Density Estimation: New Spline Approaches and a Partial Review

Martin B. Mächler
Seminar für Statistik, ETH Zentrum
CH-8092 Zürich, Switzerland
e-mail: maechler@stat.math.ethz.ch

CompStat, August 1996 (February 27, 2002)
$Id: P.tex,v 1.15 2002/02/27 17:49:20 maechler Exp

Abstract

Apart from kernel estimators, there have been quite a few different approaches of "generalized splines" for density estimation. In the present paper, *Maximum Penalized Likelihood* (MPL) approaches are reviewed. In conclusion, penalizing the *log density* seems most promising.

In my "wp" approach for semi-parametric density estimation, a novel roughness penalty is introduced. It penalizes a relative *change* of curvature which allows considering modes and inflection points. For a given number of modes, $l' = (\log f)'$ can be represented as $l'(x) = \pm(x - w_1) \cdots (x - w_m) \cdot \exp h_l(x)$, a *semi-parametric* term with parameters w_j (model order m) and nonparametric part $h_l(\cdot)$. The MPL problem is equivalently solved via a boundary value differential equation.

Key words: Nonparametric, Semi-Parametric, Density estimation; Smoothing; Roughness penalty; Maximum penalized likelihood; Inflection point; Boundary Value Problem, Differential Equation, Unimodality, Multimodality.

1 Introduction

The nonparametric estimation of distribution or density functions is well-established and a variety of methods available, with kernel estimators being the ones most researched, see (Silverman, 1986; Izenman, 1991; Wand & Jones, 1995) for some review.

Here, I will try to shed light on different approaches of density estimation. Whereas in nonparametric regression, *spline* functions have been investigated and used extensively, this is less the case for density estimation. *Maximum Penalized Likelihood* (MPL) estimation has been among the first methods ((Good & Gaskins, 1971)) and still seems appealing when "proper" smoothness is desired ((Mächler, 1995*b*) and below). The histosplines of Boneva, Kendall & Stefanov (1971) were another early approach of spline density estimation. As for the histogram however, the somewhat arbitrary choice of the knots has made the histospline less appealing.

In the present paper, I'll review different spline–like approaches in section 2, present my own MPL approach in section 3, and in section 4 shortly look at some new approaches of semiparametric density modeling.

2 Maximum Penalized Likelihood (MPL)

Here, we will mainly consider *"generalized splines"*, i.e., approaches within the framework of *"maximum penalized likelihood estimation"*, MPLE (cf. (Thompson & Tapia, 1990, ch. 3–5) which "contains" (Tapia & Thompson, 1978)), considering different *roughness* penalties.

Given independent observations $x_1, x_2, \ldots, x_n$ of $X \sim F$, with "smooth" density $f(x) = \frac{d}{dx}F(x)$ on $[a, b]$ (i.e., $P[a \le X \le b] = 1$ and therefore, $x_i \in [a, b]\ \forall i$), $-\infty \le a < b \le +\infty$, our goal is to estimate the true f, or F. Sometimes, we assume that f is (twice continuously) differentiable, and define

$$\begin{aligned} & l(x) &:=& \log f(x) && \textit{log-density}, \\ \text{and} \qquad & l'(x) &=& \frac{d}{dx} l(x) = \frac{f'(x)}{f(x)} && -\,(\textit{score function}). \end{aligned} \tag{1}$$

To estimate f (or F or l, equivalently), we Maximize the Penalized Likelihood criterion,

$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(x_i) - \lambda \tilde{\Phi}(f), \qquad \text{or} \qquad \max_{l \in \mathcal{L}} \sum_{i=1}^{n} l(x_i) - \lambda \Phi(l), \tag{2}$$

where $\mathcal{F}$ and $\mathcal{L}$ are appropriate function classes, $f : [a, b] \to \mathbb{R}_+$, i.e., $f(x) \ge 0$ for all x, or $l : [a, b] \to \mathbb{R}$, respectively, with the property

$$\int_a^b f(t)\, dt = \int_a^b e^{l(t)}\, dt = 1, \tag{3}$$

$\sum_i l(x_i)$ is the log likelihood, $\Phi : \mathcal{L} \to \mathbb{R}_+$ or $\tilde{\Phi}(f) \equiv \Phi(\log f) = \Phi(l)$ the roughness penalty, and $\lambda \ge 0$ the smoothing parameter.

Often, the null space of Φ, $\Phi_\perp := \{l \mid \Phi(l) = 0\}$, is finite dimensional. This is especially attractive, since in this situation, the limiting case, $\lambda \to \infty$ giving the "most smooth" solution, is equivalent to classical *Maximum Likelihood* estimation in $\Phi_\perp$, see below. Silverman (1982) (cf. (Silverman, 1986, ch. 5.4)), Cox & O'Sullivan (1990) and Gu & Qiu (1993) develop a nice theory, deriving existence and uniqueness results for a wide class of MPL problems, including speed of convergence and consistency in various norms.

2.1 Penalizing $\sqrt{f}$ — Good-Gaskins etc

Good & Gaskins (1971) used the roughness penalty $\tilde{\Phi}_1(f) = \int_a^b f'^2(t)/f(t)\, dt$, the Fisher information which can be written as $\tilde{\Phi}_1(f) = 4 \int_a^b u'^2(t)\, dt$ where $u := \sqrt{f}$. A second proposal was to generalize the problem to penalties $\tilde{\Phi}_2(f) = \alpha \int_a^b u'^2(t)\, dt + \beta \int_a^b u''^2(t)\, dt$.

De Montricher, Tapia & Thompson (1975) derive exact existence results for the proposed estimators of Good & Gaskins (1971), and were able to characterize the first one as "exponential spline", see also Thompson & Tapia (1990, ch. 4). However, the resulting curve has "kinks", since the derivative f' is discontinuous at every data point ((Silverman, 1986, §5.4.2)). Whereas the minimizer for the $\tilde{\Phi}_2$ problem will be smoother, it is delicate to be computed, because $u(x) \ge 0$ is necessary ((De Montricher et al., 1975)).

The "Sobolev" s-norm penalty $\tilde{\Phi}_3(f) = \int_a^b {f^{(s)}}^2(t)\, dt$ (under $f^{(j)}(a) = f^{(j)}(b) = 0$ for $j = 0, 1, \ldots, s-1$) is considered in De Montricher et al. (1975), where the authors prove existence and uniqueness, and (for $s = 1$) provide an approximating solution, using discretization.

All these approaches have the drawback that the "most smooth" solution is problematic, since the space $\{f;\ \tilde{\Phi}(f) = 0; f \ge 0\}$ is not well characterized or even degenerate.

2.2 Penalizing $\log f$

Silverman (1982) (see (Silverman, 1986, §5.4.4)) introduces the penalty $\Phi(l) = \int_a^b l'''^2(t)\,dt$ and proves consistency in three different norms. Silverman also proves that the solution of the constrained MPL problem (2, 3) is *equivalent* to solving the **un**constrained problem

$$\max_{l \in \mathcal{L}} \sum_{i=1}^{n} l(x_i) - \lambda \Phi(l) - n \int_a^b e^{l(t)}\,dt, \tag{4}$$

for a very general class of penalties. The same is seen below for the "Wp" penalty.

The choice of penalty here leads to the attractive feature that the smoothest limits ($\lambda \to \infty$) are in $\Phi_\perp = \{l''' = 0\} = \{l \text{ quadratic}\}$ which are exactly the Gaussian distributions, and $\lambda \to \infty$ corresponds to normal ML estimation. This feature is analogous to the cubic smoothing splines in regression which lead to least squares linear regression for $\lambda \to \infty$, and is a property which the vastly used kernel density estimates do *not* share (a regular "smoothest limit" does not exist there!).

A related — from several standpoints very appealing — approach is to estimate (and penalize) the *score function* $\psi = -l' = -f'/f$ (which is a straight line for a Gaussian). Cox (1985) introduced and solved a penalized "mean square error" problem for the score function, and Ng (1994) provided further properties and computational algorithms.

The *logspline* approach of Kooperberg & Stone (1991) and (1992) is an attractive practical approach of using cubic splines to model the log density. However, it is not an MPLE, but rather a ML estimation in carefully chosen space of "regression splines" (splines with knots determined by the data).

3 Penalizing Modes: 'Wp'

Very early on, the number of modes (local maxima) of the unknown density has been considered important (e.g., (Cox, 1966; Silverman, 1981)) and even the number of "bumps" (concave regions between inflection points) has been tried to be kept small ((Good & Gaskins, 1971), (1980)).

Modes are often *the* important attribute of a distribution, since they (informally) lead to conclusions about the underlying population. Tests of unimodality (vs. multimodality) have been developed (see above, (Hartigan & Hartigan, 1985), and (Minnotte & Scott, 1993), for a nice graphical tool). See Mächler (1995*a*) for more examples in the literature.

An appealing notion of smoothness is considering *local extrema* and/or *inflection points* of a curve. For nonparametric regression, I *limited the number of inflection points* ((Mächler, 1993),(1995*b*)), as a special case of a very general approach of MPLE. Here, for density estimation, I apply this approach to finding the "best" density function with a given number of local extrema, or almost equivalently, number of modes[1]. See Klonias (1984), Roussas (1991), and Cuevas & Gonzalez-Mantiega (1991) for approaches with a similar aim.

Here, our goal is to estimate the density f such that it has (at most) m^* modes, i.e. typically, $m = (2m^* - 1)_+$ local extrema. For differentiable functions f, w is a local extremum if and only if $f'(w) = 0$, or equivalently, $l'(w) = 0$. Denote the m local extrema

[1] *"mode"* is used here for maxima x where $f'(x) = 0$. For differentiable densities f, this only excludes maxima at the *end* of the support interval, e.g., $x = 0$ for the exponential distribution (which has zero modes in our definition).

of f by $w_1, w_2, \ldots, w_m$. From basic calculus we know that l' factors into

$$l'(x) = p_{\mathbf{w}}(x)\,(-1)^j\, e^{h_l(x)}, \quad \text{where} \tag{5}$$

$$p_{\mathbf{w}}(x) := (x - w_1)(x - w_2)\cdots(x - w_m), \tag{6}$$

and h_l is continuously differentiable, but otherwise "arbitrary" (in the vast class of functions with penalty $\int_a^b h_l'^2\, dt < \infty$). The sign $(-1)^j$ in (5), determined by $j = 0$ or 1, is assumed to be given (and will be negative, $j = 1$, in most situations).
Note that from (5),

$$(\log l'(x))' = (l''/l')(x) = h_l'(x) + \sum_{j=1}^{m} \frac{1}{x - w_j}, \tag{7}$$

such that h_l' can be regarded as "l''/l' minus the poles". By specifying a penalty in terms of h_l', we have a semi-parametric model approach, the parameters being the modes and antimodes w_j, $j = 1, \ldots, m$, and nonparametric component h_l.

Here, we estimate f (and h_l), by maximizing the MPL (2) where $\Phi(l) = \int_a^b h_l'^2(t)\,dt$, and $\mathcal{L}$ is the set of all functions $l : [a, b] \to \mathbb{R}$ fulfilling (5), (6) in addition to the normalizing condition (3), $\int_a^b e^{l(t)}\, dt = 1$. $\mathcal{L}$ is therefore the product of $Q^m \subset \mathbb{R}^m$ ($\mathbf{w} \in Q^m$) and the space of functions $h_{l;\mathbf{w}}$. In effect, we will minimize over the space of h_l for *fixed* $\mathbf{w}$, and "globally" over all mode locations $\mathbf{w}$.

In the following, the situation for *fixed* $\mathbf{w}$ is considered. The penalty $\int_a^b h_l'^2(t)\, dt$ contains the derivative h_l' which can be expressed as

$$h_l'(x) = \frac{\kappa_L'(x)}{\kappa_L(x)} - \sum_{j=1}^{m} \frac{1}{x - w_j} + \; l(x)l'(x)\, 3(1 + l^2)^{-1}(x), \tag{8}$$

by using (7) and $\kappa_L'/\kappa_L = l''/l' - 3ll'(1 + l^2)^{-1}$ where κ_L is the curvature of $L(x) = \int^x l(t)\, dt$. Arguments in Mächler (1989) and (1993) suggest that a natural roughness penalty for a curve $y = g(x)$ is the *"relative change of curvature"* κ'/κ where $\kappa(x) = g''(x)\left(1 + g'^2(x)\right)^{-3/2}$ is the curvature at x. Eq. (8) implies that h_l' approximates the relative change of curvature "apart from the poles" of L.

Using a Lagrange parameter α for the side condition (3) and making use of Dirac's δ distribution notation, the problem (2) is equivalent to the *variational problem*

$$\min_{l \in \mathcal{L}} \int_a^b \left\{ \alpha\, e^{l(t)} \; - \sum_{i=1}^{n} \delta(t - x_i) l(x_i) + \lambda h_l'^2(t) \right\} dt. \tag{9}$$

3.1 Differential Equation

Now we apply Theorem 1 of Mächler (1995*b*) (and (1989),(1993)) which gives the Euler-Lagrange ordinary differential equation (=: o.d.e.) and boundary conditions for a vast class of variational problems in a convenient form:

Theorem 1 (Mächler)
The minimizer f of $\int_a^b \left(S(f(t), t) + \{(d/dt)^k F(t, f^{(\nu)}(t))\}^2\right) dt$ *fulfills the o.d.e.*

$$F_g \cdot \left(\frac{d}{dt}\right)^{2k} F = \frac{1}{2}(-1)^{\nu+k+1} S_f{}^{[\nu]}(t) \quad \textit{for all } t \in [a, b], \tag{10}$$

where $F_g(t, g) := \frac{\partial}{\partial g} F(t, g)$, $S_f{}^{[0]}(x) := \frac{\partial}{\partial f} S(f, x)\Big|_{f=f(x)}$, *and* $S_f{}^{[j+1]}(x) := \int_a^x S_f{}^{[j]}(t)\, dt$ *for* $0 \le j < \nu$. *The boundary conditions of this very general problem are*

(A) $$S_f{}^{[1]}(b) = S_f{}^{[2]}(b) = \cdots = S_f{}^{[\nu]}(b) = 0, \quad \textit{and}$$

(B) $$(\tfrac{d}{dt})^j F(t_\cdot) = 0 \qquad \textit{for } t_\cdot \in \{a, b\} \quad \textit{and} \quad j \in \{k, \ldots, 2k-1\}.$$

For our problem (9), $\nu = k = 1$, and $\lambda S(l, x) = \alpha e^l - \sum_{i=1}^n \delta(x - x_i) l(x_i)$. Therefore $\lambda S_f{}^{[1]}(x) = \int_a^x \lambda \frac{\partial}{\partial l} S(l, t)\, dt = \alpha \int_a^x e^{l(t)}\, dt - \sum_{i=1}^n \mathbf{1}_{[x_i \le x]} = \alpha F(x) - n F_n(x)$, where $F_n(x)$ is the empirical distribution function defined as $F_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}_{[x_i \le x]}$. Further, $F(x, l') = h_l(x) = \log |l'/p_{\mathbf{w}}(x)|$. The boundary conditions (A) are $0 = \lambda S_f{}^{[1]}(b) = \alpha F(t) - n F_n(b)$ which entails $\alpha = n$ for the Lagrange parameter because $F(b) = F_n(b) = 1$. The conditions (B) are $0 = d/dx\ F = h_l'(x)$ at the boundary $x = a$ or b. The o.d.e. is $F_g \cdot (d/dx)^2 F = 1/2(-1)^3 S_f{}^{[1]}(x)$, and $F_g = \frac{\partial}{\partial g} F(g, x) = \frac{\partial}{\partial g} \log |g/p_{\mathbf{w}}(x)| \Big|_{g=l'} = 1/l'$, $(d/dx)^2 F = h_l''$ and $S_f{}^{[1]} = n(F - F_n)$ from above. Therefore, the characterizing Euler-Lagrange o.d.e. for problem (2) is

$$h_l''(x) = \frac{l'(x)}{2\lambda}\ n\left(F_n(x) - F(x)\right), \tag{11}$$

with boundary conditions

$$F(a) = 0,\ F(b) = 1, \quad \text{and} \quad h_l'(a) = h_l'(b) = 0, \tag{12}$$

where F_n is the empirical distribution function. This 4^{th} order o.d.e. for F can be rewritten as 1^{st} order with four unknown functions (F, l, h_l, h_l'), $F' = \exp(l)$, $l' = (-1)^j p_{\mathbf{w}} \exp(h_l)$, $h_l' = \{h_l\}'$, and h_l'' obeys (11). Note that the data only enter through F_n, and the modes (and antimodes) are specified in $p_{\mathbf{w}} = (x - w_1) \cdots (x - w_m)$.

This boundary value o.d.e. is solved numerically to obtain the MPLE for fixed $\mathbf{w} = (w_1, w_2, \ldots, w_m)^\top$. I have used a collocation method o.d.e. solver, "COLNEW" from `netlib`, see (Bader & Ascher, 1987), and, (Ascher, Mattheij & Russell, 1988). Numerical boundary value o.d.e. solvers require an initial "crude" approximate solution to be provided. The *"log-splines"* of Kooperberg & Stone (1992) have been convenient "pre-smoothers" for this. Experience indicates that the locations w_j of modes and antimodes are quite well determined by such a pre-smoother initially.

For the final estimate, the minimizer of (2), the penalized log-likelihood of the above o.d.e.-solution must be *minimized* over $\mathbf{w}$. This is relatively easy, at least as long as the number m of local extrema is not too large for the problem.

Example. The "Old Faithful geyser eruption lengths" are at least bimodal, and have been used before (e.g., (Silverman, 1986; Härdle, 1990)). Here I use the data set `dat.geyser$x` as supplied in the Statlib (`ftp lib.stat.cmu.edu`) submission `S/haerdle`. These are $n = 272$ observations (at 126 values) of consecutive eruption lengths in minutes (rounded to seconds precision). Figure 1 shows the data and a Gaussian kernel density estimate with window width $h = h^* = .3348$ which is default (3.31) from Silverman (1986) and one with smaller bandwidth $h = .1$. The first kernel density looks perfect showing two modes and no unnecessary bumps. A closer look (e.g., plot $F(x)$ and $F_n(x)$ vs. x) however reveals that the bandwidth chosen is too large (since the asymptotic for h^* is geared towards unimodal densities, (Silverman, 1986, 3.4.2, p.45ff)). The more appropriately fitting kernel ($h = .1$) is wiggly (see also Fig. 2).

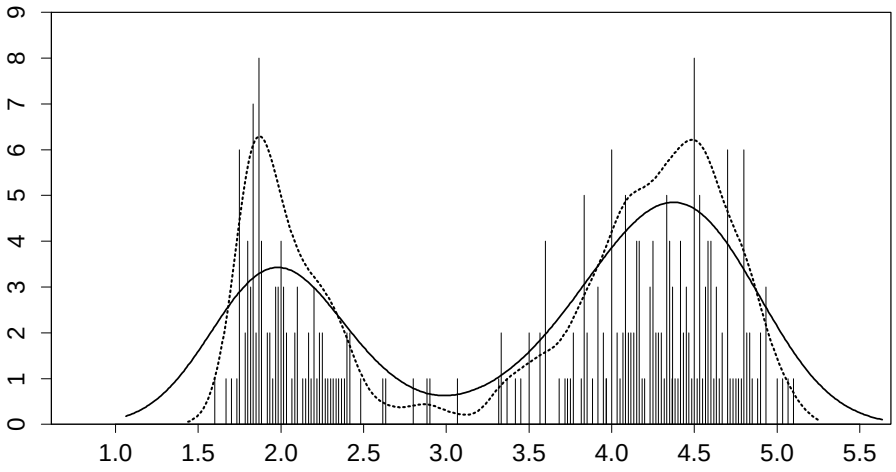


Figure 1: Old Faithful Geyser eruption lengths, $n = 272$; binned data and two (Gaussian) kernel density estimates ($\times 10$) with $h = h^* = .3348$ and $h = .1$ (dotted).

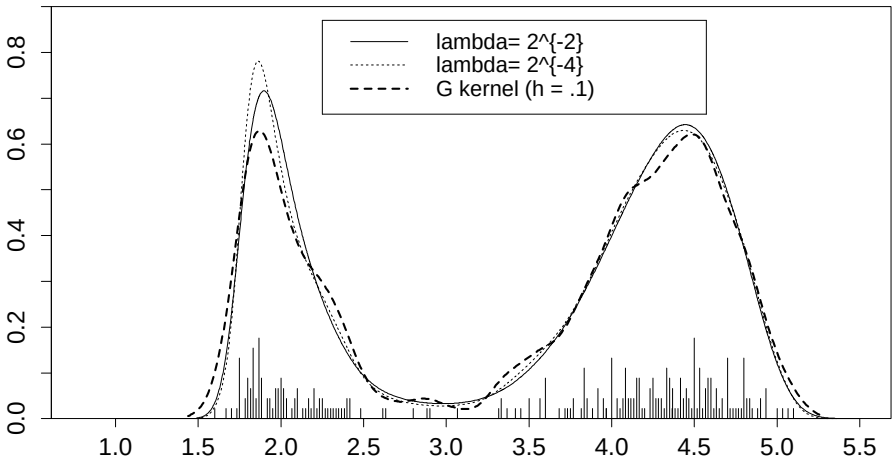


Figure 2: Estimated densities $\hat{f}(x)$ according to the new approach, (5), (6), (2), with two modes ($m = 3$), for $\lambda = 2^{-2}, 2^{-4}$ and $[a, b] = \mathbb{R}$. Both "Wp" curves are much smoother than the Gaussian kernel estimate which even fits (somewhat) less to the data than the smoother "Wp" curve ($\lambda = 1/4$).

Figure 2 shows $\hat{f}(x)$ for $\lambda = 2^{-2}, 2^{-4}$ and the Gaussian kernel from above. The estimates of our new approach are quite better (quantified in (Mächler, 1995*a*), where a more extensive comparison is made). They fit better to the data while being more smooth than traditional estimates (Gaussian kernels, parametric mixture of two Gaussian, logsplines) with comparable amount of fit.

While theoretical results are not yet available, current experience shows that the new class of density estimates provides very smooth curves while adapting to the data very well.

3.2 Properties of Solution

By this new approach, we get the distribution function with many derivatives, simultaneously, since F,f, $l' = \text{score of } f$, l'' are continuously differentiable. f''' and l''' are still "cad-lag" (right-continuous, limit from left) being continuous functions of h_l'' and therefore of the empirical distribution F_n.

The prescribed number of modes is the main smoothing parameter. The (extra) smoothing parameter λ is of less importance, since for the "roughest" situation, $\lambda \to 0$ f is still "smooth", whereas for nonparametric estimators, f would converge to the sum of δ-spikes at x_i.

"Most smooth" solution $\lambda \to \infty$. The MPL criterion is solved in the limit $\lambda \to \infty$ by making the penalty $\int h_l'^2(t)\ dt$ equal to zero, and hence, $h_l' \equiv 0$. In other words, the MPL method for $\lambda \to \infty$ consists of doing ordinary *maximum likelihood* within the class of functions satisfying $h_l'(x) \equiv 0\ \forall x \in [a, b]$. Therefore, h_l is constant, say h_0, and $l'(x) = cp_{\mathbf{w}}(x)$ (where the constant $c = (-1)^j e^{h_0}$). Consequently, l is a *polynomial* of degree $m + 1$. We consider the following situations:

- "Zeromodal" case: $m^* = m = 0$: We have $l'(x) \equiv \pm c$ and therefore $f(x) \propto e^{\pm c}$. For $[a, b] = [0, \infty]$, this is the well-known exponential distribution.

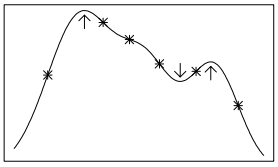
- Unimodal case: $m^* = m = 1$. If for the sign, we exclude the rare case $j = 0$, we have $l'(x) = -c(x - w_1)$, and $f(x) \propto \exp(-c(x - w_1)^2/2)$, i.e, for $[a, b] = \mathbb{R}$, f is the normal density with mean w_1 (and ML estimate $w_1 = \bar{x}_n$). Hence, the smoothest limit consists of parametrically estimating a normal density.

- General case: For simplicity, let's assume $j = 1$, i.e., minus sign in (5). We have $l'(x) = -c(x - w_1)(x - w_2) \cdots (x - w_m)$, and therefore $l(x) = l_0 - cS_{\mathbf{w}}(x)$, where $S_{\mathbf{w}}(x) := \int_0^x p_{\mathbf{w}}(t)\ dt = \frac{1}{m+1}x^{m+1} - \frac{1}{m}(\sum w_j)x^m + \frac{1}{m-1}(\sum_{i<j} w_i w_j)x^{m-1} - \ldots + (-1)^m \prod_j w_j \cdot x$. How is l_0 determined? Let $I(c, \mathbf{w}) := \int_a^b \exp(-cS_{\mathbf{w}}(t))dt$. Then, $1 = \int f(x)dx = e^{l_0} I(c, \mathbf{w})$, whence $l_0 = -\log I(c, \mathbf{w})$. Now, maximizing the likelihood means

$$\min_{c,\mathbf{w}} \quad n \log I(c, \mathbf{w}) + c \sum_{i=1}^{n} S_{\mathbf{w}}(x_i) \tag{13}$$

  This ML minimization problem is quite simpler than the general MPL problem (2).

Extension: Bumps and Dips. As in Good & Gaskins (1980), one may want to consider bumps and dips (maximal concave and convex regions) of the density. As there is at most one mode in a bump interval, we can effectively control the local extrema via inflection points.



With this approach, we can require even more "qualitative smoothness", and the MPL criterion approximately penalizes relative change of curvature of the log density l. By factoring $l''(x)$ instead of $l'(x)$, we get a very similar o.d.e. boundary value problem with an additional boundary condition ensuring $\int_a^b t \cdot f(t)\ dt = \frac{1}{n}\sum_{i=1}^n x_i$, i.e., an exact first moment property. The generalization of considering $l^{(\nu)}$ for $\nu \geq 1$ is straightforward.

For $\nu \geq 2$, the zeros of $f^{(\nu)}$ and $l^{(\nu)}$ no longer coincide which is a conceptual problem. For $\nu = 2$, e.g., $l'' = (f''f - f'^2)f^{-2}$, and we have at least $f'' \leq 0 \Longrightarrow l'' \leq 0$.

4 Semiparametric log-density modeling

The logarithm of the Gaussian density is quadratic, the simplest non-trivial concave parametric curve, and the exponential family being the most predominant class of densities, it seems natural to look at flexible modeling of the *log density* rather than the density (or its square root).

***Semi**parametric* modeling of the log density has been achieved independently by Loader (1993) and Hjort & Jones (1994) by so-called "local likelihood" approaches (modifying the likelihood by a localizing kernel). Both models have several nice properties and lie flexibly between parametric and nonparametric density estimation.

Semi-Parametric Mixture Models. With our "Wp" approach, a different kind semi-parametric density models can be envisaged. For instance, a mixture of two,

$$f(x) = pf_1(x) + (1-p)f_2(x), \tag{14}$$

where f_1 and f_2 are modeled to be "unibumpal", i.e., with only one "bump", and will be estimated by MPLE ($\nu = 2$, $m = 0$) as above. In the limit, for $\lambda_{1,2} \to \infty$, this approach comprises traditional parametric (Gaussian) mixture models.

References

Ascher, U. M., Mattheij, R. M. M. & Russell, R. D. (1988), *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall series in computational mathematics, Prentice Hall, Englewood Cliffs, New Jersey.

Bader, G. & Ascher, U. (1987), 'A new basis implementation for a mixed order boundary value ODE solver', *SIAM J. Sci. Statist. Comput.* **8**, 483–500. COLNEW.

Boneva, L. I., Kendall, D. G. & Stefanov, I. (1971), 'Spline transformation: Three new diagnostic aids for the statistical data-analyst (with discussion)', *Journal of the Royal Statistical Society B* **33**, 1–70.

Cox, D. D. (1985), 'A penalty method for nonparametric estimation of the logarithmic derivative of a density function', *Annals of the Inst. of Stat. Math.* **37**, 271–288.

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

Cox, D. R. (1966), 'Notes on the analysis of mixed frequency distributions', *J. Math. Statist. Psychol.* **19**, 39–47.

Cuevas, A. & Gonzalez-Mantiega, W. (1991), Data-driven smoothing based on convexity properties, *in* G. G. Roussas et al., eds, 'Nonparametric Functional Estimation and Related Topics', Kluwer Acad. Publ., Dordrecht, pp. 225–240.

De Montricher, G. F., Tapia, R. A. & Thompson, J. R. (1975), 'Nonparametric maximum likelihood estimation of probability densities by penalty function methods', *Annals of Statistics* **3**, 1329–1348.

Good, I. J. & Gaskins, R. A. (1971), 'Non-parametric roughness penalties for probability densities', *Biometrika* **58**, 255–277.

Good, I. J. & Gaskins, R. A. (1980), 'Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data', *JASA* **75**, 42–73. Comments & Rejoinder.

Gu, C. & Qiu, C. (1993), 'Smoothing spline density estimation: Theory', *Annals of Statistics* **21**(1), 217–234.

Härdle, W. (1990), *Smoothing Techniques With Implementation in S*, Springer (Berlin, New York).

Hartigan, J. A. & Hartigan, P. M. (1985), 'The dip test of unimodality', *Annals of Statistics* **13**, 70–84.

Hjort, N. L. & Jones, M. (1994), Locally parametric nonparametric density estimation, Stat. Res. Rep. 3, Dept. Math., Univ. Oslo.

Izenman, A. J. (1991), 'Recent developments in nonparametric density estimation', *JASA* **86**(413), 205–224.

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

Kooperberg, C. & Stone, C. J. (1991), 'A study of logspline density estimation', *Computat. Statist. Data Anal.* **12**, 327–347.

Kooperberg, C. & Stone, C. J. (1992), 'Logspline density estimation for censored data', *Journal of Computational and Graphical Statistics* **1**(4), 301–328.

Loader, C. R. (1993), Local likelihood density estimation, Tech. Rep. 91, AT&T Bell Labs.

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*a*), Estimating distributions with a fixed number of modes, *in* H. Rieder, ed., 'Robust Statistics, Data Analysis, and Computer Intensive Methods – Workshop in honor of Peter J. Huber, on his 60th birthday', Vol. 109 of *Lecture Notes in Statistics*, Springer, Berlin, pp. 267–276.

Mächler, M. B. (1995*b*), 'Variational solution of penalized likelihood problems and smooth curve estimation', *Annals of Statistics* **23**(5), 1496–1517.

Minnotte, M. C. & Scott, D. W. (1993), 'The mode tree: A tool for visualization of nonparametric density features', *Journal of Computational and Graphical Statistics* **2**(1), 51–68.

Ng, P. T. (1994), 'Smoothing spline score estimation', *SIAM J. Sci. Statist. Comput.* **15**, 1003–1025.

O'Sullivan, F. (1988), 'Fast computation of fully automated log-density and log-hazard estimators', *SIAM J. Sci. Statist. Comput.* **9**, 363–379.

Roussas, G., ed. (1991), *Nonparametric Functional Estimation and Related Topics*, Vol. 335 of *NATO ASI Series. Series C*, Kluwer Academic Publishers, Dordrecht.

Silverman, B. W. (1981), 'Using kernel density estimates to investigate multimodality', *Journal of the Royal Statistical Society B* **43**, 97–99.

Silverman, B. W. (1982), 'On the estimation of a probability density function by the maximum penalized likelihood method', *Annals of Statistics* **10**, 795–810.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Tapia, R. A. & Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, John Hopkins Univ. Press, Baltimore.

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

Wand, M. P. & Jones, M. C. (1995), *Kernel Smoothing*, Vol. 60 of *Monographs on statistics and applied probability*, Chapman & Hall, London.