Computational Statistics & Data Analysis 12 (1991) 327–347
North-Holland

A study of logspline density estimation

Charles Kooperberg and Charles J. Stone
University of California at Berkeley, Berkeley, CA 94720, USA

Received March 1990
Revised May 1990

Abstract: A method of estimating an unknown density function f based on sample data is studied. Our approach is to use maximum likelihood etimation to estimate $\log(f)$ by a function s from a space of cubic splines that have a finite number of prespecified knots and are linear in the tails. The knots are placed at selected order statistics of the sample data. The number of knots can be determined either by a simple rule or by minimizing a variant of *AIC*. Examples using both simulated and real data show that the method works well both in obtaining smooth estimates and in picking up small details. The method is fully automatic and can easily be extended to yield estimates and confidence bounds for quantiles.

Keywords: Density estimation, Exponential family, Splines, Stepwise knot deletion, AIC, Transformations.

1. Introduction

Consider data that can be thought of as arising as a random sample from a distribution having an unknown density. It is common practice to summarize the data with such statistics as the sample mean, sample standard deviation and sample quartiles. Unless the form of the density is known to be (say) normal or exponential, however, it is also very helpful to examine graphical representations of the data such as density estimates.

In a mathematical treatment of density estimation, it is convenient to use integrated squared error $\int(\hat{f}-f)^2$ as a measure of inaccuracy. But this measure does not reliably reflect qualitative fidelity. It is considerably more important that a density estimate correctly inform us about some key aspects of the underlying density – how many modes are there? how does the density behave in the tails? – Consider, for example, the bimodal density function (solid line) in fig. 1. Most people would prefer the bimodal density estimate to the unimodal estimate even though the integrated squared error of the bimodal estimate is twice that of the unimodal one.

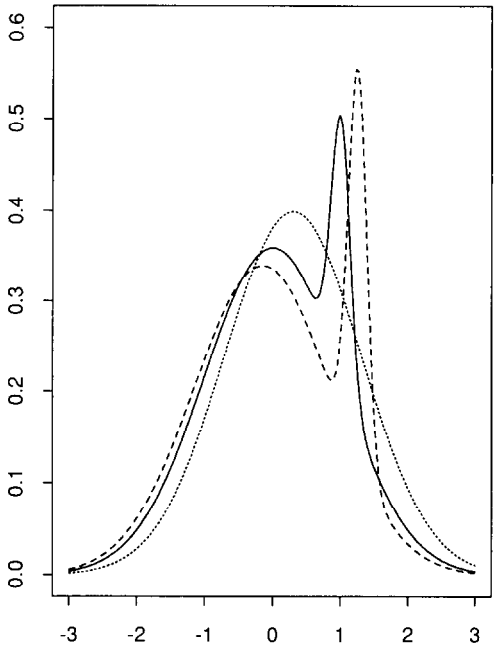

Fig. 1. Hypothetical true density (solid) and two estimates (dashed and dotted). We prefer the dashed estimate, even though the dotted estimate has half the mean squared error.

Consider the problem of estimating an unknown density function f based on sample data. One approach is to estimate $l = \log(f)$ by a function of the form $\hat{l} = \hat{s} + c(\hat{s})$, where $c(\hat{s})$ is the normalizing constant defined so that $\int \exp(\hat{l}) = 1$ and the maximum likelihood method is used to choose $\hat{s}$ from a suitably defined finite-dimensional linear space S of functions on $\mathbb{R}$. The corresponding density estimate $\hat{f} = \exp(\hat{l})$ is positive and integrates to one. This approach takes advantage of the desirable theoretical and numerical properties of maximum-likelihood estimation of the unknown parameters of an exponential family.

A well studied space S_0 is the space of the cubic splines having finitely many prespecified knots (de Boor [1]). (A cubic spline is a twice continuously differentiable, piecewise cubic polynomial and a knot is a location for a possible jump in the third derivative of the function.) The resulting models are referred to as logspline models. The mathematical theory of statistical inference based in such models, contained in Stone [11], is a blend of parametric inference and nonparametric inference that is referred to as functional inference.

In order to avoid the introduction of spurious details in the tails of the density estimate, we have restricted $\hat{s}$ to the subspace of S_0 of cubic splines that are linear to the left of the first knot and to the right of the last knot. The corresponding logspline density estimate is exponential to the left of the first knot and to the right of the last knot. The knots themselves have been placed at selected order statistics of the sample data, the first knot being placed at the minimum value and the last knot being placed at the maximum value. Some preliminary work on

the practical aspects of fitting logspline models was described in Stone and Koo [12].

Here, several refinements to logspline density estimation are proposed. In section 2 we discuss maximum-likelihood estimation of the unknown parameters of the logspline model and in section 3 we discuss a preliminary transformation having two free parameters, which are chosen to improve the exponential fit to the tails of the data. The important issue of knot placement and knot selection is discussed in section 4. First we describe a knot placement rule that depends only on the number of knots selected. Then we describe a nonadaptive rule for determining the number of knots as a function of sample size, the goal being to yield unimodal density estimates about ninety percent of the time when the true density function is suitably unimodal. Finally, we discuss an alternative procedure in which the knots are chosen from a somewhat larger number of potential knots by minimizing a variant of *AIC*. The two procedures can be made fully automatic and to yield density estimates that are smooth and yet flexible enough to reveal interesting features of the unknown density function that may be present. This is demonstrated in the later sections by applying logspline density estimation to a variety of simulated and real sets of data.

The density estimates that have been most heavily studied so far are kernel density estimates. When applied to sample data $y_1, \ldots y_n$, they have the form

$$\hat{f}(y) = \frac{1}{n} \sum \frac{1}{w_i} K\left(\frac{y - y_i}{w_i}\right), \qquad y \in \mathbb{R};$$

the positive widths w_i, $1 \le i \le n$, may or may not vary with i. For an excellent and reasonably current discussion of kernel density estimation see Silverman [9]. The main issue is the construction of such estimates is the choice of the widths: if they are too small, distracting spurious features are introduced; if they are too large, important features may be lost. When the widths do not vary with i, it may be impossible to choose the common (band)width to be large enough to avoid the introduction of spurious features in the tails of the density, but small enough to show important features in the central portion. Wand, Marron and Ruppert [13] have proposed a remedy for this deficiency in which kernel density estimation with a fixed bandwidth is applied after a data-dependent transformation. The combined procedure is similar to kernel density estimation with a variable bandwidth. O'Sullivan [7] discusses logspline estimates from the perspective of penalized likelihood density estimation.

More details about logspine density estimation can be found in Kooperberg [6]. In particular it contains more information about preliminary transformations (see section 4) and about numerical aspects of computing logspline estimates.

2. Logspline models

Let K denote an integer with $K \ge 4$ and let $t_1, \ldots, t_K$ be a (simple) knot sequence in $\mathbb{R}$; that is, such that $-\infty < t_1 < \cdots < t_K < \infty$. Let S_0 denote the collection of twice continuously differentiable functions s on $\mathbb{R}$ such that the restriction of s

to each of the intervals $(-\infty, t_1]$, $[t_1, t_3], \ldots, [t_{K-1}, t_K]$, $[t_k, \infty)$ is a cubic polynomial. We refer to the functions in S_0 as cubic splines having (simple) knots at $t_1, \ldots, t_K$. Observe that S_0 is a $(K+4)$ dimensional linear space. Let S denote the K-dimensional subspace of S_0 consisting of functions $s \in S_0$ such that s is linear on $(-\infty, t_1]$ and on $[t_K, \infty)$. Set $p = K-1$. Then S has a basis of the form 1, $B_1, \ldots, B_p$. We can choose $B_1, \ldots, B_p$ such that B_1 is a linear function with negative slope on $(-\infty, t_1]$, $B_2, \ldots B_p$ are constant on $(-\infty, t_1]$, B_p is a linear function with positive slope on $[t_K, \infty)$, and $B_1, \ldots, B_{p-1}$ are constant on $[t_K, \infty)$.

Let Θ denote the collection of all column-vectors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t \in \mathbb{R}^p$ such that $\theta_1 < 0$ and $\theta_p < 0$. Given $\boldsymbol{\theta} \in \Theta$, set

$$c(\boldsymbol{\theta}) = \log\left(\int_{\mathbb{R}} \exp\big(\theta_1 B_1(y) + \cdots + \theta_p B_p(y)\big)\, dy\right) \quad \text{and}$$

$$f(y; \boldsymbol{\theta}) = \exp\big(\theta_1 B_1(y) + \cdots + \theta_p B_p(y) - c(\boldsymbol{\theta})\big), \qquad y \in \mathbb{R}.$$

Then $\int_{\mathbb{R}} f(y; \boldsymbol{\theta})\, dy = 1$. Let $F(\cdot; \boldsymbol{\theta})$, denote the distribution function corresponding to $f(\cdot; \boldsymbol{\theta})$. We refer to the identifiable p-parameter exponential family $f(\cdot; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, of positive twice differentiable density functions on $\mathbb{R}$ as a logspline family.

For $\boldsymbol{\theta} \in \Theta$ let $\boldsymbol{I}_1(\boldsymbol{\theta})$ denote the Hessian matrix of $c(\cdot)$ at $\boldsymbol{\theta}$, which is the $p \times p$ matrix having entry $\partial^2 c(\boldsymbol{\theta})/\partial\theta_j\, \partial\theta_k$ in row j and column k for $1 \leqslant j, k \leqslant p$. This matrix is positive definite, so $c(\cdot)$ is strictly concave on Θ.

Let Y be a random variable having a continuous and positive density function. Let $Y_1, \ldots, Y_n$ be independent random variables having the same distribution as Y.

The log-likelihood function corresponding to the logspline family, given by

$$l(\boldsymbol{\theta}) = \sum \log\big(f(Y_i; \boldsymbol{\theta})\big), \quad \boldsymbol{\theta} \in \Theta,$$

is strictly concave on Θ. The maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the log-likelihood function. Since the log-likelihood function is strictly concave, the maximum-likelihood estimate is unique if it exists.

Suppose that $(-\infty, t_1]$ and $[t_K, \infty)$ each contain one or more observed values of $Y_1, \ldots, Y_n$ and the intervals $[t_1, t_2], \ldots, [t_{K-1}, t_K]$ each contain four or more of these observed values. Then the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ exists. We refer to $\hat{f} = f(\cdot; \hat{\boldsymbol{\theta}})$ as the logspline density estimate.

3. Maximum likelihood estimation of $\boldsymbol{\theta}$

Let $\boldsymbol{H}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ denote the Hessian of $c(\boldsymbol{\theta})$, the $p \times p$ matrix whose (j, k) element is

$$\frac{\partial^2 c(\boldsymbol{\theta})}{\partial\theta_j\, \partial\theta_k} = -\int_{\mathbb{R}} B_j(y) B_k(y) f(y; \boldsymbol{\theta})\, dy + \int_{\mathbb{R}} B_k(y) f(y; \boldsymbol{\theta})\, dy \int_{\mathbb{R}} B_j(y) f(y; \boldsymbol{\theta})\, dy.$$

Let $Y_1, \ldots Y_n$ be a random sample of size n from f and let $S(\boldsymbol{\theta})$ be the score function; that is, the p-dimensional vector of elements

$$\frac{\partial l}{\partial \theta_j}(\boldsymbol{\theta}) - b_j - n\frac{\partial c}{\partial \theta_j}(\boldsymbol{\theta}),$$

where the sufficient statistics $b_1, \ldots b_j$ are defined by

$$b_j = \sum B_j(Y_i).$$

The maximum likelihood equation for $\hat{\boldsymbol{\theta}}$ is $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. Let $\boldsymbol{I}(\boldsymbol{\theta}) = -n\boldsymbol{H}(\boldsymbol{\theta})$ denote the information matrix corresponding to the random sample. The Newton–Raphson method for computing $\hat{\boldsymbol{\theta}}$ is to start with an initial guess $\hat{\boldsymbol{\theta}}^{(0)}$ and iteratively determine $\hat{\boldsymbol{\theta}}^{(m)}$ from the formula

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} = \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}^{(m)})S(\hat{\boldsymbol{\theta}}^{(m)}).$$

It at some stage $l(\hat{\boldsymbol{\theta}}^{(m+1)}) \leq l(\hat{\boldsymbol{\theta}}^{(m)})$, then $\hat{\boldsymbol{\theta}}^{(m+1)}$ should be replaced by $\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} + \alpha \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}^{(m)})S(\hat{\boldsymbol{\theta}}^{(m)})$, for some constant $\alpha \in (0, 1)$. In our implementation we choose $\alpha = 2^{-k}$, where k is the smallest positive integer such that $l(\hat{\boldsymbol{\theta}}^{(m)} + 2^{-k}\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}^{(m)}) - S(\hat{\boldsymbol{\theta}}^{(m)})) > l(\hat{\boldsymbol{\theta}}^{(m)})$. When $l(\hat{\boldsymbol{\theta}}^{(m)})$ is close to $\hat{\boldsymbol{\theta}}$, the method tends to converge very fast. We found the condition that $\epsilon < 10^{-6}$, where

$$\epsilon = \sum \frac{\left(\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}^{(m)})S(\hat{\boldsymbol{\theta}}^{(m)})\right)_i^2}{\max\left(\left(\hat{\theta}_i^{(m)}\right)^2, 0.00001\right)},$$

yields a workable convergence criterion. In practice, however, the algorithm has to be modified slightly because of the requirements that $\hat{\theta}_1 < 0$ and $\hat{\theta}_p < 0$ (Kooperberg [6]). As starting values we use $\hat{\theta}_j^{(0)} = 0$ for $j = 2, \ldots p-2$, while we make a rough preliminary estimate of the rate at which $\hat{f}(y)$ tends to 0 as $y \to \pm\infty$ to obtain initial guesses for $\hat{\theta}_1^{(0)}$, $\hat{\theta}_{p-1}^{(0)}$, and $\hat{\theta}_p^{(0)}$. See Kooperberg [6] for more details.

Although it is straightforward to give confidence bounds for the logspline density estimates based on the usual asymptotic formula for the standard error of a maximum likelihood estimate, we have not done so since we have found confidence intervals in density estimation to be of limited value. In particular confidence intervals for densities give little information about the shape of the density: even when the density estimate is unimodal, the corresponding confidence bounds will contain bimodal estimates; the reverse will also often be the case.

4. Preliminary transformation

The tails of the logspline model beyond the extreme knots are exactly exponential and the tails are almost exponential close to the extreme knots. The accuracy of exponential approximation to the tails of f can be improved by means of a preliminary transformation.

In the following we assume that Y is a positive random variable that can take on arbitrary small values larger than zero. If Y has different properties, for example $Y > a$ for some $a \neq 0$, $Y \in [a, b]$, or $Y \in \mathbb{R}$ without any further restrictions, no transformation or a transformation different from the one discussed in this section might be more appropriate. Transformations are further discussed in Kooperberg [6].

First, in order to improve the accuracy of the exponential approximation to the upper tail of f, we consider choosing the positive power parameter β so that the conditional distribution of $Y^\beta - a^\beta$ given that $Y \geq a$ is approximately exponential. In practice β must be determined from the sample $Y_1, \ldots, Y_n$. We denote the corresponding fitted value by $\hat{\beta}$. We choose a to be the median $y_{.5}$. Although one could argue that the location of the last mode would be a more natural choice for a, we found that this gives no noticeable improvement, while it would require estimation of a. The largest observations will be the most influential in estimating β regardless of a.

Let $Y_{(1)}, \ldots, Y_{(n)}$ denote the increasing order statistics corresponding to $Y_1, \ldots, Y_n$. Let $m = [(n+1)/2]$ denote the greatest integer in $(n+1)/2$, so that $m \leq (n+1)/2 < m+1$. A reasonable way to choose $\hat{\beta}$ is by maximum-likelihood based on the two-parameter exponential family with the dependence of $\hat{y}_{.5} = Y_{(m)}$ on the sample data being ignored. We are led to choosing $\hat{\beta}$ based on the data $Y_{(m)}, \ldots Y_{(n)}$ to maximize the function

$$g(\beta) = (\beta - 1)\sum \log\left(Y_{(i)}\right) + (n-m)\left[\log(\beta) + \log\left(\sum\left(Y_{(i)}^\beta - Y_{(m)}^\beta\right)\right) + \log(n-m) - 1\right].$$

Observe that

$$g'(\beta) = \frac{(n-m)}{\beta}\left[1 - \frac{\sum\left(Y_{(i)}^\beta \log\left(Y_{(i)}^\beta\right) - Y_{(m)}^\beta \log\left(Y_{(m)}^\beta\right)\right)}{\sum\left(Y_{(i)}^\beta - Y_{(m)}^\beta\right)}\right] + \sum \log\left(Y_{(i)}\right).$$

Under the assumption that $0 < Y_{(m)} < Y_{(n)}$, it is straightforward to show that $g''(\beta) < 0$ for $\beta > 0$ and hence that $g'(\beta)$ is a strictly decreasing function of β. As $\beta \downarrow 0$ $g'(\beta)$ has limit $(1 - \hat{A}/2)\Sigma(Y_{(i)}/Y_{(m)})$, where

$$\hat{A} = \frac{\frac{1}{n-m}\sum \log^2\left(Y_{(i)}/Y_{(m)}\right)}{\left[\frac{1}{n-m}\sum \log\left(Y_{(i)}/Y_{(m)}\right)\right]^2};$$

as $\beta \to \infty$, $g(\beta)$ has limit $\Sigma\log(Y_{(i)}/Y_{(n)})$.

Suppose that $0 < Y_{(m)} \leq Y_{(m+1)} < Y_{(n)}$. Then $g(\beta)$ is a strictly decreasing function of β if $\hat{A} \geq 2$ and $g(\beta)$ has a unique maximum $\hat{\beta}$ if $\hat{A} < 2$. When $\hat{A} < 2$ the numerical value of $\hat{\beta}$ is easily found by solving the equation $g'(\beta) = 0$ in an iterative manner. We refer to $\hat{\beta}$ as the maximum-likelihood estimate of the power parameter. (When $\hat{A} \geq 2$, it is reasonable to consider the logarithmic transformation: $W = \log(Y)$ and $W_i = \log(Y_i)$, $1 \leq i \leq n$.) Using maximum-likelihood to

determine a preliminary power transformation was suggested in part by Box and Cox [2]. More information about the maximum likelihood estimation of a power parameter can be found in Kooperberg [6].

Next, to improve the accuracy of the exponential approximation to the lower tail of f, we consider the problem of choosing the power parameter β so that the conditional distribution of $\log^{\beta}(1 + y_{.5}/Y) - \log^{\beta}(2)$, given that $Y \le y_{.5}$, is approximately exponential. We can obtain a fitted value $\hat{\beta}$ to β by applying the maximum-likelihood estimate, as just described, to $\log(1 + Y_{(m)}/Y_i)$, $1 \le i \le n$.

Let β_1, $\beta_2 > 0$. Consider the twice continuously differentiable transformation $T(\cdot)$ on $(0, \infty)$ defined by

$$T(y) = \left(v - \log^{\beta_2}(1 + 1/v)\right) \Big|_{v=(y/y_{.5})^{\beta_1}}.$$

Observe that $T(\cdot) > 0$ is strictly increasing on $(0, \infty)$.

Let $\hat{\beta}_1$ be the maximum-likelihood estimate of the power parameter based on $Y_1, \ldots, Y_n$ and let $\hat{\beta}_2$ be the estimate of the power parameter based on $\log[1 + (Y_{(m)}/Y_i)^{\beta_1}]$, $1 \le i \le n$. Let $\hat{T}(\cdot)$ be defined on $(0, \infty)$ by

$$\hat{T}(y) = \left(v - \log^{\hat{\beta}_2}(1 + 1/v)\right) \Big|_{v=(y/y_{.5})^{\hat{\beta}_1}}.$$

We refer to $\hat{T}(\cdot)$ as the preliminary transformation.

let $\hat{f}_{W,K}$ denote the logspline density estimate obtained from the random sample $W_i = \hat{T}(Y_i)$, $1 \le i \le n$ using K knots. Let $\hat{f}_K$ denote the density estimate given by

$$\hat{f}_K(y) = \hat{T}'(y)\hat{f}_{W,K}(\hat{T}(y)), \quad y > 0.$$

We refer to $\hat{f}_K$ as the logspline density estimate based on $Y_1, \ldots, Y_n$, K knots and the preliminary transformation.

Let a denote the constant negative slope of B_1 to the left of the minimum knot. Then $a\hat{\theta}_1$ is the constant slope of $\log(\hat{f}_{W,K})$ to the left of this knot. The limiting behavior of $\hat{f}_K(y)$ as $y \downarrow 0$ is as follows.

$$\lim_{y \downarrow 0} \hat{f}_K(y) = \begin{cases} 0 & \hat{\beta}_2 < 1, \\ 0 & \hat{\beta}_2 = 1 \text{ and } a\hat{\beta}_1\hat{\theta}_1 > 1, \\ C \in (0, \infty) & \hat{\beta}_2 = 1 \text{ and } a\hat{\beta}_1, \hat{\theta}_1 = 1, \\ \infty & \hat{\beta}_2 = 1 \text{ and } a\hat{\beta}_1\hat{\theta}_1 < 1, \\ \infty & \hat{\beta}_2 > 1. \end{cases} \tag{1}$$

It is easily seen that if $\hat{f}_K(y)$ has a finite positive limit as $y \downarrow 0$ (that is if $\hat{\beta}_2 = 1$ and $a\hat{\beta}_1\hat{\theta}_1 = 1$), then $\hat{f}_K'(y) \to 0$ as $y \downarrow 0$.

5. Knots

Choosing the knots is the main problem in fitting logspline models; it is comparable to choosing a kernel-width in kernel density estimation. In choosing

the knots there are two main issues:
- How many knots should be used?
- Where should the knots be placed?

Although these issues are clearly related, they will be treated independently in the following sections.

5.1. Knot placement

The following features have arisen out of experience in fitting logspline models:
- there should be a simple and automatic knot selection rule;
- knots should be placed at or near selected order statistics;
- the corresponding indices should be approximately symmetrically distributed about $(n+1)/2$;
- there should be knots at the first and last order statistics;
- the pattern of extreme knots should be approximately independent of sample size;
- the middle knots should be approximately at equally spaced indices.

Some of these features appeared to be particularly desirable when we were interested in using logspline estimates for obtaining confidence bounds for extreme quantiles. In particular it appeared necessary to have a second large knot relatively close to the largest knot to obtain confidence bounds for extreme quantiles that both give reasonable coverage and are not extremely large (see also Breiman, Stone and Kooperberg [3]).

Let K denote the number of knots. The knot placement will be determined by a sequence of numbers $r_1, \ldots, r_K$ such that $1 \le r_1 < \cdots < r_K \le n$. Let $1 \le k \le K$ and let m denote the greatest integer in r_k, so that $m \le r_k < m+1$. Then the kth knot will be placed at

$$(m+1-r_k)\,y_{(m)} + (r_k - m)\,y_{(m+1)}.$$

(In particular if $r_k = m$, the kth knot is placed at $y_{(m)}$.)

Our symmetry condition is that

$$r_k + r_{K+1-k} = n+1, \qquad 1 \le k \le K,$$

which implies that

$$r_{k+1} - r_k = r_{K+1-k} - r_{K-k}, \qquad 1 \le k \le K.$$

In order that there be knots at the first and the last order statistics, we choose $r_1 = 1$ and $r_K = n$.

Set $g_k = r_{k+1} - r_k$ for $1 \le k \le K-1$. In order to satisfy the remaining features, we ended up by requiring that

$$g_k = 4 \cdot [(4-\epsilon) \vee 1] \cdot \ldots \cdot [(4-(k-1)\epsilon) \vee 1], \qquad 1 \le k \le K/2,$$

where $\epsilon \in \mathbb{R}$; here $a \vee b = \max(a, b)$. The constant ϵ is determined as follows: if K is an odd integer, then $2r_{(K+1)/2} = n+1$; if K is an even integer, then $r_{K/2} + r_{K/2+1} = n+1$.

We will now give two examples of our knot placement rule, in which r_k has been rounded off to the nearest integer.

Example 1. $n = 150$, $K = 7$ and $\epsilon \doteq .1881$. The rounded-off values of r_k are as follows:

k	1	2	3	4	5	6	7
r_k	1	5	20	75	131	146	150

Example 2. $n = 500$, $K = 10$ and $\epsilon \doteq .5300$. The rounded-off values of r_k are as follows:

k	1	2	3	4	5	6	7	8	9	10
r_k	1	5	19	60	158	343	441	482	496	500

The knot-placement rule described in this section will be employed from now on.

5.2. *How many knots?*

Choosing the number of knots is like selecting a bandwidth: too many knots leads to a noisy estimate; too few knots gives an estimate that is overly smoothed and thereby missing essential details. Our point of view in making this choice is that the most desirable quality of a density estimate is that it correctly inform us about key aspects of the true density, such as the number of modes.

Define a density f on $(0, \infty)$ to be (at least) bimodal if there exist positive numbers y_1, y_2, y_3 in strictly increasing order such that $f(y_2) < f(y_1)$ and $f(y_2) < f(y_3)$. Otherwise f is unimodal. For numerical convenience, we classify a logspline density estimate $\hat{f}_K$ on $(0, \infty)$ to be (at least) bimodal if there exist positive numbers y_1, y_2, y_3 in strictly increasing order such that $\hat{f}_K(y_2) < 0.99\hat{f}_K(y_1)$ and $\hat{f}_K(y_2) < 0.99\hat{f}_K(y_3)$ and as unimodal otherwise. (Due to the nature of logspline density estimation, this classification is rarely effected by changing .99 to some other value slightly less than 1.)

Let $\hat{f}_K$ denote the logspline density estimate based on $Y_1, \ldots, Y_n$, K knots and the preliminary transformation. Let $p_{f,K,n}$ denote the probability that $\hat{f}_K$ is unimodal when f is the common density of $Y_1, \ldots, Y_n$. Presumably, $p_{f,K,n}$ is a decreasing function of K. If f is a unimodal density, we would like $p_{f,K,n}$ to be close to one. If f is not a unimodal density, we would like $p_{f,K,n}$ to be close to zero. Given a unimodal f and n, let $K_{f,n}$ be the maximum value of K such that $p_{f,K,n} \geq .9$. The number of $K_{f,n}$ depends mildly on the choice of f. To obtain a specific number of knots for each given sample size, we let f be the gamma(5) density defined as the density of $\Sigma_1^5 X_i$ where X_i are 5 independent random variables having the same exponential distribution. We write $K_{f,n}$ for this choice of f as K_n and refer to the dependence of K_n on n as the knot-number rule.

Explicit (approximate) values of K_n can be obtained by Monte Carlo simulation and numerical computation of $\hat{f}_K$. We have carried out simulations of size

1000, and rounded the results to nice numbers:

n	60–124	125–199	200–349	350–499	500–650
K_n	6	7	8	9	10

The values of K_n in the above table give a reasonable number of knots to be used in logspline density estimation when not enough computing power is available to use knot deletion as described in the next subsection. We have tested the table on a wide range of gamma and lognormal distributions and found that the probability of bimodality rarely rose above .2 or fell below .05. When the underlying density function is believed to be particular irregular (smooth), however, it would be appropriate to use more (less) than K_n knots.

5.3. Stepwise knot deletion

Instead of using a fixed number of knots, determined by the knot-number rule described in the previous section, we can start out with a larger number of knots and then remove those knots that appear to be inessential for the given data.

For $1 \le j \le p$, let B_j be written in terms of the truncated power basis as

$$B_j(y) = \lambda_j + \lambda_{j0} y + \sum_k \lambda_{jk} (y - t_k)_+^3. \qquad \text{Then}$$

$$\sum_j \theta_j B_j(y) = \phi + \phi_0 y + \sum_k \phi_k (y - t_k)_+^3, \qquad \text{where}$$

$$\phi_k = \sum_j \theta_j \lambda_{jk} = \lambda_k' \boldsymbol{\theta} \quad \text{with} \quad \lambda_k = (\lambda_{1k}, \ldots, \lambda_{pk})$$

for $1 \le k \le K$. Correspondingly, $\phi_k = \lambda_k' \hat{\boldsymbol{\theta}}$ and

$$SE(\hat{\phi}_k) = \sqrt{\lambda_k' (\boldsymbol{I}(\hat{\boldsymbol{\theta}}))^{-1} \lambda_k}$$

for $1 \le k \le K$.

Consider t_1 and t_K as being permanent knots and t_k, $2 \le k \le K-1$, as being nonpermanent initial knots that may be deleted and consider stepwise knot deletion among the non-permanent initial knots. At any step we delete that knot having the smallest value of $|\hat{\phi}_k|/SE(\hat{\phi}_k)$. In this matter, we arrive at a sequence of models indexed by m, which ranges from 0 to $p-3$; the mth model has $p-m$ free parameters. Let $\hat{l}_m$ denote the loglikelihood function for the mth model evaluated at the maximum-likelihood estimate for that model. Let $AIC_{\alpha,m} = -2\hat{l}_m + \alpha(p-m)$ be the Akaike Information Criterion with parameter penalty α for the mth model. We choose the model corresponding to that value $\hat{m}$ of m that minimizes $AIC_{3,m}$. This model has $K - \hat{m}$ knots and $p - \hat{m}$ free parameters. We choose $\alpha = 3$ since, for this value of α, the probability is about .1 that $\hat{f}$ is bimodal when f is a gamma(5) density and it is also about .1 when f is the lognormal density defined as the density of $\exp(Z/2)$, where Z has the standard normal distribution. (The more traditional value of $\alpha = 2$ leads to spurious modes in the estimate with high probability, presumably because $\alpha = 2$ corresponds to

minimizing mean squared error, which is a dubious criterion in the context of density estimation (fig. 1). Schwarz [8] recommended using $\alpha = \log(n)$ in a context in which one of the fitted models is exactly valid. So far, there is no theoretical justification for choosing $\alpha = 3$ or, more generally, for choosing α independently of n.)

The idea of using stepwise knot deletion in the context of nonparametric regression is due to Smith [10].

6. Simulated examples

6.1. Unimodal examples

Figures 2 and 3 contain examples involving exponential, half-normal, gamma(5) and log-normal distributions. The random variable $|Y|$ has a half-normal distribution if Y has a normal distribution; $Y = \Sigma_1^5 X_i$ has a gamma(5) distribution if X_i, $1 \leq i \leq 5$ are independent random variables having the same exponential distribution; and $Y = \exp(Z/2)$ has a lognormal distribution if Z has a standard normal distribution. From here on we refer to these densities as the exponential, half-normal, gamma and lognormal densities respectively. It should be noted that the logspline method is scale invariant.

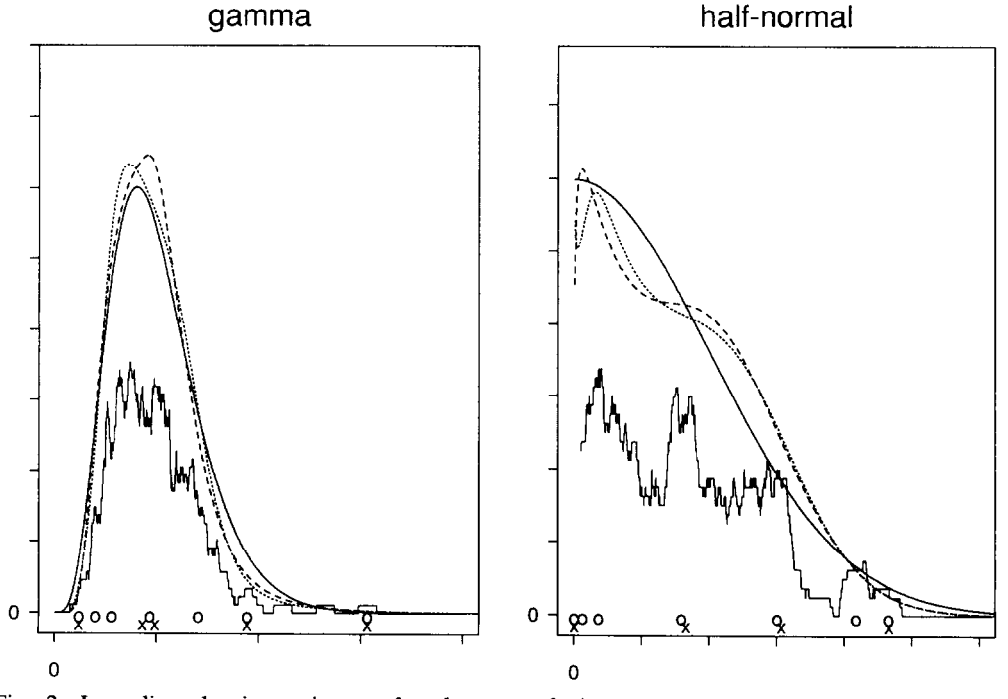

Fig. 2. Logspline density estimates for datasets of size $n = 150$ generated from gamma and half-normal densities – Solid = real, dashed = knot deletion, dotted = fixed knots.

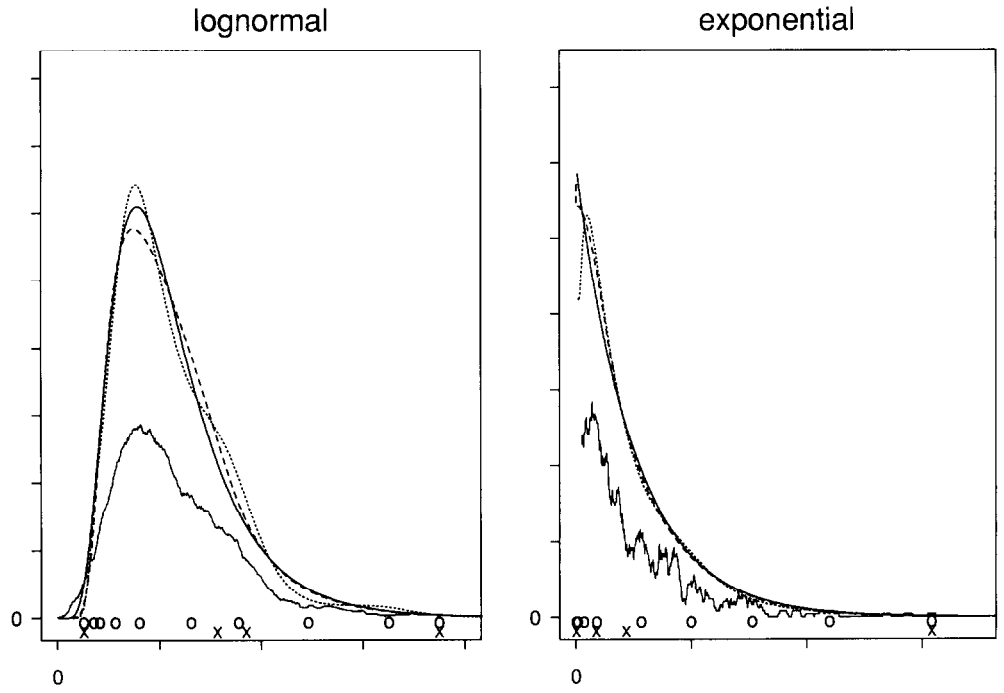

Fig. 3. Logspline density estimates for datasets of size $n = 500$ generated from lognormal and exponential densities – Solid = real, dashed = knot deletion, dotted = fixed knots.

We report here results for density estimates based on samples of size $n = 150$ and $n = 500$. The same samples were used for both the procedure with a fixed number of knots and the one with stepwise knot deletion. For the logspline density estimate with a fixed number of knots we used 7 knots for $n = 150$ and 12 knots for $n = 500$ in accordance with the knot-number rule of section 5.2. For the logspline density estimate with stepwise knot deletion we started with 12 knots for $n = 150$ and 15 knots for $n = 500$. The starting number of knots for logspline density estimation with stepwise knot deletion has relatively little influence on the final estimate, but it does have a strong influence on the amount of computing time. In particular if we choose an exceptionally large starting number of knots relative to the sample size, the computing time skyrockets since the log-likelihood function is too flat for rapid convergence.

For both the logspline density estimate with a fixed number of knots and the one with stepwise knot-deletion, we carried out the preliminary transformation and placed the knots according to the knot-placement rule of section 5.1. The placement of the knots in the final estimate (after the knot deletion for the stepwise knot deletion estimate) is indicated in the figure: "o" indicates a knot for the procedure with a fixed number of knots; "x" indicates a remaining knot for the stepwise knot deletion procedure. The solid line indicates the true density from which we generated the data, the dashed line is the estimate with stepwise knot deletion, and the dotted line is the estimate with a fixed number of knots.

In the figures we also provide a kernel density estimate. This estimate is based on a very small rectangular window. It is not included as a competitor of the logspline estimates, but as a descriptor of the raw data since, particularly for $n = 150$, the differences between the true density and the logspline density estimator are due almost entirely to sampling variation that would effect any estimate. It is scaled down to one half of its original height to prevent its graph from interfering with the other graphs. Although we only present a few (randomly selected) simulation examples, we feel that the results are fairly typical of the much larger number that we have examined.

The following numbers of knots remain in the logspline density estimates with stepwise knot deletion:

	$n = 150$		$n = 500$
gamma(5)	5	lognormal	4
half-normal	4	exponential	4

The two estimators have similar behavior. Especially for the gamma and the lognormal densities, it is clear that the amount of smoothing near the mode of the density and the amount of smoothing in the tail seem to be correct. The tails are smooth, while the mode has, approximately, the correct width and height. The half-normal density for $n = 150$ is a case where the (apparently) bad fit is due more to the sampling than to the form of the estimator. From the rectangular window estimator below the other estimators, it would (if one would ignore the few observations larger than the 4th tick mark) almost seem more believable that these data were generated from a uniform distribution than from a half-normal distribution. It is also interesting to note that the number of knots in the final fit for the procedure with knot deletion is about the same for $n = 150$ and $n = 500$ and that the procedure with knot deletion always ends up with less knots than the procedure with a fixed number of knots (which is not suprising, since the remaining knots are placed more efficiently).

Consider a density f (such as exponential or half-normal) that has a finite positive limit at zero. According to (1) the estimate $\hat{f}_K$ has a positive limit at zero if and only if $\hat{\beta}_2 = 1$ and $a\hat{\beta}_1\hat{\theta}_1 = 1$, which happens with probability zero. Thus with some probability p, $\hat{f}_K$ has limit zero at zero and, with complementary probability $1 - p$, $\hat{f}_K$ has limit infinity at zero. As a consequence, $\hat{f}_K$ has an unstable irregularity that, fortunately, is concentrated on a very small interval about 0 (and which is omitted from the plots corresponding to the half-normal density in fig. 2, the exponential density in fig. 3 and the suicide data in fig. 7). In many practical situations, the behavior of the density estimate very near the origin is not of interest. When it is of interest, however, and there is an irregularity near the origin, it probably can be reduced by recalculating the logspline density estimate under the constraints that $\hat{\beta}_2 = 1$ and $a\hat{\beta}_1\hat{\theta}_1 = 1$. This issue will be investigated in the future.

6.2. Bimodal examples

It is of particular interest to see how a density estimate behaves if the true density is of a more complex nature, for example a bimodal density. To investigate this we have generated samples from densities of the form $f(y; a_1, a_2, a_3) = (1 - a_1)g(y) + a_1 h(y; a_2, a_3)$, where $g(y)$ is the lognormal density we used before and $h(y; a_2, a_3)$ is the normal density with mean a_2 and standard deviation a_3. In figs. 4, 5 and 6 we present results for samples of size $n = 150$ and $n = 500$ where (a_1, a_2, a_3) is (.2, 2, .07), (.2, 2, .17) and (.2, 2, .27) respectively. The first of these densities is an example with a very sharp second mode, the second one has a moderately sharp second mode, while the third example only has a small second mode.

In the picture are logspline density estimates for samples sizes of $n = 150$ and $n = 500$. The computations and figures are organized in the same way as the examples in the previous section. The solid line is the true density, the dotted line is the logspline density estimate with a fixed number of knots and the dashed line is the logspline density estimate with stepwise knot-deletion.

The following number of knots remain in the logspline density esitmates with stepwise knot deletion:

density	figure	$n = 150$	$n = 500$
$(a_1, a_2, a_3) = (.2, 2..07)$	fig. 4	6	9
$(a_1, a_2, a_3) = (.2, 2..17)$	fig. 5	6	8
$(a_1, a_2, a_3) = (.2, 2..27)$	fig. 6	7	7

In the bimodal examples with $n = 500$, the main difficulty lies in estimating the secondary mode when it is relatively sharp (fig. 4). The logspline estimate with a fixed number of knots substantially underestimates the height of this mode. The method with knot deletion does a much better job: the estimated height differs significantly from the true height in both examples, but the estimate with the small rectangular window below suggests that the differences are due almost entirely to sampling variation that would effect any estimate. It is our experience, based on the examination of many more examples, that the differences between the height of the secondary peak as estimated by the logspline method with knot deletion and the true height of the mode is almost entirely due to such sampling variation. (Note that the number of observations contributing to the secondary peak is approximately binomial with parameters $n = 500$ and $p = .2$; even a window estimate with a fairly small bandwidth reduces the height of this mode somewhat when it is as sharp as in this example.)

The results for the relatively small sample size of $n = 150$ are quite different. Here there is a significant difference between the performance of the two density estimates. The logspline density estimate with a fixed number of knots clearly does not have sufficient flexibility to pick up the second mode correctly. Sometimes it misses the mode completely (fig. 5, left), while in all other cases it estimates a second mode which is considerably too wide. The method with

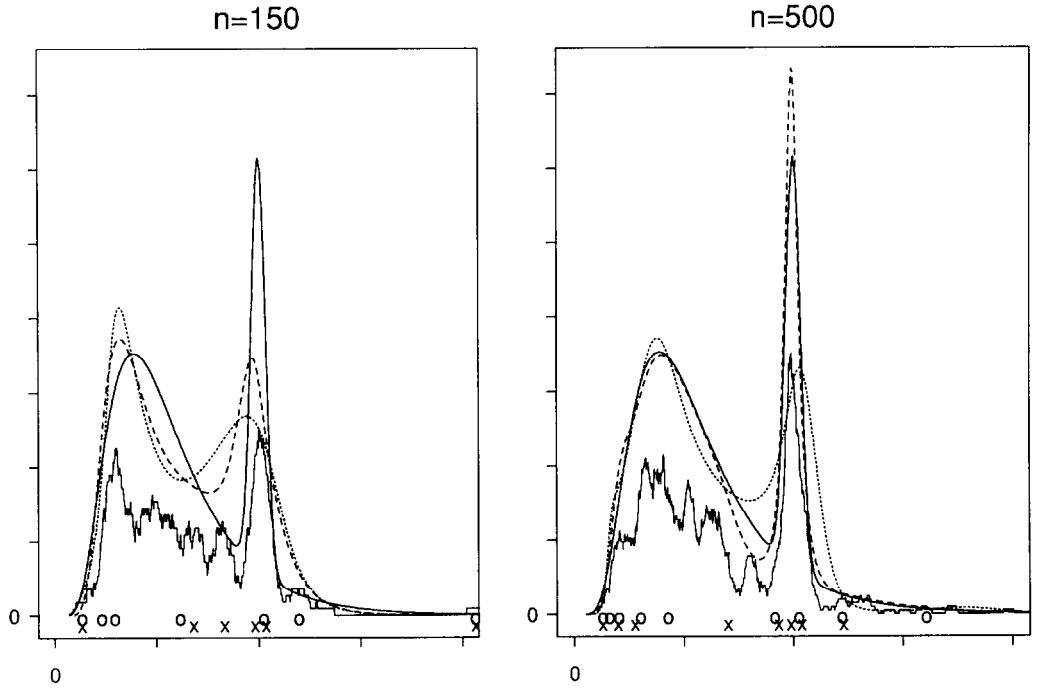

Fig. 4. Logspline density estimates for data generated from a bimodal density with a sharp second mode (.2, 2, .07) – Solid = real, dashed = knot deletion, dotted = fixed knots.

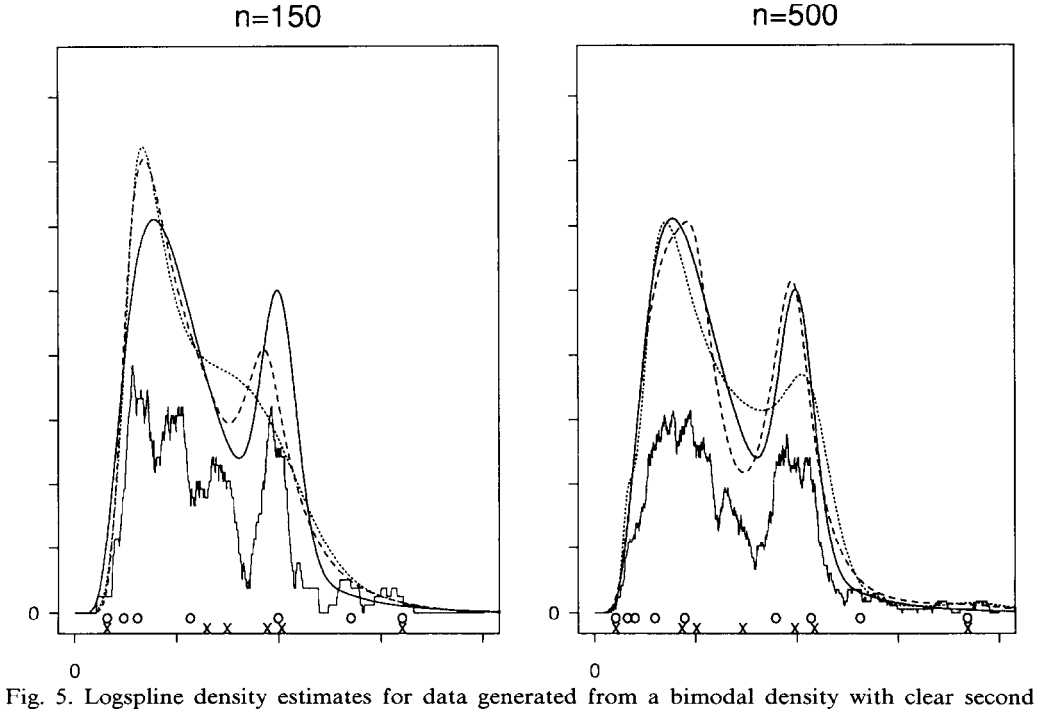

Fig. 5. Logspline density estimates for data generated from a bimodal density with clear second mode (.2, 2, .17) – Solid = real, dashed = knot deletion, dotted = fixed knots.

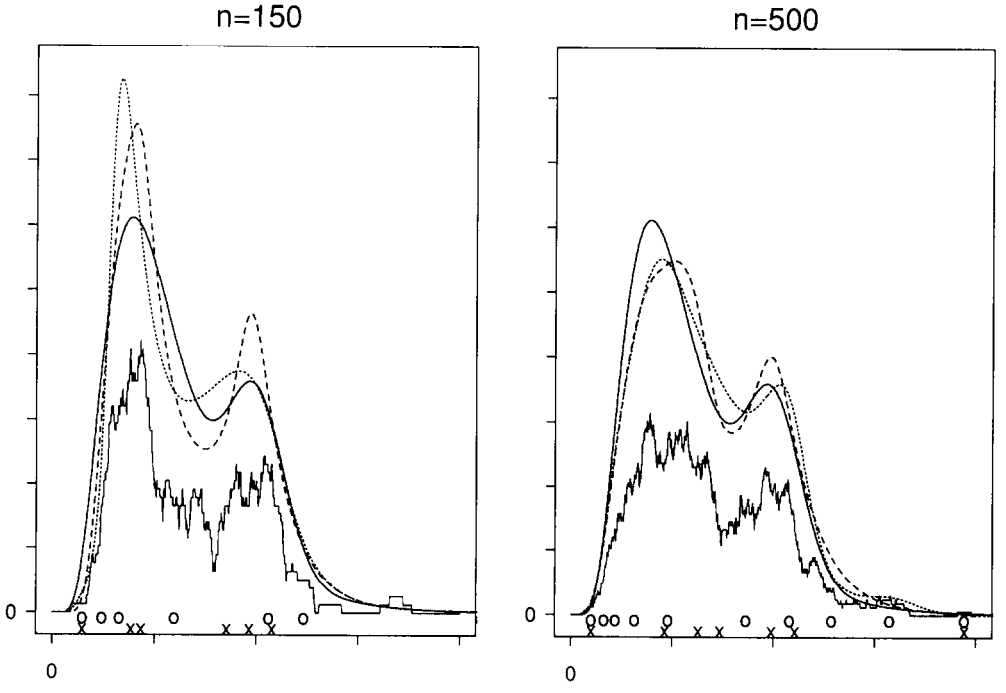

Fig. 6. Logspline density estimates for data generated from a bimodal density with small second mode (.2, 2, .27) – Solid = real, dashed = knot deletion, dotted = fixed knots.

stepwise knot deletion does a much better job. In these examples it always picks up the second mode (although we have found that it misses it sometimes too, especially when the second mode is small). It also does a good job of estimating the locations of the modes, the heights of the modes, and the general shape of the density.

To get a better idea about the performance of logspline density estimation in the case of a bimodal distribution, we carried out a small simulation in which 100 datasets of size 150 and 100 datasets of size 500 were generated from each of the three bimodal densities used in our examples. For each of these samples we computed the logspline density estimate with a fixed number of knots and with knot deletion. We then counted how often the estimate was, incorrectly, unimodal. The results can be found in the following table:

	$n = 150$		$n = 500$	
density	knot deletion	fixed knot	knot deletion	fixed knot
$(a_1, a_2, a_3) = (.2, 2..07)$	0	16	0	0
$(a_1, a_2, a_3) = (.2, 2..17)$	12	30	0	0
$(a_1, a_2, a_3) = (.2, 2..27)$	43	51	12	14

The density estimates are highly correlated: that is, the probability that both estimates are bimodal is larger than the product of the individual probabilities.

These results are in agreement with the examples. For $n = 500$ there appears to be little difference in the ability of the two methods to detect bimodalities, while for $n = 150$ the estimate with knot deletion is considerably better in detecting bimodalities than the one with a fixed number of knots.

7. Real examples

In fig. 7 we show a few density estimates based on real data, one widely used in the literature, while the others have been provided to us by two colleagues.

The data for the left top plot of fig. 7, which is labelled "suicide", consist of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks reported by Copas and Freyer [4]. The data are used extensively in Silverman [9] and they are also used by Wand, Marron and Ruppert [13]. The logspline density estimate with a fixed number of knots has 5 knots, the one with stepwise knot deletion has 4 of the original 11 knots left in the final estimate. The estimates are comparable to those reported by Wand et al.

The data for the right top plot of fig. 7, which is labelled "boston", arose in the following way. Each of 244 people gave a blood sample and a suitable treated fraction of the plasma was subjected to gradient gel electrophoresis and stained with a protein stain. The data are diameters (in Angstrom) of the major peak in the low-density lipoprotein region, the values being obtained by calibrating a densitometric scan. The data for the left bottom plot of fig. 7, which is labelled "montreal", arose in a similar way. Differences between the distributions of these two groups are due to the nature (age, sex, health status, etc.) of the two samples. This set of data was made available by Dr. Ronald M. Krauss of Lawrence Berkeley Laboratory via our colleague Terry Speed.

The logspline density estimate for boston ($n = 244$) with a fixed number of knots has 8 knots, the one with stepwise knot deletion has 7 of the original 13 knots left in the final estimate. Here the method with knot deletion clearly seems to do a better job than the one with a fixed number of knots. Looking to the rectangular window estimate, it seems that the mode, as detected by the method with stepwise knot deletion is actually there. The method with a fixed number of knots is unimodal.

The logspline density estimate for montreal ($n = 684$) with a fixed number of knots has 11 knots, the one with stepwise knot deletion has 8 of the original 16 knots left in the final estimate. Here the difference between the 2 estimates seems to be marginal. The size of the small side mode is the only observable difference.

The data for the right bottom plot of fig. 7, which is labelled "income", were provided to us by Wolfgang Haerdle. This is a dataset consisting of 7125 random samples of yearly net income in the United Kingdome (Family Expenditure Survey [5]). [1] (The data have been rescaled.) This dataset is considerably larger than the other examples that we have looked at. For such a large dataset the

[1] The calculations and investigations were made in close collaboration with the Wirtschaftstheoretische Abteilung II, University of Bonn, Bonn, West-Germany.

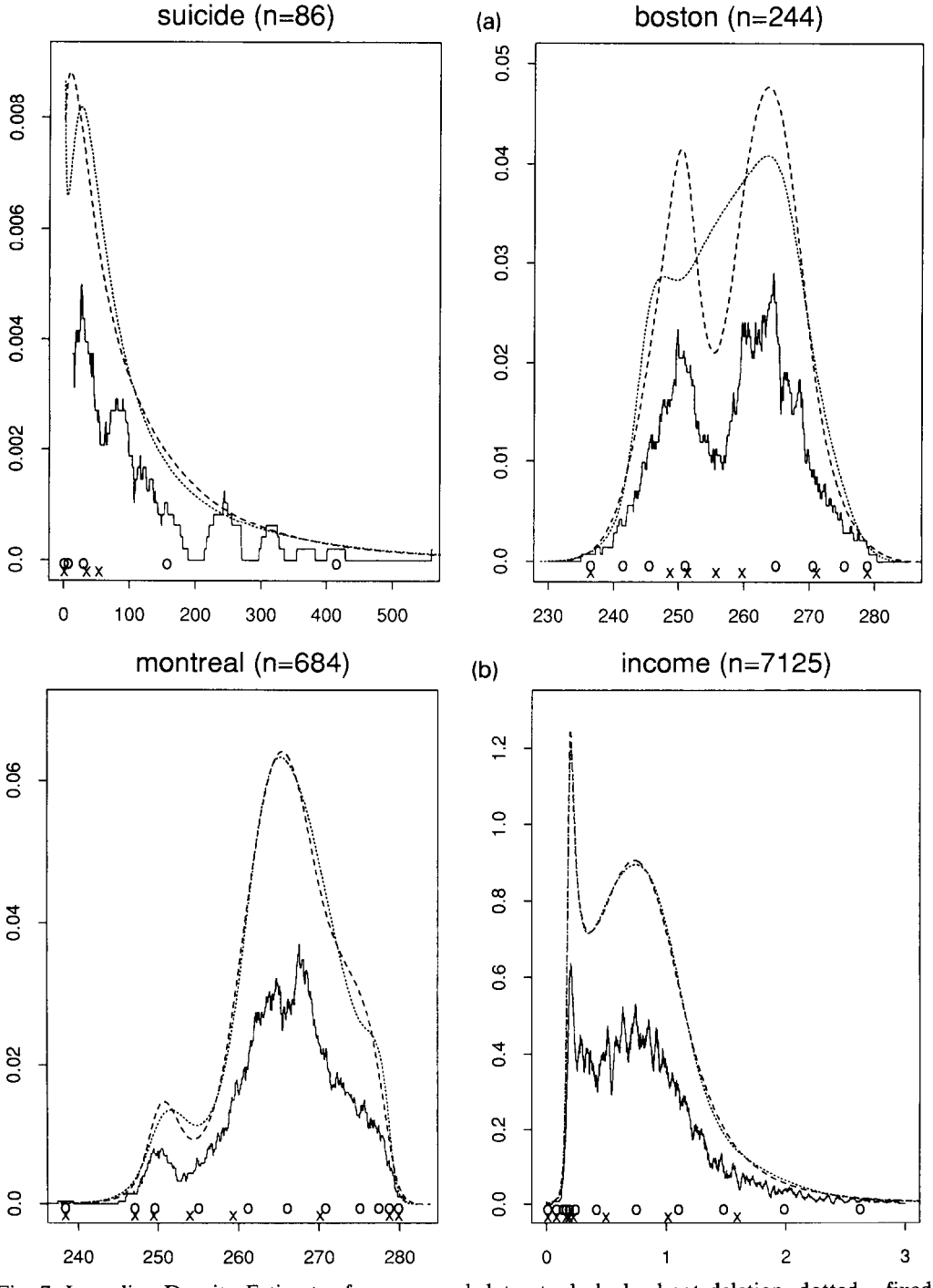

Fig. 7. Logspline Density Estimates for some real datasets dashed = knot deletion, dotted = fixed knots.

behavior of logspline density estimation is relatively insensitive to the precise number of knots. The logspline density estimate with a fixed number of knots has 15 knots and the one with stepwise knot deletion has 9 of the original 18 knots left in the final estimate. The density is clearly bimodal. The two estimates are very close, have absolutely no problem at all in picking up the sharp peak, and give almost identical estimates of the height of this peak.

Although this dataset is extremely large, so kernel estimator with a fixed window size will provide a reasonable estimate. Specifically, the sharp peak requires such a small window size that the rest of the density estimate will be too wiggly. Wand, Marron and Ruppert [13] apply modified kernel density estimates to this dataset. The modifications involve making a transformation, after which they use a global window size. The approach of Wand et al. seems to work as well as ours for a dataset of this size, but we have no comparison with their method for smaller datasets containing more details than the suicide data. It should be noted that the transformations of Wand et al. are specifically developed for skewed distributions. Presumably different transformations would have to be developed to make their method applicable to other problems. The logspline density approach is much directly applicable.

It seems appropriate here to report how much CPU-time is needed to compute logspline density estimates. The times reported below are seconds CPU-time on a Sparc-station 1 + . The code which was used had not yet been fully optimized, so it is likely that faster computations are possible.

dataset	Fixed Knots	Knot deletion
suicide	0.18	1.15
boston	0.24	1.64
montreal	0.46	2.88
income	5.02	12.03

For comparison, a kernel estimate in 100 points using the default density estimation subroutine in S uses about 0.85 seconds CPU-time for income on the same machine.

8. Conclusions

Our examination of figs. 2 through 7 (and many other similar figures) has convinced us that the logspline density estimate with stepwise knot deletion gives, automatically, a good density estimate, even for sample sizes as small as 80. The density estimate is smooth, but it also picks up those details and irregularities that seem real to the density underlying the data. Although we have not discussed any smaller datasets in this paper, we have found that even for sample sizes as small as 50 the logspline density method gives decent results. We have not tried our method on datasets of sample size less than 50.

The logspline density estimate with a fixed number of knots, performs nearly as well as the one with knot deletion when there are at least 400–500 observa-

tions. Since the method with a fixed number of knots uses less computing time it may be preferable for large datasets.

Our experience is that, in general, the spline underlying the density estimates needs about 1.5 knots close to a second mode to pick up that mode, and about 2 knots close to that mode to get a good estimate of its shape. For the estimate with stepwise knot deletion it suffices that these knot be among the knots in the initial estimate – they will rarely be deleted if there is a sizeable second mode.

The main advantages of logspline density estimation, as presented in this paper, are that it is fully automatic and produces good estimates, even for somewhat irregular densities.

Another advantage is that the logspline density estimate is a function of a relatively small number of parameters (knots, θ's), so the density can easily be used for other purposes (e.g. bootstrapping or robust regression).

A further advantage of logspline density estimation is that it can be used in a natural way to get estimates and confidence intervals for quantiles. Since logspline density estimation yields a smooth (functional) estimate of the density, it yields a smooth estimate of the quantile function too. Standard errors can be obtained using classical maximum-likelihood methods. Confidence intervals can be constructed using these standard errors and an adaption procedure similar to the one described in Breiman, Stone and Kooperberg [3].

Simulations show that the size and coverage probabilities of the confidence intervals are comparable to non-parametric bounds, where the non-parametric bounds exist, and they extend further in the tail, although for quantiles very far in the tails ($p = .1/n$) there are better methods available, which are discussed in Breiman, Stone and Kooperberg [3].

A minor disadvantage of logspline density estimation is that it is somewhat computer intensive. However with the increased availability of fast computers, we believe that this will not be a serious problem.

There is potential for further improvement to logspline density estimation and it would be worthwhile to compare this method with other methods, especially with the various forms of kernel density estimation. In the mean time, logspline density estimation, as presently constituted, works well enough to be useful as a data analytic tool. To this end, we intend to make the procedure publicly available as an "S" function and as a fortran subroutine.

Acknowledgements

We wish to thank Wolfgang Haerdle and Terry Speed for providing us with the data used in section 7 and for helpful discussions. We also wish to thank an anonymous referee for many detailed and helpful comments. This research was supported in part by National Science Foundation Grants DMS-8600409 and DMS-8902016.

References

[1] C. de Boor, *A practical guide to splines* (Springer-Verlag, New York, 1978).
[2] G.E.P. Box and D.R. Cox, An analysis of transformations (with discussion), *J. Roy. Statist. Soc.*, Ser. B **26** (1964) 211–252.
[3] L. Breiman, C.J. Stone and C. Kooperberg, Robust confidence bounds for extreme upper quantiles, *J. Statist. Comp. Simul.* **37** (1990) 127–149.
[4] J.B. Copas and M.J. Freyer, Density estimation and suicide risk in psychiatric treatment, *J. Roy. Statist. Soc.*, Ser. A **143** (1980) 167–176.
[5] Family Expenditure Survey, *Annual base tapes and reports (1968–1983)* (Department of Employment, Statistics Division – Her Majesty's Stationary Office, London, 1968–1983). (The data utilized in this paper were made available by the ESRC Data Archive at the University of Essex.)
[6] C. Kooperberg, *Smoothing images, splines and densities, Ph.D. thesis* (Department of Statistics, University of California at Berkeley, 1990).
[7] F. O'Sullivan, Fast computation of fully automated log-density and log-hazard estimators, *Siam J. Sci. Stat. Comput.*, **9** (1988) 363–379.
[8] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics*, **6** (1978) 461–464.
[9] B.W. Silverman, *Density estimation for statistics and data analysis* (Chapman and Hall, London, 1986).
[10] P.L. Smith, Curve fitting and modeling with splines using statistical variable selection methods, NASA, Langley Research Center, Hampla, VA, NASA Report 166034 (1982).
[11] C.J. Stone, Large sample inference for logspline model, *Annals of Statistics*, **18** (1990) 717–741.
[12] C.J. Stone and C.-Y. Koo, Logspline density estimation, *Contemporary Mathematics*, **59** (1986) 1–15.
[13] M.P. Wand, S.J. Marron and D. Ruppert, Transformations in density estimation (with discussion) *J.A.S.A.* (1991) to appear.