Estimating Distributions with a Fixed Number of Modes

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

October 4, 1994
Id: pap1.tex,v 1.7 1994/10/03 16:00:50 maechler Exp

Abstract

A new approach for non- or semi-parametric density estimation allows to specify modes and antimodes. The new smoother is a *Maximum Penalized Likelihood* (MPL) estimate with a novel roughness penalty. It penalizes a relative *change* of curvature which allows considering modes and inflection points. For a given number of modes, the score function, $l' = (\log f)'$ 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 variational problem is equivalent to a differential equation with boundary conditions. The exponential and normal distributions are smoothest limits of the new estimator for zero and one mode, respectively.

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

1 Motivation and Introduction

The nonparametric estimation of distribution or density functions is well-established with a variety of methods available (see Silverman, 1986; Izenman, 1991). 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 and 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 and Hartigan, 1985, and Minnotte and Scott, 1993, for a nice graphical tool). Known examples of such data are (among others)

- Percentages of silica in chondrites (Good and Gaskins, 1980): 3(or more?) modes.
- Old Faithful geyser eruption lengths: 2 or 26 modes?
- Yearly snowfalls in Buffalo: Are there sub-populations (1 or 3 modes)?
- Thickness of Mexican stamps of 1872 (Izenman and Sommer, 1988): 2, 3, 5 or 7 modes?

The present approach is within the framework of *"maximum penalized likelihood estimation"*, MPLE (cf. Thompson and Tapia, 1990, ch. 3–5 which "contains" Tapia and Thompson, 1978), using a *novel roughness* penalty. 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, 1989; Mächler, 1993), 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].

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 f and F such that f has (at most) m^* modes, i.e. typically, $m = (2m^* - 1)_+$ local extrema. We assume that f is twice (continuously) differentiable, and define

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

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

$$l'(x) = p_{\mathbf{w}}(x)\,(-1)^j\, e^{h_l(x)}, \quad \text{where} \tag{2}$$

$$p_{\mathbf{w}}(x) := (x - w_1)(x - w_2)\cdots(x - w_m), \tag{3}$$

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 (2), determined by $j = 0$ or 1, is assumed to be given and is almost always negative ($j = 1$), e.g., whenever $b = \infty$ or if there are more modes than antimodes, but $j = 0$, e.g., for a U-shaped density f ($m = 1$). Note that from (2),

$$(\log l'(x))' = (l''/l')(x) = h_l'(x) + \sum_{j=1}^{m} \frac{1}{x - w_j}, \tag{4}$$

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.

In the remaining sections, I expose the MPLE problem and derive the characterizing Euler-Lagrange differential equation. The implementation is exemplified by a real data application. Properties of the solution and possible extensions are detailed.

2 MPLE — Roughness Penalty

To estimate f (together with F and h_l), we Maximize the Penalized Likelihood criterion, or equivalently,

$$\min_{l \in \mathcal{L}, \mathbf{w}} \left\{ -\sum_{i=1}^{n} l(x_i) + \lambda \int_a^b h_l'^2(t)\, dt \right\}, \tag{5}$$

where $\mathcal{L}$ is the set of all functions $l : [a,b] \to \mathbb{R}$ fulfilling (2), (3) and

$$\int_a^b f(t)\, dt = \int_a^b e^{l(t)}\, dt = 1. \tag{6}$$

[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).

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

$$h_l'(x) = \frac{\kappa_L'(x)}{\kappa_L(x)} - \sum_{j=1}^{m} \frac{1}{x - w_j} + \; l(x)l'(x)\,3(1+l^2)^{-1}(x), \tag{7}$$

by using (4) 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. (7) 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 (6) and making use of Dirac's δ distribution notation, the problem (5) is equivalent to the *variational problem*

$$\min_{l \in \mathcal{L}} \int_a^b \left\{ \alpha\, e^{l(t)} \; - \sum_{i=1}^{n} \delta(t - x_i) l(x_i) + \lambda h_l'^2(t) \right\} dt. \tag{8}$$

3 Differential Equation

Now we apply Theorem 1 of Mächler (1993) (and Mächler, 1989) 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:
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.

$$F_g \cdot \left(\frac{d}{dt}\right)^{2k} F = \frac{1}{2}(-1)^{\nu+k+1} S_f{}^{[\nu]}(t) \quad \text{for all } t \in [a, b], \tag{9}$$

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

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

For our problem (8), $\nu = k = 1$, 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 (5) is

$$h_l''(x) = \frac{l'(x)}{2\lambda} \;\; n\left(F_n(x) - F(x)\right), \tag{10}$$

with boundary conditions

$$F(a) = 0,\;\; F(b) = 1, \quad \text{and} \quad h_l'(a) = h_l'(b) = 0, \tag{11}$$

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 (10). 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 $\mathbf{w} = (w_1, w_2, \ldots, w_M)^{\mathsf{T}}$. I have used a collocation method o.d.e. solver, "COLNEW" from `netlib`, see Bader and Ascher, 1987, and, Ascher, Mattheij and Russell, 1988. Numerical boundary value o.d.e. solvers require an initial "crude" approximate solution to be provided. The *"log-splines"* of Kooperberg and 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. Further, the numerical solution requires the boundaries a and b to be finite. It is quite clear that a limit $a \to -\infty$, $b \to +\infty$ exists and is attained to arbitrary precision for finite a and b.

For the final estimate, the minimizer of (5), 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.

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). The 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.

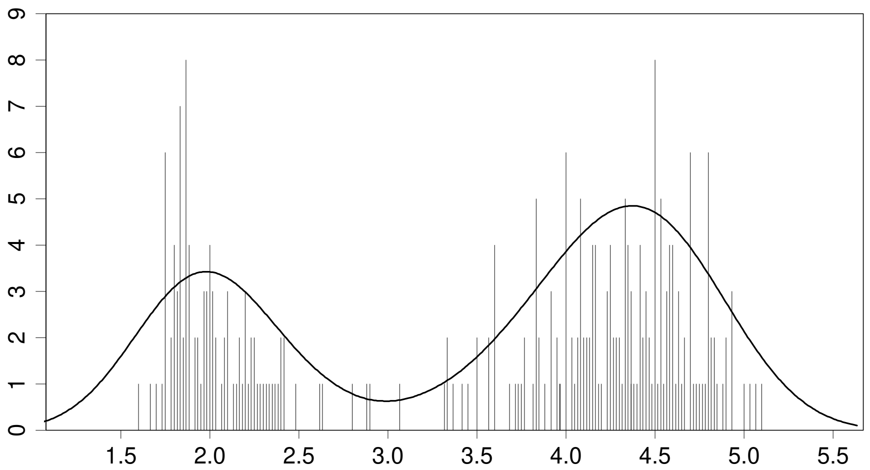


Figure 1: Old Faithful Geyser eruption lengths, $n = 272$; binned data and (Gaussian) kernel density estimate ($\times 10$) with $h = .3348$.

This is no surprise, since the asymptotic for h^* is geared towards unimodal densities (see Silverman, 1986, 3.4.2, p.45ff). For goodness of fit, we use distances of the distribution function from the empirical, $\|F - F_n\|$, where $F(x) = \int_{-\infty}^{x} f(t)\, dt$. For simplicity, we'll use the three scaled L^p norms $\|G\|_1 := r_x^{-1} \int_a^b |G(t)|\, dt$, $\|G\|_2 := \left(r_x^{-1} \int_a^b G(t)^2\, dt\right)^{1/2}$, and $\|G\|_\infty := \sup_t |G(t)|$, where $r_x :=$ range$(\{x_i\})$=max$_i(x_i) - min_i(x_i)$. Note that $\|F - F_n\|_\infty$ is the well known Kolmogorov-Smirnov distance. Further, we'll use the unit m ("milli"), i.e., thousandths, in order to have more manageable numbers. A further goodness of fit measure is the "average negative log-likelihood", i.e., $-\frac{1}{n}\sum_{i=1}^{n} l(x_i)$, denoted by ANLL. For the kernel estimate above, ANLL=1.10, and $(L^1, L^2, L^\infty)=$ (31, 33, 76) m. Another very smooth bimodal estimate is the standard Gaussian mixture model, $f(x) = \pi\sigma_1^{-1}\phi((x-\mu_1)/\sigma_1) + (1-\pi)\sigma_2^{-1}\phi((x-\mu_2)/\sigma_2)$, where ϕ is the standard normal density $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. For the geyser data, the maximum likelihood estimates for $(\pi,\ \mu_1, \mu_2,\ \sigma_1, \sigma_2)$ are $(0.3484,\ 2.019, 4.273,\ 0.2356, 0.4371)$. The goodness of fit values are ANLL=1.02, and $(L^1, L^2, L^\infty)=$ (14, 18, 49) m which are somewhat better than the above kernel but still not good enough.

More reasonable alternatives are the Gaussian kernel with $h = .1$ and the logspline estimate (with automatic knot choice) which are shown in Figure 2. The estimates of

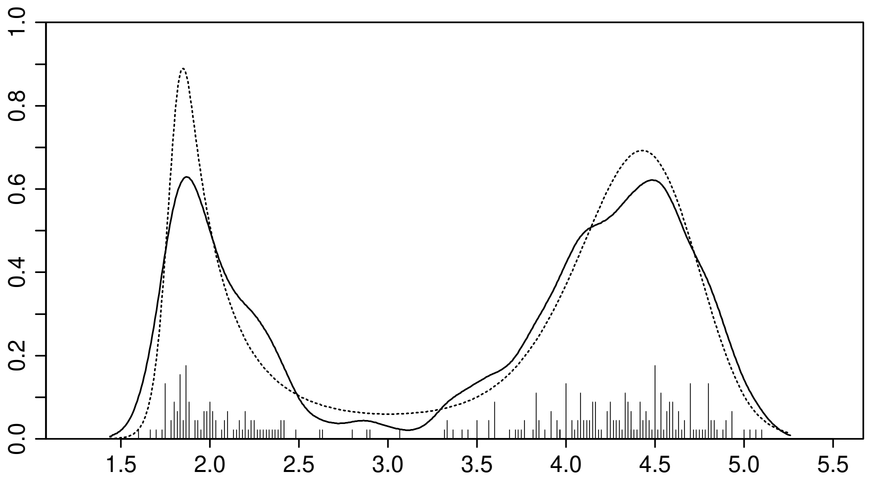


Figure 2: (Gaussian) kernel density with $h = .1$ and logspline (default $\rightsquigarrow$ 6 knots). The kernel density has a third mode which is true even for h as large as almost .13.

our new approach are much better however. Figure 3 shows $\hat{f}(x)$ for $\lambda = 2^{-2}, 2^{-4}, 2^{-6}$. The images and the goodness-of-fit quantities in Table 1 show that the new densities fit

Estimate	ANLL	L^1 [m]	L^2 [m]	L^∞ [m]
kernel $h = .335$	1.100	30.6	33.3	76.1
2 normal mixture	1.016	14.3	18.0	48.6
logspline (6 kn.)	0.957	9.1	11.1	31.9
kernel $h = .1$	0.946	6.3	8.2	29.8
$\lambda = 1/4$	0.948	5.8	8.0	39.3
$\lambda = 2^{-4}$	0.939	5.0	6.7	30.4
$\lambda = 2^{-6}$	0.929	4.7	6.1	22.2

Table 1: Goodness-of-fit values of density/distribution estimates for the geyser data.

(much) better to the data while being more smooth than the traditional estimates with comparable amount of fit.

More extensive comparisons and simulations will be done in subsequent research, but current experience shows that the new class of density/distribution estimates provides very smooth curves while adapting to the data very well.

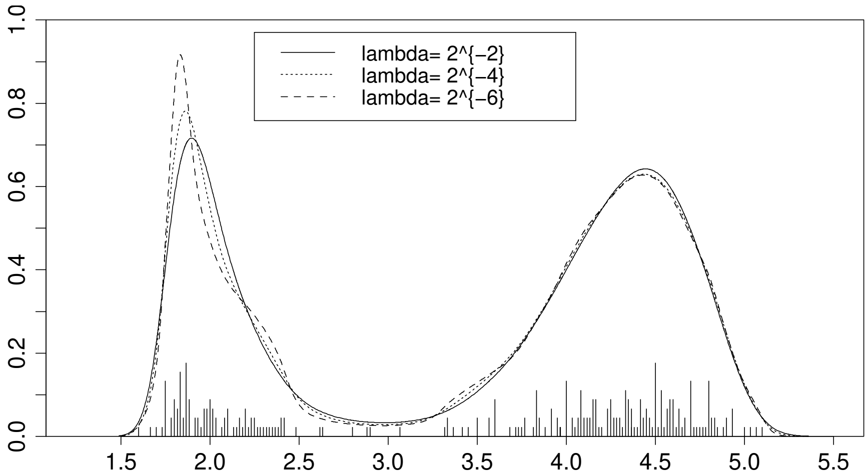


Figure 3: Estimated densities $\hat{f}(x)$ according to the new approach, (2), (3), (5), with two modes ($m = 3$), for $\lambda = 2^{-2}, 2^{-4}, 2^{-6}$ and $[a, b] = \mathbb{R}$. Even the smoothest one ($\lambda = 1/4$) fits better to the data than competitors (s. Table 1).

4 Properties of Solution

By this new approach, we get the distribution function with many derivatives, simultaneously, since F,f, l' = 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 $l'(x) = c p_{\mathbf{w}}(x)$ (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.

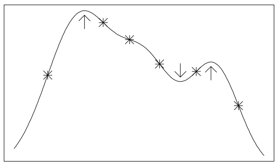
- General case: For simplicity, let's assume $j = 1$, i.e., minus sign in (2). We have $l'(x) = -c(x - w_1)(x - w_2) \cdots (x - w_m)$, and therefore $l(x) = l_0 - cS_{\mathbf{w}}(x)$, where $S_{\mathbf{w}}(x) := \int_0^x p_{\mathbf{w}}(t)\ dt = \frac{1}{m+1}x^{m+1} - \frac{1}{m}(\sum w_j)x^m + \frac{1}{m-1}(\sum_{i<j} w_i w_j)x^{m-1} - \ldots + (-1)^m \prod_j w_j \cdot x$. 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

$$\min_{c,\mathbf{w}} \quad n \log I(c,\mathbf{w}) + c \sum_{i=1}^{n} S_{\mathbf{w}}(x_i) \tag{12}$$

This ML minimization problem is quite simpler than the general MPL problem (5).

5 Extensions



Bumps and Dips. As in Good and 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.

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

Semi-Parametric Mixture Models. Semi-parametric density models can be envisaged. For instance, a mixture of two,

$$f(x) = pf_1(x) + (1-p)f_2(x), \tag{13}$$

where f_1 and f_2 are required to be unimodal or even "unibumpal" (with only one bump) and will be estimated by MPLE ($\nu = 2$) as above. In the limit, for $\lambda_{1,2} \to \infty$, this approach would comprise traditional parametric (Gaussian) mixture models.

References

Ascher, U. M., Mattheij, R. M. M. and 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. ETH Z: 748 256

Bader, G. and Ascher, U. (1987). A new basis implementation for a mixed order boundary value ODE solver, *SIAM Journal on Scientific and Statistical Computing* **8**: 483–500. — *COLNEW*

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

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

Good, I. J. and 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*

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

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

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

Izenman, A. J. and Sommer, C. J. (1988). Philatelic mixtures and multimodal densities, *JASA* **83**: 941–953.

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

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

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

Minnotte, M. C. and 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.

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

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

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