© Board of the Foundation of the Scandinavian Journal of Statistics 1999. Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA Vol 26: 87–105, 1999

Logspline Density Estimation under Censoring and Truncation

JA-YONG KOO

Hallym University

CHARLES KOOPERBERG

Fred Hutchinson Cancer Research Center

JINHO PARK

Inha University

ABSTRACT. In this paper we consider logspline density estimation for data that may be left-truncated or right-censored. For randomly left-truncated and right-censored data the product-limit estimator is known to be a consistent estimator of the survivor function, having a faster rate of convergence than many density estimators. The product-limit estimator and B-splines are used to construct the logspline density estimate for possibly censored or truncated data. Rates of convergence are established when the log-density function is assumed to be in a Besov space. An algorithm involving a procedure similar to maximum likelihood, stepwise knot addition, and stepwise knot deletion is proposed for the estimation of the density function based upon sample data. Numerical examples are used to show the finite-sample performance of inference based on the logspline density estimation.

Key words: Besov space, knot selection, left-truncation, MILE, product-limit estimator, rate of convergence, right-censoring

1. Introduction

This paper proposes a method of density estimation for left-truncated and right-censored data. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables, and let C_i and T_i denote the right-censoring variable and left-truncation variable for the ith case, respectively. We assume that $(C_1, T_1), (C_2, T_2), \ldots$ are independent and identically distributed random variables, and that they are independent of the X_is. When the X_is are right-censored, only $Y_i = \min(X_i, C_i)$ and the censoring indicator $\delta_i = I(X_i \leqslant C_i)$, where $I(X_i \leqslant C_i)$ is 1 if $X_i \leqslant C_i$ and 0 otherwise, are observed. If the X_is are also subject to left-truncation, (Y_i, δ_i, T_i) is observed only when $Y_i \geqslant T_i$. The sample data consist of $(Y_i^o, \delta_i^o, T_i^o)$, $1 \leqslant i \leqslant n$ with $Y_i^o \geqslant T_i^o$. See Fleming & Harrington (1991) and Andersen *et al*. (1993) for more discussion of right-censoring and left-truncation. An example of a data set that involves both right-censoring and left-truncation is the mortality of diabetics in the county of Fyn in Denmark (Andersen *et al*., 1993). We will discuss this example in detail in section 5.

Flexible exponential families have been used for the estimation of density functions. Stone & Koo (1986), Stone (1989, 1990), Kooperberg & Stone (1991) and Koo (1996) developed logspline density estimation, in which the logarithm of a probability density function is modeled using polynomial splines. Barron & Sheu (1991) studied density estimation procedures based on trigonometric series, polynomials, and splines. Koo & Kim (1996) considered an exponential family based on wavelets. Exponential families have been used by Koo & Park (1996) and Koo & Chung (1998) for density estimation in linear inverse problems. For an excellent discussion on density estimation see Silverman (1986).

A number of papers have dealt with density estimation based upon possibly censored data. In particular, Marron & Padgett (1987) studied bandwidth selection for kernel density estimators based upon right-censored data; Gijbels & Wang (1993) considered kernel density and hazard estimation based on a representation for the product-limit estimator of the survivor function; and Kooperberg & Stone (1992) and Kooperberg (1997) developed logspline density estimation for univariate data that may be right-censored, left-censored or interval-censored and for bivariate data that may be right-censored, respectively. Kaplan & Meier (1958) considered product-limit estimators for incomplete data. Lai & Ying (1991) addressed the problem of estimating a distribution function under truncation and censoring.

Consider logspline density estimation without truncation or censoring, so that $X_1, \ldots, X_n$ are actually observed. Let $B_1, \ldots, B_J$ be a set of basis functions that span a space of polynomial splines. The exponential family based on these basis functions has the form

$$f(x; \boldsymbol{\theta}) = \exp\{\theta_1 B_1(x) + \cdots + \theta_J B_J(x) - \psi(\boldsymbol{\theta})\},$$

where $\psi(\boldsymbol{\theta})$ is the normalizing constant. The parameters of the logspline density estimate satisfy the equation

$$\int B_k(x) f(x; \tilde{\boldsymbol{\theta}})\, dx = \int B_k\, d\tilde{F}, \quad \text{for } k = 1, \ldots, J, \tag{1.1}$$

where $\tilde{F}$ is the usual empirical distribution function.

When censoring and truncation may be present, we confine our attention to the estimation of the conditional density f_{ac} of X given that $a \leqslant X \leqslant c$, for some constants a and c. We use the product-limit estimator to find an appropriate estimator $\hat{B}_k$, whose expectation is asymptotically the same as $\int B_k f_{ac}$. The proposed density estimator has the form $f(\cdot; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ satisfies (1.1) with $\int B_k d\tilde{F}$ replaced by $\int B_k d\hat{F}$, and $\hat{F}$ is the product-limit estimator. This density estimate has many of the advantages of the usual logspline density estimates. In particular, the estimates are positive and integrate to one. The kernel density estimators by Marron & Padgett (1987) and Gijbels & Wang (1993) may not have this property when higher order kernels are used. While the approach of Kooperberg & Stone (1992) has the advantage that it models the complete density of X, rather than the conditional density given that $a \leqslant X \leqslant c$, it has the clear disadvantage that the resulting log-likelihood is not necessarily concave when censoring is present. As such, it is much harder to establish theoretical results, and in a numerical implementation one cannot guarantee that the global maximum of the likelihood function is found.

In this paper it is shown that the logspline density estimates based upon the product-limit estimator possess the rate of convergence $n^{-2\alpha/(2\alpha+1)}$, where α is the smoothness of the logarithm of the density function in a Besov space. The main idea in establishing this result is the observation that the product-limit estimator converges at a faster rate than the density estimator. Thus the rate of convergence for logspline density estimation based on the product-limit estimator is the same as the rate of convergence for logspline density estimation based upon uncensored data.

For our theoretical results, we assume that the knots are distributed regularly over the range of the data, and that the number of knots increases with the sample size. In practice, we select the knots adaptively using stepwise knot addition and stepwise knot deletion.

This paper is organized as follows. In section 2, logspline densities are defined and bounds for these densities are established. Asymptotic results are stated in section 3 and proved in section 6. Practical aspects of logspline density estimation are discussed in section 4. Numerical examples are given in section 5.

2. Logspline densities

In sections 2, 3 and 6, it will be assumed that $a = 0$ and $c = 1$. In this section we define logspline densities on the unit interval $\mathscr{T} = [0, 1]$. To simplify notation the dependence on the sample size n of various quantities will be suppressed. In the remainder of this paper $M_1, M_2, \ldots$ are positive constants, independent of n, and M is a positive constant, also independent of n, that is only used locally, and may be different in different locations.

2.1. Logspline densities based on B-splines

Let N^q be the B-spline of order q having knots at $0, 1, \ldots, q$. Thus $N^q(x) = q[0, 1, \ldots, q](\cdot - x)_+^{q-1}$, using the divided difference notation (de Boor, 1978). Let j be a positive integer, which will depend on n, and define

$$B_{j,k} = N^q(2^j x - k), \quad k \in \mathscr{Z}.$$

To approximate a function on $\mathscr{T}$, we only need those B-splines $B_{j,k}$ which do not vanish identically on $\mathscr{T}$. Let $\Lambda(j)$ denote the set of k for which this is the case and let $\mathscr{S}_j$ denote the linear span of the B-splines $B_{j,k}$, $k \in \Lambda(j)$. We refer to $\mathscr{S}_j$ as the space of dyadic splines. The dimension of $\mathscr{S}_j$ is $J = 2^j + q - 1$. Let $J \geqslant 2$ for all n.

Let Θ denote the collection of all J-dimensional vectors. Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)' \in \Theta$, set

$$s(\cdot; \boldsymbol{\theta}) = \sum_{k=1}^{J} \theta_k B_{j,k},$$

$$\psi(\boldsymbol{\theta}) = \log\left[\int_{\mathscr{T}} \exp\{s(x; \boldsymbol{\theta})\}\, dx\right],$$

and

$$f(\cdot; \boldsymbol{\theta}) = \exp\{s(\cdot; \boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\}. \tag{2.1}$$

Then $\int_{\mathscr{T}} f(x; \boldsymbol{\theta})\, dx = 1$ for $\boldsymbol{\theta} \in \Theta$. For notational convenience, let $s(\boldsymbol{\theta})$ and $f(\boldsymbol{\theta})$ denote the function $s(\cdot; \boldsymbol{\theta})$ and the density function $f(\cdot; \boldsymbol{\theta})$, respectively. The exponential family $f(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is not identifiable (Stone, 1990). Let Θ^0 denote the $(J-1)$-dimensional subspace of Θ, consisting of those vectors $\boldsymbol{\theta} \in \Theta$ whose entries add up to zero. We refer to the densities $f(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta^0$, as logspline densities.

The relative entropy (Kullback–Leibler distance: KL distance) between two densities f_1 and f_2 is defined as

$$D(f_1 \| f_2) = \int_{\mathscr{T}} f_1(x) \log\left\{\frac{f_1(x)}{f_2(x)}\right\} dx.$$

For logspline densities $f(\boldsymbol{\theta}_1)$ and $f(\boldsymbol{\theta}_2)$ write $D(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_2) = D(f(\boldsymbol{\theta}_1) \| f(\boldsymbol{\theta}_2))$. For a function h, let $\int \mathbf{B} h$ denote the J-dimensional vector of elements $\int_{\mathscr{T}} B_{j,k}(x) h(x)\, dx$, where $\mathbf{B} = (B_{j,1}, \ldots, B_{j,J})'$. Given $\boldsymbol{\beta} \in \Theta$, let $\boldsymbol{\theta}(\boldsymbol{\beta}) \in \Theta^0$ denote a solution to the equation

$$\int \boldsymbol{B} f(\boldsymbol{\theta}) = \boldsymbol{\beta} \tag{2.2}$$

2.2. Bounds for logspline densities

In this section we will describe bounds on $D(\boldsymbol{\theta}_0 \| \boldsymbol{\theta})$ in terms of the Euclidean distance $|\boldsymbol{\theta}_0 - \boldsymbol{\theta}|$ for any $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}$ in Θ^0, and give bounds on $|\boldsymbol{\theta}(\boldsymbol{\beta}_0) - \boldsymbol{\theta}(\boldsymbol{\beta})|$ in terms of $|\boldsymbol{\beta}_0 - \boldsymbol{\beta}|$, where $|\boldsymbol{\theta}| = (\sum_{k=1}^{J} \theta_k^2)^{1/2}$ for $\boldsymbol{\theta} \in \Theta$.

Let $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the usual L_2 and L_∞ norm on $\mathscr{T}$, respectively. By lem. 4.2 of DeVore & Popov (1988),

$$\frac{1}{M_1 J}|\boldsymbol{\theta}|^2 \leqslant \|s(\boldsymbol{\theta})\|_2^2 \leqslant \frac{1}{M_2 J}|\boldsymbol{\theta}|^2 \quad \text{for any } \boldsymbol{\theta} \in \Theta. \tag{2.3}$$

It follows from (2.3) and the properties of B-splines that

$$\|s(\boldsymbol{\theta})\|_\infty \leqslant \max_{1 \leqslant k \leqslant J} |\theta_k| \leqslant |\boldsymbol{\theta}| \leqslant \sqrt{M_1 J}\|s(\boldsymbol{\theta})\|_2, \quad \text{for any } \boldsymbol{\theta} \in \Theta. \tag{2.4}$$

We relate distances between logspline densities to distances between their parameters. A proof of the following lemma is similar to that of lem. 3 in Barron & Sheu (1991); it uses (2.3), (2.4) and

$$\sum_k B_{j,k}(x) = 1, \quad \text{for } x \in \mathscr{R}, \tag{2.5}$$

which is a property of B-splines (de Boor, 1978).

Lemma 1
For $\boldsymbol{\theta}_0, \boldsymbol{\theta} \in \Theta^0$,

$$\|\log\{f(\boldsymbol{\theta}_0)/f(\boldsymbol{\theta})\}\|_\infty \leqslant 2|\boldsymbol{\theta}_0 - \boldsymbol{\theta}|,$$

$$D(\boldsymbol{\theta}_0\|\boldsymbol{\theta}) \leqslant \frac{b}{2M_2 J}\exp(|\boldsymbol{\theta}_0 - \boldsymbol{\theta}|)|\boldsymbol{\theta}_0 - \boldsymbol{\theta}|^2,$$

and

$$D(\boldsymbol{\theta}_0\|\boldsymbol{\theta}) \geqslant \frac{1}{2M_1 bJ}\exp(-2|\boldsymbol{\theta}_0 - \boldsymbol{\theta}|)|\boldsymbol{\theta}_0 - \boldsymbol{\theta}|^2,$$

where $b = \exp\{\|\log f(\boldsymbol{\theta}_0)\|_\infty\}$.

We relate distances between the parameters $\boldsymbol{\theta}$ to distances between the corresponding parameters $\boldsymbol{\beta}$. The argument used to prove lem. 4 of Barron & Sheu (1991) can be used to prove lemma 2.

Lemma 2
Let $\boldsymbol{\theta}_0 \in \Theta^0$, $\boldsymbol{\beta}_0 = \int \mathbf{B} f(\boldsymbol{\theta}_0)$, *and* $\boldsymbol{\beta} \in \Theta$. *Set* $b = \exp\{\|\log f(\boldsymbol{\theta}_0)\|_\infty\}$. *If*

$$|\boldsymbol{\beta}_0 - \boldsymbol{\beta}| \leqslant \frac{1}{4M_1 beJ},$$

then the solution $\boldsymbol{\theta}(\boldsymbol{\beta})$ *to* $\int \mathbf{B} f(\boldsymbol{\theta}) = \boldsymbol{\beta}$ *exists in* Θ^0 *and satisfies*

$$|\boldsymbol{\theta}(\boldsymbol{\beta}_0) - \boldsymbol{\theta}(\boldsymbol{\beta})| \leqslant 2M_1 be^{\xi} J|\boldsymbol{\beta}_0 - \boldsymbol{\beta}|,$$

$$\|\log[f\{\boldsymbol{\theta}(\boldsymbol{\beta}_0)\}/f\{\boldsymbol{\theta}(\boldsymbol{\beta})\}]\|_\infty \leqslant 4M_1 be^{\xi} J|\boldsymbol{\beta}_0 - \boldsymbol{\beta}| \leqslant \xi,$$

and

$$D\big(\boldsymbol{\theta}(\boldsymbol{\beta}_0)\|\boldsymbol{\theta}(\boldsymbol{\beta})\big) \leqslant 2M_1 be^{\xi} J|\boldsymbol{\beta} - \boldsymbol{\beta}_0|^2,$$

for ξ *satisfying* $4M_1 be^{\xi} J|\boldsymbol{\beta} - \boldsymbol{\beta}_0| \leqslant \xi \leqslant 1$.

2.3. Information projection and approximation error

The function $\psi(\boldsymbol{\theta})$ is finite for all $\boldsymbol{\theta} \in \Theta^0$, since $B_{j,k}(x) \leqslant 1$ for $x \in \mathscr{R}$. Let $\mathscr{F}(\boldsymbol{\beta}) = \{f\colon \int \mathbf{B} f = \boldsymbol{\beta}\}$ be the hyperplane of all density functions f for which $\int \mathbf{B} f$ equals $\boldsymbol{\beta}$, with $\boldsymbol{\beta} \in \Theta$.

The linear independence of the B-splines together with (2.5) implies that if $s(\boldsymbol{\theta}) - s(\boldsymbol{\theta}')$ is constant for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta^0$, then $\boldsymbol{\theta} = \boldsymbol{\theta}'$. Using this and the argument used to show lemma 3 of Barron & Sheu (1991), we can establish the following result.

Lemma 3
Suppose that $f \in \mathscr{F}(\boldsymbol{\beta})$ and $\boldsymbol{\beta} \in \{\int \mathbf{B} f(\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta^0\}$. Then the solution $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\beta})$ to (2.2) is unique. Moreover, for all $\boldsymbol{\theta} \in \Theta^0$, the Pythagorean-like identity

$$D(f\|f(\boldsymbol{\theta})) = D(f\|f(\boldsymbol{\theta}_0)) + D(\boldsymbol{\theta}_0\|\boldsymbol{\theta})$$

holds. Consequently, $f(\boldsymbol{\theta}_0)$ is characterized as the minimizer of $D(f\|f(\boldsymbol{\theta}))$ over $\boldsymbol{\theta} \in \Theta^0$. Also, $L(\boldsymbol{\theta}) = \sum_k \theta_k \beta_k - \psi(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta}(\boldsymbol{\beta})$.

Let

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta^0} D(f_{01}\|f(\boldsymbol{\theta})),$$

and set $f^* = f(\boldsymbol{\theta}^*)$. We will refer to f^* as the information projection of f_{01} onto $\mathscr{S}_j$. It follows from lem. 1 of Stone (1990) that f^* satisfy the equation

$$\int B_{j,k} f_{01} = \int B_{j,k} f^* \quad \text{for } k = 1, \ldots, J. \tag{2.6}$$

Let $\mathbf{B}^* = (B^*_{j,1}, \ldots, B^*_{j,J})'$ where $B^*_{j,k} = \int B_{j,k} f_{01}$. It follows from lemma 3 and (2.6) that the information projection f^* exists uniquely and that $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\mathbf{B}^*)$.

Set $g_{01} = \log f_{01}$. Let

$$\Delta = \inf_{s \in \mathscr{S}_j} \|g_{01} - s\|_2$$

and

$$\gamma = \inf_{s \in \mathscr{S}_j} \|g_{01} - s\|_\infty$$

be the L_2 and L_∞ error in the approximations of g_{01} by some $s \in \mathscr{S}_j$. Theorem 1 establishes an upper bound on the approximation error $D(f_{01}\|f^*)$ in terms of Δ under the condition

(A1) $M_3^{-1} \leqslant f_{01} \leqslant M_3$.

Theorem 1
If (A1) holds, γ is bounded, and $\sqrt{J}\Delta = o(1)$, then the information projection f^* uniquely exists and satisfies

$$D(f_{01}\|f^*) \leqslant \frac{M_3}{2} \exp(\gamma)\Delta^2.$$

3. Asymptotic results

Let f_X denote the density function of X_1, and let

$$S(t) = P(X_1 \geqslant t), \quad G(t) = P(T_1 \leqslant t \leqslant C_1), \quad R(t) = P(T_1 \leqslant t \leqslant Y_1),$$

$$F(t|0) = P(X_1 \leqslant t | X_1 \geqslant 0), \quad S(t|0) = P(X_1 \geqslant t | X_1 \geqslant 0).$$

Define $\tau = \inf\{t\colon G(t) > 0\}$ and $\tau^* = \inf\{t > \tau\colon G(t)S(t) = 0\}$.

For $0 > \tau$, the product-limit estimator of $F(t|0)$ is

$$\hat{F}(t|0) = 1 - \prod_{0 \leqslant Y^o_{(i)} \leqslant t} \left(1 - \frac{1}{n_{(i)}}\right)^{\delta^o_{(i)}},$$

where $(Y^o_{(i)}, \delta^o_{(i)})$, $1 \leqslant i \leqslant n$, denote the ordered Y^o_is along with their corresponding δ^o_is, and $n_{(i)}$ is the number of js such that $T^o_j \leqslant Y^o_{(i)} \leqslant Y^o_j$. Similarly, the product-limit estimator of $S(t|0)$ is

$$\hat{S}(t|0) = \prod_{0 \leqslant Y^o_{(i)} < t} \left(1 - \frac{1}{n_{(i)}}\right)^{\delta^o_{(i)}}.$$

We confine our attention to the estimation of the conditional density f_{01} of X given $0 \leqslant X \leqslant 1$, where $\tau < 0 < 1 < \tau^*$. Note that $f_{01} = f_X / P(0 \leqslant X \leqslant 1)$. Let $\#(t)$ denote the size of the risk-set at t, which is defined by the number of js for which $T^o_j \leqslant t \leqslant Y^o_j$. For censored and truncated data, we define the incomplete likelihood function corresponding to the logspline family by

$$\begin{aligned} l(\boldsymbol{\theta}) &= \frac{1}{\hat{F}(1|0)} \int_{\mathcal{T}} \log f(\cdot; \boldsymbol{\theta})\, d\hat{F}(\cdot|0) \\ &= \frac{1}{\hat{F}(1|0)} \sum_{i=1}^{n} I(0 \leqslant Y^o_i \leqslant 1)\, d_i s(Y^o_i; \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \hat{w}^o_i s(Y^o_i; \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}), \end{aligned} \tag{3.1}$$

where the jump size $d_i = d_{ni}$ at Y^o_i of $\hat{F}(\cdot|0)$ is given by $d_i = \delta^o_i \hat{S}(Y^o_i|0)/\#(Y^o_i)$ and $\hat{w}^o_i = \hat{w}^o_{ni} = \hat{F}(1|0)^{-1} I(0 \leqslant Y^o_i \leqslant 1)\delta^o_i \hat{S}(Y^o_i|0)/\#(Y^o_i)$. Note that the incomplete likelihood function defined by (3.1) is not necessarily interpretable as a log-likelihood. We introduce $l(\boldsymbol{\theta})$ as an objective function in the definition of logspline density estimators for incomplete data under censoring and truncation. (When there is no censoring or truncation $nl(\boldsymbol{\theta})$ is the usual log-likelihood for logspline density estimation.) Let

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta^0} l(\boldsymbol{\theta}) \tag{3.2}$$

be the maximum incomplete likelihood estimator (MILE) of $\boldsymbol{\theta} \in \Theta^0$. Since the Hessian matrix of $\psi(\cdot)$ is strictly positive definite and $l(\cdot)$ is a strictly concave function on Θ^0; thus the MILE $\hat{\boldsymbol{\theta}}$ is unique if it exists. We set $\hat{f} = f(\hat{\boldsymbol{\theta}})$ and refer to $\hat{f}$ as the MILE of f_{01}.

Let $\hat{\mathbf{B}} = (\hat{B}_{1,J}, \ldots, \hat{B}_{k,J})'$ where $\hat{B}_{j,k} = \sum_i \hat{w}^o_i B_{j,k}(Y^o_i)$. From the likelihood equation, $\hat{f}$ is the logspline density (2.1) that satisfies

$$\int \mathbf{B} \hat{f}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{B}},$$

which implies that $\hat{f} = f(\boldsymbol{\theta}(\hat{\mathbf{B}}))$.

To develop upper bounds on the estimation error $D(f^* \| \hat{f})$, we assume that

(A2) $\inf_{t \in \tau} R(t) \geqslant M$; and

(A3) $\pi = P(Y_1 \geqslant T_1) > 0$.

Theorem 2 establishes the rates of convergence of $\hat{f}$ to f_{01} in KL distance.

Theorem 2

If (A1)–(A3) hold, the sequence γ *is bounded, and* $J/\sqrt{n} \to 0$, *then* $\hat{f}$ *exists, except on an event whose probability tends to zero with n, and* $D(f^{*}\|\hat{f}) = O_P(J/n)$.

As a smoothness class for f_{01}, we use the Besov space. If $h \in L_p(\mathscr{T})$, $1 \leqslant p \leqslant \infty$, let $\omega_q(h, t)_p$, $t > 0$, denote the modulus of smoothness for h: $\omega_q(h, t)_p = \sup_{|u| \leqslant t} \|\delta_u^q h(\cdot)\|_p(\mathscr{T}(qu))$, where δ_u^q is the qth order difference with step u; the norm in the above definition is the L_p norm on the set $\mathscr{T}(qu) = \{x\colon x, x + qu \in \mathscr{T}\}$. We say that h is in the Besov space $\mathscr{B}_{\alpha pp'}$ whenever

$$\|h\|_{\mathscr{B}_{\alpha pp'}} = \left[\int_0^{\infty} \{t^{-\alpha}\omega_q(h, t)_p\}^{p'} \frac{dt}{t}\right]^{1/p'}$$

is finite, where q is any integer larger than α.

The dyadic B-splines $\{B_{j,k}\colon k \in \Lambda(j), j = 0, 1, 2, \ldots\}$ give an atomic decomposition for functions in the Besov space. From DeVore & Popov (1988) we know that a function h in $\mathscr{B}_{\alpha pp'}$ can be written as

$$h = \sum_{j=0}^{\infty} \sum_{k \in \Lambda(j)} \eta_{j,k} B_{j,k}$$

and

$$M^{-1}\|h\|_{\mathscr{B}_{\alpha pp'}} \leqslant \left[\sum_{j=0}^{\infty} \{2^{j(\alpha - 1/p)} |\boldsymbol{\eta}_j|_p\}^{p'}\right]^{1/p'} \leqslant M\|h\|_{\mathscr{B}_{\alpha pp'}}, \tag{3.3}$$

with the usual modification if either p or p' equals ∞. Here $|\boldsymbol{\eta}_j|_p$ denotes $(\sum_{k \in \Lambda(j)} |\eta_{j,k}|^p)^{1/p}$. See DeVore & Popov (1988) and Donoho *et al*. (1996) for properties of Besov spaces.

Given positive numbers a_n and b_n for $n \geqslant 1$, let $a_n \asymp b_n$ mean that a_n/b_n is bounded away from zero and infinity. For theorem 3 we also assume that

(A4) $g_{01} = \log f_{01} \in \mathscr{B}_{\alpha pp'}$ such that $\|g_{01}\|_{\mathscr{B}_{\alpha pp'}} \leqslant M$; and

(A5) $\frac{1}{2} < \alpha < q - 1 + \frac{1}{p}$ and $2 \leqslant p \leqslant \infty$.

Theorem 3

If (A2)–(A5) hold, then $D(f_{01}\|\hat{f}) = O_P(n^{-2\alpha/(2\alpha+1)})$ is $J \asymp n^{1/(2\alpha+1)}$.

Remark 1. Note that that the Besov space includes the Hilbert–Sobolev space and Hölder space; spaces that are traditionally used in theoretical statistics. See DeVore & Popov (1988) and Donoho *et al*. (1996) and references therein for further properties of the Besov space.

Remark 2. When there is no censoring or truncation, the logspline density estimators achieve the optimal rate of convergence (Koo & Kim, 1996). It is anticipated that the rate of convergence in theorem 3 remains optimal when censoring and truncation are present.

Remark 3. When there is only censoring but no truncation, let $\hat{F}(1)$ be the traditional product-limit (Kaplan–Meier) estimate of $F(1|0) = P(X_1 \leqslant 1)$ and define $\tilde{f} = \hat{F}(1)\hat{f}$. Observe that $D(f_X\|\tilde{f}) = F(1|0)D(f_{01}\|\hat{f}) + F(1|0)\log\{f(1|0)/\hat{F}(1)\}$. Since under certain conditions $\hat{F}(1)$ is a $\sqrt{n}$-consistent estimate of $F(1|0)$, we can actually give an estimator of f_X with its rate of convergence given in theorem 3.

4. Practical implementation

Logspline density estimation under censoring and truncation, based upon the incomplete log-likelihood (3.1), can be implemented using a slightly modified algorithm for logspline density estimation for complete data by allowing for case weights. For our examples we used the algorithm described in Stone *et al.* (1997). In this section we give a brief description of this algorithm and discuss the modifications that make it applicable when some data may be censored or truncated. More details about the algorithm can be found in Kooperberg & Stone (1992) and Stone *et al*. (1997).

The algorithm of Stone *et al*. employs cubic splines. In particular, given the integer $K \geqslant 3$, the numbers L and U, with $-\infty \leqslant L$ and $U \leqslant \infty$, and the sequence $t_1, \ldots, t_K$, with $L < t_1 < \cdots < t_K < U$, let G be the space of twice differentiable functions s on (L, U), such that the restrictions of s to $(L, t_1]$ and $[t_k, U)$ are linear and the restrictions of s to $[t_1, t_2], \ldots, [t_{K-1}, t_K]$ are cubic polynomials. The space G is K-dimensional. Set $J = K - 1$. Let $1, B_1, \ldots, B_J$ be a basis of G. A column vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)'$ is said to be feasible if $\psi(\boldsymbol{\theta}) < \infty$. Given a feasible $\boldsymbol{\theta}$ the function $f(\cdot; \boldsymbol{\theta})$ is a positive density on $\mathscr{T} = [L, U]$. We refer to the t_i, $i = 1, \ldots, K$, as knots.

Let $(Y_i^o, \delta_i^o, T_i^o)$, $1 \leqslant i \leqslant n$, with $Y_i^o \geqslant T_i^o$ be the actual observed data based upon a random sample from an unknown distribution with density function f that was subject to right-censoring and left-truncation. Let Z_j, $1 \leqslant j \leqslant n^* \leqslant n$, be the set of unique values of the Y_i^o for which $\delta_i^o = 1$. Let $v_j = \hat{F}(Z_j|a) + \hat{S}(Z_j|a) - 1$, $1 \leqslant j \leqslant n^*$, with $a = \min T_i^o$, be the jump of the product-limit estimate of the survivor function $\hat{S}(\cdot|a)$ at Z_j. For a given set of knots, the logspline density estimate under right-censoring and left-truncation is $f(x; \hat{\boldsymbol{\theta}})$, $a \leqslant x \leqslant U$, where

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{n^*} v_j f(Z_j; \boldsymbol{\theta}),$$

is the MILE of $\boldsymbol{\theta}$ (compare with (3.2)). The Hessian matrix corresponding to this likelihood is easily established to be concave. As such, the MILE is unique when it exists, and it can be found using a suitably modified Newton–Raphson algorithm.

Initially the algorithm starts with a limited number of knots (see Kooperberg & Stone (1992) and Stone *et al*. (1997) for details). In the modified version for censored and truncated data all knots are equal to an uncensored observations in the data set. Then stepwise knot addition is employed. At each stage all uncensored observations, that are a minimum distance away from existing knots, are candidates for addition. Among these candidates, the algorithm performs a heuristic search to maximize the statistic

$$\frac{\text{Rao}(t)}{v_t'}, \tag{4.1}$$

where Rao(t) is the Rao statistic for adding a knot at the location $x = t$ of an uncensored observation to the current set of knots (see eq. (3.3) of Stone *et al*. (1997) for the exact definition of Rao(t)), and $v_t' = v_t / n_t$, where n_t is the number of uncensored observations at $x = t$. Stepwise addition of knots is employed until a prescribed maximum number of knots is reached.

Upon stopping the stepwise addition process, we carry out stepwise deletion. At each step the knot for which the statistic

$$\frac{\mathrm{Wald}(t)}{v'_t}, \tag{4.2}$$

where Wald(t) is the Wald statistic for removal of the knot at $x = t$ from the current set of knots (see eq. (3.5) of Stone *et al.*, 1997 for the exact definition), is the smallest in magnitude is removed. Stepwise deletion of knots continues as long as the statistic (4.2) is smaller than a prespecified level λ. The last fitted model is the one that we select as the logspline density estimate under right-censoring and left-truncation.

Relative to logspline density estimation, as discussed in Stone *et al.* (1997) and earlier papers, we added a division by v'_t to the equations for the Rao statistic (4.1) and Wald statistic (4.2). The reason for adding these weights is that in some examples the v'_t can vary considerably (we have observed ratios of about 100 between the largest and smallest v'_t). To see that this makes weighting necessary, consider the following example. Suppose that for all uncensored observations with $x < A$ the weight v'_t equals 100, and for all uncensored observations with $x \geqslant A$ the weight v'_t equals 1. Without weighting the logspline procedure would want to add knots t for which $t < A$, since additional detail in the log-density function in that region will have a hundred times larger influence on the incomplete log-likelihood. However, the difference in the incomplete log-likelihood does not tell us whether a knot smaller than A is any more significant than a knot larger than A. Actually, it is easily seen, that a test of significance for a knot $t \ll A$, which has virtually no influence on $\hat{f}(x)$ for $x > A$, the Wald or Rao statistic should be divided by 100 relative to the test-statistic for the significance for a knot $t \gg A$.

In Kooperberg & Stone (1992) and Stone *et al.* (1997) the final model is selected from a sequence of models using a modified version of AIC. However, since the Rao and Wald statistics in the current algorithm are scaled, the differences in the incomplete log-likelihood between successively fitted models are not comparable. Thus, we decided to select the last model before the Wald statistic would exceed a prespecified critical value. In our examples we took $\lambda = \log n'$, where n' is the number of uncensored observations. Using $\lambda = \log n'$ is comparable to choosing the penalty parameter in AIC equal to $\log n'$, as has been advocated in earlier logspline papers. In the current situation, we carried out parts of the simulation study reported in the next section with various other values of λ; we found that $\log n'$ gave the smallest integrated squared error in those simulations.

5. Examples

All the data sets that we analyse in this section involve right-censoring or left-truncation. Without censored or truncated data the algorithm described in the previous section is identical to the one of Stone *et al.* (1997); we thus refer to that paper for examples where all data are complete.

Figures 1 and 2 are based upon the Stanford heart transplant data. There exist several versions of this data; we used the data as listed in Kalbfleisch & Prentice (1980). The sample size for this data set is 103, of which 28 observations are censored. The largest uncensored observation is at 1358 days.

In Fig. 1 we compare logspline density estimates for this data based upon the current algorithm and based upon the algorithm of Kooperberg & Stone (1992). The Kooperberg–Stone algorithm, which can deal with censored data but not with truncated data, uses a full likelihood. Censoring is dealt with by integrating the density over the subset of $\mathscr{R}$ in which it is known that an observation is located. Their algorithm also employs cubic splines and stepwise deletion of

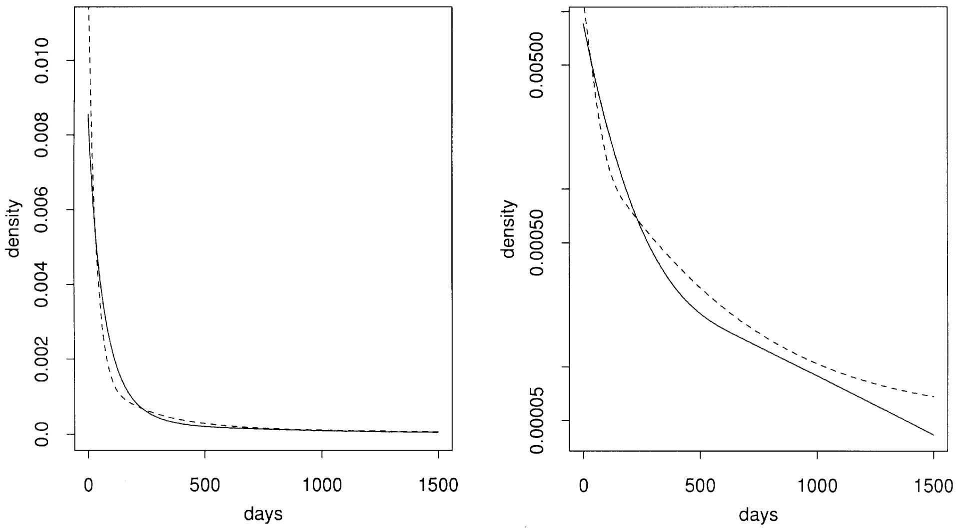

Fig. 1. Density estimates for the Stanford Heart Transplant Data ($n = 103$, 28 censored). Solid—logspline method of current paper; dashed—Kooperberg & Stone (1992).

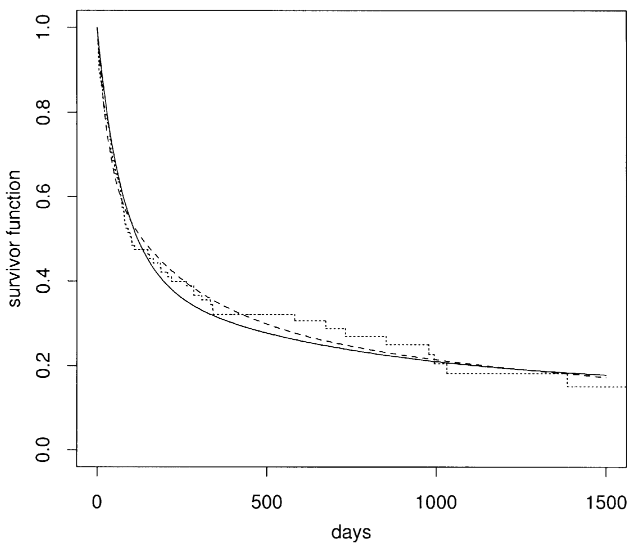

Fig. 2. Survivor function estimates for the Stanford Heart Transplant Data ($n = 103$, 28 censored). Solid—logspline method of current paper; dashed—Kooperberg & Stone (1992); dotted—Kaplan–Meier estimate.

knots, however, it does not allow for stepwise addition. In Fig. 2 we show the estimates of the survivor function corresponding to the two density estimates in Fig. 1 together with the Kaplan–Meier estimate.

As can be seen from these plots, both density estimates are similar, and the corresponding survivor curves follow the Kaplan–Meier curve closely. It is interesting to note that the knots for both estimates are at very different locations: the current estimate has three knots, located at 3, 12, and 675 days, the Kooperberg & Stone (1992) estimate has four knots, located at 1, 56, 86, and 1799 days (in that algorithm knots at the extreme observation were required). Logspline density estimation seems to be very robust when applied to this data: no matter where the knots are positioned, or what options are chosen, the estimate that is obtained is very similar.

The data set for our second example is the Fyn diabetes data, which is extensively discussed in Andersen *et al.* (1993). This data set consists of 1499 diabetes in the County of Fyn in Denmark. The data were collected by Green *et al.* (1981). For each person the data set contains the age at diagnosis (D_i) (in years) of diabetes, the age at which the person enrolled in the study (T_i), and the age at which the person left the study (Y_i), either because the participant moved (censoring), the study ended (censoring) or the participant died (uncensored). The participants who stayed in the study until the end were followed for 7.5 years. For each participants we know the gender, and whether the participant was censored or died (δ_i). There are 783 men, of whom 254 died, and 716 women, of whom 237 died. The survival data is left-truncated by the age at which the participant entered the study.

It is of interest to assess how different the survival distributions for the Fyn diabetics are from the general Danish population. In Fig. 3 we show the logspline density estimate of the survival distribution of Fyn diabetics, separately for men and women, based upon the left-truncated sample (Y_i^o, δ_i^o, T_i^o). As a comparison we have drawn the general survival distribution based upon the Danish vital statistics for 1975. The youngest deaths in the Fyn data are 19 (women) and 23 (men) years old. Formally we can thus only consider the survival distribution, conditional on surviving until age 19 or 23. However, since few people die before that age, we ignore this distinction.

In Figs 4 and 5 we draw the corresponding hazard functions (on a logarithmic scale) and survivor functions. It is interesting to note that the ratio between the fitted hazard function and the hazard function based on the Danish vital statistics is very similar for men and women: for both genders it is approximately ten for an age of about 30 years and drops to about three for ages between 40 and 60 years. We also notice that the survivor functions corresponding to the logspline estimates closely follow the Kaplan–Meier estimates.

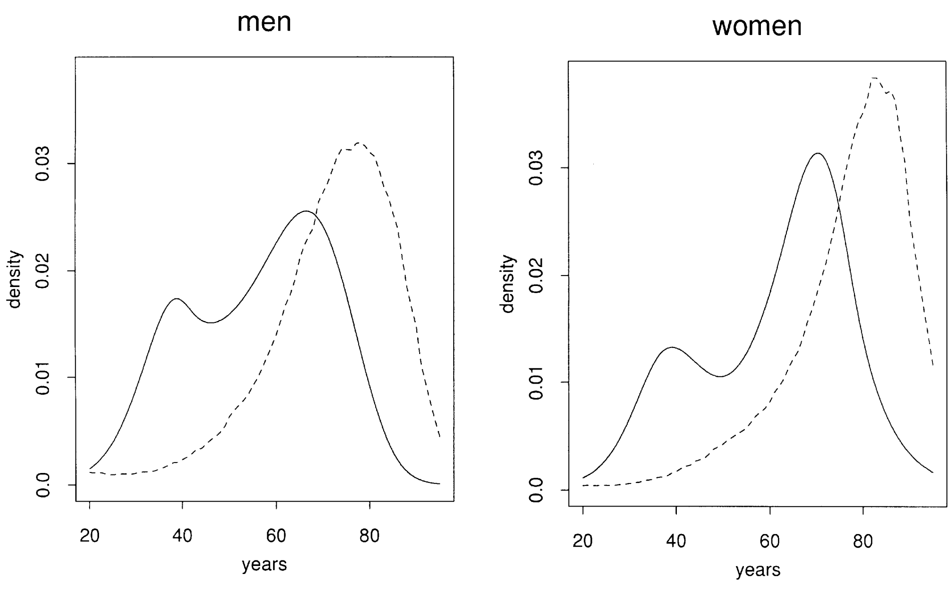

Fig. 3. Logspline density estimates for the age of participants in the Fyn diabetes study (solid), and the Danish vital statistics for 1975 (dashed).

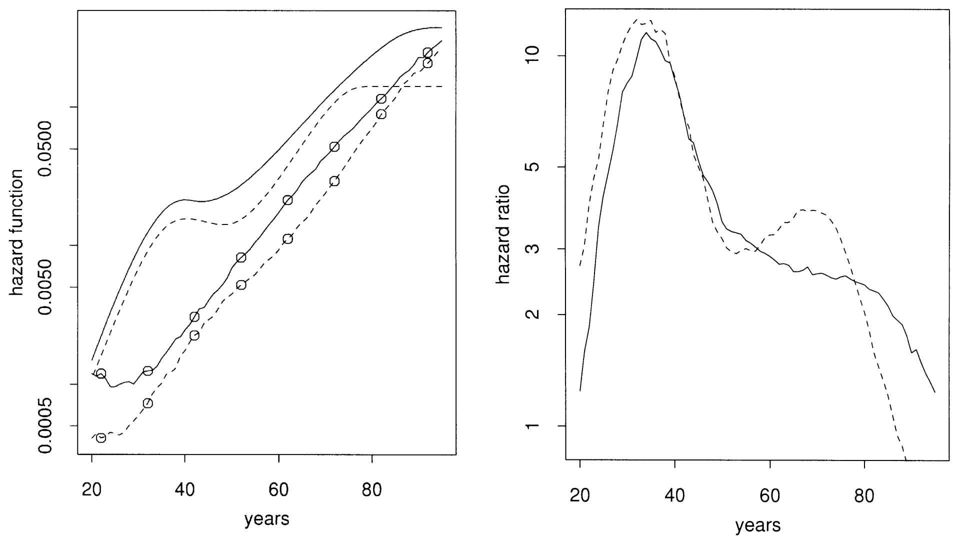

Fig. 4. Left: hazard functions corresponding to the logspline estimates for the Fyn diabetes study (no marks), and the Danish vital statistics for 1975 (marks). Right: ratio between the logspline estimate of the hazard function and the hazard function based upon the vital statistics. Solid—men; dashed—women.

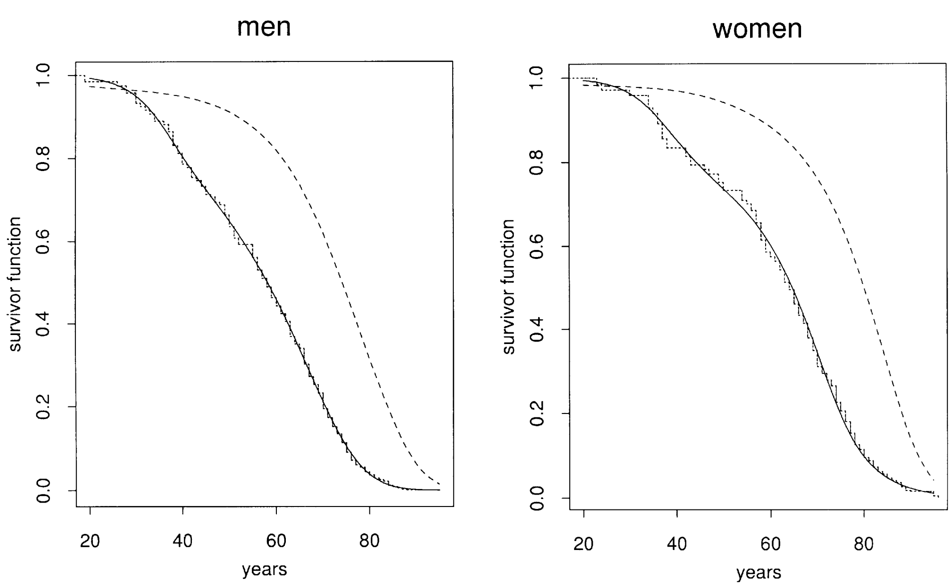

Fig. 5. Logspline estimates for the survivor function of the age of participants in the Fyn diabetes study (solid), the Kaplan–Meier estimate (dotted), and the Danish vital statistics for 1975 (dashed).

As the survival densities estimated by the logspline method in Fig. 3 look quite different from those of the general population, we carried out a small simulation study to rule out any possibility that such a difference could arise as an artefact of the logspline density estimation procedure, as well as to establish consistency of the logspline estimates. To carry out this simulation, we generated data roughly according to the same sampling procedure that was used for the Fyn data, but we used the Danish vital statistics to generate the survival times. Since the simulation procedure is fairly complicated, we describe it in detail below. All steps are carried out separately for both genders. All ages and data were truncated to the largest integer smaller than or equal to the actual age.

1. The sample of ages of entering in the study (T_i) is left-truncated by the age at diagnosis (D_i). We estimated f_T using the logspline density estimation procedure applied to the data $(T_i^o, 1, D_i^o)$. The logspline estimates $\hat{f}_T$ are shown in Fig. 6.
2. The density f_D of the age at diagnosis can be estimated by applying the logspline procedure on D_i^o. The data is not censored or truncated. The estimates for men and women are shown in Fig. 7.
3. Because of the rounding of the ages in integers, participants had a 0.5 probability of staying 7 years in the study, and a 0.5 probability of staying 8 years in the study.
4. For each simulation we proceeded as follows:
   (a) We generated an i.i.d. sample D_i^o of size n for the age of diagnosis from $\hat{f}_D$.
   (b) We generated n independent samples $T_i^o \geqslant D_i^o$ for the age of entering in the study from $\hat{f}_T$.
   (c) We generated n independent samples $X_i^o \geqslant T_i^o$ for the survival times from the Danish vital statistics.
   (d) We generated $C_i^o = T_i^o + 7 + B_i$, $i = 1, \ldots, n$, for the censoring times, where B_i are independent Bernoulli random variables with $P(B_i = 1) = 0.5$.
   (e) We set $Y_i^o = \min(X_i^o, C_i^o)$ and $\delta_i^o = I(X_i^o \leqslant C_i^o)$.
   (f) We applied the logspline density estimation algorithm to $(Y_i^o, \delta_i^o, T_i^o)$ to obtain an estimate $\hat{f}^o$ of the density of the Danish vital statistics.
   (g) We numerically computed the integrated squared difference (ISD) from age 1 year to age 90 years between $\hat{f}^o$ and the Danish vital statistics.

In Table 1 we provide the mean over 250 simulations of the ISD (the standard error of each of

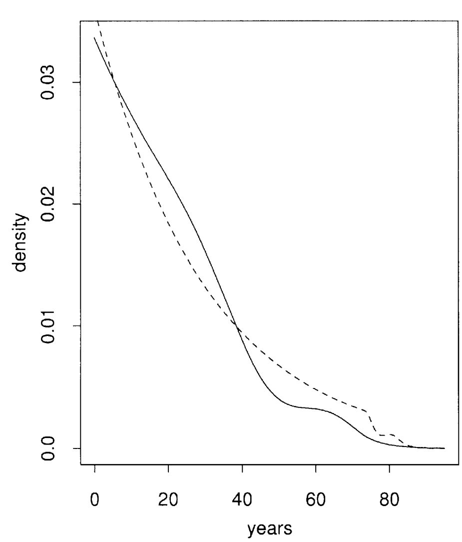

Fig. 6. Logspline density estimates for the age of entering the study for the Fyn diabetes study. Solid—men; dashed—women.

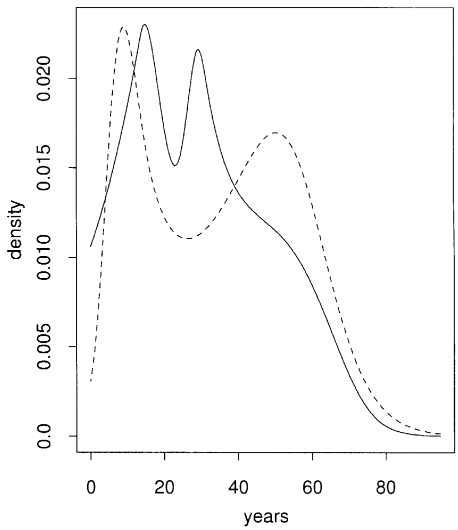

Fig. 7. Logspline density estimates for the age at diagnosis of diabetes for the Fyn diabetes study. Solid—men; dashed—women.

Table 1. *Integrated squared difference between the Danish vital statistics and the logspline estimates from that distribution, using a sampling scheme similar to that of the Fyn diabetes study*

	Men		Women	
n	MSD	$\sum I(\text{MSD} < 0.01942)$	MSD	$\sum I(\text{MSD} < 0.01395)$
200	0.00328	250	0.00408	248
716			0.00180	250
783	0.00130	250		
1000	0.00118	250	0.00145	250
5000	0.00047	250	0.00045	250

these means is approximately 0.0001), as well as how often out of these 250 simulations the ISD was smaller than the ISD of 0.01942 for the men and 0.01395 for the women between the logspline estimate of the survival density of the Fyn data and the Danish vital statistics. From this table it is clear that it is extremely unlikely that a sample from the Danish vital statistics would have given rise to ISDs as large as 0.01942 or 0.01395. The fact that the ISD goes down when the sample size increases suggests consistency of the estimates. Because of the discretization of the Danish vital statistics, we would not expect the ISD to go down to zero completely, thus the current results are indeed very promising. In Fig. 8 we show for both men and women four randomly selected density estimates from our simulation using same sample sizes of 783 for men and 716 for women from the Fyn data, as well as the underlying Danish vital statistics. From this plot it would appear as if the logspline estimates in the current situation sometimes overestimate the height of the peak. However, when we increased the sample size, we did not observe any such over estimating.

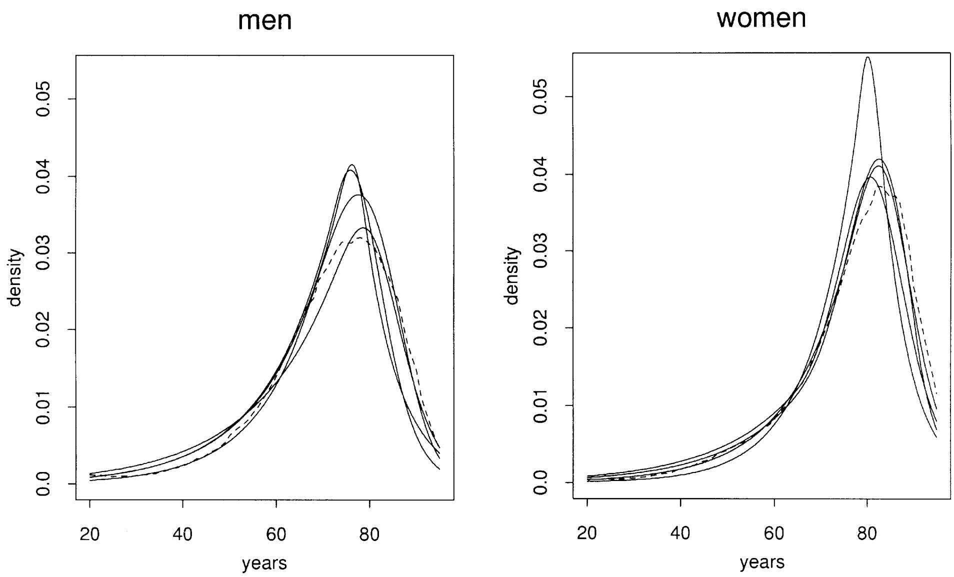

Fig. 8. Logspline density estimates for four samples of the Danish vital statistics for 1975, using a sampling mechanism similar to the Fyn diabetes study. Solid—simulations; dashed—Danish vital statistics.

6. Proof of asymptotic results

Proof of theorem 1. Let $s(\boldsymbol{\eta})$ be the approximation of g_{01}, which is assumed to satisfy the L_2 and L_∞ bounds on the error $g_{01} - s(\boldsymbol{\eta})$. Define $\tilde{\boldsymbol{\eta}} = (\eta_1 - \overline{\eta}, \ldots, \eta_J - \overline{\eta})$, where $\overline{\eta} = (1/J)\sum_k \eta_k$. Set $\boldsymbol{\beta}_0 = \int \mathbf{B} f(\tilde{\boldsymbol{\eta}})$ and $\boldsymbol{\beta} = \int \mathbf{B} f_{01}$, so that $\int \mathbf{B}\{f_{01} - f(\tilde{\boldsymbol{\eta}})\}$. It follows from (2.3), the properties of B-splines, and the Cauchy–Schwarz inequality that for any function h

$$\left|\int \mathbf{B}h\right|^2 \leqslant \frac{M_4^2}{J}\int h^2,$$

which implies that

$$|\boldsymbol{\beta} - \boldsymbol{\beta}_0| \leqslant \frac{M_4}{\sqrt{J}} \| f_{01} - f(\tilde{\eta}) \|_2. \tag{6.1}$$

Lem. 2 of Barron & Sheu (1991) and (6.1) yield

$$|\boldsymbol{\beta} - \boldsymbol{\beta}_0| \leqslant \frac{M_3 M_4}{\sqrt{J}} \exp(2\gamma)\varDelta.$$

Now apply lemma 2 with $\boldsymbol{\theta}_0 = \tilde{\eta}$ and $\boldsymbol{\beta} = \int \mathbf{B} f_{01}$ to obtain the desired result. (Compare the proof of th. 3 of Barron & Sheu, 1991.) This completes the proof of theorem 1.

Let

$$\bar{l}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \overline{w}_i^o s(Y_i^o; \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}),$$

where $\overline{w}_i^o = \{nF(1|0)\}^{-1} I(0 \leqslant Y_i^o \leqslant 1)\delta_i^o S(Y_i^o|0)/\{R(Y_i^o)/\pi\}$. Note that $\overline{f} = f(\cdot; \overline{\boldsymbol{\theta}})$ is the density in the logspline family (2.1) that satisfies

$$\int \mathbf{B}\overline{f} = \overline{\mathbf{B}},$$

where $\overline{\mathbf{B}} = (\overline{B}_{j,1}, \ldots, \overline{B}_{j,J})'$ and $\overline{B}_{j,k} = \sum_i \overline{w}_i^o B_{j,k}(Y_i^o)$.

Lemma 4

If (A2)–(A3) hold, then

$$|\hat{\mathbf{B}} - \overline{\mathbf{B}}| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Observe that

$$|\hat{\mathbf{B}} - \overline{\mathbf{B}}|^2 = \sum_k \left\{ \sum_i (\hat{w}_i^o - \overline{w}_i^o) B_{j,k}(Y_i^o) \right\}^2$$

$$\leqslant \sup_{t\in\tau} \left| \frac{1}{\hat{F}(1|0)} \frac{n\hat{S}(t|0)}{\#(t)} - \frac{1}{F(1|0)} \frac{S(t|0)}{R(t)/\pi} \right|^2 \sum_k \left\{ \sum_i \frac{\delta_i^o I(0 \leqslant Y_i^o \leqslant 1)}{n} B_{j,k}(Y_i^o) \right\}^2. \quad (6.2)$$

Using (A2), (A3), and the results in Lai & Ying (1991), we get that

$$\sup_{t\in\tau} \left| \frac{1}{\hat{F}(1|0)} \frac{n\hat{S}(t|0)}{\#(t)} - \frac{1}{F(1|0)} \frac{S(t|0)}{R(t)/\pi} \right|^2 = O_P\left(\frac{1}{\sqrt{n}}\right). \quad (6.3)$$

It follows from properties of B-splines, and the Cauchy–Schwartz inequality that

$$E \sum_k \left\{ \sum_i \frac{\delta_i^o I(0 \leqslant Y_i^o \leqslant 1)}{n} B_{j,k}(Y_i^o) \right\}^2$$

$$\leqslant \frac{1}{n^2} \sum_k \sum_{i_1}^n \sum_{i_2}^n E B_{j,k}(Y_{i_1}^o) I(0 \leqslant Y_{i_1}^o \leqslant 1) B_{j,k}(Y_{i_2}^o) I(0 \leqslant Y_{i_2}^o \leqslant 1)$$

$$\leqslant \sum_k E B_{j,k}^2(Y_1^o) I(0 \leqslant Y_1^o \leqslant 1) \leqslant 1. \quad (6.4)$$

Lemma 4 now follows from (6.2)–(6.4).

Lemma 5

If (A2)–(A3) hold, then

$$E\{\overline{w}_i^o B_{j,k}(Y_i^o)\} = \frac{1}{n} \int_0^1 B_{j,k}(t) f_{01}(t)\, dt.$$

Proof. Observe that

$$E\{\overline{w}_i^o B_{j,k}(Y_i^o)\} = E\left\{ \frac{I(T_1 \leqslant X_1 \leqslant C_1) I(0 \leqslant X_1 \leqslant 1)}{nF(1|0)} \frac{S(X_1|0)}{R(X_1)} B_{j,k}(X_1) \right\}.$$

By using a conditioning argument and the definition of $G(\cdot)$, we get that

$$E\{\overline{w}_i^o B_{j,k}(Y_i^o)\} = \frac{1}{nF(1|0)} E\left\{ I(0 \leqslant X_1 \leqslant 1) \frac{S(X_1|0)}{R(X_1)} G(X_1) B_{j,k}(X_1) \right\}. \quad (6.5)$$

Since

$$\frac{S(t|0)}{R(t)} G(t) = \frac{1}{P(X_1 \geqslant 0)} \quad (6.6)$$

and

$$F(1|0)P(X_1 \geqslant 0) = P(0 \leqslant X_1 \leqslant 1). \tag{6.7}$$

the desired result follows from (6.5)–(6.7).

Lemma 6

If (A2)–(A3) hold, then

$$|\overline{\mathbf{B}} - \mathbf{B}^*| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof. It follows from lemma 5 that

$$E|\overline{\mathbf{B}} - \mathbf{B}^*|^2 = n\sum_k \operatorname{var}\{\overline{w}_1 B_{j,k}(Y_1^o)\} \leqslant n\sum_k E\{\overline{w}_1 B_{j,k}(Y_1^o)\}^2.$$

It follows from the (A1)–(A3) and properties of B-splines that

$$E|\overline{\mathbf{B}} - \mathbf{B}^*|^2 \leqslant \frac{1}{n}\sum_k \frac{1}{F(1|0)^2} E\left\{\delta_1^o I(0 \leqslant Y_1^o \leqslant 1)\frac{S(Y_i^o|0)}{R(Y_i^o)/\pi} B_{j,k}(Y_1^o)\right\}^2 \leqslant \frac{M}{n}.$$

Hence the result of lemma 6 is valid.

Proof of theorem 2. By the triangular inequality and lemmas 4 and 6, we know that $|\hat{\mathbf{B}} - \mathbf{B}^*| = O_P(n^{-1/2})$. Now apply lemma 2 with $\boldsymbol{\beta}_0 = \int \mathbf{B} f^* = \mathbf{B}^*$ and $\boldsymbol{\beta} = \hat{\mathbf{B}}$. If $J/\sqrt{n} \to 0$ as $n \to \infty$, one can show that the MILE $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\mathbf{B}})$ exists, and that

$$D(f^*\|\hat{f}) \leqslant M\mathscr{K}\frac{J}{n}$$

except on a set of probability less than $1/\mathscr{K}$. (See the proof of th. 3 of Barron & Sheu, 1991.) This completes the proof of theorem 2.

Lemma 7

If (A4)–(A5) hold, then (i) $M_3^{-1} \leqslant f_{01}(x) \leqslant M_3$ *for* $x \in \mathscr{T}$; *and (ii)* $\Delta = O(J^{-\alpha})$ *and* $\gamma = O(J^{-\alpha+1/p})$.

Proof. Write

$$g_{01} = \sum_{j=0}^{\infty} \sum_{k \in \Lambda(j)} \eta_{j,k} B_{j,k}.$$

Using (3.3) and (A4), we get that

$$|\boldsymbol{\eta}_j|_p \leqslant M2^{-j(\alpha-1/p)}. \tag{6.8}$$

Using (2.4) and (6.8), we obtain

$$\left|\sum_{k\in\Lambda(j)} \eta_{j,k}\beta_{j,k}(x)\right| \leqslant |\boldsymbol{\eta}_j|_\infty \leqslant |\boldsymbol{\eta}_j|_p \leqslant M2^{-j(\alpha-1/P)}. \tag{6.9}$$

If follows from (6.9) and (A5) that

$$|g_{01}(x)| \leqslant \sum_j \left|\sum_k \eta_{j,k} B_{j,k}(x)\right| \leqslant M\sum_j 2^{-j(\alpha-1/p)} = O(1),$$

which implies (i). Let $A_j(g_{01}) = \sum_{m=0}^{j} \sum_{k \in \Lambda(m)} \eta_{m,k} B_{m,k}$. Note that

$$|\boldsymbol{\eta}_j|_u \leqslant |\Lambda(j)|^{(1/u-1/v)} |\boldsymbol{\eta}_j|_v, \quad \text{for } 1 \leqslant u \leqslant v \leqslant \infty, \tag{6.10}$$

where $|\Lambda(j)|$ is number of elements in $\Lambda(j)$. Using (2.3), (6.9) and (6.10) we get that

$$\Delta \leqslant \sum_{m>j} \left\| \sum_k \eta_{m,k} B_{m,k} \right\|_2 \leqslant M \sum_{m>j} 2^{-m/2} |\eta_m|_2 \leqslant M \sum_{m>j} 2^{-\alpha m} = O(J^{-\alpha}),$$

and

$$\gamma \leqslant \sum_{m>j} \left\| \sum_k \eta_{m,k} B_{m,k} \right\|_\infty \leqslant M \sum_{m>j} 2^{-m(\alpha-1/p)} = O(J^{-\alpha+1/p}).$$

This completes the proof of lemma 7.

Proof of theorem 3. Assume that (A2)–(A5) hold. Choose $J \asymp n^{1/(2\alpha+1)}$. From lemma 7 it follows that $\Delta\sqrt{J} = O(J^{-(\alpha-1/2)}) = o(1)$ and $\gamma = O(J^{-(\alpha-1/2)}) = o(1)$. Theorem 1 now implies that $D(f_{01}\|f^*) = O(J^{-2\alpha}) = O(n^{-2\alpha/(2\alpha+1)})$. On the other hand, $J/\sqrt{n} \asymp n^{(1-2\alpha)/(4\alpha+2)} = o(1)$. Theorem 2 implies that $D(f^*\|\hat{f}) = O_P(J/n) = O_p(n^{-2\alpha/(2\alpha+1)})$. Since $D(f_{01}\|\hat{f}) = D(f_{01}\|f^*) + D(f^*\|\hat{f})$ (lemma 3), the proof of theorem 3 is now complete.

Acknowledgements

Ja-Yong Koo was supported in part by the Basic Science Research Institute Program, Ministry of Education of Korea, 1996 Project No. 1418. Charles Kooperberg was supported in part by NSF grant DMS-9403371. This work was done while the first author was visiting Stanford University; he wishes to thank the faculty and staff of Stanford University's Statistics Department for their hospitality. The authors would like to express their sincere thanks to Tze L. Lai and Charles J. Stone for many helpful comments. The Fyn diabetes data were collected by Dr Anders Green. Dr Per Kragh Andersen kindly provided the Fyn data to the authors.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer, New York.

Barron, A. R. & Sheu, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19**, 1347–1369.

de Boor, C. (1978). *A practical guide to splines*. Springer, New York.

DeVore, R. & Popov, V. (1988). Interpolation of Besov spaces. *J. Amer. Math. Soc.* **305**, 397–414.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508–539.

Fleming, T. R. & Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.

Gijbels, I. & Wang, J.-L. (1993). Strong representations of the survival function estimator for truncated and censored data with applications. *J. Multivariate Anal.* **47**, 210–229.

Green, A., Hauge, M., Holm, N. V. & Rasch, L. L. (1981). Epidemiological studies of diabetes mellitus in Denmark. II. A prevalence study based on insulin prescriptions. *Diabetologia* **28**, 339–342.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data*. Wiley, New York.

Koo, J.-Y. (1996). Bivariate B-splines in tensor logspline density estimation. *Comput. Statist. Data Anal.* **21**, 31–42.

Koo, J.-Y. & Chung, H.-Y. (1998). Log density estimation in linear inverse problems. *Ann. Statist.* **26**, 335–362.

Koo, J.-Y. & Kim, W.-C. (1996). Wavelet density estimation by approximation of log-densities. *Statist. Probab. Lett.* **26**, 271–278.

Koo, J.-Y. & Park, B. U. (1996). B-spline deconvolution based on the EM algorithm. *J. Statist. Comput. Simulation* **54**, 275–288.

Kooperberg, C. (1997). Bivariate density estimation with an application to survival data. Technical report No. 296, Department of Statistics, University of Washington.

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

Kooperberg, C. & Stone, C. J. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist*. **1**, 301–328.

Lai, T. L. & Ying, Z. (1991). Estimating a distribution function with truncated and censored data. *Ann. Statist.* **19**, 417–442.

Marron, J. S. & Padgett, W. J. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *Ann. Statist.* **15**, 1520–1535.

Silverman, B. W. (1986). *Density estimation and data analysis.* Chapman & Hall, London.

Stone, C. J. (1989). Uniform error bounds involving logspline models. In *Probability statistics and mathematics: papers in honor of Samuel Karlin* (eds T. W. Anderson, K. B. Athrya & D. L. Iglehart), 335–355. Academic Press, Boston.

Stone, C. J. (1990). Large sample inference for logspline model. *Ann. Statist*. **18**, 717–741.

Stone, C. J. & Koo, C.-Y. (1986). Logspline density estimation. *Contemp. Math.* **59**, 1–15.

Stone, C. J., Hansen, M., Kooperberg, C. & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371–1470.

Received June 1996, in final form April 1998

Ja-Yong Koo, Department of Statistics, Hallym University, Chunchon, Kangwon-Do 200–702, Korea. e-mail: jykoo@sun.hellym.ac.kp.

Charles Kooperberg. Fred Hutchinson Cancer Research Center, Divisions of Public Health Sciences (MP 1002, 1100 Fairview Ave N, Seattle, WA 98109-1024, USA. e-mail: clk@bear.fhcrc.org.

Jinho Park, Department of Statistics, Inha University, Inchon, 402-751, Korea. e-mail: jhpark@stat.inha.ac.kr.