}(x) - U_{++}(x)^{-1}U_{+-}(x)$. But $U_{--}(x) = U_0(x) = w(x)$, defined before Proposition 1.

Similar remarks as in Section 5.2 now apply. In particular, in the case of weights $w_i(x) = K_h(x - X_i)$, under some assumptions, $\rho(x) \approx f(x)$, as for the Nadaraya-Watson smoother.

We describe now an algorithm for the following special case of monotone smoothing: for the local linear, do monotonization with the constraints $\beta_0(x)$ increasing, $\beta_1(x) \geq 0$. Straight forward calculus shows that this gives the minimization problem

$$\underset{\beta_0 \text{ increasing}}{\operatorname{argmin}} \int \kappa(x)dx,$$

where

$$\kappa(x) = \begin{cases} I & \text{if } x \in A \\ II & \text{otherwise} \end{cases}$$

where

$$A = \{x : U_2(x)^{-1}U_1(x)\left[\beta_0(x) - \widehat{\beta}_0(x)\right] \leq \widehat{\beta}_1(x)\},$$

$$I = \{U_0(x) - U_2(x)^{-1}U_1(x)^2\}\{\beta_0(x) - \widehat{\beta}_0(x)\}^2,$$

$$II = U_0(x)\{\beta_0(x) - \widehat{\beta}_0(x)\}^2 - 2\widehat{\beta}_1(x)U_1(x)\{\beta_0(x) - \widehat{\beta}_0(x)\} + U_2(x)\widehat{\beta}_1(x)^2.$$

This minimization can be done by the following iterative calculation. In each step the minimization is done for fixed set A. This gives a (weighted) least squares problem with monotonicity constraint (that can be solved e.g. by application of the pool adjacent violator algorithm). After each step the set A is updated by using the last solution for the minimizer.

Propsition 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 6.1 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.* (1998).

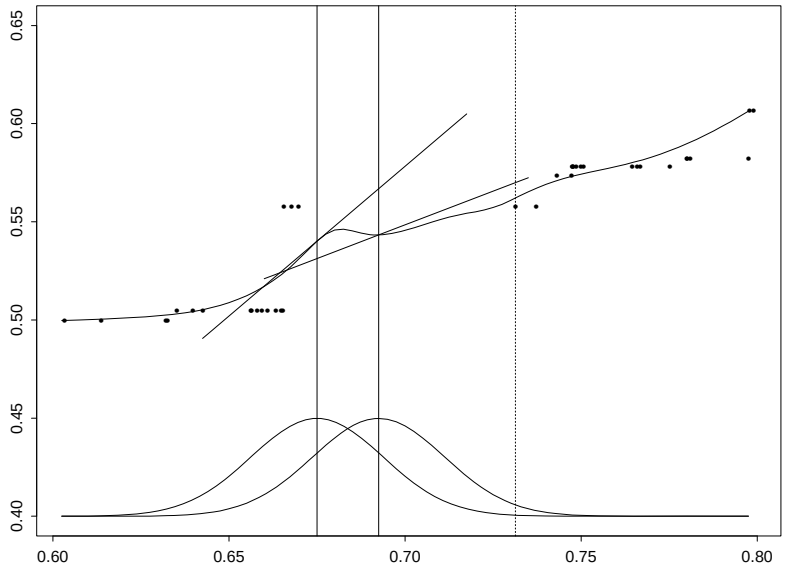


Figure 6.1: Monotone artificial data with nonmonotone local linear fit.

7. Additive models

We consider now local polynomial fitting for additive models. In this model the additive local polynomial 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 the asymptotic distribution of the additive local polynomial smoother, see Linton *et al.* (1997). We now describe how our projection framework carries over to this model. For this purpose we have to extend our approach to q dimensional covariables $X_i = (X_{i,1}, \ldots, X_{i,q})$. Our constraint on the regression function $m : \mathbb{R}^q \to \mathbb{R}$ is that

$$m(x) = m_0 + m_1(x_1) + \ldots + m_q(x_q) \quad \text{for } x = (x_1, \ldots, x_q), \tag{7.1}$$

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).

Given a set of weights $w_i(x)$, such as those of Section 3.1, the unconstrained

local polynomial of order p, can be written as

$$\widehat{m}_{AM}(x) = \widehat{\beta}_0(x),$$

where

$$\begin{aligned}
\widehat{\beta}(x) &= \begin{bmatrix} \widehat{\beta}_0(x) \\ \widehat{\beta}_{1,1}(x) \\ \vdots \\ \widehat{\beta}_{p,q}(x) \end{bmatrix} \\
&= \operatorname*{argmin}_{\beta} \int \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \beta_0(x) - \sum_{l=1}^{q} \sum_{j=1}^{p} \beta_{j,l}(x)(x_l - X_{i,l})^j \right\}^2 \\
&\qquad w_i(x)\, \nu(dx). \qquad (7.2)
\end{aligned}$$

As for $q = 1$, the integral and the weight measure ν play no role, because the minimization can be done individually for each x. In the minimization, no mixed terms of the form $(x_{l_1} - X_{i,l_1})_1^j \ (x_{l_2} - X_{i,l_2})_2^j$ are used. This reduces the number of fitted local parameters and it is natural in view of the constraint (7.1).

To use the constraint (7.1) we write $\widehat{\beta}$ as a projection. The space $\mathcal{V}_{AM}$ is now defined as a set of $n(pq+1)$ functions

$$\mathcal{V}_{AM} = \left\{ \underset{\rightarrow}{f} = \begin{bmatrix} f_{1,0}(x) \\ f_{1,1,1}(x) \\ \vdots \\ f_{1,p,q}(x) \\ \vdots \\ f_{n,0}(x) \\ f_{n,1,1}(x) \\ \vdots \\ f_{n,p,q}(x) \end{bmatrix} : \begin{array}{c} f_{i,0}, f_{i,j,l} : I\!R \to I\!R,\ i = 1, \ldots, n, \\ j = 1, \ldots, p,\ l = 1, \ldots, q \end{array} \right\}.$$

Similarly as above the data vector $Y^T = \begin{pmatrix} Y_1 & \cdots & Y_n \end{pmatrix}$ is viewed as an element $\underset{\rightarrow}{Y}$ of $\mathcal{V}_{AM}$, by putting $f_{i,0}(x) \equiv Y_i$ and $f_{i,j,l}(x) \equiv 0$ for $j = 1, \ldots, p, l = 1, \ldots, q$, $i = 1, \ldots, n$.

The subspace of such $n(pq+1)$-tuples is called $\mathcal{V}_{AM}^Y$. A candidate unconstrained smooth now involves several functions $\beta_0 : I\!R^q \to I\!R$ and $\beta_{j,l} : I\!R^q \to I\!R$. They

define an element $\underset{\rightarrow}{\beta}$ of $\mathcal{V}_{AM}$. Such elements are $n(pq+1)$-tuples where entries are common across i, and for each j and l are $f_{i,j,l}(x) = \beta_{j,l}(x)$ and $f_{i,0}(x) = \beta_0(x)$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, $l = 1, \ldots, q$. The subspace of such elements is again denoted by $\mathcal{V}_{AM}^m$. The appropriate norm on $\mathcal{V}_{AM}$ is now

$$\left\| \underset{\rightarrow}{f} \right\|^2 = \int \frac{1}{n} \sum_{i=1}^{n} \left[f_{i,0}(x) + \sum_{j=1}^{p} \sum_{l=1}^{q} f_{i,j}(x) \, (x_l - X_{i,l})^j \right]^2 w_i(x) \, \nu(dx). \tag{7.3}$$

Note that (7.2) can be rewritten as

$$\widehat{\beta}(x) = \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{V}_{AM}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \right\|^2. \tag{7.4}$$

The subset of constrained functions $\mathcal{C}_{AM}^m \subset \mathcal{V}_{AM}^m$ now consists of $n(pq+1)$ tuples of functions $(f_{i,0}, f_{i,j,l} : i = 1, \ldots, n,\ j = 1, \ldots, p,\ l = 1, \ldots, q)$ for which the following holds:

- The functions $f_{i,0}, f_{i,j,l}$ do not depend on i.
- The function $f_{i,0}$ is of additive form, i.e. there exist functions $g_1^f, \ldots, g_q^f : I\!R \to I\!R$ such that $f_{i,0}(x) = g_1^f(x_1) + \ldots + g_q^f(x_q)$.
- The functions $f_{i,j,l}$ depend only on a one dimensional argument, i.e. there exist functions $h_{1,1}^f, \ldots, h_{p,q}^f : I\!R \to I\!R$ such that $f_{i,j,l}(x) = h_{j,l}^f(x_l)$.

The additive local polynomial smoother is now defined as $\widehat{m}_{AM,C}(x) = \widehat{\beta}_{0,C}(x)$, where

$$\widehat{\beta}_C(x) = \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}_{AM}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \right\|^2, \tag{7.5}$$

Again, using the same arguments as above one can show that

$$\widehat{\beta}_C(x) = \underset{\beta:\, \underset{\rightarrow}{\beta} \in \mathcal{C}_{AM}^m}{\operatorname{argmin}} \left\| \underset{\rightarrow}{\widehat{\beta}} - \underset{\rightarrow}{\beta} \right\|^2.$$

However, in this model we do not recommend first calculating the unrestricted estimate [and then projecting this estimate on the subspace $\mathcal{C}_{AM}^m$]. The reason is

that the calculation of the unrestricted estimate involves many unknown parameters. If the data are too sparse this calculation would be instable or the estimate may not even be defined for many points. 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 $\| \underset{\rightarrow}{Y} - \underset{\rightarrow}{\beta} \|^2$. In each minimization step the norm is minimized over one additive component while letting the other components be fixed, i.e. for one $1 \leq k \leq q$ it is minimized over $g_k^{\widehat{\beta}}(x)$ and $h_{1,k}^{\widehat{\beta}}(x), \ldots, h_{p,k}^{\widehat{\beta}}(x)$ with fixed $g_l^{\widehat{\beta}}(x)$ and $h_{j,l}^{\widehat{\beta}}(x)$ for $j = 1, \ldots, p$ and $l \neq k$. 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 $\mathcal{V}_{AM}$. 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 Linton *et al.* 1997). 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 Linton *et al.* (1997) show that the local polynomial 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 (1997) and Opsomer and Ruppert (1997).

8. 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 (1998) note that US 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 (1998) consider fitting a decreasing regression function m with $m(0) = 100$ and $0 \leq m(x) \leq 100$.

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

- *Branching curves.* One observes r samples that are modeled as

$$Y_{ji} = m_j(X_{ji}) + \varepsilon_{ji}, \quad j = 1, \ldots, r; i = 1, \ldots, n_j.$$

  For the r regression functions $m_1, \ldots, m_r$ the model assumption is made that for some fixed known values τ_{jl} it holds that $m_j(x) = m_l(x)$ for $x \leq \tau_{jl}$. Smoothing splines for this model have been discussed in Silverman and Wood (1987), see also Green and Silverman (1994).

- *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).

References

[1] Barlow, R.E. and van Zwet, W.R. (1970) Asymptotic properties of isotonic estimators for generalized failure rate function. Part 1: Strong consistency. *Nonparametric techniques in statistical inference*, (M.L. Puri, ed.) 159–173, Cambridge University Press.

[2] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972) *Statistical inference under order restrictions*. Wiley, New York.

[3] Bowman, A.W. and Azzalini, A. (1997) *Applied smoothing techniques for data analysis*, Oxford Science Publications, Oxford.

[4] Cheng, K.F. and Lin, P.E. (1981) Nonparametric estimation of a regression function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57, 223–233.

[5] Chu, C.K. and Marron, J.S. (1991) Choosing a kernel regression estimator (with discussion). *Statistical Science*, 6, 404–436.

[6] Cox, D.D. (1988) Approximation of method of regularization estimators. *Annals of Statistics*, 2, 694–712.

[7] Delecroix, M. and Thomas-Agnan, C. (1997) Kernel and spline smoothing under shape restrictions. to appear in: *Smoothing and Regression: Approaches, Computation and Application* (M. Schimek, ed.), Wiley, New York.

[8] Delecroix, M., Simioni, S. and Thomas-Agnan, C. (1995) A shape constrained smoother: Simulation study. *Computational Statistics*, 10, 155–175.

[9] Delecroix, M., Simioni, S. and Thomas-Agnan, C. (1996) Functional estimation under shape restrictions. *Journal of Nonparametric Statistics* 6, 69–89.

[10] den Hertog, D. (1994) *Interior Point Approach to Linear, Quadratic and Convex Programming*. Kluwer Academic Publishers, Dordrecht.

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

[12] Dole, D. (1996) Scatterplot smoothing subject to monotonicity and convexity. Unpublished manuscript.

[13] Dykstra, R.L. (1983) An algorithm for restricted least squares regression. *Journal of the American Statistical Association* 77, 621–628.

[14] Elfving, T. and Andersson, L.E. (1988) An algorithm for computing constrained smoothing splines. *Numerische Mathematik*, 52, 583-595.

[15] Eubank, R.L. (1988) *Spline smoothing and nonparametric regression*. Marcel Dekker, New York.

[16] Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

[17] Fletcher, R. (1987) *Practical Methods of Optimization*. Second Edition, Wiley, Chichester.

[18] Földes, A. and Revesz, P. (1974) A general method for density estimation, *Studia Scientifica Mathematica Hungarica*, 9, 81-92.

[19] Friedman, J. and Tibshirani, R. (1984) The monotone smoothing of scatterplots. *Technometrics*, 26, 243–250.

[20] Fritsch, F.N. (1990) Monotone piecewise cubic data fitting. *Algorithms for Approximation II* (J.C. Mason and M.G. Cox, eds.), Chapman and Hall, London, pp. 99–106.

[21] Gaylord, C.K. and Ramirez, D.E. (1991) Monotone regression splines for smoothed bootstrapping. *Computational Statistics Quarterly*, 6, 85–97.

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

[23] Green, P.J. and Silverman, B.W. (1994) *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall, London

[24] Härdle, W. and Gasser, T. (1984) Robust nonparametric function fitting. *Journal of the Royal Statistical Society, Series B*, 46, 42–51.

[25] Härdle, W. (1990) *Applied Nonparametric Regression.* Cambrdige University Press, Cambridge, UK.

[26] Härdle, W. and Mammen, E. (1993) Testing parametric versus nonparametric regression. *Annals of Statistics* 21, 1926–1947.

[27] Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer-Verlag, New York.

[28] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[29] He, X. and Ng, P. (1998) COBS: Qualitatively Constrained Smoothing via Linear Programming. *Computational Statistics.* To appear

[30] Irvine, L.D., Marin, S.P. and Smith, P.W. (1986) Constrained interpolation and smoothing. *Constructive Approximation*, 2, 129–151.

[31] Kelly, C. and Rice, J.R. (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46, 1071–1086.

[32] Linton, O., Mammen, E. and Nielsen, J. (1997) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Unpublished manuscript.

[33] Mammen, E. (1991a) Estimating a smooth monotone regression function. *Annals of Statistics*, 19, 724–740.

[34] Mammen, E. (1991b) Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics*, 19, 741–759.

[35] Mammen, E. and Marron, J.S. (1997) Mass recentered kernel smoothers. *Biometrika*, 84, 765–778.

[36] Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. (1998b) Monotone local polynomial smoothers. Forthcoming manuscript.

[37] Mammen, E. and Thomas-Agnan, C. (1998) Smoothing splines and shape restrictions. *Scandinavian Journal of Statististics*, to appear.

[38] McCormick, G.P. (1983) *Nonlinear programming: theory, algorithms and applications.* Wiley, New York.

[39] Micchelli, C.A. and Utreras, F.I. (1988) Smoothing and interpolation in a convex subset of a Hilbert space. *SIAM Journal of Scientific and Statistical Computing*, 9, 728–746.

[40] Müller, H.G. (1988), *Nonparametric regression analysis of longitudinal data.* Springer-Verlag, New York.

[41] Mukerjee, H. (1988) Monotone nonparametric regression. *Annals of Statistics*, 16, 741–750.

[42] Nash, S.G. and Sofer, A. (1996) *Linear and Nonlinear Programming.* McGraw–Hill, New York.

[43] Opsomer, J.D. (1997) On the existence and asymptotic properties of backfitting estimators. *Preprint.*

[44] Opsomer, J.D. and Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186–211.

[45] Ramsay, J.O. (1988) Monotone regression splines in action (with discussion). *Statistical Science*, 3, 425–461.

[46] Ramsay, J.O. (1998) Estimating smooth monotone functions. *Journal of the Royal Statistical Society*, 60, 365–375.

[47] Ramsay, J.O. and Silverman, B.W. (1997) *Functional data analysis.* Springer-Verlag, New York.

[48] Robertson, T., Wright, F.T. and Dykstra, R.L. (1988) *Order restricted statistical inference.* Wiley, New York.

[49] Rudin, W. (1987) *Real and Complex Analysis.* McGraw–Hill, New York.

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

[51] Schmidt, J.W. and Scholz, I. (1990) A dual algorithm for convex-concave data smoothing by cubic C^2-splines. *Numerische Mathematik*, 57, 333–350.

[52] Schwetlick, H. and Kunert, V. (1993) Spline smoothing under constraints on derivatives. *Bit*, 33, 512–528.

[53] Silverman, B.W. (1986) *Density estimation for statistics and data analysis.* Chapman and Hall, London.

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

[55] Simonoff, J. S. (1996) *Smoothing methods in statistics*, Springer-Verlag, New York.

[56] 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) *Annals of Statistics*, 25, 1371–1424.

[57] Tantiyaswasdikul, C. and Woodroofe, M.B. (1994) Isotonic smoothing splines under sequential designs. *Journal of Statistical Planning and Inference*, 38, 75–88.

[58] Tsybakov, A.B. (1986) Robust reconstruction of functions by the local approximation method. *Prob. Info. Transmission*, 22, 133–46.

[59] Utreras, F. (1985) Smoothing noisy data under monotonicity constraints: existence, characterization and convergence rates. *Numerische Mathematik*, 47, 611–625.

[60] van de Geer, S. (1990) Estimating a regression function. *Annals of Statistics*, 18, 907–924.

[61] Villalobos, M. and Wahba, G. (1987) Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *Journal of the American Statistical Association*, 82, 239–248.

[62] Wahba, G. (1990) *Spline models for observational data*, Philadelphia: SIAM.

[63] Walter, G.G. and Blum, J. (1979) Probability density estimation using delta sequences, *Annals of Statistics*, 7, 328-340.

[64] Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. Chapman and Hall, London.

[65] Wright, F.T. (1982) Monotone regression estimates for grouped observations. *Annals of Statistics*, 10, 278–286.