16

Confidence Intervals for Logspline Density Estimation

Charles Kooperberg and Charles J. Stone[1]

Summary

Several ways to obtain pointwise confidence intervals corresponding to logspline density estimation are studied. These methods include a variety of approaches based on estimation using free knot splines, a couple of approaches based on the bootstrap, and a Bayesian approach. It is concluded that a variation of the bootstrap, in which only a limited number of bootstrap simulations are used to estimate standard errors that are combined with standard normal quantiles, seems to perform the best, especially when coverages and computing time are both taken into account.

16.1 Introduction

Getting confidence intervals corresponding to function estimates that are obtained using an adaptive polynomial spline method is a notoriously hard problem. After model selection has been carried out, the estimated function has a simple parametric form [12]. However, treating the final model as a fixed parametric model, ignoring the large amount of model selection that may have occurred, yields confidence intervals with too low coverage.

Recently, Kooperberg and Stone [9] described an algorithm for logspline density estimation with free knots. This is a modification to previous

[1]Charles Kooperberg is Member, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024 (E-mail: clk@fhcrc.org). Charles J. Stone is Professor, Department of Statistics, University of California, Berkeley, CA 94720-3860 (E-mail: stone@stat.berkeley.edu). Charles Kooperberg was supported in part by National Institutes of Health grant CA74841. Charles J. Stone was supported in part by National Science Foundation grant DMS-9802071.

logspline density algorithms [7, 8, 12], in which the knots are not selected by a greedy stepwise algorithm, but are viewed as additional parameters. Two reasons for studying logspline density estimation with free knot splines are that (i) stepwise selection algorithms can be seen as crude approximations to the free knot algorithm and (ii) coverages of (pointwise) confidence intervals based on the free knot algorithm may be more accurate since they reflect uncertainty in the knot placement. It was concluded that the coverages of nominal 95% confidence intervals using the free knot algorithm, while closer to 95% than the coverages ignoring knot selection, are still well below 95%.

In the current chapter we investigate alternative methods for obtaining confidence intervals corresponding to logspline density estimation. In particular, we investigate whether an expansion of the free knot intervals improves the coverage, and we also discuss bootstrap and Bayesian methods for obtaining confidence (credible) intervals.

In Section 16.2 we briefly review logspline density estimation in general and the procedure with free knots in particular. In Section 16.3 we discuss the various approaches to obtaining confidence intervals. The approaches based on expansion of the standard errors for free knot splines and on the bootstrap are compared in a simulation study in Section 16.4. In Section 16.5, the various approaches are applied to a real example. We end with a brief discussion.

16.2 Logspline density estimation with free knots

Given the free knots $-\infty < L < \gamma_1 < \cdots < \gamma_J < U < \infty$, set $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ and let $\mathbb{G}_{\boldsymbol{\gamma}}$ denote the space of cubic splines on $[L, U]$ corresponding to the knot sequence $\boldsymbol{\gamma}$ and satisfying the usual tail linear constraints. Thus a function g on $[L, U]$ is a member of $\mathbb{G}$ if and only if it is twice continuously differentiable on $[L, U]$, its restriction to each of the intervals $[L, \gamma_1]$, $[\gamma_1, \gamma_2]$, $\ldots$, $[\gamma_{J-1}, \gamma_J]$, $[\gamma_J, U]$ is a cubic polynomial, $g''(L) = 0$, and $g''(U) = 0$. Observe that $\mathbb{G}_{\boldsymbol{\gamma}}$ is a $(J+2)$-dimensional linear space. Set $p = J + 1$, and let $1, B_{\boldsymbol{\gamma}1}, \ldots, B_{\boldsymbol{\gamma}p}$ be a basis of $\mathbb{G}_{\boldsymbol{\gamma}}$. Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta = \mathbb{R}^p$, set

$$\eta_{\boldsymbol{\gamma}}(y; \boldsymbol{\theta}) = \theta_1 B_{\boldsymbol{\gamma}1}(y) + \cdots + \theta_p B_{\boldsymbol{\gamma}p}(y) - C_{\boldsymbol{\gamma}}(\theta), \qquad L \le y \le U,$$

where

$$C_{\boldsymbol{\gamma}}(\theta) = \log\Big(\int_L^U \exp(\theta_1 B_{\boldsymbol{\gamma}1}(y) + \cdots + \theta_p B_{\boldsymbol{\gamma}p}(y))\, dy \Big).$$

Note that $\exp \eta_{\boldsymbol{\gamma}}(y; \boldsymbol{\theta})$ is a positive density function on $[L, U]$ for every $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$.

Let $Y_1, \ldots, Y_n$ be a random sample of size n from a distribution having density f and log-density $\eta = \log f$. Consider the log-likelihood

$$\ell_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \eta_{\boldsymbol{\gamma}}(Y_i; \boldsymbol{\theta}) = \sum_{j=1}^{p} \theta_p \sum_{i=1}^{n} B_{\boldsymbol{\gamma} j}(Y_i) - nC_{\boldsymbol{\gamma}}(\boldsymbol{\theta}).$$

Let $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\theta}}$ denote the maximum likelihood estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, so that

$$\widehat{\ell} = \ell_{\widehat{\boldsymbol{\gamma}}}(\widehat{\boldsymbol{\theta}}) = \operatorname*{argmax}_{\boldsymbol{\gamma}, \boldsymbol{\theta}} \ell_{\boldsymbol{\gamma}}(\boldsymbol{\theta}).$$

Observe that for the free knot procedure [9] the positive integer parameter J must also be chosen. Let $\widehat{\boldsymbol{\gamma}}_J$, $\widehat{\boldsymbol{\theta}}_J$, and $\widehat{\ell}_J$ now indicate the dependence of $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\theta}}$, and $\widehat{\ell}$, respectively, on J. For choosing J, we will employ the Akaike Information Criterion $\mathrm{AIC}_{J,a} = -2\widehat{\ell}_J + (2J+1)a$ [1], which depends on the complexity parameter a. (Note that $\widehat{\boldsymbol{\gamma}}_J$ has J free parameters and $\widehat{\boldsymbol{\theta}}_J$ has $p = J+1$ free parameters.) We select the value $\widehat{J}$ of J that minimizes $\mathrm{AIC}_{J,2}$. Set $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}_{\widehat{J}}$ and $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{\widehat{J}}$. We refer to $\widehat{\eta}(y) = \eta_{\widehat{\boldsymbol{\gamma}}}(y; \widehat{\boldsymbol{\theta}})$ as the maximum (penalized) likelihood estimate of the log-density η at y and to $\widehat{f}(y) = \exp \widehat{\eta}(y)$ as the logspline estimate with free knots of the density f at y.

Computing the maximum likelihood estimates with free knots is a highly nontrivial numerical problem, as the likelihood function $\ell_{\widehat{\boldsymbol{\gamma}}}(\widehat{\boldsymbol{\theta}})$ is severely multimodal, and degenerate solutions exist when too many of the knots γ_j get close together.

For a procedure with fixed knots we would have to specify a set of knots. Rather than specifying a complete set in advance, we select knots by a stepwise procedure. Such a procedure can be either a stepwise deletion procedure or a stepwise addition and deletion procedure. For the former, we initially position a large number of knots and remove the "least significant" knot one at a time [8]. For the later we add knots one at a time, to increase the log-likelihood as much as possible, until a maximum number of knots is reached, after which we carry out a stepwise deletion procedure [12]. For either of the two stepwise procedures we use the AIC criterion with $a = \log n$, as in the Bayesian Information Criterion [10] to select the number J of knots.

16.3 Confidence intervals

16.3.1 Free knot splines

In [9] we proposed obtaining confidence intervals for the density using a "standard" maximum likelihood approach. In particular, let $\widehat{\nabla}_J \eta(y)$ denote the $(2J+1)$-dimensional gradient of $\eta_{\boldsymbol{\gamma}}(y; \boldsymbol{\theta})$ at the maximum likelihood

estimate, and let $\widehat{H}_J$ denote the $(2J+1) \times (2J+1)$ Hessian matrix of the log-likelihood at the maximum likelihood estimate when there are J free knots. Set $\widehat{\nabla}\eta(y) = \widehat{\nabla}_{\widehat{J}}\eta(y)$ and $\widehat{H} = \widehat{H}_{\widehat{J}}$. The standard error in the estimate $\widehat{\eta}(y)$ is given by

$$\mathrm{SE}(\widehat{\eta}(y)) = \sqrt{\left[\widehat{\nabla}\eta(y)\right]^T \left(-\widehat{H}\right)^{-1} \widehat{\nabla}\eta(y)}. \tag{16.1}$$

This leads to the nominal 95% confidence interval

$$\Big(\exp\left(\widehat{\eta}(y) - 1.96\mathrm{SE}(\widehat{\eta}(y))\right), \exp\left(\widehat{\eta}(y) + 1.96\mathrm{SE}(\widehat{\eta}(y))\right)\Big) \tag{16.2}$$

for $f(y)$.

The distribution function corresponding to f is given by $F(y) = \int_L^y \exp\eta(z)\,dz$ for $L \le y \le U$, which can be estimated by $\widehat{F}(y) = \int_L^y \exp\widehat{\eta}(z)\,dz$. The corresponding standard error is given by

$$\mathrm{SE}(\widehat{F}(y)) = \sqrt{[\widehat{\nabla}F(y)^T](-\widehat{H})^{-1}\widehat{\nabla}F(y)},$$

where $\widehat{\nabla}F(y) = \int_L^y \widehat{\nabla}\eta(z)\exp\widehat{\eta}(z)\,dz$.

Simulation studies were carried out, which suggested that the actual coverage of nominal 95% confidence intervals using these standard errors is about 87% for the density and 93% for the distribution function. This coverage is, however, much better than when the uncertainty in the knots is ignored. Let $\mathrm{SEFX}(\widehat{\eta}_i(y))$ and $\mathrm{SEFX}(\widehat{F}_i(y))$ be the standard errors assuming that the knots are fixed (so that they make use only of the $(\widehat{J}_i+1) \times (\widehat{J}_i+1)$ Hessian matrix for the coefficients). The coverage of the nominal 95% confidence intervals using these standard errors was only about 81% for the density and 92% for the distribution function. To improve the coverage we will also investigate confidence intervals

$$\Big(\exp\left(\widehat{\eta}(y) - 1.96\alpha\mathrm{SE}(\widehat{\eta}(y))\right), \exp\left(\widehat{\eta}(y) + 1.96\alpha\mathrm{SE}(\widehat{\eta}(y))\right)\Big), \tag{16.3}$$

in which the standard errors are expanded by the factor α for some $\alpha > 1$ and using similarly expanded confidence intervals for the distribution function in this chapter.

16.3.2 The bootstrap

Alternatively, we can employ the bootstrap in combination with either the stepwise knot deletion algorithm of [8] or the stepwise addition and deletion algorithm of [12] to obtain confidence intervals corresponding to logspline density estimates. In this chapter we use the former algorithm and examine the coverage of bootstrap percentile intervals [3] for both the log-density and the distribution function. That is, we take B (we used $B = 1000$) samples $\mathbf{Y^i}$ with replacement of size n from the data $Y_1, \ldots, Y_n$, and for each sample $\mathbf{Y^i}$ we obtain the logspline density estimate. The 95% pointwise

confidence interval for $\widehat{\eta}(y)$ $(F(y))$ is then from the 2.5th to the 97.5th percentile of the B bootstrap estimates for the log-density (distribution function).

Clearly, the bootstrap is a computationally time consuming procedure for getting confidence intervals, as we need to fit B logspline densities. However, it is still slightly faster than using the algorithm developed in [9] for fitting logspline densities with free knots.

A considerably cheaper approach is to hope that the logspline estimates of the log-density and distribution function have approximately a normal distribution, but that the estimates of the standard errors that are obtained using standard techniques are too small. If so, we can get by with a much smaller number B of bootstrap estimates (say $B = 25$) by using these estimates to obtain pointwise bootstrap estimates of $\mathrm{SE}^B(\widehat{\eta}(y))$ and $\mathrm{SE}^B(\widehat{F}(y))$ and then using equation (16.2) or the equivalent to obtain confidence intervals for η and F.

16.3.3 A Bayesian approach

Hansen and Kooperberg [6] describe a Bayesian approach to logspline density estimation, which involves a prior $p(J)$ on the dimension of the model, a prior $p(\gamma \mid J)$ on the location of the knots, and a prior $P(\boldsymbol{\theta} \mid \gamma, J)$ on the coefficients. Given the data $Y_1, \ldots, Y_n$, the posterior distribution of $(J, \gamma, \boldsymbol{\theta})$ is explored using a reversible jump Markov chain Monte Carlo [5] algorithm. At each step of the algorithm a new density is proposed by either adding a knot, deleting a knot, moving a knot, or updating the coefficients. This new proposed density is always accepted if the posterior probability is higher than the previous density; otherwise it still has a positive probability of being accepted. The acceptance probability is governed by the reversible jump algorithm. The algorithm of [6] for logspline density estimation is similar to algorithms for univariate regression using polynomial splines proposed by [2] and [11].

To make (pointwise) 95% credible intervals about the logspline density estimate obtained from this Bayesian procedure, we use as endpoints the 2.5th and 97.5th percentiles of all Markov chain Monte Carlo simulations. Credible intervals have a different interpretation from (frequentist) confidence intervals. For confidence intervals we are 95% confident that the confidence interval will cover the true value of the density; for a 95% credible interval, there is 95% (posterior) probability that the density falls within the interval.

Hansen and Kooperberg [6] point out that, depending on how priors are selected, a Bayesian procedure can be similar in performance to a greedy stepwise procedure using AIC to select the number of knots when a geometric prior on the number of knots is used, or it can be similar to a

smoothing spline approach when a uniform prior on the number of knots and a particular multivariate normal prior on the coefficients are used.

16.4 A simulation study

In this section we augment the results of the simulation study in [9], in which we generated 250 samples of size 250 and 250 samples of size 1000 from each of four distributions:

Normal 2 A mixture of two normal distributions, so that the true density of Y is given by

$$f(y) = c\Big(\frac{1}{3}f_{Z_1}(y) + \frac{2}{3}f_{Z_2}(y)\Big)\mathrm{ind}(-4, 8),$$

where Z_1 has a normal distribution with mean 0 and standard deviation 0.5, Z_2 has a normal distribution with mean 2 and standard deviation 2, $\mathrm{ind}(\cdot)$ is the usual indicator function, and c is the multiplier to correct for the truncation to $(-4, 8)$.

Normal 4 As in example 1, but the mean of Z_2 is 4 and Y is truncated to $(-2, 10)$.

Normal 6 As in example 1, but the mean of Z_2 is 6 and Y is truncated to $(-1.5, 12)$.

Gamma 2 A gamma distribution with shape parameter 2 and mean 1, with Y truncated to the interval $(0, 9)$.

The Normal 2 density has one mode, but a clear second hump; Normal 4 has two, not very well separated, modes; Normal 6 has two well separated modes; and the Gamma 2 density is unimodal.

In Table 16.1 we compare the coverages of four approaches for getting confidence intervals using the free knot spline methodology. The first two columns are taken from [9]. These columns are the coverages obtained by using the regular SE (see equation (16.1)) or SEFX, for which it is assumed that the knots are fixed. As can be seen from this table, the coverages are well below the nominal 95% level. For the third and fourth columns, these standard errors are expanded by $\alpha = 1.34$ for SE and $\alpha = 1.55$ for SEFX, respectively (see equation (16.3)). These expansion factors were chosen so that the average coverage over these eight simulations is exactly 95%; thus, it could be argued that these columns do not provide a completely fair comparison. The last two columns are using bootstrap samples for the logspline density estimation procedure of [8]. The fifth column is based on 1000 bootstrap samples, and the confidence intervals are from the 2.5th through the 97.5th (pointwise) percentiles. For the sixth column we generated only 25 bootstrap samples, computed the pointwise standard errors for the log-density, and then used (16.2) to obtain the confidence intervals.

Table 16.1. Coverages for six different approaches to obtaining confidence intervals for a log-density, estimated using logspline.

	Free Knot Standard Error					
	Nominal		Expanded		Bootstrap	
Density	SE	SEFX	1.34SE	1.55SEFX	Percentiles	SE
$n = 250$						
Normal 2	84.0	77.4	91.9	91.9	97.4	95.2
Normal 4	88.8	82.5	95.5	94.9	97.4	96.4
Normal 6	89.0	84.0	96.9	97.2	96.5	94.6
Gamma 2	86.2	81.2	94.8	96.0	97.8	97.3
$n = 1000$						
Normal 2	89.2	79.6	95.8	93.9	96.8	94.4
Normal 4	89.3	82.7	97.4	97.0	98.0	94.7
Normal 6	86.2	81.4	95.2	96.2	96.3	92.9
Gamma 2	84.0	77.3	92.6	92.8	97.4	95.4
Average	87.1	80.7	95.0	95.0	97.2	95.1

It is clear from this table that the confidence intervals based on the free knot spline standard error or the fixed knot spline standard error have too low coverage. With an appropriate expansion factor it is possible to get the coverage to be about 95%. With this approach the problem is, naturally, to find the right expansion factor α. If we had chosen separate expansion factors for each density and each sample size, we would have had factors that for SE varied between $\alpha = 1.2$ for Normal 4 with $n = 1000$ to $\alpha = 1.63$ for Normal 2 with $n = 250$. On the other hand, the expansion factor that gives an overall coverage of about 95% for the four distributions being studied essentially does not depend on n for the two sample sizes being studied. Actually, there seems to be little advantage of the expanded SE over the expanded SEFX in this case, except that the expansion factors are larger for SEFX.

Surprisingly, the coverages for the bootstrap percentile intervals are consistently too high. It is our impression that this is due to some instability in the stepwise logspline algorithm when there are many repeat observations, causing the intervals to be occasionally too large. That is in line with what we will see for the income data in the next section. Interestingly, the coverages in the sixth column of Table 16.1, corresponding to the bootstrap SE approach, not only are very close to 95% on average, but have considerably less variation than those in columns 3 and 4 based on expanded SE's.

For the distribution function all approaches yielded somewhat better results (coverage closer to nominal, less variation between different distri-

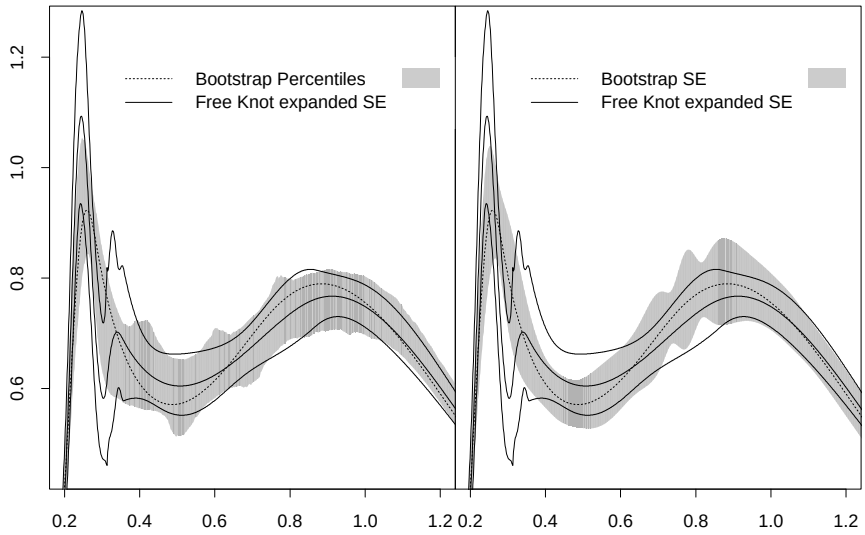

Figure 16.1. Comparison of the expanded free knot spline pointwise confidence intervals and bootstrap pointwise confidence intervals for the income data. The solid lines are estimate and confidence bounds for the free knot procedure, the dashed line is the estimate for the stepwise procedure and the grey area are the bootstrap intervals (left side percentiles, right side SE).

butions) than for the (log-)density, except for the bootstrap SE approach, for which the average coverage was down to 93.6%. This is not surprising, since the logspline estimate of the distribution function is presumably not approximately normal in the tails. A logistic transformation may improve the results here.

16.5 An example

In this section we further analyze the income data, which was also discussed in [9] and [6]. In Figure 16.1 we show the 95% free knot (pointwise) confidence intervals, expanded by $\alpha = 1.34$ as in the previous section, together with the corresponding logspline density estimate (solid lines for the estimate, the lower and upper confidence bounds). We also show the stepwise logspline density estimate with knot deletion (dashed line) along with the 95% bootstrap percentile confidence intervals (left side, grey area) and the 95% bootstrap confidence intervals using 25 samples to estimate the standard error (right side, grey area). As can be seen from these plots, the bootstrap SE approach and the bootstrap percentile approach yield intervals that are approximately the same size as the expanded free knot approach, but which are slightly less smooth. Averaged over the region shown, the average size of all three intervals are within 5% of each other.

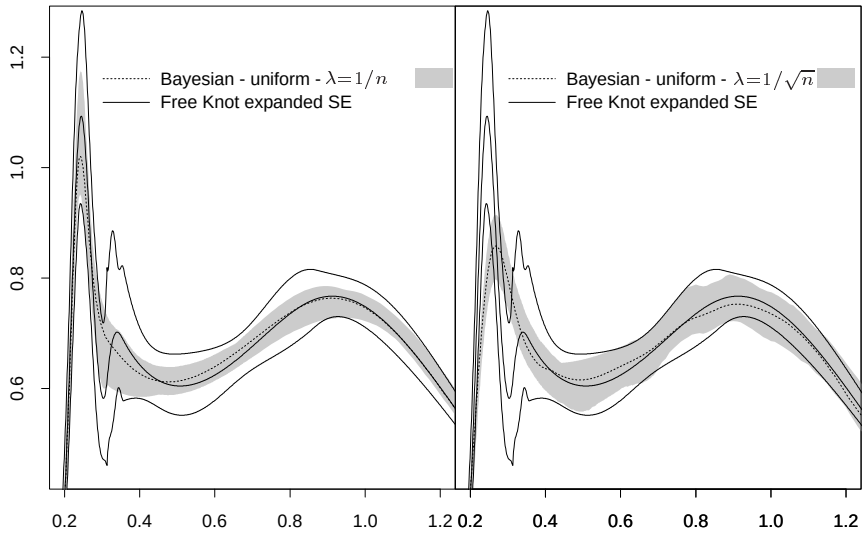

Figure 16.2. Comparison of the expanded free knot spline pointwise confidence intervals and Bayesian pointwise credible intervals for the income data. The solid lines are the same as in Fig. 16.1, the dashed lines are the estimates for the Bayesian procedures and the grey areas are the corresponding credible intervals.

Overall, these intervals agree with the conclusion from the previous section: the bootstrap SE approach yields reasonable confidence intervals at a computing price that is mush smaller than free knot splines or a full bootstrap approach.

In Figure 16.2 we show the same expanded free knot intervals as in Figure 16.1, but this time we added 95% credible intervals from the Bayesian algorithm described in [6] (dashed lines and grey area). The algorithms shown have a uniform prior for the number of knots and a multivariate normal prior on the coefficients. The variance of this later prior (proportional to the λ parameter indicated in these plots) plays a role as a smoothing parameter. The results shown in this figure are based on a run of 100,000 MCMC iterations, which takes a cpu time that is comparable to the bootstrap percentile approach, and which is considerably larger than what is needed to obtain good point estimates. The estimates with $\lambda = 1/n$ were the ones with the largest value of λ that gave a reasonable estimate for the height of the peak, as argued in [6]. The corresponding 95% credible intervals are still considerably smaller than the 95% expanded free knot intervals, suggesting that the coverages of the former intervals may be significantly under 95%. Even when $\lambda = 1/\sqrt{n}$, so large that the height of the peak gets reduced to about 0.86, rather than the "correct" height of between 1.00 and 1.10, the credible intervals still appear too small.

16.6 Discussion

Several ways for obtaining confidence or credible intervals for logspline density estimates were studied here. Free knot and fixed knot confidence intervals that are not expanded yield substantially too low coverages. These intervals can be expanded to give reasonable coverage, but it is not obvious how well the expansion factors used in the simulation study reported here would work for other choices of the underlying density or sample size. Bayesian credible intervals for density estimates that look reasonable appear too small, while those intervals that are wide enough seem to correspond to density estimates that smooth too much. Bootstrap percentile intervals appear ragged, suggesting that very large numbers of bootstrap samples are needed, and their coverages are too high. The bootstrap SE approach—estimating the standard error based on a limited number of bootstrap estimates and using "1.96" to obtain 95% confidence intervals—seems to have the best performance. The coverage is about right, the computational expense is low, and the pointwise confidence intervals are fairly smooth. This performance came as a pleasant surprise to us and suggests that the bootstrap SE approach deserves a more thorough investigation.

References

[1] Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle", in *Second International Symposium on Information Theory* (eds B. N. Petrov and F. Csáki), Budapest: Akademia Kiadò, 267–281.

[2] Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian curve fitting," *J. Roy. Statist. Soc.*, Ser. B., **60**, 333–350.

[3] Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

[4] Family Expenditure Survey (1968–1983), *Annual Base Tapes and Reports (1968–1983)*, London: Department of Employment Statistics Division, Her Majesty's Stationary Office.

[5] Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

[6] Hansen, M. H. and Kooperberg, C. (2002). "Spline adaptation in extended linear models (with discussion)", *Statist. Science*, **17**, to appear.

[7] Kooperberg, C. and Stone, C. J. (1991), "A study of logspline density estimation," *Comp. Statist. Data Anal.*, **12**, 327–347.

[8] Kooperberg, C. and Stone, C. J. (1992), "Logspline density estimation for censored data," *J. Comp. Graph. Statist.*, **1**, 301–328.

[9] Kooperberg, C. and Stone, C. J. (2002), "Logspline density estimation with free knots," manuscript.

[10] Schwarz, G. (1978), "Estimating the dimension of a model", *Ann. Statist.*, **6**, 461–464.

[11] Smith, M. and Kohn R. (1996), "Nonparametric regression using Bayesian variable selection," *J. Environ.*, **75**, 317–344.

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