Density Estimation via Hybrid Splines

Ronaldo Dias

Universidade Estadual de Campinas *

Abstract

The Hybrid Spline method (H-spline) is a method of density estimation which involves regression splines and smoothing splines methods. Using basis functions (B-splines), this method is much faster than Smoothing Spline Density Estimation approach (Gu, 1993). Simulations suggest that with more structured data (e.g., several modes) H-spline method estimates the modes as well as Logspline (Kooperberg and Stone, 1991). The H-spline algorithm is designed to compute a solution to the penalized likelihood problem. The smoothing parameter is updated jointly with the estimate via a cross-validation performance estimate, where the performance is measured by a proxy of the symmetrized Kullback-Leibler. The initial number of knots is determined automatically based on an estimate of the number of modes and the symmetry of the underlying density. The algorithm increases the number of knots by

*Postal address: Departamento de Estatística, IMECC, Cidade Universitária "Zeferino Vaz", Caixa Postal 6065, 13.081-970 - Campinas, SP - BRAZIL

1 until the symmetrized Kullback-Leibler distance, based on two consecutives estimates satisfies a condition which was determined empirically.

Key words: density estimation, penalized loglikelihood, B-splines, logspline, smoothing parameter, Kullback-Leibler distance.

1 Penalized Loglikelihood for H-splines

In this section we describe the Hybrid spline method for density estimation which combines ideas from logspline density estimation (Kooperberg and Stone, 1991) and smoothing splines density estimation (Wahba, 1990; Gu, 1993). In logspline density estimation the choice of the number and location of knots are optimized and it is combined with maximum likelihood estimation, while smoothing spline density estimation employs the penalized likelihood method.

Let $X_1, \ldots, X_n$ be i.i.d. random variables with an unknown probability density function f on a finite domain $\mathcal{X}$. The goal is to estimate the density f from the data X_i. Since f is a probability density function, any estimate must be positive and integrate to one. To enforce the positivity and unity constrains on f, we use the logistic transformation (Leonard, 1978) $f = e^g/(\int e^g)$. It is easy to see that this transformation is not one-to-one, however, a side condition on g, such as $g(x_0) = 0, x_0 \in \mathcal{X}$ or $\int_{\mathcal{X}} g = 0$ is needed to determine this transformation uniquely. Under this scheme a density estimate is defined to be the minimizer of the penalized likelihood equation.

$$-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \log \int_{\mathcal{X}} e^{g} + \lambda \int (g'')^2 \tag{1.1}$$

The minimization is over the class $\mathcal{H}$ of the absolutely continuous functions on $\mathcal{X}$ whose second derivative is square integrable.

In the smoothing spline density estimation approach the class of functions $\mathcal{H}$ is assumed to be a reproducing kernel in Hilbert space ($\mathcal{RKHS}$) and the basis functions are the ones used in (Wahba, 1990)). Here, a different approach is suggested and B-splines (de Boor, 1978) are used as basis functions, and their computational advantages make this algorithm faster, for univariate densities, than Gu's algorithm ((Gu, 1993)). Differently than other methods, this approach takes into account the number of basis functions and the smoothing parameter λ to control the trade-off between goodness of fitting and smoothness of the density estimate. For this, assume that a density function f_0 is in $\mathcal{H}$. Then, we approximate g_0, the solution of the penalized likelihood equation, by linear combinations of basis functions that span $\mathcal{H}_K = \{g \in \mathcal{NCS}_4 : g = \sum_{i=1}^{K} c_i B_i\}$, where $\mathcal{NCS}_4$ stands for natural cubic spline of order 4. That is, $g_0 \approx g = \sum_{i=1}^{K} c_i B_i$, where the notation $g_0 \approx g$ means g_0 is approximately equal to g. Later, we measure this approximation by taking the symmetrized Kullback-Leibler distance.

Consider $\mathcal{X}$, the domain of the density function f_0, and the logistic transformation $f_0 = e^{g_0} / \int_{\mathcal{X}} e^{g_0}$ with side condition $\int_{\mathcal{X}} g_0 = 0$, then we obtain,

$$\int_{\mathcal{X}} g_0(x)dx \approx \int_{\mathcal{X}} g(x)dx = \sum_{j=1}^{K} c_j \int_{\mathcal{X}} B_j(x)dx. \tag{1.2}$$

Letting $p_j = \int_{\mathcal{X}} B_j(x)dx$, we have $\sum_{j=1}^{K} c_j p_j = 0$, or $c_K = -\sum_{j=1}^{K-1} c_j p_j / p_K$. Therefore for any function g, such that $g \in \mathcal{H}_{\mathcal{K}}$ can be written as

$$g(x) \quad = \quad \boldsymbol{R^T c} \tag{1.3}$$

where $\boldsymbol{R} = (R_1, \ldots, R_{K-1})^T$ and $R_j = B_j - \frac{p_j}{p_k} B_K$ for $j = 1, \ldots, K-1$.

Now the numerical problem becomes to minimize:

$$A_\lambda(\boldsymbol{c}) = \frac{\boldsymbol{-1^T}}{n}(Q\boldsymbol{c}) + \log \int_{\mathcal{X}} \exp(\ \boldsymbol{R^T c}) + (\lambda/2)\boldsymbol{c^T}\Omega\boldsymbol{c} \tag{1.4}$$

where Q is $n \times (K-1)$ matrix with entry $R_i(X_j) = R_{ij}$, the penalty matrix has entries $\Omega_{ij} = \int R_i'' R_j''$ for $i, j = 1 \ldots, K-1$, where R'' is the second derivative of the function R. The vector $\frac{\boldsymbol{-1^T}}{n} = (\frac{1}{n}, \ldots, \frac{1}{n})$ is of dimension $1 \times n$, and $\text{rank}(\Omega) = K-2$, since two of the basis functions are linear.

A standard procedure to minimize $A_\lambda(\boldsymbol{c})$ is to apply Newton-Raphson iteration. For that, let us consider two densities f and h and denote their mean and covariance with respect to g as, $\mu_g(h) = \int he^g / \int e^g$, $V_g(f,h) = \mu_g(fh) - \mu_g(f)\mu_g(h)$ with $V_g(h) = V_g(h,h)$. Write $\tilde{g} = R^T\tilde{c}$ as the current iteration of g. Then, letting $\mu_R = \mu_{\tilde{g}}(R)$ and $V_{R,R} = V_{\tilde{g}}(R, R^T)$, we have

$$\frac{\partial A_\lambda(c)}{\partial c} \mid_{\tilde{g}} = -Q^T\mathbf{1}/\text{n} + \mu_R + \lambda\Omega\tilde{c}\ ,$$

and,

$$\begin{aligned} \frac{\partial^2 A_\lambda(c)}{\partial c \partial c^T} \mid_{\tilde{g}} \quad &= \quad \mu_{\tilde{g}}(R, R^T) - \mu_{\tilde{g}}(R)\mu_{\tilde{g}}(R^T) + \lambda\Omega \\ &= \quad V_{\tilde{g}}(R, R^T) + \lambda\Omega = V_{R,R} + \lambda\Omega. \end{aligned}$$

Hence, the Newton updating equation is:

$$(V_{R,R} + \lambda\Omega)^T c = Q^T \mathbf{1}/n - \mu_{\tilde{g}}(R) + V_{R,g}. \tag{1.5}$$

Since $V_{R,R}^T \tilde{c} = V_{R,g}$, where $V_{R,g} = V_{\tilde{g}}(R, \tilde{g})$.

Consequently, to compute the minimizer of (1.4) for a fixed λ, one iterates (1.5) not directly but making use of the symmetrized Kullback-Leibler distance between $g = \log f$ and $g_0 = \log f_0$ given by

$$\begin{aligned} L(g, g_0) &= KL(f, f_0) + KL(f_0, f) && (1.6)\\ &= \int (g - g_0)e^g + \int (g_0 - g)e^{g_0} && (1.7)\\ &= \mu_g(g - g_0) + \mu_{g_0}(g_0 - g). && (1.8) \end{aligned}$$

Note that equation (1.5) defines a class of estimates with variable λ, and a better use of this equation is to start from a current $\tilde{g}$, instead of calculating the next iterate based on any prespecified λ, one may choose λ which delivers a small $L(g, g_0)$ for g among the class of estimates defined by (1.5). Now, compute an update using such a λ. This iteration scheme tries to minimize the loss function $L(g, g_0)$ instead of minimizing the penalized likelihood equation (1.4). When such a performance-oriented iteration converges at λ_*, and g_* is apparently the fixed point of the Newton iteration for minimizing A_{λ_*}, and hence it is the minimizer of A_{λ_*}.(See (Gu, 1993) for details in the case of reproducing kernel in Hilbert space)

It can be noticed that such a L is not easily computable for a class of estimates defined by (1.5), even with a known g_0. Thus, approximation is needed. Similarly to

Gu (1993) a proxy of L is given by, $L_{\tilde{g}}(g, g_0) = V_{\tilde{g}}(g)/2 - V_{\tilde{g}}(g, \tilde{g}) + \mu_{\tilde{g}}(g) - \mu_{g_0}(g)$. The idea is to try to minimize L by calculating an update from $\tilde{g}$ using (1.5) with variable λ. Observe that the first three terms of $L_{\tilde{g}}$ are readily computable, but the fourth needs estimation. For that, define: $H = V_{R,R} + \lambda\Omega$, with $u_R = Q^T\mathbf{1}/n - \mu_R + V_{R,g}$ and $v_R = V_{R,g} - \mu_R$. Note that H is $K-1 \times K-1$ matrix and not $n \times n$ as in (Wahba, 1977; Gu, 1993), with $K \ll n$.

Since $V_{R,R}$ is positive definite matrix, Ω is nonnegative definite matrix, and $\lambda > 0$, the estimate defined by (1.5) has the expression,

$$g = R^T c = R^T H^{-1} Q^T \mathbf{1}/n + R^T H^{-1} v_R, \tag{1.9}$$

substituting $c = H^{-1}u_R$. Moreover, straight but tedious calculations give $\mu_{\tilde{g}}(g) - V_{\tilde{g}}(g, \tilde{g}) = -v_R^T H^{-1} u_R$ and $V_{\tilde{g}}(g) = u_R^T H^{-1} u_R - \lambda u_R^T H^{-1} \Omega H^{-1} u_R$.

To estimate $\mu_{g_0}(g)$, the only source of information is the empirical distribution of the data. For the last term of (1.9), sample means simply give $\mathbf{1^T}/nQH^{-1}v_R$. (For a comparison to reproducing kernel in Hilbert space see Gu (1993) eq. 3.4). For the first term,

$$\mu_{g_0}(R^T H^{-1} Q^T \mathbf{1}/n) = \frac{1}{n}\sum_{i=1}^{n} \mu_{g_0}(R^T(X_i) H^{-1} R(X_i)).$$

It can be seen that the problem with the naive sample mean is the use of X_i itself in the estimation of $\mu_{g_0}(R^T(X_i)H^{-1}R(X_i))$. Using the empirical distribution of the remaining $n-1$ data to estimate $\mu_{g_0}(R^T(X_i)H^{-1}R(X_i))$, we have:

$$
\begin{aligned}
&\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} R^T(X_j) H^{-1} R(X_i) = \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} R^T(X_j) H^{-1} R(X_i) - \frac{1}{n(n-1)} \sum_{i=1}^{n} R^T(X_i) H^{-1} R(X_i) \\
&= \frac{n}{n-1} (Q^T \mathbf{1}/n)^T H^{-1} Q^T\ \mathbf{1}/n - \frac{1}{n(n-1)} trace(QH^{-1}Q^T). \qquad (1.10)
\end{aligned}
$$

Letting $tr(\cdot) = trace(\cdot)$ and noting that

$$
\frac{V_{\tilde{g}}(g)}{2} - \mu_{g_0}(g) = \frac{tr(QH^{-1}Q^T)}{n(n-1)} - \frac{(Q^T\mathbf{1}/n)^T H^{-1} Q\mathbf{1}/n}{n-1} - \frac{u_R^T H^{-1} u_R}{2} + v_R^T H^{-1} u_R,
$$

we have a proxy of L as:

$$
\hat{L_\lambda}(g, g_0) = \frac{tr(QH^{-1}Q^T)}{n(n-1)} - \frac{(Q^T\mathbf{1}/n)^T H^{-1} Q\mathbf{1}/n}{n-1} - \frac{u_R^T H^{-1} u_R}{2} - \frac{\lambda u_R^T H^{-1} \Omega H^{-1} u_R}{2}. \qquad (1.11)
$$

Since (1.11) is a cross-validation of $\mu_{g_0}(R^T H^{-1} Q^T \mathbf{1}/n)$, is natural to think of it as cross-validation score under the context of Hybrid splines for density estimation. Moreover, the performance-oriented iteration can then be conducted by minimizing (1.11) in each iteration, similarly to (Gu, 1993) in the reproducing kernel in Hilbert space.

2 Choosing the Initial Number of Knots

It is well know that the number of knots is, somewhat, a function of the structure of the data, such as the number of modes and asymmetry. For example, if the underlying density has 2 modes then we need to get an estimate based on more knots than if we had a "true" density with a single mode. Even when the true density has a single mode but it is asymmetric we need larger number of basis functions than in the case of symmetric densities. Considering this, the algorithm estimates the number of modes and determines whether the distribution is symmetric.

First a naive, but extremely fast estimate of the unknown density is provided by the histogram of the data. The number of equal intervals of this histogram is given by the integer part of the logarithm of the number of observations, denoted by $N = \lceil \log n \rceil$. To avoid overcounting the number of modes a smoothing procedure is applied on the histogram. To describe such procedure, let $freq(j)$ be the frequency of jth class (interval) of the histogram, with $j = 1, \ldots, N$. Below we describe an algorithm to estimate the number of modes.

Algorithm: Estimating the number of modes

`For` $j = 1, \ldots, N-1$

(a) `if` $(freq(j) - freq(j+1) < 0)$, $sign(j) = -1$.

(b) `if` $(freq(j) - freq(j+1) > 0)$, $sign(j) = 1$.

(c) `if` $freq(j) = freq(j+1)$ `and` $freq(j) = 0$ `then` $sign(j) = 0$.

(d) `if`$freq(j) = freq(j+1)$ `and` $freq(j) \neq 0$ `then` generate u a uniform random number on $(0,1)$ and check: `if` $u < .5$, $sign(j) := -1$, `else` $sign(j) := 1$.

`end for` j.

`For` $k = 1, \ldots, N-2$

(e) `if` $sign(k) * sign(k+2) > 0$ `then` $sign(k) = sign(k+1)$.

`end for` k.

`For` $l = 2, \ldots, N-1$

(f) `if` $sign(l) * sign(l+1) = -1$ `then` $count = count + 1$

`end for` l.

(g) Compute the number of modes m

$$m = \lceil .5(count + 1) \rceil$$

For asymmetric densities more basis functions are necessary, but this implies in higher computational cost. It would be interesting to detect densities belonging to asymmetric families, e.g., Gamma and Beta (and their scale translations) and to apply well known transformations in such way that makes the underlying density approximately symmetric, and hence reducing the computational cost. See for details (Hinkley, 1975; Hinkley, 1977). By default asymmetry is tested as proposed by (Hinkley, 1977), and if the distribution is asymmetric and unimodal, the usual power transformation is applied to $(X_i - X_{(1)})$ (in order to scale the data to the interval $(0, \infty)$), where $X_{(1)} = \min\{X_1, \ldots, X_n\}$. The best optimal value of the power is obtained based on the data set (Hinkley, 1975; Hinkley, 1977). However the transformation is optional.

The symmetry test used can be described as the following, let $X_1, \ldots, X_n$ be i.i.d. nonnegative random variables, $\tilde{X}$ the sample median, $\bar{X}$ the sample mean, and S_n the sample standard deviation. if

$$T = \frac{\sqrt{n}(\tilde{X} - \bar{X})}{S_n} \geq c \; ,$$

we say the density is asymmetric. Otherwise the density is considered to be symmetric. The constant c depends on the $P(T \geq c) = \alpha$, where α is prespecified. Chernoff, Gastwirth and Jones (1967) proved that $\tilde{X} - \bar{X}$ divided by an appropriate constant converges in distribution to a standard normal. Our empirical study with one thousand replications of normal densities with sample sizes 100, 500 and 1000 showed that T is approximately normal with mean zero and standard deviation 0.75. Hence, $T = \frac{\sqrt{n}(\tilde{X} - \bar{X})}{0.75 S_n}$ is approximately normal distributed with mean zero and variance one. Thus the value of c in our study is 1.96×0.75

Simulations have shown that when the shape parameter of Gamma density is 2, 3 and 4 with sample size 200, Hybrid splines is about 3 times faster than when it does not apply a transformation.

The Hybrid splines code estimates the true density of the transformed data and back-transforms the estimate to original scale. As mentioned (Scott, 1992), we also observed that when the estimate is transformed back, it returns a smoother estimate than if it was computed in the original scale. An explanation for that fact would be that the H-splines code for transformed data needs fewer knots than for the untransformed data (for $n \leq 300$, roughly $\log(n)$ and $\sqrt{n} - 4$ respectively). Consequently we obtain a smoother curve in the original scale.

Figure 2.1 shows the effect of power transformation on the original data points. One observes that the result is quite interesting because the estimate obtained after transformation, visually, is very close to the true density. In fact the symmetrized

Kullback-Leibler distance between the true density and the estimate is 0.017 while, without transforming the data the symmetrized Kullback-Leibler distance is 0.358.

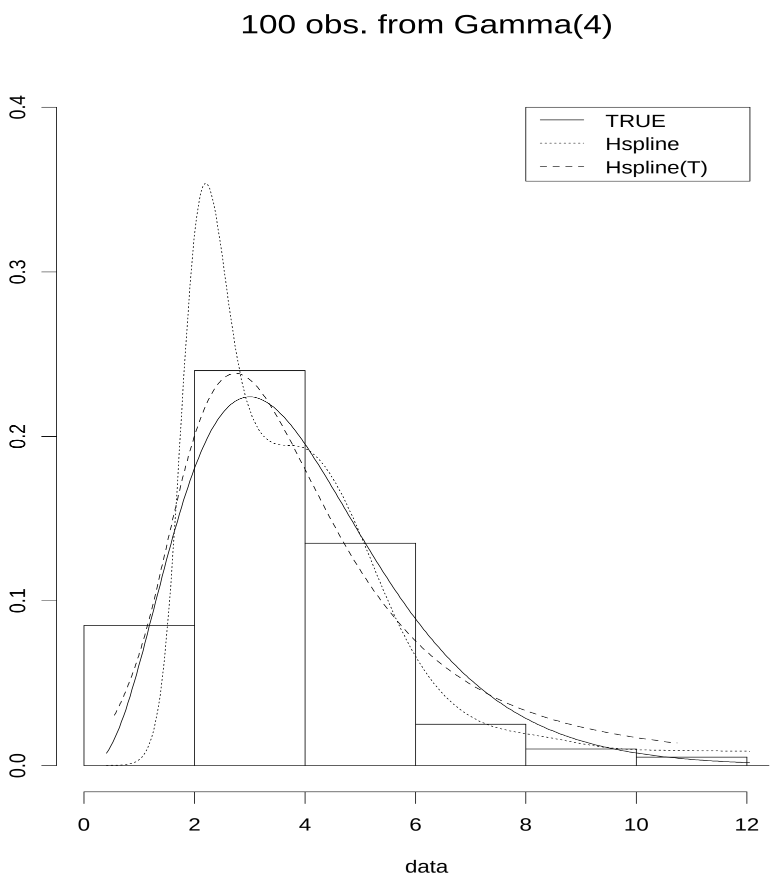

Figure 2.1: Comparison of True density (solid line), H-splines estimate using original data (dotted line) and H-splines estimate using transformed data (dashed line)

In the general case, simulations were performed in order to find approximately the number of basis functions that the algorithm needs to start computing a density estimate. This empirical study includes the following unimodal distributions: gamma with shape parameters 1, 2, 3 and 4, normal with means 0, 3, 5, 7 and variance

1, normal with means 30 and 50, 80 and variances 5, 10, 20. In the situation of multimodality we used mixtures of normal distributions, except in the case when the underlying density has 8 modes. A typical example of such a density can be seen in Figure 4.4 , although the number of simulations for that particular type of density were smaller than the other densities mentioned before. The samples sizes in those simulations were 100, 200, 300 and 500 with a very small study involving samples with 50 and 80 data points. Given a sample with n observations, we noticed that for single mode symmetric distributions, the symmetrized Kullback-Leibler distance between the true density and the estimates based on $\log n$ or more were approximately 1% and estimates based on more than $3 + \log n$ basis functions could not be distinguished visually. Similar conclusions were obtained for the other densities included in this study. However, the number of basis functions required for densities other than the those with a single mode and symmetric is approximately $\sqrt{n}$. Since the algorithm increases the number of basis functions by one, we decided to start the algorithm with fewer knots than the ones indicated by simulations. Based on this empirical study, we suggest the following table for the initial number of knots when $n \leq 300$. For larger values of n the algorithm initiates with $\log n$ knots.

Table 2.1. Initial number of knots.

no. of modes	Other criteria	no. of knots
1	symmetric	$\log(n) - 1$
1	asymmetric	$\sqrt{n} - 4$
2 or 3	-	$\sqrt{n} - 4$
> 4	-	$\sqrt{n} - 2$

Figure 2.2 exhibits the curves of the true density and the estimates based on 5, 7 and 15 knots. Note that there is considerable difference between estimates with 5 and 7 knots but this difference is relatively small between 7 and 15 knots, and so it seems reasonable to choose initial number of knots as 5 to start computing the estimates. We notice that, if the number of knots K is too small the algorithm will need more basis functions to catch the structure of the data and comparing the distance between density estimates with K and $K + 1$ it could stop too early because both curves are too alike but not good estimates.

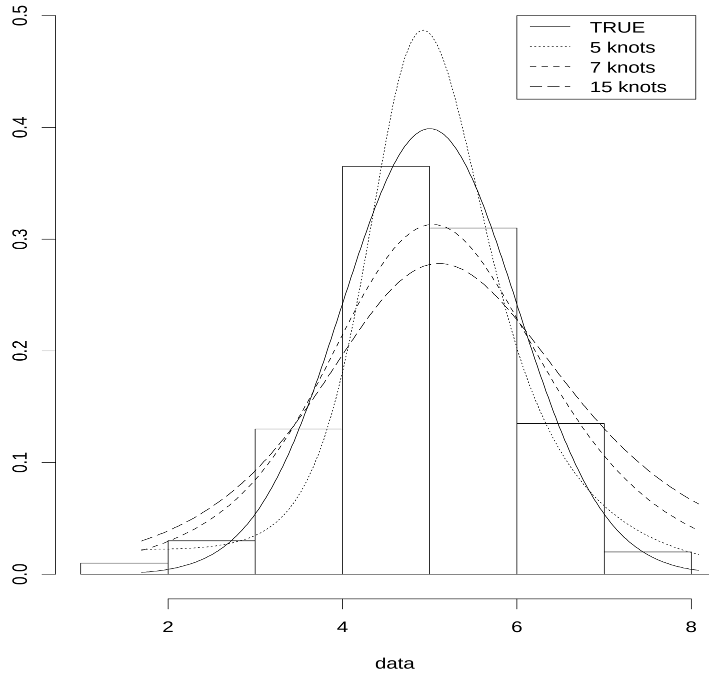

Figure 2.2: Density N(5,1) and 3 estimates for number of knots=5, 7 and 15

It is well known that knots placement plays an important role to get good density estimates. (Kooperberg and Stone, 1991) designed a knot placement where knots are placed on the order statistics with fixed knots on $X_{(1)}$ and $X_{(n)}$. This attempt was tried but simulations show that equally spaced design for H-splines is much faster and performs equally better. Except on pathological cases with outliers, not equally space knots placement is more appropriate.

3 Finding the optimal smoothing parameter and the number of knots.

In this section we describe the algorithm to compute Hybrid splines which is going to be called H-splines. The algorithm provides a fully automatic procedure to estimate a density based on $K < n$ basis functions. Given the data $X_1, \ldots, X_n$, the procedure starts with a initial number of knots and places them equally spaced in the range of the observations. To compute an estimate it fixes a smoothing parameter λ_0 large and computes an iterate $\tilde{g}_{\lambda_0}$ by solving the Newton updating equation (1.5). Now, instead of computing the next iteration using the same λ_0, the algorithm searches for λ_i that minimizes a proxy of symmetrized Kullback-Leibler distance $\hat{L}_{\lambda_i}$. Then it iterates the Newton updating equation (1.5) using $\tilde{g}_{\lambda_i}$. The algorithm repeats previous steps until it attains convergence. When convergence is achieved we obtain a density estimate $\hat{f}_K$ based on K knots. Now, the algorithm computes an estimate $\hat{f}_{K+1}$ based on $K+1$ knots and evaluates the symmetrized Kullback-Leibler distance between $\hat{f}_K$ and $\hat{f}_{K+1}$. If the computed distance is greater than a prespecified number δ, then it increases the number of knots by 1 and starts the procedure again. Otherwise the procedure stops and provides an estimate based on $K+1$ knots.

Similar simulations as mentioned before but with a small number of replications (100) were performed. The symmetrized Kullback-Leibler distance between density estimates based on K and $K+1$, for $K = 5, \ldots, 35$ were computed. The plot of medians showed that the symmetrized Kullback-Leibler distances between densities become approximately constant when the number of knots is greater than K_f, where K_f depends on the the structure (curvature, modes) of the underlying density f. In general, this empirical study showed that K_f is greater than $K + 2$. Similar evidence was found when measuring the symmetrized Kullback-Leibler distance between the true density and the estimates. However, the number of replications in this study was small, suggesting some caution in interpretation. Moreover, based on these simulations the difference among estimates whose distance does not exceed .05 seem not to be distinguished by eyeball.

4 Numerical Simulations

In this section we present various examples of the performance of Hybrid Splines. Since the examples deals with density estimation in the one dimensional case, the SSDE (smoothing splines density estimation (Gu, 1993)) is not going to be considered because Gu's algorithm is of order of $O(n^3)$ while logspline is of order of $O(n)$. One of the goals of H-splines method is to provide a code that it will be possible to extend to the multivariate case keeping the same speed of logspline. Logspline and SSDE are state of art codes and any new code should have a performance comparable to them.

The data sets were obtained by simulations either by using RATFOR programs based on algorithms found in (Kennedy, Jr and Gentle, 1980) or by making use of S-language (Becker, Chambers and Wilks, 1988) routines.

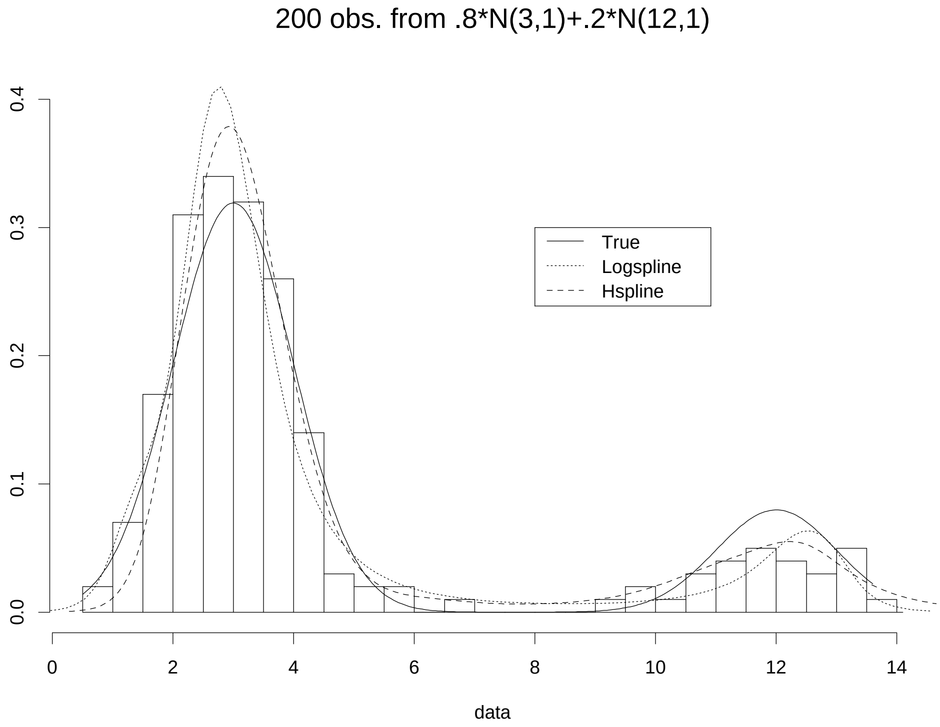

Figure 4.1: True density, Logspline and Hybrid Spline methods are compared

Figure 4.1 show us that even under obvious case when two normal distributions with different means are combined, Hybrid Spline method has a very satisfactory performance.

In Figure 4.2, we can see that H-splines produces a good estimate of the density. Also we notice that in the tails Logspline with deletion procedure provides a more precise estimate than H-splines.

Comparing Hybrid Spline method with Logspline using deletion procedure we

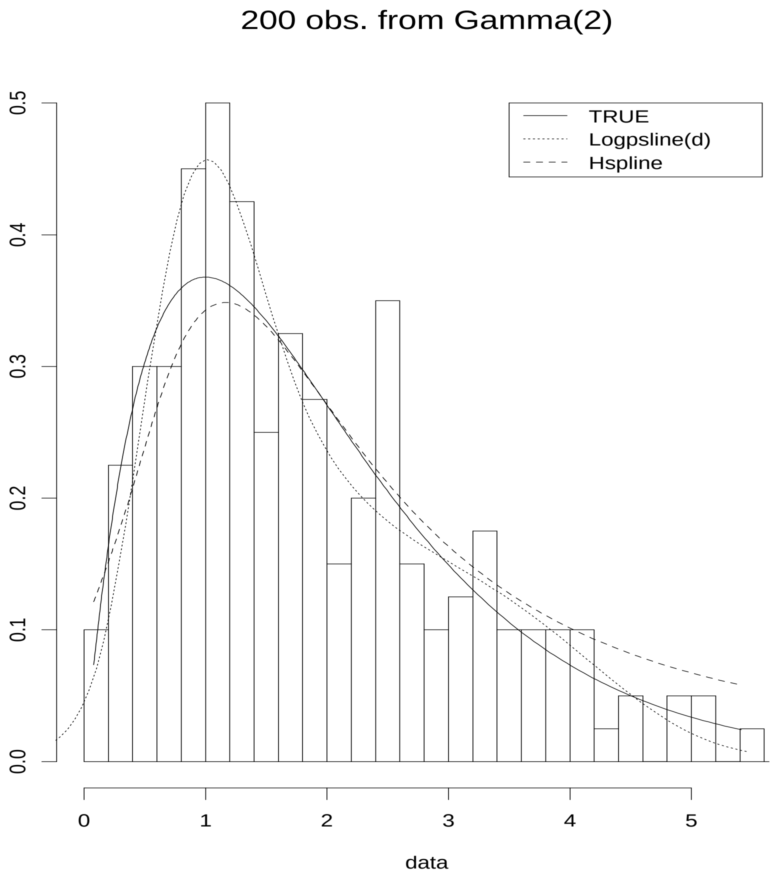

Figure 4.2: True density of Gamma(2), estimates given by Logspline, Hybrid Splines

see in Figure 4.3 that Hybrid Spline does a better job identifying peaks and valleys of the underlying density. Deletion procedure of Logspline allows it to have the position of the knots that is more significant for the fitting and hence provides the best estimate that Logspline method can produce. Note that logspline adapts using deletion procedure while H-spline adapts using addition.

Figure 4.4 shows that the kernel method produces the best and fastest estimate of the underlying density. The bandwidth was chosen by eyeball and since that

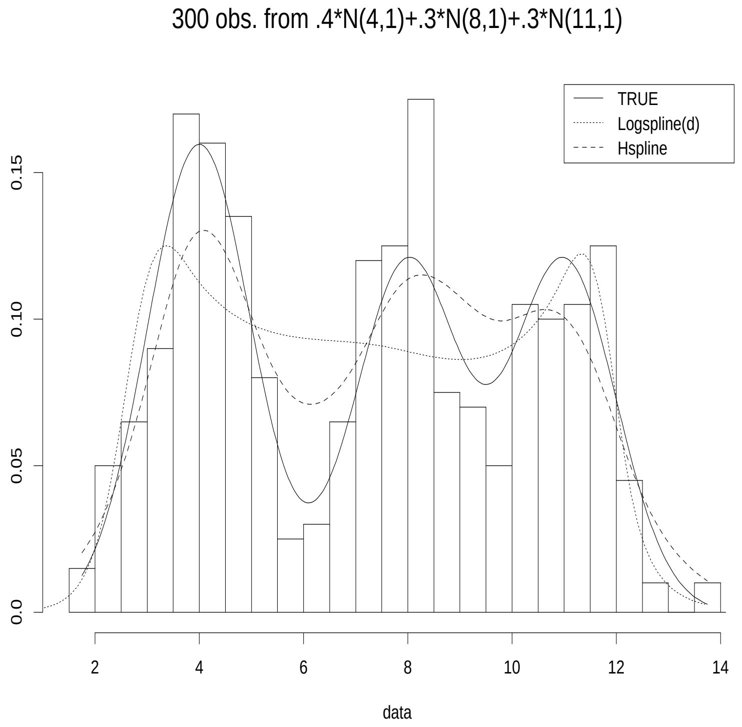

Figure 4.3: True density, Logspline and Hybrid Spline estimates

one might have a prior information about the true density, this choice can be well accepted. We notice that Hybrid Spline, again, has a superior performance over the other methods when identifying 5 of the 6 peaks and 6 out of 7 valleys. Hybrid spline has shown in simulations that it is suitable in situation where the data have a lot of structure.

In figure 4.5 we have the age in years of 256 patients under treatment for diabetes. Comparing Hybrid spline and Logspline with deletion procedure, we see that the density estimate obtained by Hybrid spline is able to detect the same number of modes that the histogram indicates while Logspline estimate seems not to detect the

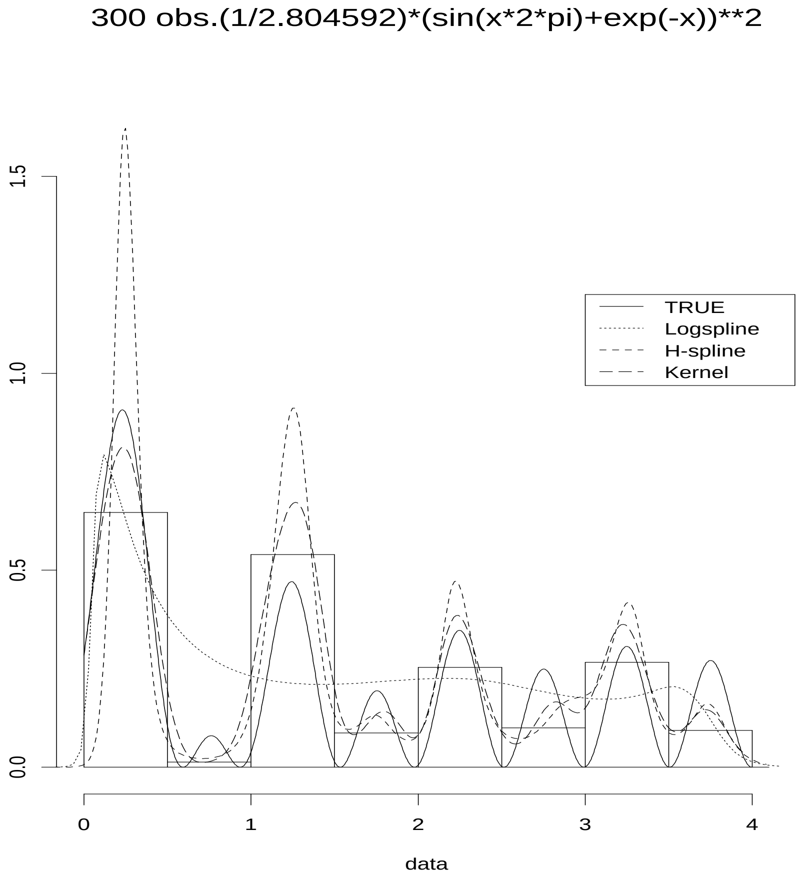

Figure 4.4: True density and estimates given by Logspline, Hybrid spline and kernel methods

second bump. Also, as histogram shows, there is some evidence of another bump between 9 and 15 years which was not detect, in this data set, by Logspline estimate. Since we don't know the true density of this data we cannot tell for sure which of those estimates is more appropriate to reproduce the underlying density but it seems that the histogram and Hybrid spline provide very similar estimates although the number of intervals of the histogram given here is not the default of most of the

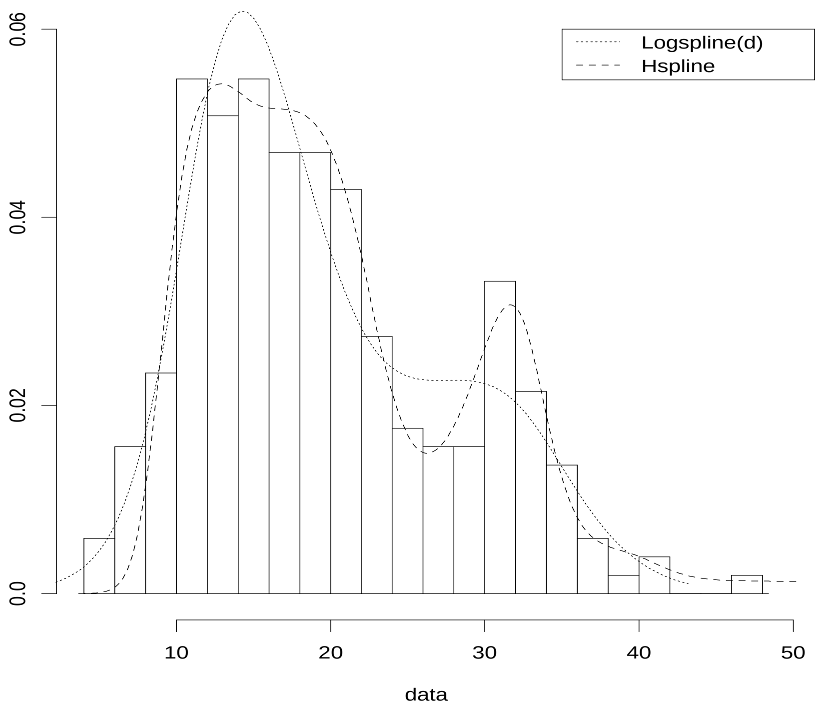

Figure 4.5: Age in years of 256 patients in diabetes study

softwares available. Under the default the histogram would have 7 intervals and hence it wouldn't be able to identify the second mode. That is, asking for more details, the histogram gives an estimate where one could see the fine structure of this data set. Such requirement, increasing the number of intervals, is not necessary if ones uses Hybrid spline procedure.

5 Conclusion

Comparing H-spline, Logspline and SSDE one may observe that H-spline yields good estimates when the data have more complicated structure, such as several modes, or when the observations come from a gamma family of distributions. However, this empirical study performed simulations using a restricted range of shape parameters of gamma densities. For symmetric densities, H-spline performs as well as Logspline and SSDE, but it is faster than SSDE. However, in some cases H-splines obtains a smoother density estimate. Comparing performance, Logspline is superior in terms of speed or computational cost because it assumes that a density can be written as linear combinations of the basis functions (B-splines) and estimates the coefficients of this linear combination by Maximum Likelihood Method (a parametric procedure), while H-spline and SSDE minimize a penalized loglikelihood equation in such way that the optimal smoothing parameter and "optimal" density estimate are iteratively computed based on a loss function which is the symmetrized Kulback-Leibler distance. Smoothing of the Logspline estimate involves the choice of the numbers and the locations of knots in adaptive manner (fewer knots, smoother curve). The initial number of knots is determined from a table according to the number of observations, based on the experience of Kooperberg and Stone (1991). When the number and placement of knots of Logspline and H-spline approaches are the same, H-spline produces smoother estimates. An explanation for this situation would be that H-spline controls the trade-off between goodness-of-fit and smoothness. Consequently, H-spline appears to be useful for estimating densities with more complicated structure. Density estimation is not recommended for small samples and hence H-spline approach should be used for large samples (> 80).

Acknowledgements: This work is part of the author's Ph.D. thesis and was financially supported by CAPES, grant no. 10405/88. The author would like to thank Prof. Grace Wahba for her advice and help during his Ph.D. He is also grateful to Prof. Brian Yandell for useful discussions and Prof. C. Kooperberg for kindly sending data sets to test this algorithm.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth Publishing Co (Belmont CA).

Chernoff, H., Gastwirth, J. L. and Jones, M. V. (1967). Asymptotic distribution of linear combinations of function of order statistics with applications to estimation, *Ann. of Mathematical Stat.* **38**: 52–72.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators, *Ann. of Statistics* **18**: 1676–1695.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, New York.

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

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm, *J. of the Amer. Stat'l. Assn.* **88**: 495–504.

Gu, C., Bates, D. M., Chen, Z. and Wahba, G. (1989). The computation of generalized cross-validation functions through Householder tridiagonalization with applications to the fitting of interaction spline models, *SIAM J. on Matrix Analysis and Application* **10**: 457–480.

Hinkley, D. V. (1975). On power transformations to symmetry (corr: V63 p218), *Biometrika* **62**: 101–112.

Hinkley, D. V. (1977). On quick choice of power transformations, *Applied Statistics* **26**: 67–68.

Kennedy, W. J., Jr and Gentle, J. E. (1980). *Statistical Computing*, Dekker:NY.

Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation, *Computational Stat. and Data Analysis* **12**: 327–347.

Leonard, T. (1978). Density estimation, stochastic processes and prior information, *JRSS-B, Methodological* **40**: 113–146.

O'Sullivan, F. (1990). An iterative approach to two-dimensional Laplacian smoothing with application to image restoration, *J. of the Amer. Stat'l. Assn.* **85**: 213–219.

Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice, and Visualization*, John Wiley Sons (New York, Chichester).

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality, *JRSS-B, Methodological* **43**: 97–99.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method, *Ann. of Statistics* **12**: 898–916.

Wahba, G. (1971). On the regression design problem of Sacks and Ylvisaker, *Ann. of Mathematical Stat.* **42**: 1035–1053.

Wahba, G. (1975a). Interpolating spline methods for density estimation. I: Equispaced knots, *Ann. of Statistics* **3**: 30–48.

Wahba, G. (1975b). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation, *Ann. of Statistics* **3**: 15–29.

Wahba, G. (1977). Optimal smoothing of density estimates, *Classification and Clustering* pp. 423–458.

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates, *Ann. of Statistics* **9**: 146–156.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM:PA.