ELSEVIER

Statistics & Probability Letters 46 (2000) 133–147

STATISTICS & PROBABILITY LETTERS

www.elsevier.nl/locate/stapro

Logspline density estimation for binned data

Ja-Yong Koo[a], Charles Kooperberg[b, *]

[a] *Department of Statistics, Hallym University, Chunchon, Kangwon-Do 200-702, South Korea*
[b] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue, N/MP-1002, Seattle, WA 98109-1024, USA*

Received December 1998; received in revised form March 1999

Abstract

In this paper we consider logspline density estimation for binned 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 binned data. Numerical examples are used to show the finite-sample performance of inference based on the logspline density estimation. © 2000 Elsevier Science B.V. All rights reserved

Keywords: Besov space; Binning; Knot selection; MILE; Optimal rate of convergence

1. Introduction

This paper proposes a method of density estimation for binned data. Let $X_1,\dots,X_n$ be a random sample from a distribution with density f. In some experiments, the data are reported in the form of a histogram. The observed random variables are

$$Y_q = \#\{X_i\colon X_i \in I_q\},$$

where I_q are bins. We want to estimate the unknown density function f based on Y_q's.

Flexible exponential families have been used for the estimation of density functions. Stone and Koo (1986), Stone (1989, 1990), Kooperberg and Stone (1991), Koo (1996), and Stone et al. (1997) developed logspline density estimation, in which the logarithm of a probability density function is modeled using polynomial splines. Koo et al. (1998, [KKP] hereafter) studied logspline density estimation under truncation and censoring. Barron and Sheu (1991) studied density estimation procedures based on trigonometric series, polynomials, and splines. Koo and Kim (1996) considered an exponential family based on wavelets. Exponential families have been used by Koo and Park (1996) and Koo and 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 on binned data. Kooperberg and Stone (1992) developed logspline density estimation for univariate data that may be right-censored, left-censored or

[*] Corresponding author.

PII: S0167-7152(99)00097-8

interval-censored. Antoniadis et al. (1998) considered a wavelet method for density and hazard rate estimation based on binning the data. Minnotte (1998) discusses the histospline procedure for binned data, in which the unknown density is estimated by a spline, such that the integral of the density estimate equals the mass of the histogram for any bin. Interestingly, Minnotte (1998) obtains the same rate of convergence for the mean-squared error as those obtained in this paper. Minnotte, however, assumes smoothness of the true underlying density, whereas we assume smoothness for the log-density. There have been a number of papers that consider kernel density estimation for binned data. See the introduction of Minnotte (1998) for details. Hall and Wand (1996) find that in the context of kernel density estimation, linear binning yields more efficient kernel density estimates. In this paper we only consider constant binning, since we see our methodology as a procedure for data that has already be binned (constantly). This contrasts with linear binning for kernel density estimation, which is considered part of a more computationally efficient kernel density estimation procedure for unbinned data.

Consider logspline density estimation without binning, so that $X_1,\dots,X_n$ are actually observed. Let $B_1,\dots,B_J$ 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

$$\int B_k(x) f(x;\tilde{\boldsymbol{\theta}})\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int B_k(X_i) \tag{1}$$

for $k=1,\dots,J$.

We use binned data to find an appropriate estimator $\hat{B}_k$, whose expectation is asymptotically the same as $EB_k(X)$. The proposed density estimator has the form $f(\cdot;\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ satisfies (1) with $n^{-1}\sum_{i=1}^{n}\int B_k(X_i)$ replaced by $\hat{B}_k$. 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 histospline estimate of Minnotte (1998) may be negative in small regions. The wavelet estimates of Antoniadis et al. (1998) may have negative values so that they consider $\hat{f}_+(t)=\max(\hat{f}_n(t),0)$ as a simple way of guaranteeing positivity, where $\hat{f}_n$ is the wavelet density estimator having the form of an orthogonal series estimator. The approach of Kooperberg and Stone (1992) has the advantage that the estimate of the parameters can be justified in the context of maximum likelihood, but it has the disadvantage that the resulting log-likelihood is not necessarily concave when binning is present. While with a moderate amount of binning the lack of concavity does not seem many problems in practice, it is much harder to establish theoretical results.

In this paper it is shown that the logspline density estimates based upon binning possess the rate of convergence $n^{-2\sigma/(2\sigma+1)}$, where σ 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 estimator $\hat{B}_k$ may converge sufficiently fast when the length of each bin decreases to zero fast enough. Thus the rate of convergence for logspline density estimation based on the binning is the same as the rate of convergence for logspline density estimation based upon unbinned 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. 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 and information projection

In this section we define logspline densities on the unit interval $I=[0,1]$. To simplify notation the dependence on the sample size n of various quantities will be suppressed. In the remainder of this paper $M, M_1, M_2, \dots$ are positive constants, independent of n, and C is a positive constant, also independent of n, that is only used locally and may be different in different locations.

Let B^r be the B-spline of order r having knots at $0, 1, \dots, r$ so that

$$B^r(x) = r[0,1,\dots,r](\cdot - x)_+^{r-1}$$

using the divided difference notation (de Boor (1978)). Let j be a positive integer, which will depend on n, and define

$$B_{j,k} = B^r(2^j x - k), \quad k \in \mathbb{Z}.$$

To approximate a function on I, we only need those B-splines $B_{j,k}$ which do not vanish identically on I. 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. (Our implementation, discussed in Section 4, does not assume that the number of bins is a power of 2.) The dimension of $\mathscr{S}_j$ is $J = 2^j + r - 1$. Let $J \geqslant 2$ for all n.

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

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

$$\psi(\boldsymbol{\theta}) = \log \left[\int_I \exp\{s(x;\boldsymbol{\theta})\}\,\mathrm{d}x \right]$$

and

$$f(\cdot\,;\boldsymbol{\theta}) = \exp\{s(\cdot\,;\boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\}. \tag{2}$$

Then $\int_I f(x;\boldsymbol{\theta})\,\mathrm{d}x = 1$ for $\boldsymbol{\theta} \in \Theta$. For notational convenience, let $f(\boldsymbol{\theta})$ denote the function $f(\cdot\,;\boldsymbol{\theta})$. 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 and g is defined as

$$D(f\|g) = \int_I f \log \frac{f}{g}.$$

For a function h, let $\int \boldsymbol{B} h$ denote the J-dimensional vector of elements $\int_I B_{j,k}(x)h(x)\,\mathrm{d}x$, where $\boldsymbol{B} = (B_{j,1}, \dots, B_{j,J})'$. Given $\boldsymbol{\beta} \in \Theta$, let $\boldsymbol{\theta}(\boldsymbol{\beta}) \in \Theta^0$ denote a solution to the equation

$$\int \boldsymbol{B} f(\boldsymbol{\theta}) = \boldsymbol{\beta}. \tag{3}$$

Let

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

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

$$\int B_{j,k} f = \int B_{j,k} f^* \quad \text{for } k = 1, \dots, J. \tag{4}$$

Let $\boldsymbol{B}^* = (B^*_{j,1},\dots,B^*_{j,J})'$ where $B^*_{j,k} = \int B_{j,k} f$. From [KKP], we know that the information projection f^* is unique if it exits and that $\boldsymbol{\theta}^* = \boldsymbol{\theta}(\boldsymbol{B}^*)$.

Let

$$\gamma_2 = \inf_{s\in\mathscr{S}_j} \|\log f - s\|_2$$

and

$$\gamma_\infty = \inf_{s\in\mathscr{S}_j} \|\log f - s\|_\infty$$

be the L_2 and L_∞ error in the approximations of g by some $s \in \mathscr{S}_j$. Theorem 1 establishes an upper bound on the approximation error $D(f\|f^*)$ in terms of γ_2 under the condition

(A1) $M_1^{-1} \leqslant f \leqslant M_1$.

Theorem 1. *If* (A1) *holds*, γ_∞ *is bounded, and* $\sqrt{J}\gamma_2 = \mathrm{o}(1)$, *then the information projection* f^* *uniquely exists and satisfies*

$$D(f\|f^*) \leqslant \frac{M_1}{2} \mathrm{e}^{\gamma_\infty} \gamma_2^2.$$

Proof. Refer to the proof of Theorem 1 in [KKP]. □

3. Asymptotic results

Let N be an integer that may depend on n and let $Q = 2^N$. Define the dyadic intervals

$$I_q = \left[\frac{q-1}{Q}, \frac{q}{Q}\right) \quad \text{for } 1 \leqslant q < Q-1 \quad \text{and} \quad I_Q = \left[\frac{Q-1}{Q}, 1\right]. \tag{5}$$

Let x_q be the center point of each subinterval defined by

$$x_q = \frac{q - 1/2}{Q}, \quad q = 1,\dots,Q. \tag{6}$$

As an estimator $\hat{B}_{j,k}$ of $EB_{j,k}(X)$, consider

$$\hat{B}_{j,k} = \sum_{q=1}^{Q} B_{j,k}(x_q) Y_q, \tag{7}$$

where $Y_q = \#\{X_i\colon X_i \in I_q\}$. Define the incomplete likelihood function

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{J} \theta_k \hat{B}_{j,k}(x_q) - \psi(\boldsymbol{\theta}). \tag{8}$$

Note that the incomplete likelihood function defined by (8) 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 binned data. Let

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta^0} l(\boldsymbol{\theta}) \tag{9}$$

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.

Let $\hat{\boldsymbol{B}} = (\hat{B}_{j,k})_{k=1}^{J}$. From the incomplete likelihood equation, $\hat{\boldsymbol{\theta}}$ is the parameter that satisfies

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

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

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

Theorem 2. *If the sequence γ_∞ is bounded, $J/Q^2 = \mathrm{O}(n^{-1})$, and $J/\sqrt{n} \to 0$, then $\hat{f}$ exists, except on an event whose probability tends to zero with n, and*

$$D(f^* \| \hat{f}) = \mathrm{O_P}\left(\frac{J^2}{Q^2} + \frac{J}{n}\right).$$

As a smoothness class for f, we use the Besov space. If $h \in L_p(I)$, $1 \leqslant p \leqslant \infty$, let $\omega_r(h,t)_p$, $t > 0$, denote the modulus of smoothness for $h : \omega_r(h,t)_p = \sup_{|u| \leqslant t} \|\Delta_u^r h(\cdot)\|_p(I(ru))$, where Δ_u^r is the rth order difference with step u; the norm in the above definition is the L_p norm on the set $I(ru) = \{x\colon x, x + ru \in I\}$. We say that h is in the Besov space $B_{\sigma pq}$ whenever

$$\|h\|_{B_{\sigma pq}} = \left[\int_0^\infty \{t^{-\sigma}\omega_r(h,t)_p\}^q \frac{\mathrm{d}t}{t}\right]^{1/q}$$

is finite, where r is any integer larger than σ.

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 and Popov (1988) we know that a function h in $B_{\sigma pq}$ can be written as

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

and

$$C^{-1}\|h\|_{B_{\sigma pq}} \leqslant \left[\sum_{j=0}^{\infty} \{2^{j(\sigma - 1/p)} |\boldsymbol{\eta}_j|_p\}^q\right]^{1/q} \leqslant C\|h\|_{B_{\sigma pq}}, \tag{10}$$

with the usual modification if either p or q equals ∞. Here $|\boldsymbol{\eta}_j|_p$ denotes $(\sum_{k \in \Lambda(j)} |\eta_{j,k}|^p)^{1/p}$. See DeVore and Popov (1988) and Donoho et al. (1996) for properties of Besov spaces. The Besov space includes the Hilbert–Sobolev space and Hölder space; spaces that are traditionally used in theoretical statistics. In particular, Stone (1990) studied logspline inference when the logarithm of the density function is in a Hölder space.

Let $\mathscr{F}_{\sigma pq}(M)$ be the set of functions defined by

$$\mathscr{F}_{\sigma pq}(M) = \{f : \log f \in B_{\sigma pq} \text{ and } \|\log f\|_{B_{\sigma pq}} < M\}.$$

Consider an unknown distribution P_f depending on the density function $f \in \mathscr{F}_{\sigma pq}(M)$ and suppose $\{b_n\}$ is some sequence of positive numbers. This sequence is called a *lower bound* for f if

$$\lim_{c \to 0} \liminf_{n} \inf_{\hat{T}} \sup_{f \in \mathscr{F}_{\sigma pq}(M)} P_f(D(f \| \hat{T}) \geqslant c b_n) = 1, \tag{11}$$

where the infimum is over all possible estimators f based on $Y_1, \ldots, Y_Q$. Alternatively, the sequence in question is said to be an *upper bound* for f if there is a sequence of estimators $\{\hat{f}_n\}$ of f such that

$$\lim_{c \to \infty} \limsup_{n} \sup_{f \in \mathscr{F}_{\sigma pq}(M)} P_f(D(f \| \hat{f}_n) \geqslant c b_n) = 0. \tag{12}$$

The sequence of numbers $\{b_n\}$ is called the *optimal rate of convergence* for f if it is both a lower bound and an upper bound with the associated estimators $\{\hat{f}_n \geqslant 1\}$, being called *asymptotically optimal.*

For the following theorems, we assume that

(A2) $f \in \mathscr{F}_{\sigma pq}(M)$.

Theorem 3 shows $n^{-2\sigma/(2\sigma+1)}$ is a lower bound for f in the metric D.

Theorem 3. *Under* (A2), *if* $s > 1/2$ *and* $1 \leqslant p, q \leqslant \infty$, *then*

$$\lim_{c\to 0} \inf_{\hat{T}} \sup_{f\in\mathscr{F}_{\sigma pq}(M)} P_f(D(f\|\hat{T}) \geqslant cn^{-2\sigma/(2\sigma+1)}) = 1,$$

where $\hat{T}$ *is any estimator of* f *based on* $Y_1, \ldots, Y_Q$.

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 4 we also assume that

(A3) $\frac{1}{2} < \sigma < r - 1 + \frac{1}{p}$ and $2 \leqslant p \leqslant \infty$.

The condition $\frac{1}{2} < \sigma$ is a sufficient condition for deriving that $\log f$ is bounded away from zero and infinity. The condition $\sigma < r - 1 + (1/p)$ was used in DeVore and Popov (1988). For fixed σ, r can be chosen large enough to derive bounds on γ_2 and γ_∞.

Theorem 4. *If* (A2) *and* (A3) *hold, then*

$$D(f\|\hat{f}) = \mathrm{O}_{\mathrm{P}}(n^{-2\sigma/(2\sigma+1)})$$

if $J \asymp n^{1/(2\sigma+1)}$ *and* $Q \asymp n^{(\sigma+1)/(2\sigma+1)}$.

According to Theorem 4, the logspline estimate achieves the lower bound $n^{-2\sigma/(2\sigma+1)}$, which means that the logspline estimators are asymptotically optimal.

4. Practical implementation

Logspline density estimation for binned data, based upon the incomplete log-likelihood (9), can be implemented using a logspline density estimation algorithm for complete data. 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 the data is binned. More details about the algorithm can be found in Kooperberg and 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 S 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$ 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 $\mathscr{I} = [L, U]$. We refer to the t_i, $i = 1, \ldots, K$, as knots.

Let $Z_1, \ldots, Z_n$ be the pseudo-sample consisting of Y_q copies of x_q, for $q = 1, \ldots, Q$. Thus, $n = \sum_{q=1}^{Q} Y_q$. For a given set of knots, the logspline density estimate for binned data is $f(x; \hat{\boldsymbol{\theta}})$, $L \leqslant x \leqslant U$, where

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{j=1}^{n} f(Z_j; \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{q=1}^{Q} Y_q f(x_q; \boldsymbol{\theta}),$$

is the MILE of $\boldsymbol{\theta}$ (compare with (9)). 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 and Stone (1992), and Stone et al. (1997) for details). Then stepwise knot addition is employed. At each stage every t that is a minimum number of order statistics from existing knots, is a candidate for addition. Among these candidates, the algorithm performs a heuristic search to maximize the Rao statistic for adding a knot $x = t$ to the current set of knots. Stepwise addition of knots is employed until a prespecified maximum number of knots is reached, the default of which is $K_{\max} = \min(8n^{0.2}, n/4, N - 1)$, where N is the number of bins for which $Y_q > 0$. Stone et al. (1997) used as the maximum number of knots $K_{\max} = \min(4n^{0.2}, n/4, N, 30)$ where N is the number of distinct data points. The change from N to $N - 1$ is the correction of a small error; the other changes are based on more recent experiences with large data sets.

Upon stopping the stepwise addition process, we carry out stepwise deletion. At each step the knot for which the Wald statistic for removal of the knot at $x = t$ from the current set of knots is the smallest in magnitude is removed.

During the combination of stepwise addition and stepwise deletion, we get a sequence of models indexed by v, with the vth model having J_v parameters. The (generalized) Akaike information criterion (AIC) can be used to select one model from this sequence. Let $\hat{l}_v$ denote the fitted log-likelihood for the vth model, and let $\mathrm{AIC}_{a,v} = -2\hat{l}_v + aJ_v$ be the Akaike information criterion with penalty parameter a for this model. We select the model corresponding to the value $\hat{v}$ of v that minimizes $\mathrm{AIC}_{a,v}$. In light of practical experience, we generally recommend choosing $a = \log n = \log \sum_q Y_q$ as in the Bayesian information criterion (BIC).

The possibly large number of repeated pseudo-observations when the data are binned can cause some numerical difficulties when positioning the knots. Clearly, it no longer suffices that knots are a minimum distance apart in order statistics: if a particular Y_q is large two potential knots t' and t'' could easily satisfy the rule of being several order statistics apart, while $t' = t'' = x_q$. Thus, we require that knots are located in different bins.

Hansen and Kooperberg (1999) consider a number of other approaches to knot selection for (unbinned) logspline density estimation and Triogram regression (Hansen et al. 1998) including a simulated annealing approach to minimize BIC, and a variety of Bayesian approaches using Markov chain Monte Carlo algorithms. They conclude that for logspline density estimation the stepwise procedure performs quite good compared to the other approaches that consider many more models during the model selection stage, but require much more cpu time.

The algorithm described in this section can be applied whether or not the assumptions made in Sections 2 and 3 (e.g. (A1)–(A3)) hold or not. In our experience, the logspline density estimation procedure works well, even if f is (very close to) zero: the density estimate will typically be very small, indistinguishable from zero, and even if f is (locally) not smooth, since the stepwise algorithm will position more knots close to the nonsmooth regions.

5. Numerical examples

Good and Gaskins (1980) give data on a mass-spectrum histogram that was gathered at the Lawrence Radiation Laboratory in Berkeley. It contains $n = 25,752$ events that are binned in 172 bins. In Fig. 1 we show the logspline density estimate for this data, as well as the raw counts in the various cells. The logspline density in Fig. 1 is based on 18 knots; the largest number of knots considered when fitting this data was 60. Modes I, IV, V, VI, VII, VIII, X and XIII of Good and Gaskins (1980) are clearly visible in this estimate. Of the five modes that Good and Gaskins (1980) identified that are not distinguishable, three were identified by

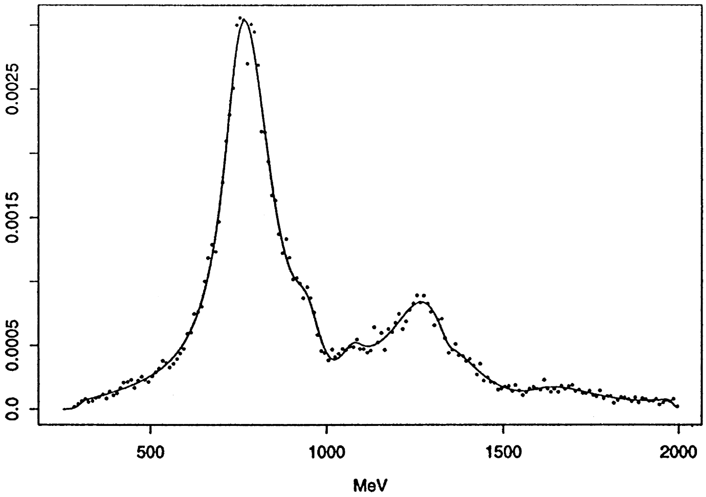

Fig. 1. Logspline density estimate for the LRL data of Good and Gaskins (1980).

them as not substantiated, the remaining two they indicated in the “flat” area between 1150 and 2000 MeV. We feel that this is a reasonable estimate of the density based upon the binned data. It took about 7 s of CPU time on a Sun ULTRA 2 workstation to compute the logspline density estimate.

To assess the asymptotic performance of logspline density estimation with binned data, we carried out a simulation study. We generated data from the bimodal densities that were considered in Kooperberg and Stone (1991):

$$f(y;\sigma) = 0.8g(y) + 0.2h(y;\sigma),$$

where $g(y)$ is the lognormal density of $Y = \exp(Z/2)$ and Z has a standard normal distribution, and $h(y)$ is the normal density with mean 2 and standard deviation σ. We use $\sigma = 0.03, 0.07, 0.17$, and 0.27. All but $\sigma = 0.03$ were also used in Kooperberg and Stone (1992); we expect that the density corresponding to $\sigma = 0.03$ has details that will be lost during estimation because of binning of the data. For each of the four distributions we generated 100 independent samples of size $n = 100, 300, 1000$, and $n = 3000$. For each sample we computed density estimates with bin widths of of 0.3, 0.1, and 0.03. To prevent artifacts that may be caused by bins starting or ending at fixed locations, we choose a uniform random origin. For each density estimate we computed the integrated squared error between the logspline density estimate and the true density, $\int(f-\hat{f})^2$. For each sample we also computed the ISE for the logspline density estimate for unbinned data, as we cannot expect the density estimate for binned data to do better than the one for unbinned data. The mean of the integrated squared error, MISE, for each of the simulation set-ups is shown in Table 1.

From Table 1 we notice that, as predicted by the asymptotic theory, the ISE does go down when the bin width decreases and when the sample size increases. It appears that for $\sigma = 0.03$ and $\sigma = 0.07$ and the larger bin sizes the MISE does not get real small for large n, essentially because with such wide bins to much signal is “lost” that it is impossible to recover the complete density no matter how large the sample size. For the densities with the larger values of sigma, the density is sufficiently regular that some amount of binning does not seem to hurt the estimation at all. The MISE for the estimates based on unbinned data is often as small as the MISE for the estimates based on binned data with a small bin width.

Summary statistics, such as those in Table 1 only tell part of the story. As pointed out in Kooperberg and Stone (1991), two estimates can have a similar integrated squared error, but they can qualitatively differ

Table 1
Mean-integrated squared error for simulations from three bimodal distributions

		Bin width				Bin width		
n	Not binned	0.3	0.1	0.03	Not binned	0.3	0.1	0.03
	$\sigma = 0.03$				$\sigma = 0.07$			
100	0.1602	0.3229	0.2212	0.1960	0.0633	0.1095	0.0663	0.0635
300	0.0378	0.2943	0.1364	0.0445	0.0218	0.0848	0.0251	0.0203
1000	0.0125	0.2915	0.1139	0.0205	0.0077	0.0796	0.0114	0.0069
3000	0.0060	0.1067	0.1166	0.0112	0.0033	0.0808	0.0083	0.0030
	$\sigma = 0.17$				$\sigma = 0.27$			
100	0.0434	0.0374	0.0333	0.0406	0.0392	0.0315	0.0340	0.0370
300	0.0161	0.0177	0.0134	0.0147	0.0152	0.0114	0.0120	0.0131
1000	0.0055	0.0127	0.0052	0.0051	0.0049	0.0092	0.0041	0.0040
3000	0.0022	0.0116	0.0022	0.0019	0.0019	0.0072	0.0016	0.0015

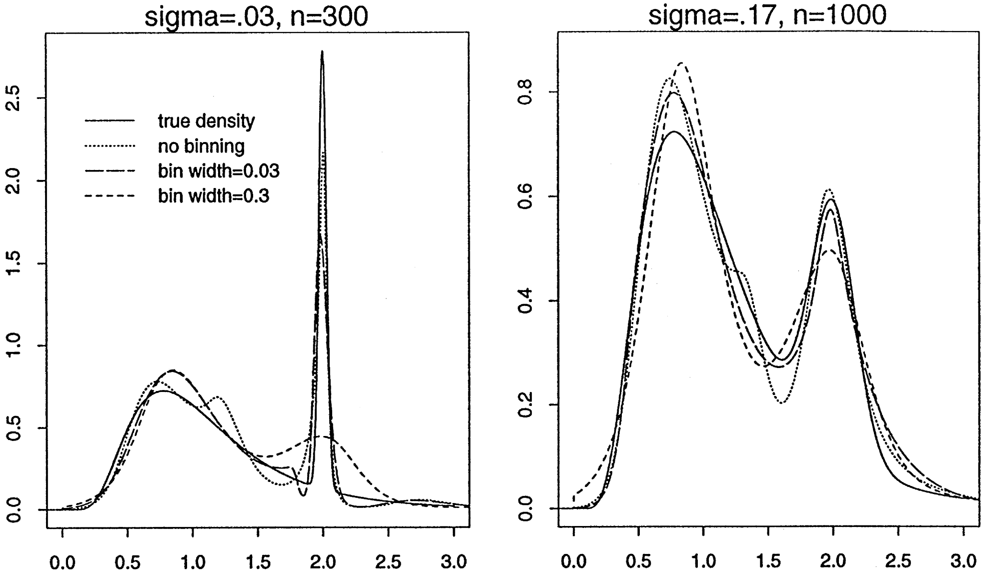

Fig. 2. Logspline density estimate for simulated data sets with various amounts of binning.

considerably. In Fig. 2 we show a typical estimate for $\sigma = 0.03$ with $n = 300$ and one for $\sigma = 0.017$ with $n = 1000$. (To keep these plots interpretable, we omitted the estimates with bin width 0.1.) The histograms width bin-width 0.03 and 0.3 for this data are shown in Fig. 3. The estimates in Fig. 2 have approximately average integrated squared error. We note that for the plot on the left, the estimate with bin width 0.3 really does not properly summarize the density near the peak; the estimate with the smaller bin width looks much better, but still underestimates the peak considerably compared to the estimate based upon the data without binning. On the other hand, all density estimates on the right seem reasonable, although they vary in integrated squared error from 0.041 to 0.114.

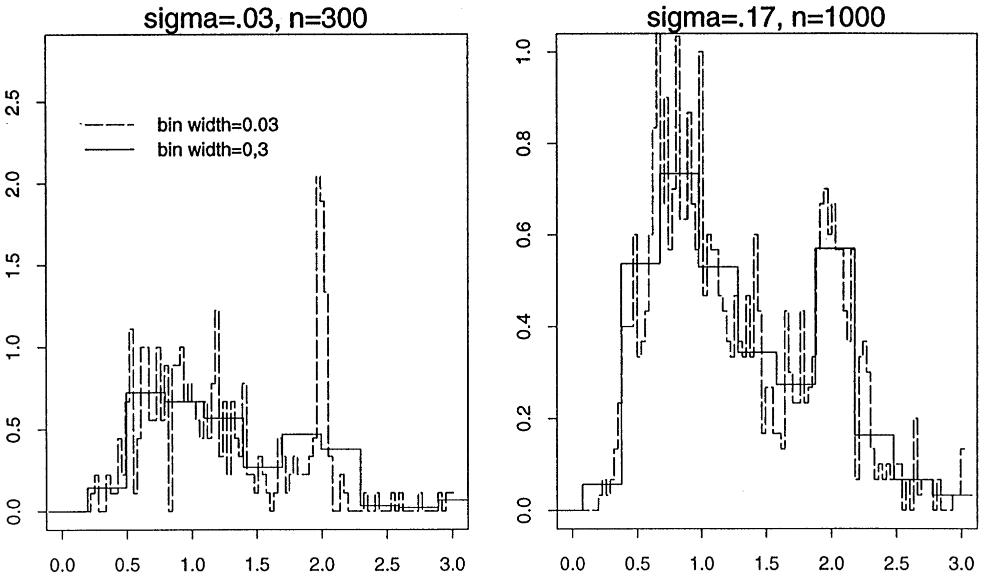

Fig. 3. Histograms on density scale for the logspline density estimates shown in Fig. 2.

6. Proofs of asymptotic results

For a subinterval $\mathscr{I}$ of I and a function h which is integrable on $\mathscr{I}$, write

$$\mathscr{I}[h] = \int_{\mathscr{I}} h.$$

Note that $Y_1, \ldots, Y_Q$ have a multinomial distribution with

$$EY_q = nI_q[f]$$

and

$$\operatorname{Cov}(Y_{q_1}, Y_{q_2}) = \begin{cases} nI_q[f](1 - I_q[f]), & q_1 = q_2 = q, \\ -nI_{q_1}[f]I_{q_2}[f], & q_1 \neq q_2. \end{cases}$$

6.1. Proof of Theorem 2

For the proof of Theorem 2, we need the following lemma.

Lemma 1.

$$E|\hat{\boldsymbol{B}} - \boldsymbol{B}^*|^2 = \mathrm{O}\left(\frac{J}{Q^2} + \frac{1}{n}\right).$$

Proof. Let

$$\overline{B}_{j,k} = E\hat{B}_{j,k} = \sum_{q=1}^{Q} B_{j,k}(x_q)I_q[f]. \tag{13}$$

It follows from the properties of B-splines that

$$\sup_{x,y\in I_q} |\overline{B}_{j,k}(x) - B_{j,k}(y)| \leqslant 2 \sup_{x,y\in I_q} 2^j|x-y| = \mathrm{O}(2^{j-N}) \tag{14}$$

and

$$\#\{m\colon B_{j,k} \not\equiv 0 \text{ on } I_q\} \leqslant r2^{N-j} + 2 = \mathrm{O}(2^{N-j}). \tag{15}$$

By (14) and (15), we have that under (A1),

$$\begin{aligned} |\overline{B}_{j,k} - B^*_{j,k}| &\leqslant \sum_q \int_{I_q} |f(x)||B_{j,k}(x_q) - B_{j,k}(x)|\,\mathrm{d}x \\ &= \mathrm{O}(M_1 2^{N-j}2^{j-N}2^{-N}) \\ &= \mathrm{O}(2^{-N}). \end{aligned}$$

This implies that

$$\sum_{k=1}^{J} |\overline{B}_{j,k} - B^*_{j,k}|^2 = \mathrm{O}\left(\frac{J}{Q^2}\right). \tag{16}$$

Now we derive a bound on the variance $\mathrm{Var}(\hat{B}_{j,k})$ of $B_{j,k}$. Let

$$\mathrm{V} = \frac{1}{n^2}\sum_q B^2_{j,k}(x_q)\mathrm{Var}(Y_q)$$

and

$$\mathrm{CV} = \frac{1}{n^2}\sum_{q_1\neq q_2} B_{j,k}(x_{q_1})B_{j,k}(x_{q_2})\mathrm{Cov}(Y_{q_1}, Y_{q_2}).$$

Then

$$\mathrm{Var}(\hat{B}_{j,k}) = \mathrm{V} + \mathrm{CV}.$$

Under (A1), we have

$$\mathrm{Var}(Y_q) = nI_q[f](1 - I_q[f]) = \mathrm{O}\left(\frac{n}{Q}\right), \tag{17}$$

$$\mathrm{Cov}(Y_{q_1}, Y_{q_2}) = nI_{q_1}[f]I_{q_2}[f] = \mathrm{O}\left(\frac{n}{Q^2}\right). \tag{18}$$

From (15), (17) and (18), we have

$$\mathrm{V} = \mathrm{O}\left(\frac{1}{n^2}\frac{n}{M}\sum_q B^2_{j,k}(x_q)\right) = \mathrm{O}\left(\frac{1}{nJ}\right) \tag{19}$$

and

$$\mathrm{CV} = \mathrm{O}\left(\frac{1}{n^2}\frac{n}{Q^2}\sum_{q_1\neq q_2} B_{j,k}(x_{q_1})B_{j,k}(x_{q_2})\right) = \mathrm{O}\left(\frac{1}{nJ^2}\right). \tag{20}$$

Combining (19) and (20), we have

$$\sum_{k=1}^{J} \operatorname{Var}(\hat{B}_{j,k}) = \mathrm{O}\left(\frac{1}{n}\right). \tag{21}$$

The desired result follows from (16) and (21). □

Now we can prove Theorem 2. By Lemma 1,

$$|\hat{\boldsymbol{B}} - \boldsymbol{B}^*|^2 = \mathrm{O}_{\mathrm{P}}\left(\frac{1}{n} + \frac{J}{Q^2}\right).$$

Now apply Lemma 2 in [KKP] with $\boldsymbol{\beta}_0 = \int \boldsymbol{B} f^* = \boldsymbol{B}^*$ and $\boldsymbol{\beta} = \hat{\boldsymbol{B}}$. If $J/\sqrt{n} \to 0$ as $n \to \infty$ and $J/Q^2 = \mathrm{O}(n^{-1})$, then the MILE $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{B}})$ exists and

$$D(f^* \| \hat{f}) \leqslant M\mathscr{K}\left(\frac{J}{n} + \frac{J^2}{M^2}\right)$$

except on a set of probability less than $1/\mathscr{K}$. This completes the proof of Theorem 2. □

6.2. Proof of Theorem 3

To show that $n^{-2\sigma/(2\sigma+1)}$ is a lower bound, we follow the approach:

- specify a subproblem;
- use Fano's lemma to calculate the difficulty of the subproblem.

For any two probability measures P and Q, their Kullback–Leibler information

$$K(P,Q) = E_P \log(\mathrm{d}P/\mathrm{d}Q)$$

if P is absolutely continuous with respect to Q; otherwise, $K(P,Q) = +\infty$. Observe that

$$K(P_f, P_g) = \sum_{q=1}^{n} n I_q[f] \log\frac{I_q[f]}{I_q[g]},$$

where P_f and P_g denote the distributions of $Y_1, \ldots, Y_Q$ with densities f and g, respectively.

Lemma 2. *For any two positive density functions f and g on I,*

$$K(P_f, P_g) \leqslant nD(f\|g).$$

Proof. Define f^q and g^q by $f^q(x) = f(x)/I_q[f]$ and $g^q(x) = g(x)/I_q[g]$, respectively. Since f and g are positive functions on I_q, f^q and g^q are density functions on I_q. By the information inequality (1.e.66) of Rao (1973),

$$I_q\left[f \log\frac{f}{g}\right] - I_q[f] \log\frac{I_q[f]}{I_q[g]} = I_q[f] I_q\left[f^q \log\frac{f^q}{g^q}\right] \geqslant 0.$$

Since

$$nD(f\|g) - K(P_f, P_g) = n\sum_{q=1}^{n}\left\{I_q\left[f \log\frac{f}{g}\right] - I_q[f]\log\frac{I_q[f]}{I_q[g]}\right\},$$

we have the desired result. □

Here we prove that $n^{-2\sigma/(2\sigma+1)}$ is a lower bound. For notational convenience, choose r to be an odd integer such that $J = 2^j + r - 1$ is even. Let $K = J/2$ and $\mathscr{K}_j = \{k\colon k = 1, \ldots, K\}$. For $\boldsymbol{\tau} = (\tau_k)_{k \in \mathscr{K}_j} \in \{0,1\}^K$, define

$$g_{\boldsymbol{\tau}} = M_2 2^{-\sigma j} \left(\sum_{k=1}^{K} \tau_k B_{j,k} - \sum_{k=1}^{K} \tau_k B_{j,k+K} \right).$$

where M_2 is a constant to be chosen below. Let us observe that

$$\|g_{\boldsymbol{\tau}}\|_{B_{\sigma pq}} \leqslant C M_2 2^{-\sigma j} 2^{j(\sigma - 1/p)} 2 \left(\sum_{k=1}^{K} \tau_k^p \right)^{1/p} \leqslant C M_2$$

and

$$\|g_{\boldsymbol{\tau}}\|_\infty \leqslant M_2 2^{-\sigma j}. \tag{22}$$

Define

$$\psi(\boldsymbol{\tau}) = \log \int_I \exp(g_{\boldsymbol{\tau}}).$$

From (22), we have

$$|\psi(\boldsymbol{\tau})| \leqslant M_2 2^{-\sigma j}. \tag{23}$$

For $j \to \infty$ as $n \to \infty$, we consider the set of vertices of a cube

$$\mathscr{F}_j = \{ f_{\boldsymbol{\tau}} = \exp(g_{\boldsymbol{\tau}} - \psi(\boldsymbol{\tau}))\colon \tau \in \{0,1\}^K \}.$$

By (22), one can choose M_2 such that $\mathscr{F}_j$ is a subset of $\mathscr{F}_{\sigma pq}(M)$ and that $\|\log f\|_\infty \leqslant C$, which implies together with (23) that

$$C^{-1} \leqslant f \leqslant C \quad \text{for all } f \in \mathscr{F}_j. \tag{24}$$

Now let $f_i = f_{\boldsymbol{\tau}^i} \in \mathscr{F}_j$ with $\boldsymbol{\tau}_1 \neq \boldsymbol{\tau}_2$. By Lemma 4.2 of DeVore and Popov (1988) and the properties of B-splines, we have

$$\|\log f_1 - \log f_2\|_2^2 \geqslant C 2^{-j} \left(\sum_{k=1}^{K} (\tau_k^1 - \tau_i^2 - d)^2 + \sum_{k=1}^{K} (\tau_k^1 - \tau_i^2 + d)^2 \right) \geqslant C 2^{-j} \sum_{k=1}^{K} (\tau_k^1 - \tau_k^2)^2,$$

where $d = \psi(\boldsymbol{\tau}^1) - \psi(\boldsymbol{\tau}^2)$. By Lemma 3.1 of Koo (1993), there is a subset $\mathscr{F}_j^*$ of $\mathscr{F}_j$ such that

$$\|\log f - \log g\|_2 \geqslant C 2^{-\sigma j}, \quad f \neq g \in \mathscr{F}_j^* \text{ and } \log(|\mathscr{F}_j^*| - 1) > 0.272(2^j), \tag{25}$$

when n is sufficiently large and $2^j > 8$. It follows from Lemma 1 in Barron and Sheu (1991), (24) and (25) that

$$D(f\|g) > C 2^{-2\sigma j} \quad \text{for } f \neq g \in \mathscr{F}_j^*. \tag{26}$$

Using Lemma 1 in Barron and Sheu (1991) and (24), we obtain

$$D(f\|g) \leqslant C 2^{-2\sigma j} \qquad \text{for all} \quad f, g \in \mathscr{F}_j. \tag{27}$$

By Fano's lemma, see for example Yatracos (1988), if $\hat{T}$ is any estimator of f based on $Y_1,\dots,Y_Q$, then

$$\begin{aligned}\sup_{f\in\mathscr{F}_{\sigma pq}(M)} P_f(D(f\|\hat{T})>c2^{-2\sigma j}) &\geqslant \sup_{f\in\mathscr{F}_j^*} P_f(D(f\|\hat{T})>c2^{-2\sigma j})\\ &\geqslant 1-\frac{\sup_{f,g\in\mathscr{F}_j^*}K(P_f,P_g)+\log 2}{\log(|\mathscr{F}_j^*|-1)}\\ &\geqslant 1-\frac{n\sup_{f,g\in\mathscr{F}_j^*}D(f\|g)+\log 2}{\log(|\mathscr{F}_j^*|-1)}.\end{aligned}$$

Apply (26) to the second line, and Lemma 2 to the third line. Finally, let $2^j \asymp n^{1/(2\sigma+1)}$ as $n\to\infty$. Then the desired result of Theorem 3 follows for the smoothness class from (26), (27) and (28).

6.3. *Proof of Theorem 4*

The following lemma is Lemma 4 in [KKP].

Lemma 3. *If* (*A*2) *and* (*A*3) *hold, then* (i) $M_1^{-1}\leqslant f(x)\leqslant M_1$ *for* $x\in I$; *and* (ii) $\gamma_2=\mathrm{O}(J^{-\sigma})$ *and* $\gamma_\infty=\mathrm{O}(J^{-\sigma+1/p})$.

Now we prove Theorem 4. Assume that (A2) and (A3) hold. Choose $J\asymp n^{1/(2\sigma+1)}$ and $Q\asymp n^{(\sigma+1)/(2\sigma+1)}$. From Lemma 3 it follows that $\gamma_2\sqrt{J}=\mathrm{O}(J^{-(\sigma-1/2)})=\mathrm{o}(1)$ and $\gamma_\infty=\mathrm{O}(J^{-(\sigma-1/p)})=\mathrm{o}(1)$. Theorem 1 now implies that $D(f\|f^*)=\mathrm{O}(J^{-2\sigma})=\mathrm{O}(n^{-2\sigma/(2\sigma+1)})$. On the other hand, $J/\sqrt{n}\asymp n^{(1-2\sigma)/(4\sigma+2)}=\mathrm{o}(1)$ and $J/Q^2=\mathrm{O}(n^{-1})$. Theorem 2 implies that

$$D(f^*\|\hat{f})=\mathrm{O}_\mathrm{P}\left(\frac{J}{n}+\frac{J^2}{Q^2}\right)=\mathrm{O}_\mathrm{P}\left(\frac{J}{n}\right)=\mathrm{O}_\mathrm{P}(n^{-2\sigma/(2\sigma+1)}).$$

Since $D(f\|\hat{f})=D(f\|f^*)+D(f^*\|\hat{f})$ (see Lemma 3 in [KKP]), the proof of Theorem 4 is now complete. □

Acknowledgements

Ja-Yong Koo was supported in part by The Hallym Academy of Sciences, Hallym University. Charles Kooperberg was supported in part by NIH grant R29 CA 74841.

References

Antoniadis, A., Grégorie, G., Nason, G., 1998. Density and hazard rate estimation for right censored data using wavelet methods, manuscript.

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. 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.

Good, I.J., Gaskins, R.A., 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J. Amer. Statistic. Assoc. 75, 42–73.

Hall, P., Wand, M.P., 1996. On the accuracy of binned kernel density estimators. J. Multivariate Anal. 56, 156–184.

Hansen, M.H., Kooperberg, C., 1999. Strategies for spline adaptation, manuscript.

Hansen, M.H., Kooperberg, C., Sardy, S., 1998. Triogram models. J. Amer. Statist. Assoc. 93, 101–119.

Koo, J.-Y., 1993. Optimal rates of convergence for nonparametric statistical inverse problems. Ann. Statist. 21, 590–599.

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., Kooperberg, C., Park, J., 1998. Logspline density estimation under censoring and truncation. Scandinavian J. Statist., to appear.

Koo, J.-Y., Park, B.U., 1996. B-spline deconvolution based on the EM algorithm. J. Statist. Comput. Simul. 54, 275–288.

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. Graphical Statist. 1, 301–328.

Minnotte, M.C., 1998. Achieving higher-order convergence rates for density estimation with binned data. J. Amer. Statist. Assoc. 93, 663–672.

Rao, C.R., 1973. Linear Statistical Inference and its Applications 2nd Edition. Wiley, New York.

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

Stone, C.J., 1989. Uniform error bounds involving logspline models, in: T.W. Anderson, K.B. Athrya, D.L. Iglehart (Eds.), In Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin. Academic Press, Boston, pp. 335–355.

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

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.

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

Yatracos, Y.G., 1988. A lower bound on the error in nonparametric regression type problems. Ann. Statist. 16, 1180–1187.