Very Smooth Nonparametric Curve Estimation by Penalizing Change of Curvature

by

Martin Mächler

Research Report No. 71
May 1993

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Very Smooth Nonparametric Curve Estimation by Penalizing Change of Curvature

Martin Mächler
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

13 May 1993

Abstract

Usual non-parametric regression estimators such as smoothing splines or kernel estimators are good tools for optimal mean squared error approximation of many smooth functions. However, they often show many little wiggles which do not appear to be necessary for a good description of the data.

The new "Wp" smoother is a *Maximum Penalized Likelihood* estimate with a novel roughness penalty. It penalizes a relative *change* of curvature. This leads to disjoint classes of functions, each with a given number, n_w, of inflection points. For a "Wp" estimate, $f''(x) = \pm(x - w_1) \cdots (x - w_{n_w}) \cdot \exp h_f(x)$, which is *semi-parametric* with parameters w_j and nonparametric part $h_f(\cdot)$.

For a very general class of M.P.L. estimators, a convenient form of the characterizing differential equation is derived.

If the main smoothing parameter n_w is specified correctly, this approach yields very smooth functions (including derivatives) while apparently not suffering from erosion, the principal bias problem of traditional smoothers.

Key words: Roughness penalty; Inflection point; Maximum penalized likelihood; Robust smoothing.

Contents

1 Introduction

In the last decades, non-parametric regression methods have been developed to gain flexibility in regression problems of data analysis. The parametric models used in diverse areas have been too restrictive for many applications. The usual non-parametric regression curves such as smoothing splines (Silverman, 1985; Wahba, 1990; Eubank, 1988), kernel estimators (Härdle and Gasser, 1984; Müller, 1988; Härdle, 1990; Chu and Marron, 1991), or locally weighted regression 'LOWESS' (Cleveland, 1979) have the nice property to fit a vast class of smooth functions well. But they still may show many little wiggles which do *not* appear to be necessary for a good description of the data. A statistician using the "naked eye" would draw a curve with only as few inflection points as possible for a low-bias fit to the data.

Since "wiggles" are characterized by inflection points, one may ask for a smooth curve with as few inflection points as reasonably possible. This idea is made precise in the present paper in a more general concept of *change of curvature*.

Let us consider an example with real data. The 'housing starts' series from the software

Figure 1: Trend component (hs$trend) of the Housing Starts series as computed by the SABL procedure in S. The curve has 19 inflection points.

package S was de-seasonalized using 'SABL', and the resulting data (including the noise part) taken as raw data (in S statements: `hs <- sabl(hstart); data <- hs$trend + hs$irregular`). The trend component computed by `sabl` is a smooth of this data (figure 1) with 19 inflection points and a residual sum of squares of 9.70. Figure 2 suggests that a smooth curve which fits the data reasonably well only needs three inflection points. The smooth solid line is the result of the "Wp" procedure to be defined below. A cubic spline and a LOWESS curve are also shown. The smoothness parameters were chosen to produce the same residual sum of squares as the "Wp" smoother. The two competitors

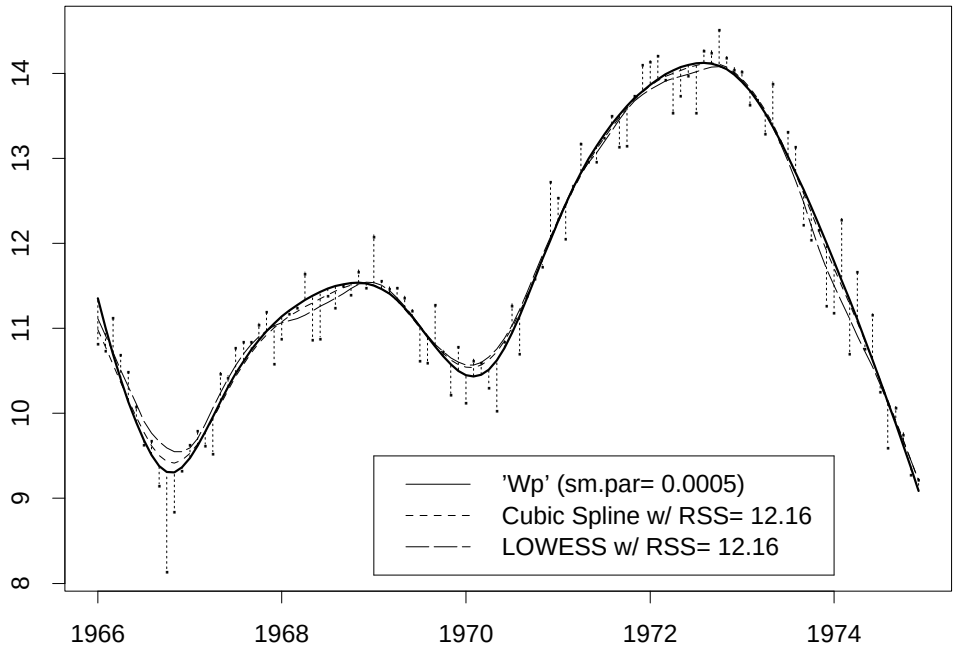


Figure 2: De-seasonalized 'housing starts', a times-series of length 108. A "Wp" smoother, restricted to 3 inflection points, with residual sum of squares = 12.16. Cubic splines and LOWESS are tuned to have the same residual sum of squares.

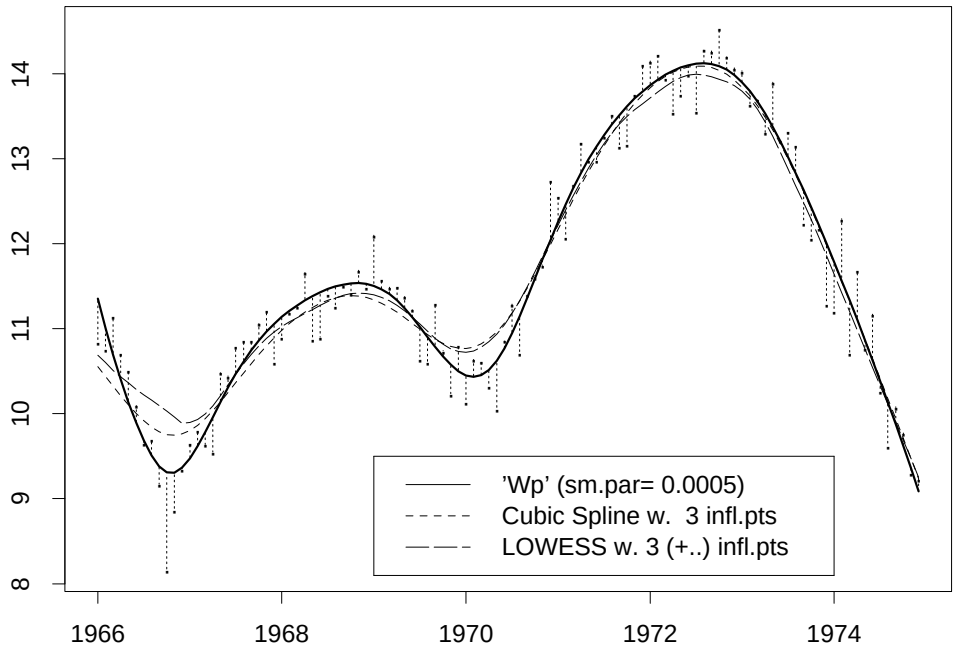
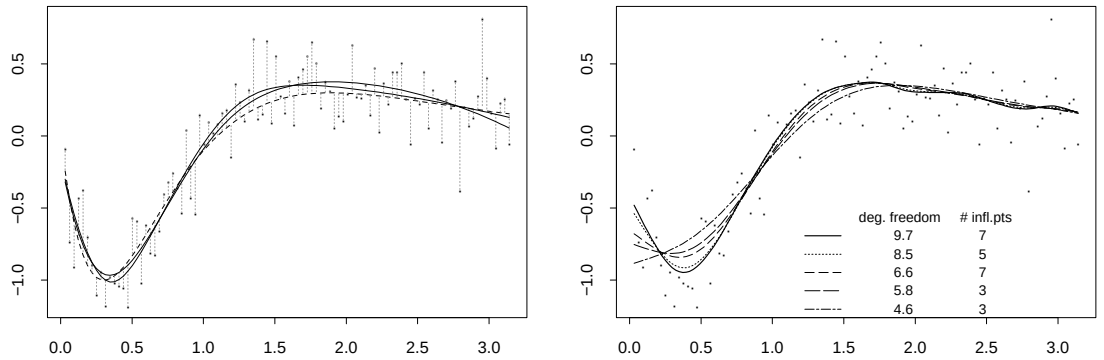


Figure 3: 'Housing starts' data. Cubic splines and LOWESS with the best fits for 3 inflection points (for LOWESS only if wiggles near both ends are omitted).

both show rather unsmooth behavior. In figure 3, the competitors were tuned to produce

only just 3 inflection points. Comparable to "Wp" in terms of smoothness, they now suffer from "erosion", i.e., bias near local extrema.

A second example for comparing splines and the "Wp" smoothers stems from Wahba and Wold (1975) (also cited in Wahba (1990, p. 45–46)). $n = 100$ data points (x_i, y_i) are generated as $x_i = \pi i/n$, and $y_i = f(x_i) + \frac{1}{5}U_i$, where U_i are i.i.d. standard normal variates, and the "true" function is $f(x) = 4.26e^{-x}(1 - e^{-x})(1 - 3e^{-x})$. In figure 4, the data clearly



(a) Data, generating function $f(x)$ (dashed line), and "Wp" smoothers with 1 inflection point and λ= .01 and .0004.

(b) Smoothing Splines with different smoothing parameters λ. The GCV curve is slightly thicker.

Figure 4: The example of Wahba and Wold (1975). (a) shows $f(x)$, the data and "Wp" smoothers, whereas (b) confirms that smoothing splines sometimes are either not smooth or suffer from considerable erosion.

suggest that one inflection point should suffice for a smooth curve fitting the data well. The cubic smoothing splines and other classical smoothers produce spurious wiggles in the right half, the GCV (generalized cross-validated) smoothing spline has as many as seven inflection points. Further smoothing reduces this extra variability in the right part (still leaving two spurious inflection points) but leads to an unacceptable amount of erosion near the minimum.

Maximum Penalized Likelihood

The approach which leads to the "Wp" procedure is based on the idea of maximizing a penalized likelihood. If

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \quad \varepsilon_i \sim H_i \text{ with density } h_i, \text{ independently,} \tag{1}$$

then the negative log-likelihood equals $\sum_{i=1}^{n} \rho_i (y_i - f(x_i))$, where $\rho_i = -\log h_i$. Typically, $\rho_i(x) = W_i \rho(x)$, where the weights W_i are given. For convenience, we assume that $x_1 \leq x_2 \leq \ldots \leq x_n$.

It is natural to ask for the function $\hat{f}$ which minimizes this sum subject to a bound B on a roughness measure of $R[f]$ to be defined below. The minimum under the restriction $R[f] \leq B$ will be attained at the boundary $R[f] = B$ (whenever $B < R[f_{\text{interpolating}}]$). As Reinsch (1971) proved for spline smoothing, this restricted variational problem can be restated using a Lagrange multiplier λ, as

$$\min_f \left\{ \sum_{i=1}^{n} \rho_i (y_i - f(x_i)) \; + \; \lambda R[f] \right\}.$$

Instead of the bound B, the multiplier λ can be fixed. It is then called *'smoothing parameter'* since higher values lead to smoother curves by giving more weight to the roughness penalty $R[f]$.

Let us briefly discuss the role of ρ. The usual least-squares choice $\rho(x) = x^2/2$ corresponding to normally distributed errors leads to estimators which are not robust. This means that they are highly influenced by only few outlying observations. From the theory of robustness, it is well known that one gets robust 'M-estimators' if $|\rho'| \leq c$ for some $c \in \mathbb{R}$ (Huber, 1979). The choice of Huber's ρ,

$$\rho_c(x) \stackrel{\text{def}}{=} \frac{1}{2} \left(x^2 - (|x| - c)_+^2 \right), \qquad \text{where } a_+ \stackrel{\text{def}}{=} \max(0, a), \tag{2}$$

gives the optimal 'minimax' estimators in the case of linear regression, minimizing the maximal variance over a full neighborhood of the normal distribution. Therefore, such a ρ leads to true 'non-parametric' estimation, whereas for $\rho(x) = ax^2$, the quality depends too much on the error distribution. In Mächler (1989), I show for the dataset above how a smooth based on the latter ρ is distorted by a single gross error, even for the "Wp" procedure restricted to three inflection points.

Spline approaches for reducing unnecessary fluctuations

The cubic smoothing splines approach uses integrated curvature as the roughness penalty R. As seen in the example above this may produce curves with spurious wiggles. This problem has been approached in several ways within the splines framework. Most of these fall into the two classes of *restricted* and *generalized splines*.

Restricted splines restrict the class of spline functions used, to monotonicity (Ramsay, 1988), piecewise monotonicity and/or convexity (Wright and Wegman, 1980; Mammen, 1991). Mammen shows that for a class of piecewise (m-)convex functions, restricted splines form a sufficient basis. Numerical analysts (Dierckx, 1980; Cox, 1973) have devised sophisticated algorithms allowing to prescribe the sign of the curvature at each data x_i and thus preventing inflection points in given data intervals. Dierckx uses regression splines and implements an automatic ad-hoc algorithm to choose the knots. My experience with his program and example show that the curves are sometimes nearly piecewise linear with

brisk changes of the slopes. This unsatisfactory behavior arises also in other restricted spline approaches.

It is known that interpolating splines applied to monotone data are not monotone in general. Nor are they convex for convex data. These bad properties carry over to smoothing as a generalization of interpolation. For interpolation, Micchelli, Smith, Swetits and Ward (1985) have shown that *constrained* polynomial splines uniquely minimize the usual splines criterion in the set of functions with restricted sign of the k-th derivative. Irvine, Marin and Smith (1986) have implemented the solution in an algorithm which subsequently was put into IMSL (subroutine "CSCON"). It gives a cubic spline which 'preserves shape', i.e., it is convex and concave where the data are. On the other hand, it is not co-monotone with the data in general. The number of knots is 'unpredictable' and often higher than the number of data. A similar tendency as above is observed: The 'smooth' functions, though being in C^2, are sometimes almost piecewise linear.

Generalized splines arise in approximation theory. They are piecewise very smooth functions with continuity conditions at the knots. Some are based on more general roughness penalties, e.g., the exponential splines. These are piecewise of the form $E(x) = a + bx + c/p^2(\cosh(px) - 1) + d/p^3(\sinh(px) - px)$. For $p \to 0$, $E(x)$ converges to a cubic polynomial. Hence, the cubic splines are a special limit case. It can be shown (Hess and Schmidt, 1986) that the knots and for each knot interval the 'weight' p (a completely free parameter) may be chosen such that convex data is interpolated by a convex exponential spline. This approach is appealing as a generalization of cubic splines, and properties and algorithms are established (Rentrop, 1980; Pruess, 1976), but ad-hoc methods are necessary to choose the many parameters and the application to non-interpolating smoothing is not clear.

2 Change of Curvature as a Roughness Penalty

The *smoothing splines* approach was originally based on the roughness measure of integrated squared curvature, $R[f] = \int_{x_1}^{x_n} \kappa(t)^2 \, dt$. The curvature can be expressed as $\kappa(x) = f''(x) \, (1 + f'(x)^2)^{-3/2}$. Traditionally, for computational and mathematical convenience, κ has been approximated by $\kappa(x) \approx c \cdot f''(x)$, but Glass (1966) indicates that using (exact) κ leads to more satisfactory results in the case of interpolation.

The present approach is based on measuring roughness by *relative* or *standardized change of curvature*

$$\kappa'/\kappa \;=\; f'''/f'' - 3f'f''/(1 + f'^2) \;\approx f'''/f''. \tag{3}$$

As a motivation, consider the following daily life situation: A bicycle or car driver has no difficulties in a curve with constant high curvature (which would be a sector of a circle). But there is high danger if the curvature is changing too rapidly, e.g., at the end or beginning of the curve. The most dangerous are the S-shaped curves, where the road's curvature changes sign.

The measure κ'/κ equals the "change of log curvature", $\frac{d}{dx} \log \kappa(x)$. The approximation $\kappa'/\kappa \approx f'''/f''$ follows from differentiating $\kappa(x) \approx cf''(x)$, the approximation used in the spline case. As (3) shows, it holds exactly at all the local extrema and inflection points which can be considered as the most interesting points of the curve. On the other hand, the approximation is clearly too small for example if $f(x) = e^{\lambda x}$, for large x. This unfavorable example also applies in the splines case. Since the corresponding curve segment does not need substantial smoothing, the problem is not important. The approximation leads to

the preliminary penalty

$$\tilde{R}[f] := \int_{x_1}^{x_n} \left(\frac{f'''(t)}{f''(t)} \right)^2 dt.$$

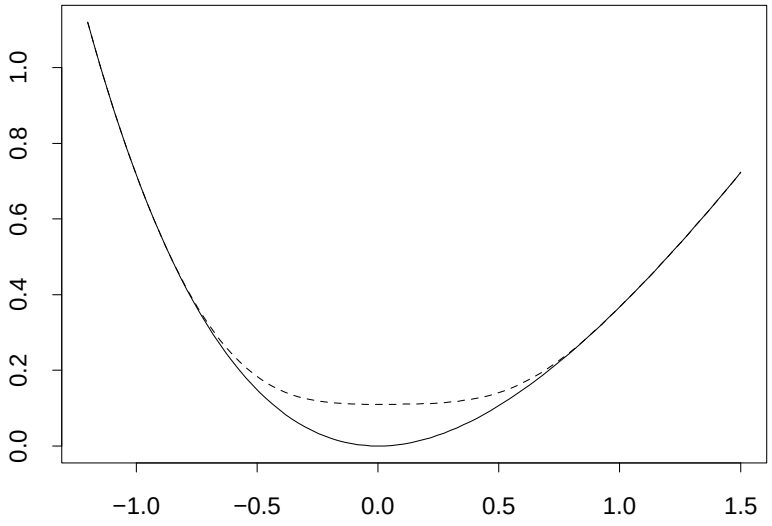


Figure 5: The functions (4) for $q = 0$ and $q = .11$. Both are convex.

For comparing the 'curvature' with the 'change of curvature' approach and their respective approximations consider the following smooth curves as an example. We take the sum of a linear and an exponential function and contaminate its smoothness, still retaining a convex curve,

$$f(x) := e^{-x} - 1 + x \; + \; q(1 - x^2)_+^4. \tag{4}$$

We have $f''(x) = e^{-x} + 8q(7x^2 - 1)(1 - x^2)_+^2$ which is positive for $q \leq .121$. Figure 5 shows the simple exponential (plus linear) for $q = 0$ and the function for $q = .11$. Note that the exponential looks much smoother in an intuitive way though both curves are convex. The curve for $q = .11$ "nearly" has an inflection point at $x = x^* = .060$ (i.e., $\min_x f''(x) = f''(x^*) = .090$). The curvature κ (figure 6(i)) is rapidly coming down near $x = 0$. This leads to large change of curvature (figure 6(ii)). The integrated *change* of curvature clearly distinguishes the roughness of the two curves, while the integrated curvature is similar in size: In the upper-right corner of each figure, we give the ratio of the integrals $\int_{-1.2}^{1.5} g_q(x)\, dx$ for the corresponding curves $y = g_q(x)$, $q = .11, 0$. Figure 7 shows that the approximation f'''/f'' is quite close to the intended κ'/κ of figure 6, but the "approximation" $c \cdot f''$ to κ is rather inadequate.

If f has the inflection points $w_1, w_2, \ldots, w_{n_w}$, then f'''/f'' has first order poles at these locations and "$\tilde{R}[f]$ contains n_w times ∞". This means that $n_w + 1$ inflection points are infinitely less smooth than n_w, and hence the number of inflection points is the principal roughness measure. In order to measure roughness for functions with the same (given) number of inflection points, we can rescale the problem appropriately, and define a roughness of the form $R[f] = \int_{x_1}^{x_n} \left(\frac{f'''(t)}{f''(t)} - \text{ "poles" } \right)^2 dt$.

More precisely, $w_1, \ldots, w_{n_w}$ shall be all the zeros of f''. Multiple zeros are enumerated explicitly, i.e., $w_j = w_k$ for $j \neq k$. The zeros of odd order are the inflection points of f.

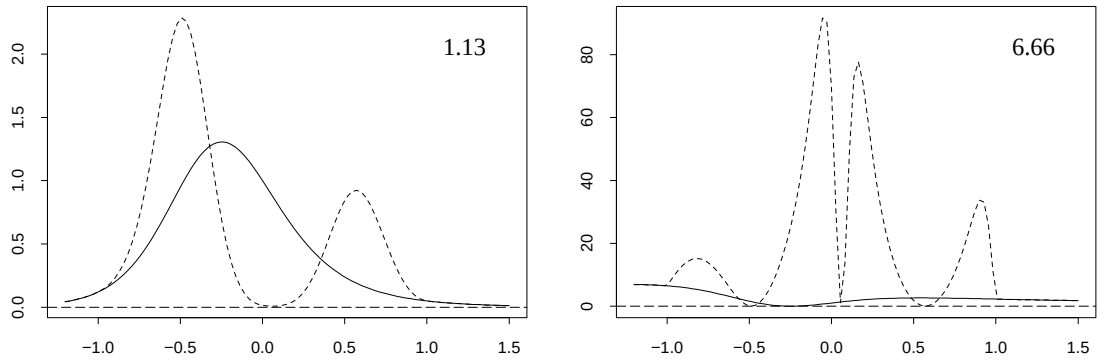


Figure 6: 'Ideal' penalty curves, using curvature κ, for functions (4) with $q = .11$ (dotted line) and $q = 0$ (full line). The figures (1.13 and 6.66) are the ratio of the areas beneath these two curves.

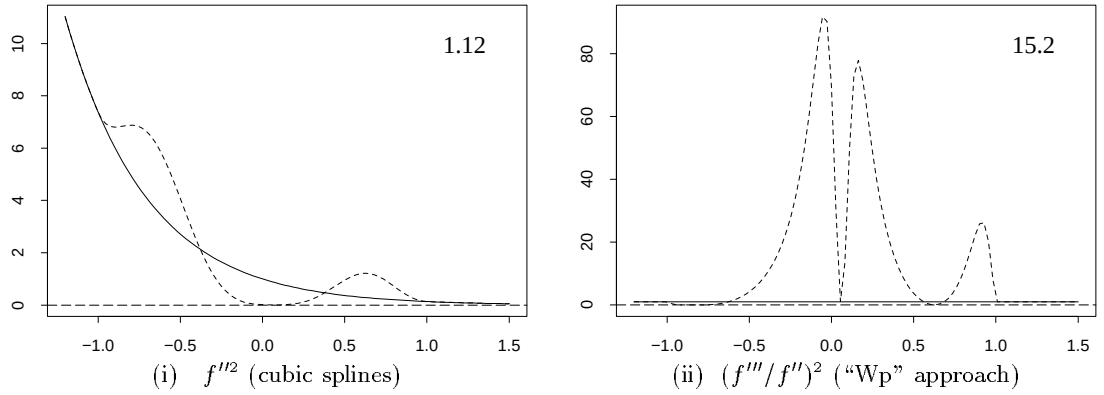


Figure 7: Penalty curves for functions (4); "approximations" of figure 6.

Note that multiple zeros will rarely arise for reasonable models for f and real data would hardly suggest them.

By elementary calculus, under weak regularity conditions, e.g., $f'''(w_j) \neq 0$ for simple zeros, f'' is of the form $f''(x) = (x-w_1)(x-w_2)\cdots(x-w_{n_w})\cdot q_f(x)$ where q_f has no zero and is of the same differentiability as f''. Hence, it can be written as $q_f(x) = s_f e^{h_f(x)}$, where $s_f = \pm 1$ and the function h_f is as many times differentiable as f''. More conveniently, we define the degree n_w polynomial,

$$p_{\mathbf{w}}(x) \stackrel{\text{def}}{=} s_f(x - w_1)(x - w_2)\cdots(x - w_{n_w}), \tag{5}$$

and have

$$f''(x) = p_{\mathbf{w}}(x)\, e^{h_f(x)}, \quad \text{or} \tag{6}$$

$$h_f = \log(f''/p_{\mathbf{w}}). \tag{7}$$

Hence, $f'''/f'' = \frac{d}{dx}\log f'' = (\log p_{\mathbf{w}})' + h_f{}'$, or

$$h_f{}' = \frac{f'''}{f''} - \sum_{j=1}^{n_w} \frac{1}{x - w_j}. \tag{8}$$

Note that the sum containing the singularities is independent of f. This allows to "discount" the inflection points in a way which is independent of all other aspects of f. Thus, the penalty

$$R[f] = \int_{x_1}^{x_n} h_f{}'(t)^2 \, dt, \tag{9}$$

is suitable for measuring the change of curvature "apart from the inflection points".

The number of inflection points, n_w, is the main smoothing parameter of this approach which we will call "Wp", abbreviating the German word 'Wendepunkt' for inflection point. The smoothing parameter λ controlling the weight of $R[f]$ is of less importance. Here, the limit $\lambda \to 0$ exists and corresponds to a smooth function (with only n_w inflection points) whereas for classical smoothers such as splines, one would get an interpolating curve.

To determine λ algorithmically, we look at the autocorrelations (ACF) of the residuals. It is intuitively clear that oversmoothing leads to positive autocorrelations at small lags. Therefore, start with a "big" λ, decrease it (about exponentially, i.e., linear in log scale) until the residual ACF doesn't show relevant structure anymore.

Generalizations of this approach are possible in two directions: First, we can work with $f^{(\nu)}$ instead of f''. For $\nu = 1$ this means considering local minima and maxima instead of inflection points, for $\nu > 2$, the inflection points of f' or higher derivatives. The second generalization consists in penalizing change of change of curvature instead of 'simple' change, i.e., using a general k-th derivative of h_f instead of $h_f{}'$. The present approach appears to be the most natural one to interpret. The generalizations on the other hand follow easily and one of their limiting cases for 'infinite smoothing' may seem somewhat more attractive than ours (see Corollary 4).

3 Variational Problem and Differential Equation

Given the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we want to determine the function f minimizing

$$\sum_{i=1}^{n} \rho_i \left(y_i - f(x_i)\right) + \lambda \int_{x_1}^{x_n} h_f{}'(t)^2 \, dt. \tag{10}$$

Equivalently, using Dirac's δ–distribution notation, the problem is to *solve*

$$\min_f \int_{x_1}^{x_n} \left(\sum_{i=1}^{n} \delta(t - x_i) \, \rho_i(y_i - f(t)) + \lambda \left(h_f{}'(t)\right)^2 \right) dt. \tag{11}$$

This section shows that f is necessarily a solution of the Euler-Lagrange differential equation (24). The basic result, Theorem 1 below, applies to a more general problem

"Minimize the functional $J[f]$ over $f \in \mathcal{F} \subset C^{2k+\nu}$" with the following specifications: Let $k \geq 0$ and $\nu \geq 1$ be given integers. We define

$$J[f] \quad \stackrel{\text{def}}{=} \quad \int_a^b \left\{ S(f(x), x) + \lambda \left(\frac{d^k}{dx^k} F(x, f^{(\nu)}(x)) \right)^2 \right\} dx, \tag{12}$$

$$F_g(x, g) \quad \stackrel{\text{def}}{=} \quad \frac{\partial}{\partial g} F(x, g), \tag{13}$$

$$S_f{}^{[0]}(x) \quad \stackrel{\text{def}}{=} \quad \frac{\partial}{\partial f} S(f, x) \Big|_{f=f(x)},$$

$$S_f{}^{[j+1]}(x) \quad \stackrel{\text{def}}{=} \quad \int_a^x S_f{}^{[j]}(t) \, dt, \quad \textit{for } 0 \leq j < \nu, \tag{14}$$

where we assume that $\tilde{F}(x) = F(x, f^{(\nu)}(x))$ *is* $2k$ *times differentiable and* $F_g(x, f^{(\nu)}(x))$ *is continuous. Further, for all* $f, \eta \in \mathcal{F}$ *(*$\mathcal{F} \subset C^{2k+\nu}$ *is specified later),* S *must fulfill*

$$\frac{d}{d\epsilon} \int_a^b S(f(x) + \epsilon\eta(x), x)\, dx \Big|_{\epsilon=0} = \int_a^b S_f{}^{[0]}(x)\, dx. \tag{15}$$

Note that $S_f{}^{[j]}$ *is the* j*-th principal function of* $S_f{}^{[0]}$*, i.e.,* $\frac{d^k}{dx^k} S_f{}^{[j]}(x) = S_f{}^{[j-k]}(x)$ *for* $0 \le k \le j$.

Note that this general formulation encompasses a vast class of maximum penalized likelihood problems, not only in nonparametric regression, but also density estimation, for example. The scatter term, which may not even be a log-likelihood, has the general form $\int_a^b S(f(x), x)\, dx$, satisfying the conditions (14)–(15). Typically, $S(f(x), x) = \sum_{i=1}^n \delta(x - x_i)\ell(f(x_i))$ and $S_f{}^{[1]}(x) = c + \sum_{i=1}^n \mathbf{1}_{[x \ge x_i]}\ell'(f(x_i))$.

In Mächler (1989), I proved the existence of a minimizer f of the $J[f]$ of Corollary 3, applying Tonelli's theorem (a "direct method") of variational calculus. One then sees that the function h_f (7) belongs to a decent (Sobolev) Hilbert space and the problem is well posed.

To find this optimal f, one can use the Euler-Lagrange (ordinary) differential equation ('o.d.e.') which asserts a *necessary* condition for f. Here, we also must determine the *"natural"* boundary conditions.

Theorem 1. *Using definitions and assumptions (12) to (15), a minimizer* f *of* $J[f]$ *fulfills the*

differential equation (i) $$F_g \cdot \frac{d^{2k}}{dx^{2k}} F = \frac{1}{2}(-1)^{\nu+k+1} S_f{}^{[\nu]}(x) \quad \forall x \in [a, b]$$

and, if $\frac{d^j}{dx^j} F_g(x, f^{(\nu)}(x)) \neq 0$ *for* $x \in \{a, b\}$ *and* $0 \le j \le k-1$ *, the*

boundary conditions (ii) *a)* $$S_f{}^{[1]}(b) = S_f{}^{[2]}(b) = \cdots = S_f{}^{[\nu]}(b) = 0,$$

b) $$\frac{d^j}{dx^j} F = 0 \quad \forall j \in \{k, \ldots, 2k-1\} \quad \text{for } x \in \{a, b\},$$

where we used the short forms F *and* F_g *for* $F(x, f^{(\nu)}(x))$ *and* $F_g(x, f^{(\nu)}(x))$.

The proof of this theorem relies on one hand on the classical result about the Euler-Lagrange differential equation, Lemma 7 which is not quite standard if S is allowed to contain δ distributions. On the other hand, our case allows a very nice and convenient re-expression of the usual general form of the Euler-Lagrange differential equation. This is given in Lemma 2.

Our aim is to apply the general Euler-Lagrange result to our general class of maximum penalized likelihood problems (12). For a proof of the theorem, we further need identities for higher derivatives of a composite function. These identities do not seem to appear in the literature, though they are quite pretty. In Appendix B, we prove the following

Lemma 2. *For* $k \in \mathbb{N}_0$*, let* $g : D \to \mathbb{R}$ *and* $F : D \times g(D) \to \mathbb{R}$ *both be* $2k$ *times differentiable. If* $\left(\frac{d^k}{dx^k} F(x, g(x))\right)^2$ *is represented as* $\mathcal{V}(x, g(x), \ldots, g^{(k)}(x))$*, and* F_g *as in (13), then*

$$\text{(i)} \quad \sum_{j=0}^{k} (-1)^j \frac{d^j}{dx^j} \mathcal{V}_{g^{(j)}}(x) \;=\; 2\,(-1)^k F_g(x, g(x)) \cdot \frac{d^{2k}}{dx^{2k}} F(x, g(x)) \quad \forall x \in D,$$

(ii) *If* $\frac{d^j}{dx^j}F_g(x) \neq 0$ *for* $j = 0, \ldots, k-1$, *the following sets of equations are equivalent:*

$$(1) \quad \sum_{j=j_0}^{k} (-1)^j \frac{d^{j-j_0}}{dx^{j-j_0}} \mathcal{V}_{g^{(j)}}(x) = 0 \qquad \forall j_0 \in \{1, \ldots, k\}.$$

$$(2) \quad \frac{d^{k+m}}{dx^{k+m}} F(x, g(x)) = 0 \qquad \forall m \in \{0, \ldots, k-1\}.$$

Proof of Theorem 1: We write the term to be minimized as $\int_a^b \mathcal{U}(x)\, dx$, where we let $\mathcal{U}\left(x; f(x), f^{(\nu)}(x), \ldots, f^{(\nu+k)}(x)\right) := S(f(x), x) + \left(\frac{d^k}{dx^k} F(x, f^{(\nu)}(x))\right)^2$.

(i): Applying Lemma 7 for $m = \nu + k$ and noting that here $\mathcal{U}_{f^{(j)}} = 0$ for $j \in \{1, \ldots, \nu - 1\}$, we have $\sum_{j=0}^{k} (-1)^{\nu+j} \frac{d^{\nu+j}}{dx^{\nu+j}} \mathcal{U}_{f^{(\nu+j)}} = -\mathcal{U}_f = -{S_f}^{[0]}$. Integrating ν times "$\int_a^x \ldots dt$", we get

$$\sum_{j=0}^{k} (-1)^j \frac{d^j}{dx^j} \mathcal{U}_{f^{(\nu+j)}} = (-1)^{\nu+1} {S_f}^{[\nu]}(x) \; + \; c_\nu + c_{\nu-1}(x-a) + \cdots + c_1 (x-a)^{\nu-1}.$$

Now, we apply Lemma 2 to the l.h.s. for $g = f^{(\nu)}$, and get (i), if we can show that $c_1 = c_2 = \cdots = c_\nu = 0$. This happens iff the l.h.s. and its first $\nu - 1$ derivatives vanish at $x = a$, which in turn follows from ${S_f}^{[j]}(a) = 0$ for $j \in \{\nu, \nu - 1, \ldots, 1\}$, exactly half of the boundary conditions derived in (ii)a) below.

(ii): The boundary conditions from Lemma 7 (with the same remark as above) can be re-expressed as $0 = \sum_{j=(i-\nu)_+}^{k} (-1)^j \frac{d^{\nu-i+j}}{dx^{\nu-i+j}} \mathcal{U}_{f^{(\nu+j)}}$, $\forall i \in \{1, \ldots, \nu + k\}$. We consider the cases

a) $i \in \{1, \ldots, \nu\} : (i - \nu)_+ = 0$, and this (for $x = a$) completes the proof of (i), by application of Lemma 2(i). The boundary conditions are, applying Lemma 2 first, and then setting $l = \nu - i$:

$$0 = \frac{d^l}{dx^l} \left\{ F_g(x, f^{(\nu)}(x)) \cdot \frac{d^{2k}}{dx^{2k}} F(x, f^{(\nu)}(x)) \right\} \quad \forall l \in \{0, \ldots, \nu - 1\},$$

for $x \in \{a, b\}$. Using the diff.eq. (i) at $x = b$, the remaining boundary conditions are equivalent to $0 = \frac{d^l}{dx^l} {S_f}^{[\nu]}(x)\Big|_{x=b} = {S_f}^{[\nu-l]}(b)$ or simply ${S_f}^{[1]}(b) = {S_f}^{[2]}(b) = \cdots = {S_f}^{[\nu]}(b) = 0$.

b) $i \in \nu + \{1, \ldots, k\}$: Let $j_0 = i - \nu \in \{1, \ldots, k\}$. We see that $0 = \sum_{j=j_0}^{k} (-1)^j \frac{d^{j-j_0}}{dx^{j-j_0}} \mathcal{U}_{f^{(\nu+j)}}$, and these conditions are proved equivalent to Theorem (ii) b) by Lemma 2(ii) (for $g = f^{(\nu)}$).

■

We now return to the special case of our "Wp"-approach, but still generalized for an arbitrary $\nu \geq 1$ and $k \geq 0$. Note that for the moment, the number and locations of the inflection points, n_w, and $w_1, \ldots, w_{n_w}$, are considered as fixed. For our problem it is feasible to minimize $J[f]$ only over functions f which do have the 'generalized inflection points' $w_1, \ldots, w_{n_w}$, since all others give infinite penalty. As in (6) for f'', we factorize $f^{(\nu)}$ as $f^{(\nu)}(x) = p_{\mathbf{w}}(x) e^{h_f}(x)$:

Corollary 3 (Generalized "Wp"). *The necessary equation system for a minimizer of*

$$J[f] = \sum_{i=1}^{n} \rho_i \left(y_i - f(x_i)\right) + \lambda \int_{x_1}^{x_n} \left(\frac{d^k}{dx^k} h_f(x)\right)^2 dx,$$

among all $2\nu+k$ times differentiable functions f with the "ν-inflection points" $w_1, \ldots, w_{n_w}$ (i.e., $f^{(\nu)}$ has exactly the zeros $w_1, \ldots, w_{n_w}$) is

$$h_f{}^{(2k)} / f^{(\nu)} = \frac{(-1)^{\nu+k}}{2\lambda(\nu-1)!} \sum_{i=1}^{n} (x - x_i)_+^{\nu-1} \psi_i(y_i - f(x_i)), \tag{16}$$

where $h_f(x) \stackrel{\text{def}}{=} \log \dfrac{f^{(\nu)}}{p_{\mathbf{w}}}$, $p_{\mathbf{w}}$ is defined analogously to (5), and $\psi_i(x) = \frac{d}{dx}\rho_i(x)$. The natural boundary conditions are

$$a) \quad \sum_{i=1}^{n} {x_i}^m \psi_i(y_i - f(x_i)) = 0 \qquad \forall m \in \{0, \ldots, \nu-1\} \tag{17}$$

$$b) \quad h_f{}^{(k)} = h_f{}^{(k+1)} = \cdots = h_f{}^{(2k-1)} = 0 \quad \text{for } x \in \{x_1, x_n\}, \tag{18}$$

Proof: We apply the theorem with $F(x, f^{(\nu)}(x)) = \log(f^{(\nu)}(x)/p_{\mathbf{w}}(x))$ and

$$\lambda S(f(x), x) = \sum_{i=1}^{n} \delta(x - x_i)\, \rho_i(y_i - f(x)). \tag{19}$$

We have $F_g(x, f^{(\nu)}(x)) = 1/f^{(\nu)}(x)$, $\lambda {S_f}^{[0]}(x) = -\sum_{i=1}^{n} \delta(x - x_i)\, \psi_i(y_i - f(x))$, and $\lambda {S_f}^{[1]}(x) = -\sum_{i=1}^{n} \mathbf{1}_{[x-x_i \geq 0]} \psi_i(y_i - f(x_i)) + c = -\sum_{i=1}^{n} (x - x_i)_+^0 \psi_i(y_i - f(x_i))$, where $c = 0$ because $0 = \lambda {S_f}^{[1]}(x_1) = c$. And, generally, for $m = 0, 1, \ldots$, ${S_f}^{[m+1]}(x) = -1/(\lambda m!) \sum_{i=1}^{n} (x - x_i)_+^m \psi_i(y_i - f(x_i))$. This S fulfills equation (15), since $\lambda \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int_{x_1}^{x_n} S(f(x) + \epsilon\eta(x), x) dx = \sum_{i=1}^{n} \frac{d}{d\epsilon} \rho_i(y_i - (f(x_i) + \epsilon\eta(x_i)))\Big|_{\epsilon=0} = \sum_{i=1}^{n} \psi_i(y_i - f(x_i)) = \lambda \int_{x_1}^{x_n} {S_f}^{[0]}(x)\, dx$. Only the equivalence of the boundary conditions a) remains to be seen: ${S_f}^{[m+1]}(b) = {S_f}^{[m+1]}(x_n) = 0$ is equivalent to $\sum_{i=1}^{n} (x_n - x_i)^m \psi_i(y_i - f(x_i)) = 0$. If this is true $\forall m \in \{0, \ldots, \nu-1\}$, then also (by induction, formally) $\sum_{i=1}^{n} {x_i}^m \psi_i(y_i - f(x_i)) = 0$ for all those m. ■

It is of interest to consider the "most smooth" generalized "Wp" smoother, i.e., the solution for $\lambda \to \infty$ of Corollary 3:

Corollary 4 (Smoothest Limit). *For $\lambda \to \infty$, we have*

$$f^{(\nu)}(x) \to p_{\mathbf{w}}(x)\ \exp(a_0 + a_1 x + \cdots + a_{k-1} x^{k-1}). \tag{20}$$

For

- *$k = 0$: $f^{(\nu)} \to p_{\mathbf{w}}$ and f minimizes $\sum_i \rho_i(y_i - f(x_i))$ among all degree $n_w + \nu$ polynomials with $f^{(\nu)} = p_{\mathbf{w}}$.*
- *$k = 1$ [**"Wp"**]: $f^{(\nu)} \to A \cdot p_{\mathbf{w}}$ (for some $A \in \mathbb{R}$) and f is the least-ρ degree $n_w + \nu$ polynomial with ν-inflection points $w_1, \ldots, w_{n_w}$ (i.e., $f^{(\nu)}(w_j) = 0 \quad \forall j$).*
- *$k = 2$: $f^{(\nu)}(x) \to A \cdot p_{\mathbf{w}}(x)\, e^{Bx}$ and $f(x) \to P_{\nu-1}(x) + P_{n_w}^*(x) e^{Bx}$, where P_k^* is a polynomial of degree k and $B \in \mathbb{R}$.*

Remark: We see that k indicates extra-degrees of freedom for our function, where ν gives the "order of inflection points" to penalize, yielding degrees of freedom, too.

Proof: From the differential equation (16), we have $h_f{}^{(2k)} \to 0$ (uniformly) for $\lambda \to \infty$, and because of the boundary conditions b) also $h_f{}^{(k)} \to 0$, such that $h_f \to$ polynomial of degree $k - 1$. $f^{(\nu)} = p_{\mathbf{w}} e^{h_f}$ completes the proof. ■

From Theorem 1, we get the 'classical' result about robust smoothing splines of order m:

Corollary 5 (Splines). *The necessary equation system for a minimizer of*

$$J[f] = \sum_{i=1}^{n} \rho_i \left(y_i - f(x_i)\right) + \lambda \int_{x_1}^{x_n} \left(f^{(m)}(x)\right)^2 \, dx$$

is

$$f(x) = c_0 + c_1 x + \ldots + c_{m-1} x^{m-1} + \frac{(-1)^m}{2\lambda(2m-1)!} \sum_{i=1}^{n} (x - x_i)^{2m-1}_{+} \psi_i(y_i - f(x_i)) \qquad (21)$$

with conditions

$$\sum_{i=1}^{n} x_i^k \; \psi_i(y_i - f(x_i)) = 0 \qquad \text{for } k = 0, \ldots, m-1, \qquad (22)$$

where $\psi_i(x) = \frac{d}{dx}\rho_i(x)$.

Remarks:

1. Greville (1969), theorem 4.1, gives an integrated version of this corollary for weighted least-squares splines. Huber (1979) discusses the problem of robustifying (discrete penalty) cubic splines and the choice of ρ or ψ functions. Cox (1983) considers the general robust "M-type splines" and proves asymptotic properties.

2. Equation (21) is a robustified 'truncated power' representation of a so-called *natural* spline of order $2m$, i.e., f is a degree $m-1$ polynomial outside $[x_1, x_n]$.

   The conditions (22) are the orthogonality relations of the "huberized residuals" ψ_i.

3. Note that the 'robustification' (i.e., the introduction of ρ_i's instead of $()^2$ which gives robust splines only if $\psi_i(x) = \rho_i'(x) \le C$ for all x) hardly complicates the variational problem. In the least squares case, $\psi_i(x) = W_i x$, the conditions (22) are equivalent to a linear equation system for $(c_0, \ldots, c_{m-1})$.

Proof: We use the theorem as for Corollary 3 with $F(x, g(x)) = g(x)$ whence $F_g \equiv 1$, and $m \equiv \nu + k$ where ν and k are not unique, say $\nu = m$ and $k = 0$. The differential equation is

$$f^{(2m)} = \frac{(-1)^m}{2\lambda} \sum_{i=1}^{n} \delta(x - x_i) \; \psi_i(y_i - f(x_i)) \qquad (23)$$

with natural boundary conditions $f^{(m)} = f^{(m+1)} = \cdots = f^{(2m-1)} = 0$ for $x \in \{x_1, x_n\}$. The boundary conditions are identical to those of Corollary 3, but by integration of (23), the conditions a) are seen to be equivalent to $f^{(m+k)} = f^{(m+k+1)} = \cdots = f^{(2m-1)} = 0$ for $x \in \{x_1, x_n\}$, which give — together with b) — the boundary conditions above. Now, we integrate the differential equation (23) m times, each time making use of one of the conditions $f^{(m+k)}(x_1) = 0$. This yields $f^{(m)} = (-1)^m/(2\lambda(m-1)!) \sum_{i=1}^{n} (x - x_i)^{m-1}_{+} \psi_i(y_i - f(x_i))$. Note that the remaining conditions, $f^{(m+k)}(x_n) = 0$, are equivalent to (22). Further m-fold integration concludes the proof of (21). ∎

A special case of corollary 3, namely $\nu = 2$ and $k = 1$, is basic for of the smoothing algorithm implementing our special "Wp" problem:

Corollary 6 ("Wp"). *The solution of problem (10) has to fulfill the following ordinary differential equation*

$$h_f'' = p_{\mathbf{w}}\, e^{h_f} \cdot L_f\,, \tag{24}$$

where L_f is a piecewise linear function, defined as

$$L_f(x) := -\frac{1}{2\lambda}\sum_{i=1}^{n}(x - x_i)_+\, \psi_i(y_i - f(x_i)).$$

Furthermore, it has to satisfy the **"multi-boundary"** *conditions*

$$h_f'(x_1) = h_f'(x_n) \;=\; \sum_i \psi_i(y_i - f(x_i)) = \sum_i x_i\, \psi_i(y_i - f(x_i)) = 0.$$

Because these conditions involve $f(x_i)$ $\forall i$, they are not simple boundary, but *multi-boundary* conditions.

In Mächler (1989), an algorithm for this non-standard problem is devised. A *multiple-shooting* Runge-Kutta method (Keller, 1976), adapted to this multi-boundary situation, is used to solve (24). The algorithm, a Newton-type iteration, needs a starting approximation (for f, f', h_f, h_f' and the $w_1, \ldots, w_{n_w}$). Finally, the overall procedure needs to minimize the penalized log likelihood over all possible $w_1, \ldots, w_{n_w}$.

4 Summary

The new "Wp" procedure has the following properties

- Differentiability: The functions $\hat{f}$, $\hat{f}'$, $\hat{f}''$ (via $\hat{h_f}$ and $\hat{p}_{\mathbf{w}}$), and $\hat{f}^{(3)}$ (from $\hat{h_f}'$) are continuously differentiable, and $\hat{f}^{(4)}$ is continuous. Note that $\hat{f}''$ for a "Wp" smoother is much smoother than for a spline.

  The generalized "Wp" approach yields even more derivatives. For $\nu \geq 1$ and $k \geq 0$, one gets derivatives of f up to order $\nu + 2k$.

- Guaranteed number of inflection points (or local extrema when $\nu = 1$, or higher order inflections for $\nu > 2$).

  The factorization $f^{(\nu)}(x) = \pm(x - w_1)\cdots(x - w_{n_w}) \cdot \exp h_f(x)$ is of *semi-parametric* nature with parameters w_j and nonparametric part $h_f(\cdot)$. The main smoothing parameter is n_w, the "order" of the parametric part.

  Note that the restriction on n_w, the number of sign changes of $f^{(\nu)}$, automatically limits the number of zeros of the lower derivatives: $f^{(\nu-j)}$ cannot have more than $n_w + j$ zeros.

- The (extra) smoothing parameter λ is of minor importance: Note that for $\lambda \to 0$, the number of [generalized] inflection points is still restricted, and a limit $\lim_{\lambda\to 0} \hat{f}(x)$ exists everywhere. For splines, the limit for $\lambda \to 0$ is a trivial *interpolating* function whereas here, the limit is still smooth, namely "the best fitting function" for a given number of inflection points, n_w.

- The "most smooth" generalized "Wp" smoothers (for $\lambda \to \infty$) gives 'natural' parametric curves (Corollary 4).

- *Exact Fit Property:* From the boundary conditions (17) of Corollary 3, the "orthogonality conditions", it is easily seen that the [generalized] "Wp" smoother fits the data exactly if they lie on a straight line [polynomial of degree $\nu - 1$].

  It also follows that the smoother is regression equivariant under superposition of linear functions [degree $\nu - 1$ polynomials].

- Bias – Erosion problem: For linear smoothers such as splines, kernel estimators, or LOWESS, the bias problem is mostly characterized by "erosion": The bias is predominant near local extrema, i.e., peaks and valleys.

  The "Wp" smoother is *not linear*, and from the examples considered, it appears that the problem of erosion is hardly present in situations where the number of inflection points, n_w, is specified correctly.

Appendix A The Euler-Lagrange Differential Equation

The following lemma is classical in the usual case when $\mathcal{U}(x; f, \ldots)$ is defined and twice differentiable for all $f \in C^m[a,b]$. In our situation, it is still valid but by slightly different reasoning. We use the function class $\mathcal{F}_m^\nu(\mathbf{w})$ of Corollary 3 in order to show the principle. For many other subsets of $C^m[a,b]$, the theorem will be valid by an analoguous proof.

Lemma 7 (Euler-Lagrange differential equation and natural boundary conditions). *Given integers $m \geq \nu \geq 0$, let*
$\mathcal{F}_m^\nu(\mathbf{w}) \stackrel{\text{def}}{=} \{f \in C^m[a,b]\,;\ f^{(\nu)}$ *has exactly the zeros* $w_1, \ldots, w_{n_w}\}$, *and*
$J[f] = \int_a^b \mathcal{U}(x; f(x), f'(x), \ldots, f^{(m)}(x))\, dx$,
where $\mathcal{U}(x; f_0, f_1, \ldots, f_m)$ is twice differentiable with respect to $f_0, \ldots, f_m$ and is "smooth" as integrand of $J[f]$ (fulfilling $(*)$ below), then a function f minimizing $J[f]$ among all $f \in \mathcal{F}_m^\nu(\mathbf{w})$ necessarily fulfills:

$$\text{(i)} \qquad \sum_{j=0}^{m} (-1)^j \frac{d^j}{dx^j} \mathcal{U}_{f^{(j)}} \equiv 0 \qquad \forall x \in [a,b] \qquad \textit{'differential equation'}$$

$$\text{(ii)} \quad \forall i \in \{1, \ldots, m\} : \sum_{j=0}^{m-i} (-1)^j \frac{d^j}{dx^j} \mathcal{U}_{f^{(i+j)}} = 0 \qquad \textit{for } x \in \{a,b\} \quad \textit{'boundary condition'},$$

where $\mathcal{U}_{f^{(j)}} = \frac{\partial \mathcal{U}}{\partial f^{(j)}}$.

Proof: The first part is the *standard variational argument* of calculus of variation: Looking for the optimal f, we consider the trial functions $f + \epsilon\eta$, where $\eta(\cdot)$ is any function $\in \mathcal{F}_m^\nu(\mathbf{w})$, and $|\epsilon|$ small enough such that $f + \epsilon\eta \in \mathcal{F}_m^\nu(\mathbf{w})$. A necessary condition for f to be extreme among the $f + \epsilon\eta$ is then $\delta J(f;\eta) \stackrel{\text{def}}{=} \frac{d}{d\epsilon} J[f + \epsilon\eta]\Big|_{\epsilon=0} = 0$, where the "Gâteaux-variation" δJ is the Gâteaux-derivative of the functional J (in direction of η) and corresponds to the directional derivative of $\mathbb{R}^n$ calculus.

We have to show that the condition $\delta J(f;\eta) = 0\, \forall \eta$ is equivalent to the stated Lemma. Under weak smoothness conditions on $\mathcal{U}$, we can interchange integration and differentiation and get

$$(*) \quad \delta J(f;\eta) = \int_a^b (\eta \mathcal{U}_f + \eta' \mathcal{U}_{f'} + \ldots + \eta^{(m)} \mathcal{U}_{f^{(m)}})\, dx.$$

We integrate the terms $\int_a^b \eta^{(j)} \mathcal{U}_{f^{(j)}}$ partially j times to see that

$$\begin{aligned} \delta J(f;\eta) &= \sum_{j=0}^{m} \left(\int_a^b \eta(x) (-1)^j \frac{d^j}{dx^j} \mathcal{U}_{f^{(j)}}\, dx \;+\; \left[\sum_{i=0}^{j-1} (-1)^i \eta^{(j-1-i)} \frac{d^i}{dx^i} \mathcal{U}_{f^{(j)}} \right]_a^b \right) \\ &= \int_a^b \left(\sum_{j=0}^{m} (-1)^j \frac{d^j}{dx^j} \mathcal{U}_{f^{(j)}} \right) \eta(x) dx + \sum_{j=1}^{m} \left[\eta^{(j-1)} \left(\sum_{i=0}^{m-j} (-1)^i \frac{d^i}{dx^i} \mathcal{U}_{f^{(i+j)}} \right) \right]_a^b, \end{aligned}$$

where the two sums in the second term have been rearranged by the substitution $j' = j - i$, and we used the notation $[H(x)]_a^b := H(b) - H(a)$. From $\delta J = 0$, for all η (with ν - inflection points $w_1, \ldots, w_{n_w}$), we conclude that both terms have to vanish, because otherwise we might vary the integral part while fixing all $\eta^{(j)}|_{x=a \text{ or } b}$. We easily conclude that the inner sums (over i) must all vanish. These are the boundary conditions (ii).

The classical way to get the differential equation (i) is applying the 'fundamental lemma of variational calculus' which states that from G continuous, and $\int_a^b G(x)\eta(x)dx = 0 \quad \forall$ continuous η, one concludes that $G(x)$ has to vanish on $[a,b]$. This lemma is proved indirectly, assuming, e.g., that $G(x_0) > 0$, and therefore $G > 0$ on a whole neighborhood of x_0. Then one takes $\eta > 0$ on this same neighborhood and zero outside, such that $\int G(x)\eta(x)dx > 0$ which is a contradiction.

Here, this fundamental lemma may not be applied directly, since we must not have $\eta \equiv 0$ on any interval. Let us use a special class of varying functions η, namely

$$\eta(x) = e^{-\alpha x}Q(x),$$

where $Q(x)$ is a polynomial $Q(x) = \sum_{k=0}^{n_w} q_k x^k$ such that $\eta^{(\nu)}(x) = 0$ is equivalent to $x \in \{w_1, \ldots, w_{n_w}\}$. By the product rule of differentiation and reversing the order of summation, we see that

$$\eta^{(\nu)}(x) = e^{-\alpha x}\sum_{k=0}^{n_w}\left(\sum_{j=0}^{n_w-k}\binom{\nu}{j}(-\alpha)^{\nu-j}(k+j)_j\ q_{k+j}\right)x^k, \tag{25}$$

where $(n)_k = n(n-1)\ldots(n-k+1)$, as in definition (28). If we require that $\eta^{(\nu)}(x) \stackrel{!}{=} c(x-w_1)(x-w_2)\ldots(x-w_{n_w})$ and compare the coefficients of the two polynomials for $\eta^{(\nu)}$, we see that factors of x^k in (25) form a linear system for $(q_0, \ldots, q_{n_w})$ with an upper triangular matrix. This matrix is regular with constant diagonal elements $(-\alpha)^\nu$, such that the coefficients q_j and the polynomial Q always exist with the required property. Now we have $\int_a^b G(x)Q(x)e^{-\alpha x}dx = 0, \forall\alpha > 0$ which means that the Laplace transform of $G(x)Q(x)$ is identically zero, and therefore $G(x)$ must vanish every where, since the polynomial $Q(x)$ is not identically zero. ■

Appendix B Higher Chain-Rule Identities

The goal of this appendix is to prove Lemma 2 in section 3. To this end, we have to consider formulas which are connected with the chain-rule for higher derivatives of a composite function $F(x, g(x))$. In the standard books, we have not found those which we use in the following. A well-known formula of a similar naturse is "Faà di Bruno's formula" which uses multinomial coefficients in sums with combinatorial indices (Abramowitz and Stegun, 1972, chapter 24).

Our goal is to re-express the partial derivatives of Euler's differential equation for the penalty part $\frac{d^n}{dx^n}F(x, g(x))$. It is of the form

$$\frac{d^n}{dx^n}F(x, g(x)) = F_n(x; g(x), g'(x), \ldots, g^{(n)}). \tag{26}$$

Lemma 8 (Higher chain-rule identity 1).
Let $g()$ and $F()$ be as in Lemma 2 and F_n the n-th derivative of F as above. Then

$$\frac{\partial}{\partial g^{(j)}}F_n = \binom{n}{j}\frac{\partial}{\partial g}F_{n-j} \quad \forall j \in \mathbb{N}_0 \quad \forall n \in \mathbb{N}_0 \tag{27}$$

Proof: We denote the partial derivatives as $F_x := \frac{\partial F(x,*)}{\partial x}$ and $F_g := \frac{\partial F(*,g)}{\partial g}$. The equality is trivially fulfilled for $j = 0$ and $j > n$. It remains to be proved $1 \le j \le n$. For $j = n$, (27) follows from the identity

$$F_n(x, g(x)) = g^{(n)}F_g(x, g(x)) + R_n(g, \ldots, g^{(n-1)}; F_g, \ldots, F_g^{(n)}, F_x, \ldots, F_x^{(n)}),$$

where $R_n(\ldots)$ is a "remainder" *not* containing $g^{(n)}$, which is proved by induction: $n = 1$ is the simple chain rule with $R_1 = F_x$. For $n \geq 1$ we have $F_{n+1} = \frac{d}{dx}\, F_n = g^{(n+1)}\, F_g + g^{(n)}(F_g{}'g' + F'_x) +$
$\frac{d}{dx}R_n(g, \ldots, g^{(n-1)}; F_g, \ldots, F^{(n)}, F'_x, \ldots, F_x^{(n)})$, using the result for n. We see that $R_{n+1} := g^{(n)}(F_g{}'g' + F'_x) + \frac{d}{dx}R_n(\ldots)$ does not depend on $g^{(n+1)}$.

To complete the proof of the lemma, doing induction $n \to n+1$, we may assume its truth for n, and have to show it for $j = 1, \ldots, n$ ($j = n+1$ was done above !). We again apply the chain rule to F_n: $F_{n+1} = \frac{d}{dx}F_n = \sum_{i=0}^{n} \frac{d}{dx}g^{(i)}\frac{\partial}{\partial g^{(i)}}F_n + \frac{\partial}{\partial x}F_n$. Therefore, the l.h.s. of equation (27) is $= \frac{\partial}{\partial g^{(j)}}F_{n+1} = \sum_{i=0}^{n} \frac{\partial}{\partial g^{(j)}} \left(g^{(i+1)}\frac{\partial}{\partial g^{(i)}}F_n\right) + \frac{\partial}{\partial g^{(j)}}\frac{\partial}{\partial x}F_n$. Note that $\frac{\partial}{\partial g^{(j)}}\frac{\partial}{\partial x} = \frac{\partial}{\partial x}\frac{\partial}{\partial g^{(j)}}$, such that we have $\frac{\partial}{\partial g^{(j-1)}}F_n + \sum_{i=0}^{n} g^{(i+1)}\frac{\partial}{\partial g^{(i)}}\frac{\partial}{\partial g^{(j)}}F_n + \frac{\partial}{\partial x}\frac{\partial}{\partial g^{(j)}}F_n$, which is, using the result for n, $= \binom{n}{j-1}\frac{\partial}{\partial g}F_{n-j+1} + \sum_{i=0}^{n} g^{(i+1)}\frac{\partial}{\partial g^{(i)}}\binom{n}{j}\frac{\partial}{\partial g}F_{n-j} + \binom{n}{j}\frac{\partial}{\partial g}\frac{\partial}{\partial x}F_{n-j}$. Note that, in $\sum_i$, the last j summation terms vanish, because $\frac{\partial}{\partial g^{(i)}}F_{n-j} = 0$ for $i > n-j$. This sum is therefore equal to $\binom{n}{j}\frac{\partial}{\partial g}\sum_{i=0}^{n-j} g^{(i+1)}\frac{\partial}{\partial g^{(i)}}F_{n-j}$. We have

$$\frac{\partial}{\partial g^{(j)}}F_{n+1} = \binom{n}{j-1}\frac{\partial}{\partial g}F_{n-j+1} + \binom{n}{j}\frac{\partial}{\partial g}\left\{\sum_{i=0}^{n-j} g^{(i+1)}\frac{\partial}{\partial g^{(i)}}F_{n-j} + \frac{\partial}{\partial x}F_{n-j}\right\},$$

and the $\{\ldots\}$ is $\frac{d}{dx}F_{n-j}$, by the chain rule. Therefore, $\frac{\partial}{\partial g^{(j)}}F_{n+1} = \left(\binom{n}{j-1} + \binom{n}{j}\right)\frac{\partial}{\partial g}F_{n-j+1} = \binom{n+1}{j}\frac{\partial}{\partial g}F_{n-j+1}$. ■

In the following, we will also make use of 'elementary' identities for binomial coefficients. Let us define $\forall k \in \mathbb{N}_0 \quad \forall a \in \mathbb{R}$:

$$\begin{aligned} (a)_k &\stackrel{\text{def}}{=} a(a-1)\cdots(a-k+1) \quad \text{with} \quad (a)_0 = 1, \\ \binom{a}{k} &\stackrel{\text{def}}{=} \frac{(a)_k}{k!}, \quad \text{and for } k < 0 : \binom{a}{k} = 0. \end{aligned} \tag{28}$$

The special case, $\binom{-n}{k} = (-1)^k\binom{n+k-1}{k}$, will be used in the next proof.

The following binomial identities will be used later, and are (to our knowledge) not available in the standard literature:

Lemma 9 (Binomial identities). $\forall n \in \mathbb{N}_0 \quad \forall m \in \{0, \ldots, n\}$:

(i) $\forall a \in \mathbb{R} \quad \sum_{j=0}^{n}(-1)^{n-j}\binom{n}{j}\binom{j+a}{m} = \delta_{m,n} \stackrel{\text{def}}{=} 1_{[m=n]}$

(ii) $\forall k \in \mathbb{N}_0 \quad \sum_{j=0}^{k}(-1)^j\binom{n}{k-j}\binom{m+j}{j} = \binom{n-m-1}{k}$

(iii) $c_{n,m,J} := \sum_{j=m+J}^{n}(-1)^j\binom{n}{j}\binom{j-J}{m}$ *fulfills* $\forall J \in \{0, \ldots, n-m\}$

1) $c_{n,m,J} = (-1)^n\delta_{m,n} + (-1)^{m+J}\binom{n-m-1}{J-1}$

2) $c_{n,m,J} = 0 \iff J = 0 \wedge n \neq m$

Proof: (i): Consider the forward difference operator $\Delta_x : f \mapsto f(x+1) - f(x)$. We apply the well-known formula $\Delta_x{}^n f = \sum_{j=0}^{n}(-1)^{n-j}\binom{n}{j}f(x+j)$ at $x = 0$ to the polynomial

$f(t) = \binom{t+a}{m}$ which gives the l.h.s. of (i). Applying the mean value theorem to the n-th derivative, we have $\Delta_0{}^n f = f^{(n)}(\xi)$ for a $\xi \in [0, n]$. Because here $f(t) = t^m/(m!)+O(t^{m-1})$, $m \le n$, we have $f^{(n)}(\xi) \equiv \delta_{n,m}$.

(ii): A well-known formula, $\sum_{j=0}^{n} \binom{n}{j}\binom{a}{k-j} = \binom{n+a}{k}$, is seen by comparison of coefficients of the binomial theorem, applied to $(1+x)^n(1+x)^a = (1+x)^{n+a}$, and is valid for any real a. We apply it to $a = -m-1$, using $\binom{-N}{j} = (-1)^j\binom{N+j-1}{j}$ to get $\sum_{j=0}^{n}\binom{n}{j}(-1)^{k-j}\binom{m+k-j}{k-j} = \binom{n-m-1}{k}$. Here, the adding terms are zero whenever $j > k$ or (also) $j > n$, such that we may sum from $j = 0$ to k instead of n. Reversing the order of summation, we have (ii).

(iii): $c_{n,m,J} = \sum_{j=m+J}^{n} \ldots = \sum_{j=0}^{n} \ldots - \sum_{j=0}^{m+J-1}(-1)^j\binom{n}{j}\binom{j-J}{m}$, where the first sum is $(-1)^n\times$ the l.h.s. of (i) for $a = -J$, and therefore equal to $(-1)^n\delta_{n,m}$. In the second sum, the terms are zero whenever $0 \le j - J < m$, such that we may sum only to $J-1$. For these j, $j - J < 0$, so that we can apply $\binom{j-J}{m} = (-1)^m\binom{J-j+m-1}{m}$ to get $c_{n,m,J} - (-1)^n\delta_{n,m} = \sum_{j=0}^{J-1}(-1)^{m-j+1}\binom{n}{j}\binom{m+J-j-1}{m}$. This is seen to be $(-1)^{m+J}\times$ the sum in (ii) and by setting $j' = J-1-j$ and $k := J-1$, 1) is proved. 2) is an immediate consequence if we remember that $0 \le m + J \le n$. ∎

Lemma 10 (Higher chain-rule identity 2).
Let $g()$ and $F()$ be as general as in Lemma 2, and F_n defined by (26). Then

$$\forall n \in \mathbb{N}_0 \quad \forall m \in \{0, \ldots, n\} \quad \forall m' \in \{m, \ldots, n\}:$$

$$\begin{aligned} C_{n,m,m'} &:= \sum_{j=m'}^{n} (-1)^j \binom{j-m'+m}{m} \frac{d^{j-m'}}{dx^{j-m'}} \frac{\partial}{\partial g^{(j)}} F_n(x, g(x)) \\ &\equiv \frac{\partial}{\partial g} F_{n-m'} \cdot \left(\delta_{m,n} + (-1)^{m'}\binom{n-m-1}{n-m'}\right) \\ &= c_{n,m,m'-m} \cdot \frac{\partial}{\partial g} F_{n-m'}, \end{aligned}$$

where $c_{n,m,J}$ is defined as in Lemma 9.

Proof: We apply Lemma 8 to $C_{n,m,m'}$ above, and get $C_{n,m,m'} = \sum_{j=m'}^{n}(-1)^j\binom{j-m'+m}{m}\binom{n}{j}$ $\frac{d^{j-m'}}{dx^{j-m'}}\frac{\partial}{\partial g}F_{n-j}(x, g(x))$. Now we make use of the basic rule

$$\text{For any function } G(x, u(x), u'(x), \ldots, u^{(n)}(x)): \quad \frac{\partial}{\partial u}\frac{d}{dx}G = \frac{d}{dx}\frac{\partial}{\partial u}G \tag{29}$$

which is a simple consequence of the chain rule for several arguments. (Note that this rule would be wrong with $u^{(j)}, j \ge 1$, instead of u.) Apply this rule $j - m'$ times to see that $\frac{d^{j-m'}}{dx^{j-m'}}\frac{\partial}{\partial g}F_{n-j}(g(x)) = \frac{\partial}{\partial g}F_{n-m'}(g(x))$ which is *independent* of j so that we can apply directly Lemma 9 (iii) 1) with $J = m' - m$ to complete the proof. ∎

Proof of Lemma 2: Because of $\mathcal{V} = F_k(x, g(x))^2$, we have $\mathcal{V}_{g^{(j)}} = 2\,F_k\frac{\partial}{\partial g^{(j)}}F_k$. In the l.h.s. of (i) and (ii) 1), we take $(j - j_0)$-th derivatives of this product, applying Leibniz' rule $\frac{d^j}{dx^j}(a(x)\cdot b(x)) = \sum_{m=0}^{j}\binom{j}{m}a^{(m)}b^{(j-m)}$ such that this l.h.s. becomes
$2\sum_{j=j_0}^{k}(-1)^j\sum_{m=0}^{j-j_0}\binom{j-j_0}{m}F_{k+m}\cdot\frac{d^{j-j_0-m}}{dx^{j-j_0-m}}\frac{\partial}{\partial g^{(j)}}F_k$, or, switching the order of summation (keeping $0 \le m \le j - j_0 \le k - j_0$):

$$\sum_{j=j_0}^{k}(-1)^j\frac{d^{j-j_0}}{dx^{j-j_0}}\mathcal{V}_{g^{(j)}}(x) = 2\sum_{m=0}^{k-j_0} C_m(x)F_{k+m}(x), \tag{30}$$

where $C_m = \sum_{j=m+j_0}^{k} (-1)^j \binom{j-j_0}{m} \frac{d^{j-(m+j_0)}}{dx^{j-(m+j_0)}} \frac{\partial}{\partial g^{(j)}} F_k$. Remark that $C_m = C_{n,m,m'}$ of Lemma 10 if we let $n := k$ and $m' := m + j_0$. Therefore $C_m = c_{k,m,j_0} \frac{\partial}{\partial g} F_{k-m-j_0}$.

(i): For $j_0 = 0$, we see (i), because $c_{k,m,0} = 0$ for all m but $m = k$ (Lemma 9 (iii)2)), where $C_k = (-1)^k \cdot \frac{\partial}{\partial g} F_{k-k-0} = (-1)^k F_g(x, g(x))$.

(ii): Because of equation (30), the equivalence of (1) and (2) is proved if we can show that here, $C_m \not\equiv 0$ for $m \in \{0, \ldots, k - j_0\}$. We have $c_{k,m,j_0} \not\equiv 0$ from Lemma 9 (iii)2), since $j_0 \geq 1$. And, by the basic rule, also $\frac{\partial}{\partial g} F_{k'} = \frac{d^{k'}}{dx^{k'}} \frac{\partial}{\partial g} F$ which are non-zero by the assumption in (ii).

■

Acknowledgements

This paper presents main results from the author's Ph.D. thesis (Mächler, 1989), which was supervised by Frank Hampel. His intuition of considering inflection points and the pre-penalty $\int (f'''/f'')^2$ were crucial for this work. I am also indebted to several colleagues, notably Werner Stahel and Hans R. Künsch who helped to improve earlier versions of this manuscript.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*, Dover, New York.

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator, *Statistical Science* **6**(4): 404–436. (with discussion).

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**: 828–836.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines, *Annals of Statistics* **11**: 530–551.

Cox, M. G. (1973). Cubic spline fitting with convexity and concavity constraints, *Technical report*, Report NAC 23, National Physical Laboratory.

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

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, Dekker:NY.

Glass, J. M. (1966). Smooth-curve interpolation: A generalized spline-fit procedure, *BIT* **6**: 277–293.

Greville, T. N. E. (1969). Introduction to spline functions, *in* T. N. E. Greville (ed.), *Theory and Application of Spline Functions*, Academic Press, pp. 1–35.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge U. Press, Cambridge, UK.

Härdle, W. and Gasser, T. (1984). Robust non-parametric function fitting, *Journal of the Royal Statistical Society B* **46**: 42–51.

Hess, W. and Schmidt, J. W. (1986). Convexity preserving interpolation with exponential splines, *Computing* **36**: 335–342.

Huber, P. J. (1979). Robust smoothing, *in* R. L. Launer and G. N. Wilkinson (eds), *Robustness in Statistics*, Acad. Press, New York, pp. 33–47.

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

Keller, H. B. (1976). *Numerical Solution of Two Point Boundary Value Problems*, Regional conf. series in appl. math., SIAM, Philadelphia.

Mächler, M. B. (1989). *'Parametric' Smoothing Quality in Nonparametric Regression: Shape Control by Penalizing Inflection Points*, Ph.d. thesis, no 8920, ETH Zurich, Switzerland.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions, *Annals of Statistics* **19**(2): 741–759.

Micchelli, C. A., Smith, P. W., Swetits, J. and Ward, J. D. (1985). Constrained L_p approximation, *Constructive Approximation* **1**: 93–102.

Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Vol. 46 of *Lecture Notes in Statistics*, Springer.

Pruess, S. (1976). Properties of splines in tension, *Journal of Approximation Theory* **17**: 86–96.

Ramsay, J. O. (1988). Monotone regression splines in action (with discussion), *Statistical Science* **3**: 425–459.

Reinsch, C. H. (1971). Smoothing by spline functions II, *Numerische Mathematik* **16**: 451–454.

Rentrop, P. (1980). An algorithm for the computation of the exponential spline, *Numerische Mathematik* **35**: 81–93.

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

Wahba, G. (1990). *Spline Models for Observational Data*, Vol. 59 of *CBMS-NSF Regional conference series in applied mathematics*, SIAM.

Wahba, G. and Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation, *Communications in Stat.* **4**: 1–18.

Wright, I. W. and Wegman, E. J. (1980). Isotonic, convex and related splines, *Annals of Statistics* **8**: 1023–1035.