Variational Solution of Penalized Likelihood Problems and Smooth Curve Estimation.

Martin B. Mächler
ETH Zurich, Switzerland

April 24, 1995
$Header: /u/maechler/tex/wp/diss-edited/RCS/m2.tex,v 1.12 1995/01/13 14:22:24 maechler Exp maechler

Abstract

Usual non-parametric regression estimators 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* (MPL) 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 $h_f(\cdot)$.

The main mathematical result is a convenient form of the characterizing differential equation for a very general class of MPL estimators.

1 Introduction

In the last decades, non-parametric regression methods have been developed to gain flexibility in regression problems of data analysis. 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.

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 with a more general concept, using *change of curvature* as roughness measure.

Let us consider an example with real data. The 'housing starts' series from the software package S was de-seasonalized using 'SABL', and the resulting data (including the noise part) taken as raw data (in S: `hs <- sabl(hstart); data <- hs$trend + hs$irregular`). The trend component computed by `sabl` is a smooth of this data with 19 inflection points and a residual sum of squares of 9.70. Figure 1 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. Two cubic splines are shown for comparison. The smoothness parameter of the first was chosen to produce the same residual sum of squares as the "Wp" smoother. It results in unsmooth behavior — close inspection exhibits 11 inflection points (marked in the figure) of which 6 are

AMS 1991 subject classifications. Primary 62G07; secondary 34B10,41A29,65D07,65D10.
Key words and phrases. Nonparametric Regression; Roughness penalty; Maximum penalized likelihood; Inflection point; Robust smoothing; Change of Curvature.

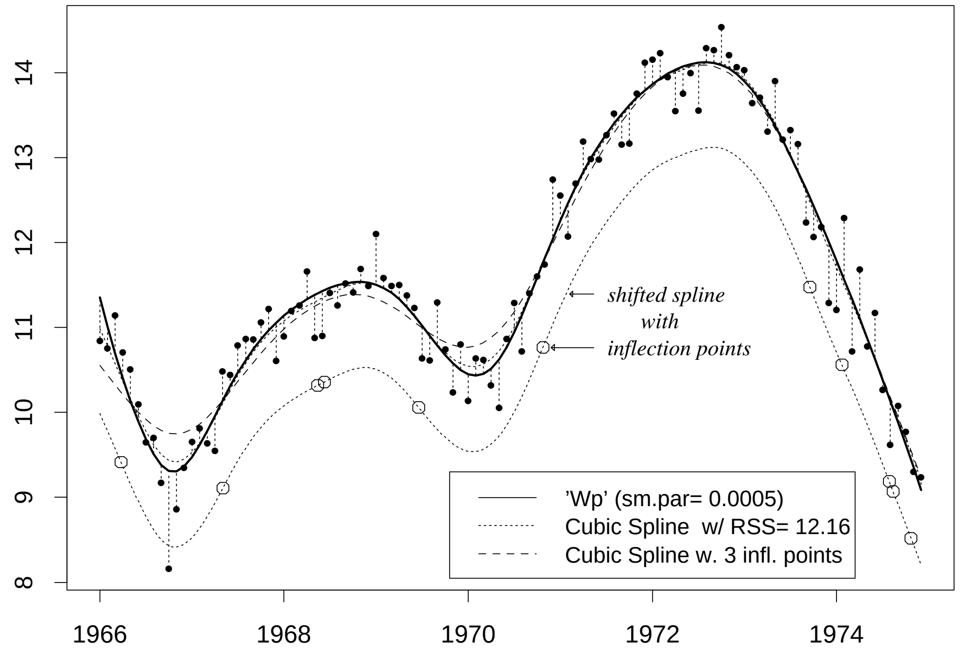

Figure 1: De-seasonalized 'housing starts', a times-series of length 108. A "Wp" smoother, restricted to 3 inflection points, with residual sum of squares = 12.16. The first cubic spline is tuned to have the same residual sum of squares and also shown shifted downwards by 1 with marked inflection points. The second spline is the best fitting one for 3 inflection points.

"significant". The second spline was tuned to produce only just 3 inflection points. Comparable to "Wp" in terms of smoothness, it now suffers from "erosion", i.e., bias near local extrema. For more details and a second example, see Mächler, 1993.

There have been other approaches to deal with the spurious "wiggles" seen in traditional nonparametric curves. In the domain of splines, this problem has been approached traditionally by either *restricting* or *generalizing* splines (Ramsay, 1988; Wright and Wegman, 1980; Mammen, 1991; Dierckx, 1993; Ansley, 1993), see Mächler (1993) for some discussion and further references.

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 $B < \inf\{R[g] \mid g(x_i) = y_i\}$). As Reinsch (1971) proved for spline smoothing, this restricted variational problem can be restated using a Lagrange multiplier λ, as

$$\min_f \left\{ \sum_{i=1}^n \rho_i\,(y_i - f(x_i)) \;+\; \lambda R[f] \right\}. \tag{2}$$

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, i.e., they are highly influenced by only few outlying observations. This fits in poorly with the idea of using a nonparametric curve. From the theory of robustness, it is well known that one gets robust 'M-estimators' if $|\rho'| \le c$ for some $c \in \mathbb{R}$ (Huber, 1979). The choice of Huber's ρ,

$$\rho_c(x) \stackrel{\text{def}}{=} \frac{1}{2}\left(x^2 - (|x| - c)_+^2\right), \qquad \text{where} \quad a_+ \stackrel{\text{def}}{=} \max(0, a), \tag{3}$$

gives the optimal 'minimax' estimators in the case of linear regression, minimizing the maximal variance over a full neighborhood of the normal distribution.

Roughness Penalties $R[f]$. In section 3, we will define the roughness penalty $R[f]$ that leads to our "Wp" approach and ensures no unnecessary "wiggles". Note that for density estimation, the choice of different roughness penalties has a long tradition, see, e.g., Thompson and Tapia, 1990; Klonias, 1984, whereas for regression, this is less so. We will only consider roughness penalties of the form $R[f] = \int_a^b G(x, f(x), f'(x), \ldots)(x)\, dx$ for some function G. Kimeldorf and Wahba (1971) and others subsequently (e.g., Cox and O'Sullivan, 1990; Ansley, 1993) have considered "general spline" smoothing problems where $G(..) = (L\, f)^2(x)$ and L is a *linear* (differential) operator such that $\int_a^b (L\, f)^2(x)\, dx$ is the norm $\|f\|^2$ in a Hilbert space with reproducing kernel. Then, the solution of (2) is well characterized using a basis and the reproducing kernel of the Hilbert space.

Here, we consider a more general situation where G may be *non*-linear of the form $G(x, f, \ldots) = \left(\frac{d^k}{dx^k} F(x, f^{(\nu)}(x))\right)^2$. Using Dirac's δ–distribution notation, we can restate (2) as an integral, the usual form of variational calculus,

$$\min_f \int_{x_1}^{x_n} \left(\frac{1}{\lambda} \sum_{i=1}^{n} \delta(t - x_i)\, \rho_i(y_i - f(t)) \;+\; \left(\frac{d^k}{dx^k} F(x, f^{(\nu)}(x))\right)^2 \right) dt. \tag{4}$$

In section 2, we will solve this general problem by reducing it to an explicit ordinary differential equation boundary value problem, and apply it to special cases like robust polynomial splines. In section 3, introducing change of curvature as a new roughness measure, we develop our "Wp" approach considering inflection points, also applying Theorem 1 for its solution ("Wp" stands for an abbreviation of the German word 'Wendepunkt' for inflection point).

2 Variational Problem and Differential Equation

This section shows that the minimizing f of (4) is necessarily a solution of the Euler-Lagrange differential equation (24). The basic result, Theorem 1 below, applies to the following general problem.

For given $k, \nu \in \mathbb{N}$, $k, \nu \ge 0$, $k + \nu \ge 1$, "Minimize the functional $J[f]$ over $f \in \mathcal{F}_{k,\nu,F}$" with the following specifications:

$$J[f] \stackrel{\text{def}}{=} \int_a^b \left\{ S(f(x), x) + \left(\frac{d^k}{dx^k} F(x, f^{(\nu)}(x))\right)^2 \right\} dx, \tag{5}$$

$$\mathcal{F}_{k,\nu,F} \stackrel{\text{def}}{=} \left\{ f \in C^{2k+\nu}[a,b]\,;\; \int_a^b \Big(\frac{d^k}{dx^k} F(x, f^{(\nu)}(x))\Big)^2 dx < \infty \right\}, \tag{6}$$

$$F_g(x,g) \stackrel{\text{def}}{=} \frac{\partial}{\partial g} F(x,g), \tag{7}$$

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

$$S_f{}^{[j+1]}(x) \stackrel{\text{def}}{=} \int_a^x S_f{}^{[j]}(t)\, dt, \quad \textit{for } 0 \le j < \nu, \tag{8}$$

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. Furthermore, for all $f, \eta \in \mathcal{F}_{k,\nu,F}$, S must allow for interchanging of integration and differentiation, fulfilling

$$\frac{d}{d\epsilon} \int_a^b S(f(x) + \epsilon\eta(x), x)\, dx \Bigg|_{\epsilon=0} = \int_a^b \eta(x) \cdot S_f{}^{[0]}(x)\, dx. \tag{9}$$

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 $0 \le k \le j$.

This general formulation encompasses a vast class of maximum penalized likelihood problems, not only in nonparametric regression, but also density estimation. The scatter term, which may not even be a log-likelihood, has the general form $\int_a^b S(f(x), x)\, dx$, satisfying (8)–(9). Typically, $S(f(x), x) = \sum_{i=1}^n \delta(x - x_i)\, \ell_i(f(x))$ and $\int_a^b S(f(x), x)\, dx = \sum_{i=1}^n \ell_i(f(x_i))$, see Lemma 3. Convexity of $S(u, x)$ in u is often sufficient for uniqueness of the solution.

The *existence* of a minimizer f of the $J[f]$ is verified in Mächler (1989, Appendix A, p. 68–73), for the case of Corollary 6. Tonelli's theorem (a "direct method") of variational calculus is applied and one sees that the function h_f (19) 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 (5) to (9), 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) $$\frac{d^j}{dx^j} F = 0 \quad \forall j \in \{k, \ldots, 2k-1\} \quad \textit{for } x \in \{a,b\},$$

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

This theorem is proven in two steps: First, a version of the classical result about the Euler-Lagrange differential equation, Lemma 8 in Appendix A gives the differential equation and boundary conditions for our case where S is allowed to contain δ distributions. Then, we can re-express the usual general form of the Euler-Lagrange differential equation into a more convenient form as given in the following Lemma 2 which is proven in Appendix B.

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 (7), then*

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

(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:*

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

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

Proof of Theorem 1: We write the term to be minimized as $\int_a^b \mathcal{U}(x)\, dx$, where we let

$$\mathcal{U}(x; f(x), f^{(\nu)}(x), \ldots, f^{(\nu+k)}(x)) := \; S(f(x), x) + \left(\tfrac{d^k}{dx^k} F(x, f^{(\nu)}(x))\right)^2.$$

(i): Applying Lemma 8 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 8 (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 $g = f^{(\nu)}$).

$\blacksquare$

The following Lemma shows how the assumptions and results of Theorem 1 apply for a wide class of MPL problems.

Lemma 3. *For the general ("log likelihood") scatter*

$$S(f(x),x) = \sum_{i=1}^{n} \delta(x-x_i)\,\ell_i(f(x)), \tag{10}$$

$a \le x_i \le b$, $\ell_i'(x) \stackrel{\text{def}}{=} \frac{d}{dx}\ell_i(x)$, *and using (8), we have*

i. $\int_a^b S(f(x),x)\,dx = \sum_{i=1}^n \ell_i(f(x_i))$, *a "log likelihood",*

ii. ${S_f}^{[0]}(x) = \sum_{i=1}^n \delta(x-x_i)\,\ell_i'(f(x))$,

iii. ${S_f}^{[m+1]}(x) = \frac{1}{m!}\sum_{i=1}^n (x-x_i)_+^m \ell_i'(f(x_i))$, *m=0,1,....*

iv. *This* $S(\cdot)$ *(10) fulfills condition (9).*

v. *The boundary conditions* (ii) *a) of Theorem 1,* ${S_f}^{[m+1]}(b) = 0$, *are equivalent to* $\sum_{i=1}^n x_i^m \ell_i'(f(x_i)) = 0$, *for* $m = 0,1,\ldots,\nu-1$.

Proof: ${S_f}^{[0]}(x) = \sum_{i=1}^n \delta(x-x_i)\,\ell_i'(f(x_i))$, and by definition (8), ${S_f}^{[1]}(x) = \sum_{i=1}^n \int_a^x \delta(t-t_i)\,\ell_i'(f(t))\,dt = \sum_{i=1}^n \mathbf{1}_{[x-x_i\ge 0]}\ell_i'(f(x_i)) = \sum_{i=1}^n (x-x_i)_+^0 \ell_i'$ (writing ℓ_i' for $\ell_i'(f(x_i))$). And for $m \ge 1$, by induction, ${S_f}^{[m+1]}(x) = \sum_{i=1}^n \ell_i' \frac{1}{m-1!}\int_a^x (t-x_i)_+^{m-1}\,dt = \frac{1}{m!}\sum_{i=1}^n (x-x_i)_+^m \ell_i'$, since $\int (t-c)_+^m\,dt = \frac{1}{m+1}(t-c)_+^{m+1}$ and $a \le x_i$. iv: This S fulfills equation (9), since $\frac{d}{d\epsilon}\Big|_{\epsilon=0}\int_{x_1}^{x_n} S(f(x)+\epsilon\eta(x),x)\,dx = \sum_{i=1}^n \frac{d}{d\epsilon}\ell_i(f(x_i)+\epsilon\eta(x_i)))\Big|_{\epsilon=0} = \sum_{i=1}^n \eta(x_i)\ell_i'(f(x_i)) = \int_{x_1}^{x_n}\eta(x)\cdot {S_f}^{[0]}(x)\,dx$. v: The boundary conditions (ii) a) $0 = {S_f}^{[m+1]}(b) = 1/m!\sum_{i=1}^n (b-x_i)^m \ell_i'$ (since $x_i \le b\,\forall i$) $\forall m \in \{0,\ldots,\nu-1\}$, are seen by induction $m \to m+1$. For $m=0$, ${S_f}^{[1]}(b) = \sum_{i=1}^n (b-x_i)_+^0\ell_i' = \sum_{i=1}^n \ell_i'$, since $x_i \le b \forall i$. For the induction step $m \to m+1$, we assume that $\sum_{i=1}^n x_i^j \ell_i' = 0$ is true for $j = 0,\ldots,m-1$. Now, $0 = m! {S_f}^{[m+1]}(b) = \sum_{i=1}^n (b-x_i)^m \ell_i' = \sum_{i=1}^n \left(\sum_{j=0}^m \binom{m}{j}(-x_i)^j b^{m-j}\right)\ell_i' = \sum_{j=0}^m \binom{m}{j}(-1)^j b^{m-j}\sum_{i=1}^n x_i^j \ell_i'$, where the inner sum is zero for all j but $j=m$. Hence, $0 = \binom{m}{m}(-1)^m b^0 \sum_{i=1}^n x_i^m \ell_i' = \pm\sum_{i=1}^n x_i^m \ell_i'$ which was to be seen. ∎

The above theorem can be applied to many special situations. Many of them are MPL problems as our (23) and other problems of nonparametric curve estimation for which Lemma 3 applies, see below. An application to density estimation is presented in Mächler (1995).

The theorem also entails the main results of Huber (1974), reformulated as following

Corollary 4 (Huber's Spline). *Consider the problem of*

"Given n (≥ 2) values of an unknown distribution function F*, determine the distribution with minimal Fisher information. Equivalently, given* $\mathrm{F}(x_i) = t_i$, $i = 1,\ldots,n$, *(and* $\mathrm{F}(-\infty) = 0$, $\mathrm{F}(\infty) = 1$*), find* $f(x) = \frac{d}{dx}\mathrm{F}(x) \ge 0$ *such that* $\int_{-\infty}^{\infty}\left(\frac{f'}{f}(x)\right)^2 f(x)\,dx$ *is minimal."*

The solution is characterized in each interval $[x_i, x_{i+1}]$ *by*

$$\sqrt{f}\,'' = \lambda_i\sqrt{f}\,, \tag{11}$$

for constants λ_i, *and* $\lim_{x\to\pm\infty} f'(x) = 0$. *For* $\lambda_i > 0$, *e.g.,* $\sqrt{f}\,(x) = a_i e^{\sqrt{\lambda_i}x} + b_i e^{-\sqrt{\lambda_i}x}$.

Remark: This corrects (iv) and (11) in Huber (1974) which have λ_i instead of $\sqrt{\lambda_i}$.

Proof: We want to show that (11) follows from Theorem 1. First, we see that $\left(\frac{f'}{f}(x)\right)^2 f(x) = f'^2(x)/f(x) = \left(f^{-\frac{1}{2}}f'\right)^2(x) = \left(\frac{d}{dx}F(x,f(x))\right)^2$, for $F(x,f(x)) := 2\sqrt{f}\,(x)$. We will apply the Theorem for $k=1$ and $\nu = 0$.

Using Lagrange parameters μ_i, the interpolation conditions $\mathrm{F}(x_i) = \int_{-\infty}^{x_i} f(x)\,dx = t_i$, $i = 0,1,\ldots,n+1$ with $x_0 := -\infty$, $t_0 = 0$, $x_{n+1} := \infty$, $t_{n+1} = 1$ (as in Huber (1974)) are incorporated into the minimization problem, as terms $\mu_i\left(\int_{-\infty}^{x_i} f(x)\,dx - t_i\right) = \int_{-\infty}^{\infty} \mu_i \cdot \Big(\mathbf{1}_{[x\le x_i]}(x) f(x) - \delta(x-x_i)\,t_i\Big)dx$, resulting in

$$\text{Min.} \int_{-\infty}^{\infty}\left\{\underbrace{\Big(\sum_{i=1}^{n+1}\mu_i \mathbf{1}_{[x\le x_i]}(x)\Big)f(x) - \sum_{i=1}^{n+1}\delta(x-x_i)\,\mu_i t_i}_{S(f(x),x)} + \Big(\frac{d}{dx}F(x,f(x))\Big)^2\right\}dx.$$

Now, we apply Theorem 1. The differential equation (i) (for $k=1, \nu=0$) is $F_g \cdot \frac{d^2}{dx^2}F = \frac{1}{2}(-1)^{0+1+1}S_f{}^{[0]}$, where $S_f{}^{[0]}(x) = \sum_{i=1}^{n+1}\mu_i \mathbf{1}_{[x\le x_i]}(x) = \sum_{i=1}^{n+1} 4\lambda_i \mathbf{1}_{[x_{i-1}<x\le x_i]}(x)$ for constants λ_i. Since $F_g = \frac{\partial}{\partial g}(2g^{\frac{1}{2}})|_{g=f(x)} = 1/\sqrt{f}$, (i) is equivalent to $2\sqrt{f}\,''/\sqrt{f} = \frac{1}{2}S_f{}^{[0]}$, or $\sqrt{f}\,'' = \sqrt{f}\sum_{i=1}^{n+1}\lambda_i \mathbf{1}_{[x_{i-1}<x\le x_i]}(x)$ which is $\sqrt{f}\,\lambda_i$ for $x_{i-1} < x \le x_i$ and $i = 1,2,\ldots,n+1$.

The boundary conditions (ii) a) are empty and (ii) b) for $j=k=1$ is $\frac{d}{dx}F(x^*) = 0$ for $x \in \{-\infty,\infty\}$, or $\lim_{x\to\pm\infty} f'(x) = 0$. ■

From Theorem 1, we get the 'classical' result about robust smoothing splines of order m. See Greville (1969), theorem 14.1, for the case of weighted least-squares splines, Huber (1979) discussing the problem of robustifying *discrete penalty* cubic "splines" and the choice of ρ or ψ functions, and Cox (1983) who considers general robust "M-type splines".

Corollary 5 (Splines). *For $m \ge 1$, the 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 of the form

$$f(x) = c_0 + c_1 x + \ldots + c_{m-1}x^{m-1} + \frac{(-1)^m}{2\lambda(2m-1)!}\sum_{i=1}^{n}(x-x_i)_+^{2m-1}\psi_i(y_i - f(x_i)) \qquad (12)$$

with conditions

$$\sum_{i=1}^{n} x_i^k\ \psi_i(y_i - f(x_i)) = 0 \qquad \textit{for } k = 0,\ldots,m-1, \qquad (13)$$

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

Note that equation (12) is a robustified 'truncated power' representation of so-called *natural* spline of order $2m$, i.e., f is a degree $m-1$ polynomial outside $[x_1, x_n]$. 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 (13) — orthogonality relations of the "huberized residuals" ψ_i — are equivalent to a linear equation system for $(c_0,\ldots,c_{m-1})$.

Proof: We will apply the theorem with

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

and can apply Lemma 3, since S is of the form (10) with $\ell_i(f(x)) = \frac{1}{\lambda}\rho_i(y_i - f(x))$. Hence, ${S_f}^{[m+1]}(x) = -1/(\lambda m!) \sum_{i=1}^{n} (x - x_i)_+^m \psi_i(y_i - f(x_i))$ for $m \geq 0$.

Further, $F(x, g(x)) = g(x)$ whence $F_g \equiv 1$, and $m \equiv \nu + k$ where ν and k are most conveniently set to $\nu = m$ and $k = 0$, we can apply the theorem and the differential equation (i) becomes

$$f^{(m)}(x) = (-1)^m / (2\lambda(m-1)!) \sum_{i=1}^{n} (x - x_i)_+^{m-1}\, \psi_i(y_i - f(x_i)), \tag{14}$$

and the boundary conditions are (13) from Lemma 3 v. Integrating (14) m times results in (12). ∎

3 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 $\kappa(x) \approx c \cdot f''(x)$. Glass (1966) however 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*

$$\kappa'/\kappa \;=\; f'''/f'' - 3f'f''/(1 + f'^2) \;\approx f'''/f''. \tag{15}$$

This shows that the approximation $\kappa'/\kappa \approx f'''/f''$ holds exactly at all the local extrema and inflection points which can be considered as the most interesting points of the curve, such that this approximation seems to be less problematic than $\kappa(x) \approx cf''(x)$ for the splines (Mächler, 1993). The approximation leads to the preliminary penalty

$$\tilde{R}[f] := \int_{x_1}^{x_n} \left(\frac{f'''(t)}{f''(t)}\right)^2 dt. \tag{16}$$

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 a curve with $n_w + 1$ inflection points is infinitely less smooth than one with 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 $R[f] = \int_{x_1}^{x_n} \left(\frac{f'''(t)}{f''(t)} - \text{“poles”}\right)^2 dt$, see (21).

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. 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 $s_f = 1$ or $s_f = -1$. More conveniently, we define the degree n_w polynomial,

$$p_{\mathbf{w}}(x) \stackrel{\text{def}}{=} s_f(x - w_1)(x - w_2)\cdots(x - w_{n_w}), \tag{17}$$

and have

$$f''(x) = p_{\mathbf{w}}(x)\, e^{h_f(x)}, \quad \text{or} \tag{18}$$
$$h_f = \log(f''/p_{\mathbf{w}}), \tag{19}$$

where by definition of $p_{\mathbf{w}}$, h_f is as many times differentiable as f'' (at least once). Hence, $f'''/f'' = \frac{d}{dx}\log f'' = (\log p_{\mathbf{w}})' + {h_f}'$, or

$${h_f}' = \frac{f'''}{f''} - \sum_{j=1}^{n_w} \frac{1}{x - w_j}. \tag{20}$$

Note that the continuous function ${h_f}'$ is well defined for all x, whereas the r.h.s. of (19) is "$\infty - \infty$" for $x = w_j$. Further, 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

$$R[f] = \int_{x_1}^{x_n} {h_f}'(t)^2\, dt, \tag{21}$$

is suitable for measuring the change of curvature "apart from the inflection points". Note that for $n_w = 0$, $R[f]$ generalizes $\tilde{R}[f]$ from (16) which is not defined for $n_w > 0$.

The number of inflection points, n_w, is the main smoothing parameter of the "Wp" approach. 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.

We will consider a straightforward generalization of this approach. First, we will 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. Another generalization consists in penalizing higher order 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 ($\nu = 2$, $k = 1$) corresponds to the change of curvature roughness measure. The generalizations on the other hand follow easily, and $k = 2$ may seem more attractive for its limit of 'infinite smoothing' (Corollary 7).

Given the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and n_w, the maximal number of (generalized) inflection points, we want to determine the function f minimizing $\sum_{i=1}^n \rho_i\,(y_i - f(x_i)) + \lambda \int_{x_1}^{x_n} \frac{d^k}{dx^k} h_f(x)^2\, dx$. In the following, $\mathbf{w} = (w_1, \ldots, w_{n_w})^\mathsf{T}$ and s_f are fixed, i.e., $p_{\mathbf{w}}$ is fully specified. An "outer" minimization over $\mathbf{w}$ is needed to find the global optimum.

As in (18) for f'', we factorize $f^{(\nu)}$ as $f^{(\nu)}(x) = p_{\mathbf{w}}(x)e^{h_f}(x)$. If we let

$$\mathcal{F}_m^{(\nu)}(\mathbf{w}) \stackrel{\text{def}}{=} \{f \in C^m[a, b]\ ;\ f^{(\nu)} \text{ has exactly the zeros } w_1, \ldots, w_{n_w}\},\ (m \geq \nu \geq 0), \tag{22}$$

the function class $\mathcal{F}_{k,\nu,F}$ (6) is here equal to $\mathcal{F}_{2k+\nu}^{(\nu)}(\mathbf{w})$, since for our problem, $\int_{x_1}^{x_n} \frac{d^k}{dx^k} h_f(x)^2\, dx$ is only finite for functions f which do have the 'generalized inflection points' $w_1, \ldots, w_{n_w}$ and no others.

Corollary 6 (Generalized "Wp"). *The necessary equation system for a minimizer of*

$$J[f] = \sum_{i=1}^{n} \rho_i\,(y_i - f(x_i)) + \;\lambda \int_{x_1}^{x_n} \left(\frac{d^k}{dx^k} h_f(x)\right)^2 dx, \qquad (23)$$

among all functions $f \in \mathcal{F}^{(\nu)}_{2k+\nu}(\mathbf{w})$*, i.e.,* $f^{(\nu)}(x) = \pm(x-w_1)(x-w_2)\cdots(x-w_{n_w})\cdot e^{h_f(x)}$*, or* $f^{(\nu)}(x) = p_{\mathbf{w}}(x)e^{h_f(x)}$ *is*

$$h_f{}^{(2k)}/f^{(\nu)} = \frac{(-1)^{\nu+k}}{2\lambda(\nu-1)!}\sum_{i=1}^{n}(x-x_i)_+^{\nu-1}\psi_i(y_i - f(x_i)), \qquad (24)$$

where $h_f(x) = \log\frac{f^{(\nu)}}{p_{\mathbf{w}}}$ *and* $\psi_i(x) = \frac{d}{dx}\rho_i(x)$*. The natural boundary conditions are*

$$a) \qquad \sum_{i=1}^{n} x_i{}^m \psi_i(y_i - f(x_i)) = 0 \qquad \forall m \in \{0,\ldots,\nu-1\} \qquad (25)$$

$$b) \quad h_f{}^{(k)} = h_f{}^{(k+1)} = \cdots = h_f{}^{(2k-1)} = 0 \quad \text{for } x \in \{x_1, x_n\}, \qquad (26)$$

Proof: We use the theorem and Lemma 3 as for Corollary 5 with $F(x, f^{(\nu)}(x)) = \log(f^{(\nu)}(x)/p_{\mathbf{w}}(x))$ and $F_g(x, f^{(\nu)}(x)) = 1/f^{(\nu)}(x)$. Applying the theorem, (i) becomes (24), the boundary conditions a) are Lemma 3,v, and the b) conditions are those of the theorem. ∎

Our smoother satisfies an *exact fit property*. From the boundary conditions (25) of Corollary 6, the "orthogonality conditions", it is easily seen that the generalized Wp smoother fits the data exactly if they lie on a polynomial of degree $\nu - 1$ (straight line for $\nu = 2$). It also follows that the smoother is regression equivariant under superposition of such polynomials.

It is of interest to consider the "most smooth" generalized "Wp" smoother, i.e., the solution for $\lambda \to \infty$ of Corollary 6:

Corollary 7 (Smoothest Limit). *For* $\lambda \to \infty$*, we have*

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

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* $f^{(\nu)} = p_{\mathbf{w}}$*. For* $k = 1$ *["Wp"],* $f^{(\nu)} \to A\cdot p_{\mathbf{w}}$ *(for some* $A \in \mathbb{R}$*) and* f *is the least-*ρ *polynomial of degree* $n_w + \nu$ *with* ν*-inflection points* $w_1, \ldots, w_{n_w}$ *(i.e.,* $f^{(\nu)}(w_j) = 0 \quad \forall j$*). For* $k = 2$*,* $f^{(\nu)}(x) \to A\cdot p_{\mathbf{w}}(x)\ e^{Bx}$ *and* $f(x) \to P_{\nu-1}(x) + P^*_{n_w}(x)e^{Bx}$*, where* $P^\cdot_k$ *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 (24), 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. ∎

The special case of corollary 6 for $\nu = 2$ and $k = 1$ is basic for the algorithm implementing the Wp smoother: The ordinary differential equation (24) is equivalent to

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

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, f has to satisfy the 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 $w_1, \ldots, w_{n_w}$.

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.

4 Summary

We have derived the general Theorem 1, useful for characterizing many MPL problems for curve estimation, including polynomial splines.

The main application however is for our new roughness penalty of "change of curvature". 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 $\lambda \to \infty$) give 'natural' parametric curves (Corollary 7).

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 $\mathcal{F}_m^{(\nu)}(\mathbf{w})$ (22) to show the principle. For many other subsets of $C^m[a,b]$, the theorem will be valid by an analogous proof.

Lemma 8 (Euler-Lagrange differential equation and boundary conditions). *Given integers $m \geq \nu \geq 0$, $\mathcal{F}_m^{(\nu)}(\mathbf{w})$ the set of m times differentiable functions where $f^{(\nu)}$ 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

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

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, the fundamental lemma may not be applied directly, since we have $\eta \in \mathcal{F}_m^{(\nu)}(\mathbf{w})$ and can not have $\eta \equiv 0$ on any interval. We consider a subset of $\mathcal{F}_m^{(\nu)}(\mathbf{w})$, namely functions η of the form

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

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

$$\eta^{(\nu)}(x) = e^{-\alpha x} \sum_{k=0}^{n_w} \left(\sum\nolimits_{j=0}^{n_w-k} \binom{\nu}{j} (-\alpha)^{\nu-j} (k+j)_j \; q_{k+j} \right) x^k, \tag{27}$$

where $(n)_k = n(n-1)\ldots(n-k+1)$, as in definition (30). 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 (27) 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 2. 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 nature 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

$$\frac{d^n}{dx^n} F(x, g(x)) = F_n(x; g(x), g'(x), \ldots, g^{(n)}). \tag{28}$$

Lemma 9 (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

$$\frac{\partial}{\partial g^{(j)}} F_n = \binom{n}{j} \frac{\partial}{\partial g} F_{n-j} \quad \forall j \in \mathbb{N}_0 \quad \forall n \in \mathbb{N}_0 \tag{29}$$

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 for $j > n$ (where both sides vanish). It remains to be proved $1 \le j \le n$. For $j = n$, (29) 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 (29) 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 $\forall k \in \mathbb{N}_0 \quad \forall a \in \mathbb{R}$:

$$\begin{aligned} (a)_k &\stackrel{\text{def}}{=} a(a-1)\cdots(a-k+1) \quad \text{with } (a)_0 = 1, \\ \binom{a}{k} &\stackrel{\text{def}}{=} \frac{(a)_k}{k!}, \quad \text{and for } k<0: \binom{a}{k} = 0. \end{aligned} \tag{30}$$

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 10 (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 $f(t) = t^m/(m!) +$ (lower power terms of t), $m \le n$, 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 and $|x| < 1$ (e.g. Abramowitz and Stegun, 1972, 3.6.8). 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): $c_{n,m,J} = \sum_{j=m+J}^{n} \cdots = \sum_{j=0}^{n} \cdots - \sum_{j=0}^{m+J-1} (-1)^j \binom{n}{j}\binom{j-J}{m}$, 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 $0 \le j - J < m$ ($\binom{k}{m} = 0$ for $0 \le k < m$), 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 11 (Higher chain-rule identity 2).
Let $g()$ and $F()$ be as general as in Lemma 2, and F_n defined by (28). 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 10.

Proof: We apply Lemma 9 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

$$\text{For any function } \; G(x, u(x), u'(x), \ldots, u^{(n)}(x)): \; \frac{\partial}{\partial u}\frac{d}{dx}G = \frac{d}{dx}\frac{\partial}{\partial u}G \tag{31}$$

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 10 (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 $0\le m\le j-j_0\le k-j_0$):

$$\sum_{j=j_0}^{k}(-1)^j\frac{d^{j-j_0}}{dx^{j-j_0}}\mathcal{V}_{g^{(j)}}(x)=2\sum_{m=0}^{k-j_0}C_m(x)F_{k+m}(x), \tag{32}$$

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 11 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 10 (iii)2)), where $C_k=(-1)^k\cdot\frac{\partial}{\partial g}F_{k-k-0}=(-1)^kF_g(x,g(x))$.

(ii): Because of equation (32), the equivalence of (1) and (2) is proved if we can show that here, $C_m\not\equiv 0$ for $m\in\{0,\dots,k-j_0\}$. We have $c_{k,m,j_0}\not\equiv 0$ from Lemma 10 (iii)2), since $j_0\ge 1$. 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 the main theoretical 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, and also Steve Marron, who helped to improve earlier versions of this manuscript. The comments of the editor Michael Woodroofe and the others in the refereeal process have lead to further improvement of the present paper.

References

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

Ansley, C. F. (1993). Nonparametric spline regression with prior information, *Biometrika* **80**(1): 75–88.

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, *JASA* **74**: 828–836.

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

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators, *Annals of Statistics* **18**(4): 1676–1695.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*, Monographs on numerical analysis, Clarendon Press, Oxford.

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.

Huber, P. J. (1974). Fisher information and spline interpolation, *Annals of Statistics* **2**: 1029–1033.

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.

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

Kimeldorf, G. and Wahba, G. (1971). Some results on Tschebycheffian spline functions, *Journal of Analysis and Applications* **33**: 82–95.

Klonias, V. K. (1984). On a class of nonparametric density and regression estimators, *Annals of Statistics* **12**: 1263–1284.

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

Mächler, M. B. (1993). Very smooth nonparametric curve estimation by penalizing change of curvature, *Res. Report 71*, Seminar für Statistik, ETH Zurich.

Mächler, M. B. (1995). Estimating distributions with a fixed number of modes, *in* H. Rieder (ed.), *Festschrift on the occasion of Peter Huber's 60th birthday*, Springer, Berlin. *to appear*

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

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

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.

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.

Thompson, J. R. and Tapia, R. A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation*, SIAM, Philadelphia.

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

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

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich
Switzerland

Figure Caption:

Figure 1: 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. The first cubic spline is tuned to have the same residual sum of squares and also shown shifted downwards by 1 with marked inflection points. The second spline is the best fitting one for 3 inflection points.