

A note on maximum likelihood density estimation using a proxy of the Kullback-Leibler distance.

Ronaldo Dias

*Universidade Estadual de Campinas **

Abstract

Given a random sample from a continuous and positive density f , the logistic transformation is applied and a log density estimate is provided by using B-splines. The log density estimate maximizes the likelihood function which has equivalent solution when subject to a constraint that guarantees identifiability of the model. An finite approximation is provided and the number of basis functions which acts as the smoothing parameter is estimated by minimizing a proxy of the Kullback-Leibler distance.

Keywords: non-parametric density estimation; B-splines; partitions of unity.

1 Introduction

Let X_1, \dots, X_n be i.i.d. random variables with an unknown continuous and po-

*Postal address: Departamento de Estatística, IMECC, Cidade Universitária "Zeferino Vaz", Caixa Postal 6065, 13.081-970 - Campinas, SP - BRAZIL, e-mail address: dias@ime.unicamp.br

sitive probability density function f on a compact interval \mathcal{X} . The goal of this paper 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 constraints on f , we use the logistic transformation (Leonard1978) $f = e^g / (\int e^g)$. It is easy to see that this transformation is not one-to-one, since $g_1 = g + c$ implies $f = e^{g_1} / (\int e^{g_1}) = e^g / (\int e^g)$. Gu (1993) and Dias (1998a) have used a side condition on g , such as $g(x_0) = 0, x_0 \in \mathcal{X}$ or $\int_{\mathcal{X}} g = 0$, needed to determine this transformation uniquely. The search for the function g will be in a infinite dimensional space which is, in general, not computable. Therefore, a finite approximation is necessary.

A well known method to find a good estimate for f is the penalized log-likelihood (Silverman (1982) and Gu (1993)), which has been used in problems of smoothing and requires a high computational cost, in general $O(n^3)$ where n is the sample size. Dias (1998a) and Dias (1998b) suggested adaptive procedures to reduce the computational cost by introducing the H-splines method that combines ideas from regression and smoothing splines approaches. Both approaches rely on the estimation of the smoothing parameters and require computational intensive methods. Differently from other smoothing methods, the proposed procedure in this paper does not penalize the likelihood function and the general solution for this optimization problem, in a finite dimensional space, is equivalent to optimization with constraint that enforces identifiability to the model. Section 2 introduces a method completely based on Basis functions where the number of basis acts as the smoothing parameter. Section 3 shows how the procedure determines K , the number of basis functions, by a proxy of the Kullback-Leibler distance. The coefficients of the expansion are found by maximizing the likelihood. Section 4 presents some simu-

lation results using a code written in Splus and R that does not require dynamic loading, therefore regarding portability it can be used in any personal computer.

2 Finite Approximation

Let \mathcal{G} be the set of real function g on \mathcal{X} for which:

1. $\log \int e^g < \infty$;
2. The $(m - 1)$ th derivatives of g exist and are piecewise differentiable, for $m=1,2,\dots$

Consider \mathcal{N} , the set of functions g which are linear combinations of cubic B-splines, that is, $g = \langle \theta, M \rangle_K = \sum_{j=1}^K \theta_j M_j$, where M_j are the well known normalized cubic B-splines (see de Boor (1978)). It is easy to check that $\mathcal{N} \subset \mathcal{G}$. Given the data X_1, \dots, X_n i.i.d. random variables on a compact interval \mathcal{X} , assume that $f = e^g / \int e^g$, $g \in \mathcal{N}$, that is, f can be written as

$$f(x|\theta, K) = \frac{e^{\langle \theta, M(x) \rangle_K}}{\int e^{\langle \theta, M(x) \rangle_K} dx}.$$

Hence, the log-likelihood equation is,

$$L_K(\theta) = \frac{1}{n} \sum_{i=1}^n \langle \theta, M(X_i) \rangle_K - \log \int e^{\langle \theta, M(x) \rangle_K} dx. \quad (2.1)$$

Observe that, by considering $g \in \mathcal{N}$, we reduce the problem of choosing g from an infinite-dimensional class of functions to a finite class of functions \mathcal{N} since the dimension of \mathcal{N} is finite (see de Boor (1978) for details). The optimization problem is to find, for a fixed K , a vector $\hat{\theta} = \hat{\theta}^{(K)} = (\theta_1, \dots, \theta_K) \in \Theta \subseteq \mathbb{R}^K$ the maximizer of $L_K(\theta)$. Our estimate will be $\hat{f}_K = e^{\hat{g} - \log \int e^{\hat{g}}}$ such that, $\hat{g} = \langle \hat{\theta}, M \rangle_K$. To make

the logistic transformation one-to-one we have to enforce the side condition $\int g = 0$ which implies that $\sum_{j=1}^K \theta_j = 0$, since $\int M_j = 1$ (normalized cubic B-splines) for $j = 1, \dots, K$. Let $\Theta_0 = \{\theta \in \Theta \subset \mathbb{R}^K : \sum_j^K \theta_j = 0\}$

Lemma 2.1 *For a fixed K , $L_K(\theta)$ is concave in θ . Moreover, $L_K(\theta)$ is strictly concave for $\theta \in \Theta_0$. Hence there exists at most one maximizer on Θ_0*

Proof. It is enough to show that $-\log \int e^{\langle \theta, M(x) \rangle_K}$ is concave in θ . For this, take $\theta_1, \theta_2 \in \Theta$ an open set in \mathbb{R}^K , and $\alpha, \beta > 0$, $\alpha + \beta = 1$. We have, by applying Holder's inequality,

$$\log \int e^{\alpha \langle \theta_1, M \rangle_K + \beta \langle \theta_2, M \rangle_K} \leq \alpha \log \int e^{\langle \theta_1, M \rangle_K} + \beta \log \int e^{\langle \theta_2, M \rangle_K} < \infty. \quad (2.2)$$

Note that, the equality holds in (2.2) if and only if, $e^{\langle \theta_1, M \rangle_K} = |\gamma| e^{\langle \theta_2, M \rangle_K}$ for some γ which amounts to $\theta_2 = \theta_1 + c$, where c is a constant. Therefore, if $\theta_1, \theta_2 \in \Theta_0$ with $\theta_1 \neq \theta_2$, we have strict inequality in (2.2) and $L_K(\theta)$ is strictly concave if we restrict θ to the subspace Θ_0 . Moreover, it is not difficult to show that $L_K(\theta)$ is continuous and at least twice differentiable in θ for a fixed K . Thus, restrict to Θ_0 one may guarantee a unique density estimate.

The next theorem shows the relationship between the maximizers $\hat{\theta}$ in Θ and θ^* in Θ_0 .

Theorem 2.1 *If the vector $\hat{\theta}$ maximizes $L_K(\theta)$ then $\theta^* = \hat{\theta} - \frac{1}{K} \sum_{j=1}^K \hat{\theta}_j$ maximizes $L_K(\theta)$ subject to $\sum_{j=1}^K \theta_j = 0$. Moreover, θ^* is unique.*

Proof. For all $c_\theta : \Theta \subset \mathbb{R}^K \rightarrow \mathbb{R}$, $K \geq 1$,

$$L_K(\theta + c_\theta) = \frac{1}{n} \sum_{i=1}^n \langle (\theta + c_\theta, M(X_i)) \rangle_K - \log \int e^{\langle \theta + c_\theta, M(x) \rangle_K} dx$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n \langle \theta, M(X_i) \rangle_K + c_\theta \langle 1, M(X_i) \rangle_K - \log \int e^{\langle \theta, M(x) \rangle_K + c_\theta \langle 1, M(x) \rangle_K} dx \\
&= L_K(\theta),
\end{aligned} \tag{2.3}$$

since, by the partition of unity property, we have

$$\langle 1, M(x) \rangle_K = \sum_{j=1}^K M_j(x) = 1, \quad \forall x.$$

As a consequence of (2.3), we have,

$$\max_{\theta} L_K(\theta) = \max_{\theta: c_\theta=0} L_K(\theta).$$

Therefore, if $\hat{\theta}$ is such that $L_K(\hat{\theta}) = \max_{\theta} L_K(\theta)$, then $L_K(\theta^*) \geq L_K(\theta)$, for all θ so that, $\sum_{j=1}^K \theta_j = 0$. In fact, by (2.3)

$$\begin{aligned}
L_K(\theta^*) &= L_K\left(\hat{\theta} - \frac{1}{K} \sum_{j=1}^K \hat{\theta}_j\right) \\
&= L_K(\hat{\theta}) \geq L_K(\theta) \quad \forall \theta.
\end{aligned} \tag{2.4}$$

To see that θ^* is the unique, we have to recall the properties of B-splines and the fact that the maximum log likelihood function is strictly concave on Θ_0 , hence if the maximum likelihood estimator exists then it is unique. Let t_1, \dots, t_p be sequence of knots such that $-\infty, < t_1, < t_2 < \dots < t_{p-1} < t_p < \infty$. Suppose $(-\infty, t_1]$ and $[t_p, \infty)$ have at least one observed value and four or more observations lie in the compact sets $[t_1, t_2], \dots, [t_{p-1}, t_p]$, then $\hat{\theta}$ exists.

For fixed K , let $\hat{\theta}_n^{(K)}$ be defined as

$$\hat{\theta}_n^{(K)} = \arg \max_{\theta \in \Theta_0} L_K(\theta),$$

Notice that, in fact,

$$L_K(\theta) = \langle \theta, \bar{M} \rangle_K - \log \int e^{\langle \theta, M(x) \rangle_K} dx,$$

then $\hat{\theta}_n^{(K)}$ is the unique solution of the equation where \bar{M} is a K -dimensional vector with j -th components given by

$$\frac{1}{n} \sum_{i=1}^n M_j(X_i) = \bar{M}_j, \quad j \in \{1, \dots, K\}. \quad (2.5)$$

Since $L_K(\theta)$ is at least twice differentiable we have $\hat{\theta}_n^{(K)}$ as the unique solution of the equation,

$$\frac{\partial L_K(\theta)}{\partial \theta} := h(\theta, \bar{M}) = 0, \quad (2.6)$$

where, $\bar{M} = (1/K) \sum_{j=1}^K \bar{M}_j$ and $h : \Theta_0 \times [0, \infty)^K \rightarrow \mathbb{R}^K$ with j -th entry,

$$h_j(\theta, \mathbf{u}) = u_j - \frac{\int \exp(\langle \theta, M(z) \rangle_K) M_j(z) dz}{\int \exp(\langle \theta, M(z) \rangle_K) dz}, \quad (2.7)$$

for $j \in \{1, \dots, K\}$.

Lemma 2.2 *Let θ_0 be the unique solution of*

$$h(\theta, \int f(x)M(x)dx) = 0$$

in Θ_0 , then for fixed K , $\hat{\theta}_n^{(K)} \rightarrow \theta_0$ almost surely as $n \rightarrow \infty$.

Proof. The map $h(\theta, \mathbf{u})$ is of the class $C^1(\mathbb{R}^K)$, therefore by the implicit function theorem there exists a function $\phi : \mathbb{R}^K \rightarrow \Theta_0$ of the class $C^1(\mathbb{R}^K)$ such that $h(\theta, \mathbf{u}) = 0 \iff \phi(\mathbf{u}) = \theta$ (by (2.1) ϕ is one-to-one map). Since $\hat{\theta}_n^{(K)} = \phi(\bar{M})$ and by the strong law of large numbers $\bar{M} \rightarrow \int f(x)M(x)dx$ we have that

$$\hat{\theta}_n^{(K)} \rightarrow \phi\left(\int f(x)M(x)dx\right)$$

and defining $\theta_0 = \phi\left(\int f(x)M(x)dx\right)$ completes the proof.

3 Computing the number of basis functions

One may notice the density estimate \hat{f}_K strongly depends on the number of basis functions K which regularizes the optimization problem (2.1). In order to provide an appropriate K , this approach is equivalent to minimize the Kullback-Leibler distance (not a metric, for using Hellinger pseudo metric see Dias (1999)) between the true f and the random function \hat{f}_K .

$$d(f, \hat{f}_K) = \int (\log f - \log \hat{f}_K) f \quad (3.1)$$

Of course, we cannot compute $d(f, \hat{f}_K)$ from the data, since it requires the knowledge of f . But theoretically we can investigate this distance for the choice of an appropriate K . Then, one may define the best K as

$$\hat{K} = \arg \min_{K \in \{1, \dots, K_{max}\}} d(f, \hat{f}_K),$$

for $K_{max} < n$. Define

$$D_n(K) = \int f \log \hat{f}_K.$$

Notice that $D_n(K)$ is a random function of K and also can be approximate by

$$Z_n(K) = \frac{1}{n} \sum_{i=1}^n \log \hat{f}_K(X_i).$$

Proposition 3.1 *For any fixed K ,*

$$D_n(K) - Z_n(K) = \sum_{j=1}^K \hat{\theta}_{n_j}^{(K)} \left(\int f(x) M_j(x) dx - \frac{1}{n} \sum_{i=1}^n M_j(X_i) \right) \longrightarrow 0 \quad (3.2)$$

$n \longrightarrow \infty$ almost surely.

Proof. By lemma (2.2), $\hat{\theta}_n^{(K)} \rightarrow \theta_0$ almost surely and applying the Strong Law of Large Number to term inside of the parenthesis of the equation (3.2) completes the proof.

Thus, in fact, the optimization relies on a proxy of the Kullback-Leibler. To compute the proposed density estimate I suggest the following algorithm.

Algorithm 3.1

1. For $K \in \{1, \dots, K_{max}\}$, where $K_{max} < n$, get $\hat{\theta}_n^{(K)}$ the maximizer of $L_K(\theta)$ and compute \hat{f}_K .
2. choose \tilde{K} that maximizes $Z_n(K)$.
3. Go to step 1 and compute $L_{\tilde{K}}(\theta)$, $\hat{\theta}_n^{(\tilde{K})}$ and $\hat{f}_{\tilde{K}}$. Deliver $\hat{f}_{\tilde{K}}$.

Knots' Placement: Placement knots is a very important issue in density estimation via polynomial splines. In this procedure, the fitting can be done either by equally spaced knots (cardinal splines) or by putting the knots at the order statistics as it was first recommended by (de Boor1978). There are several other algorithms which make use of placement knots at order statistics (see also O'Sullivan (1988), Kooperberg and Stone (1991) and Dias (1998a)).

4 Monte Carlo Simulation

In this section we verify the performance of the proposed procedure through typical examples of the simulations. All the simulated data were generated by Splus and R functions. The entire code is written in Splus and R. It does not require any dynamic loading and it can be implemented in any personal computer where Splus (or R) is running. Although the procedure supposes that the density is continuous

and positive on a compact set \mathcal{X} our simulations include test functions which do not have a compact support, e.g., normal and gamma distributions. Nevertheless, for practical purposes a density with a infinite domain can be approximate by a density with an appropriate compact support. For example, a normal density ϕ_{μ,σ^2} with mean μ and variance σ^2 do not differ significantly from a density on $[\mu - 5\sigma, \mu + 5\sigma]$ proportional to ϕ_{μ,σ^2} . Similarly, for densities in the gamma family. Estimation of the support of a density is a very difficult problem (for details see Hall, Nussbaum and Stern (1997)) which has not been answered appropriately and it will not be addressed in this work.

100 obs from N(5,1)

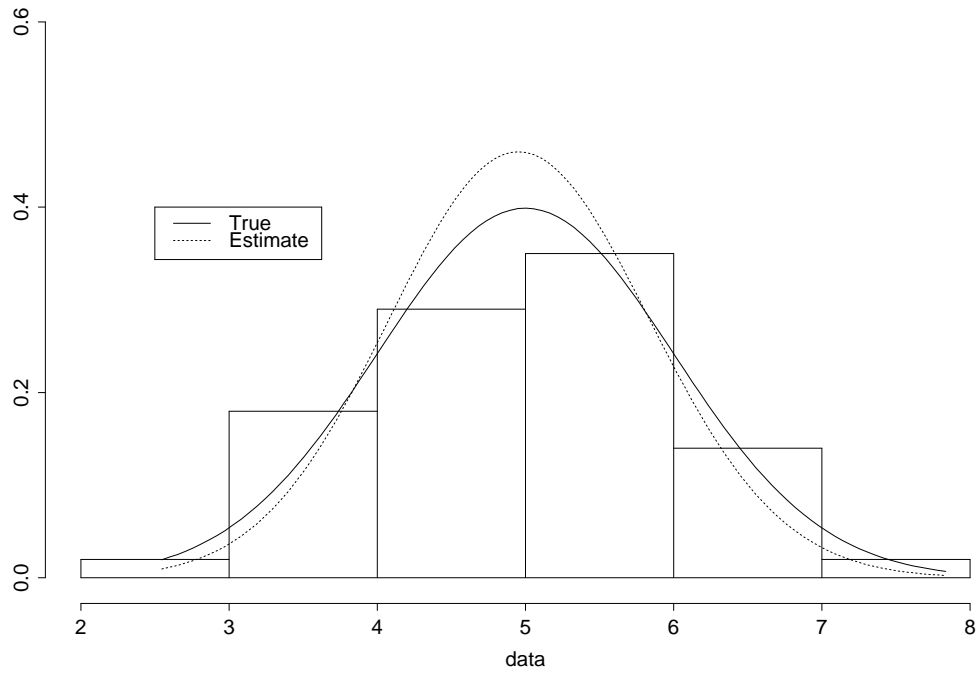


Figure 4.1: One hundred observations from N(5,1) and the optimal K is 3.

Figure 4.1 exhibits a comparison between a true density $N(5, 1)$ and the estimate using 3 basis functions. Visually, we can see this estimate does a good job in estimating the underlying density.

Figure 4.2 and Figure 4.3 show typical example of data coming from a Gamma density with shape parameter equal 3 and Beta density with parameters 5 and 3 respectively. Note that the estimates are very close to the true densities in both cases.

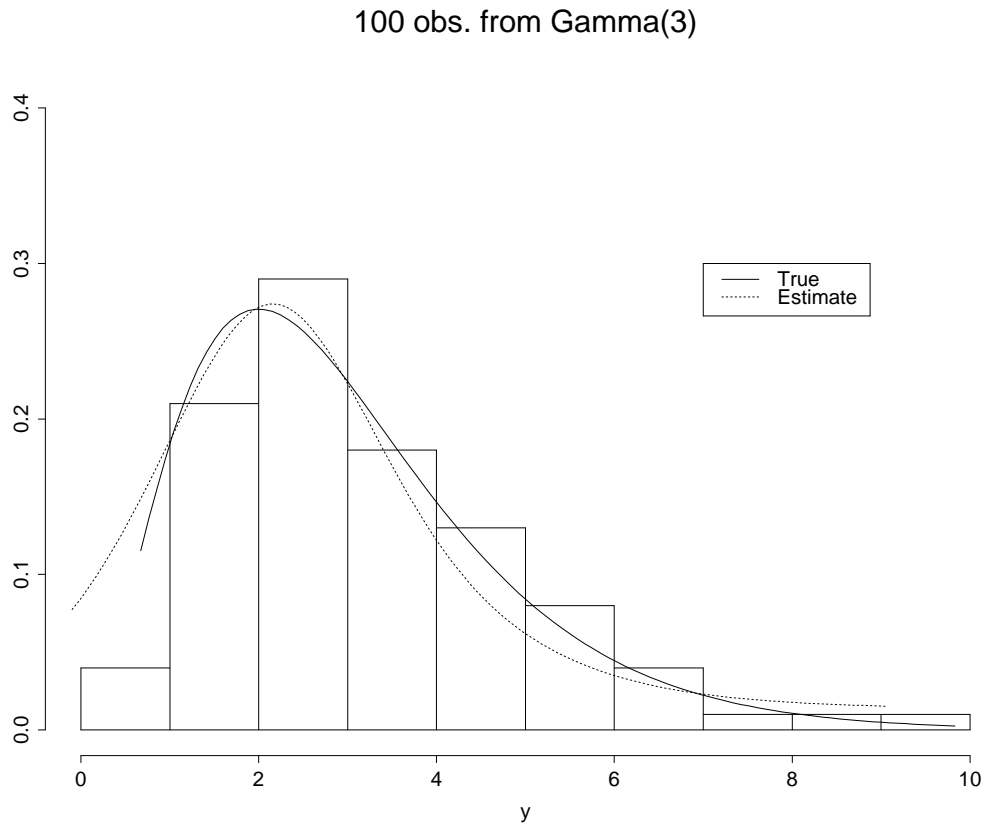


Figure 4.2: two hundred observations from Gamma(3) and the optimal K is 4.

200 obs. from Beta(5,3)

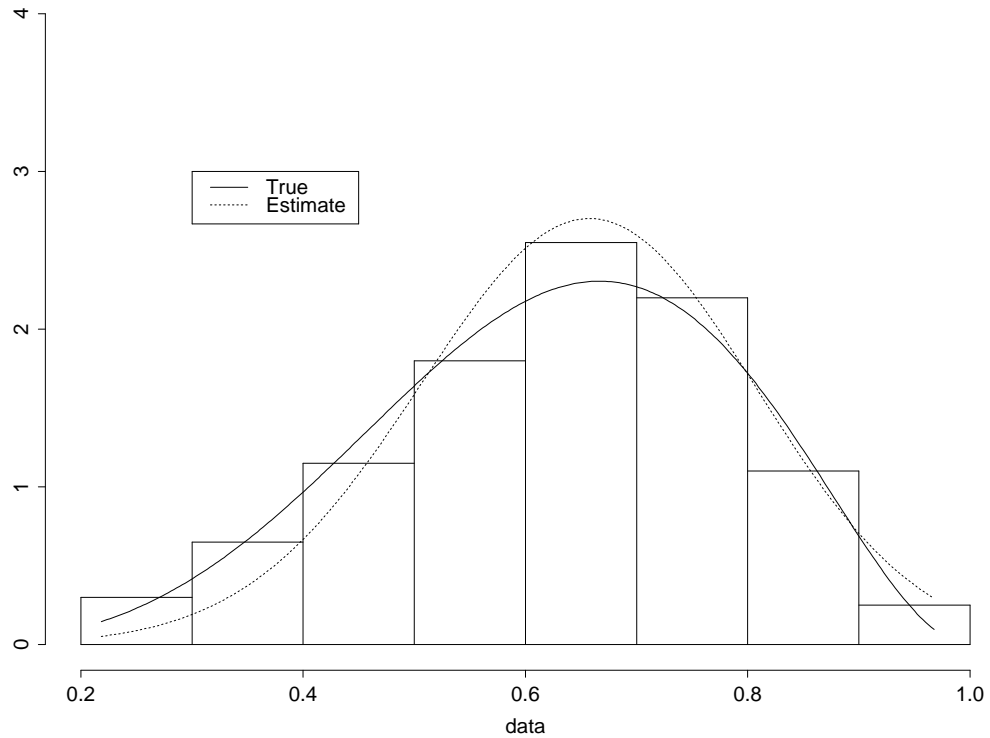


Figure 4.3: The data contain two hundred observations from Beta(5,3) and the optimal K is 4.

Figure 4.4 shows a mixture of two normal distributions with the same variance but different means. As we can see, the fitting seems to be very good.

200 obs from $.6*N(.4,.1)+.4*N(.8,.1)$

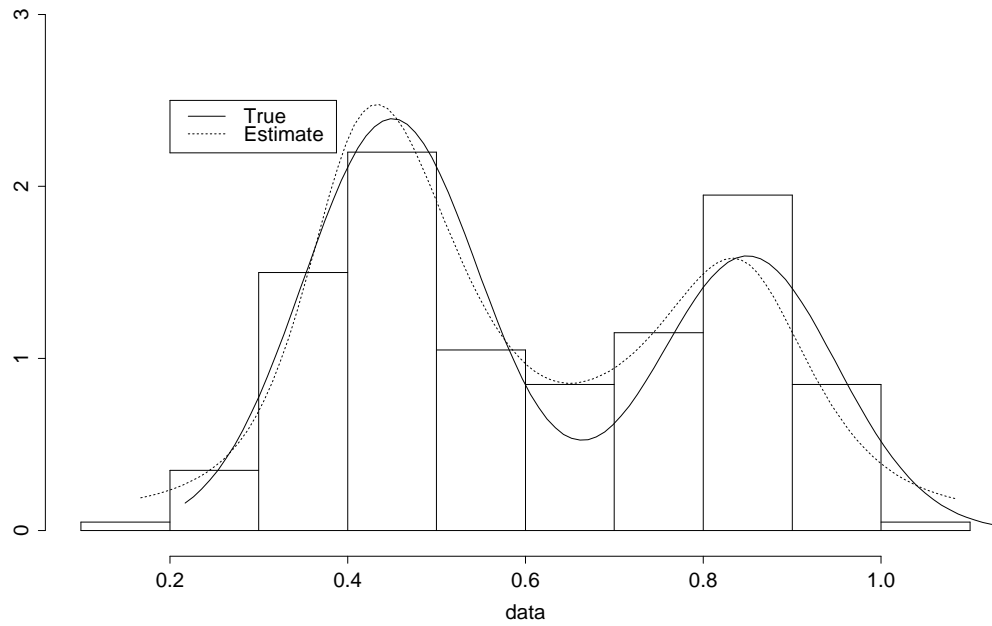


Figure 4.4: The data contain two hundred observations from mixed normal distributions and K is 7.

Figure 4.5 exhibits a real data example where 7126 magnitude of some stars were measured.

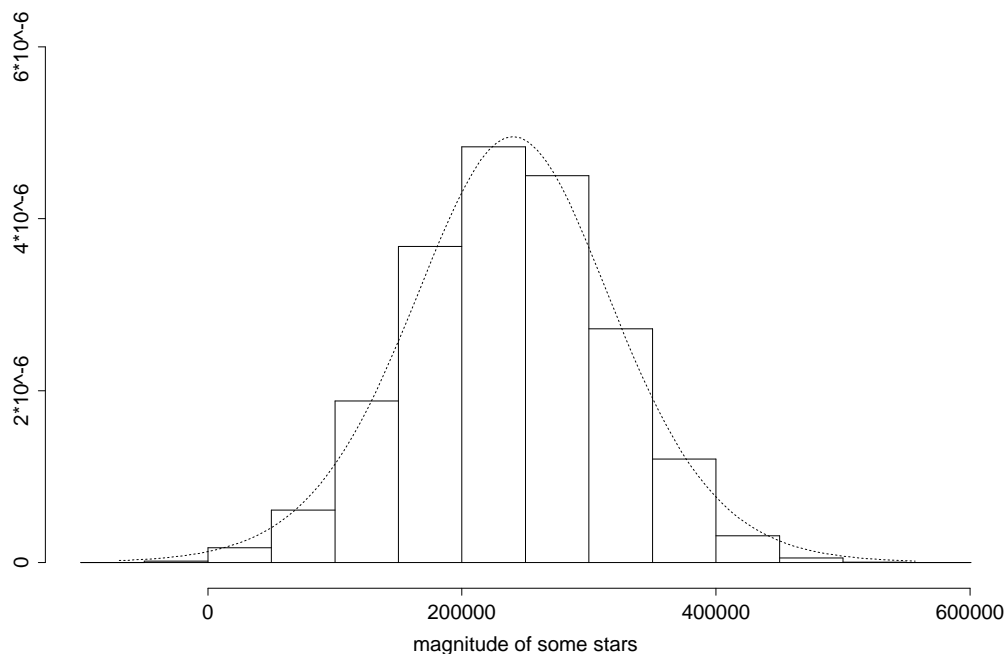


Figure 4.5: Comparison of the density estimate provided by this method and the histogram for the data Magnitude of stars.

On the left of Figure 4.6 we have the histogram (Splus and R default) of 300 observations from the mixture of four normal distributions, $X \sim (.3 \times N(.2, .05) + .2 \times N(.4, .05) + .25 \times N(.6, .05) + .25 \times N(.8, .05))$. It seems to suggest a parametric model from the family of densities $Beta(a, b)$, most likely uniform distribution. On the right, we notice how useful a non-parametric estimate is, it can be a starting point to postulate a mixture of normal distributions for a possible parametric model.

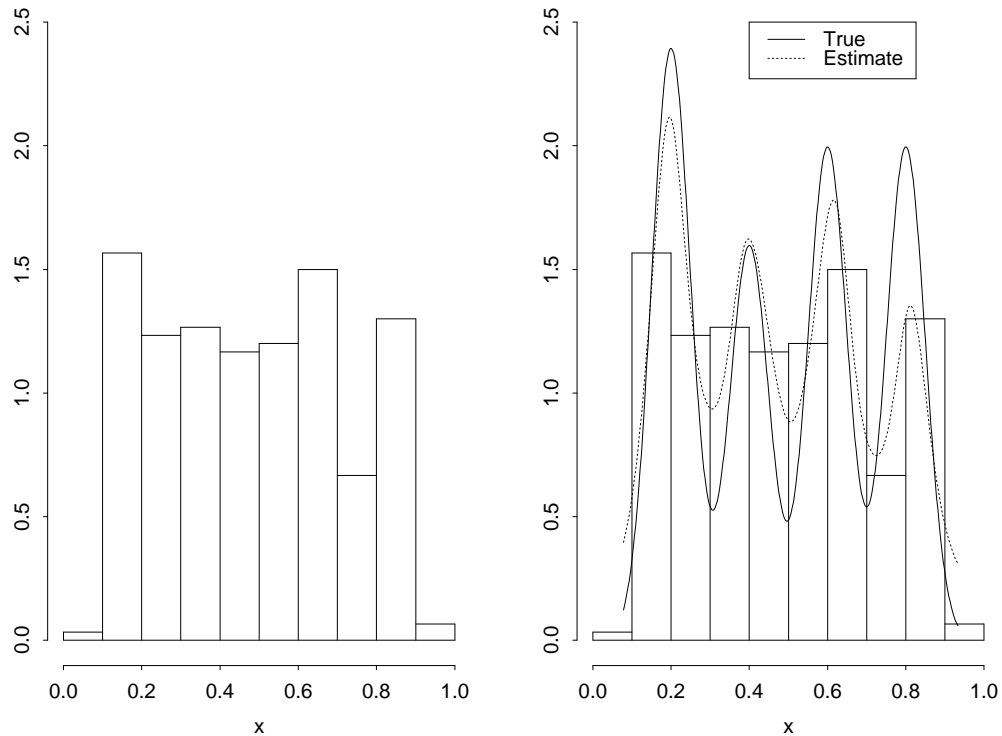


Figure 4.6: Comparison of the density estimate provide by this method and the histogram.

In conclusion, this procedure is easy to use, easy to implement and easy to understand. It provides a unique solution for the optimization problem without having to penalize the likelihood function. Very practical for data analysis when the non-parametric estimation can be a starting guess for a parametric model. Simulations have shown that $K=3$ is appropriate to estimate a single mode. More structured data require a larger number of knots. However, in general, this procedure do not use more than 20 knots as typical value of K_{max} . Also, placement knots on the order

statistics is much better when one knows a priori that the data might have outliers. Non-parametric density estimation, usually, is not recommended for small samples. Nevertheless, this procedure can perform very well for sample size as small as 30 points, if the underlying density is symmetric. For non-symmetric and/or multi-modal densities, the performance of this method can vary from poor to fair when the sample size is smaller than 50 points. Hence for relatively small samples caution must be taken. Comparing with H-spline density estimation (Dias1998a) and smoothing splines density estimation (Gu1993) this procedure can have a good performance for relatively small sample sizes. It is much faster since it uses features of parametric and non-parametric estimation (functional inference, Stone (1990)). Speedwise, it is slower than logspline density estimation (Kooperberg and Stone1991). However, it does not require dynamic memory and hence can be implemented in any personal computer that runs S-plus or R. Moreover, its algorithm is easy to program in any computer language, such as Fortran, that has interface with public libraries where B-splines and optimization routines are available (see details in de Boor (1978)). Comparing with parametric techniques we have, for the non-parametric approach, more flexibility since it allows one to choose the infinity dimensional class of functions that the underlying density belongs. In general, this type of choice depends on the unknown smoothness of the true density. But for the most of the cases one can assume mild restrictions such that a density has an absolutely continuous first derivative and a square integrable second derivative. Nevertheless, non-parametric estimators are less efficient than the parametric ones when a parametric model is valid. For many parametric estimators the mean square error goes to zero with rate of n^{-1} , while non-parametric estimators have rate of $n^{-\alpha}$, $\alpha \in [0, 1]$, and α depends on the smoothness of the underlying curve. When the postulate parametric model

is not valid, many parametric estimators cannot have, *ad hoc*, rate n^{-1} . In fact, those estimators will not converge to the true curve. Consequently, non-parametric estimators are good candidates when one does not know the form of the underlying density. This procedure does a very good job estimating non pathological data sets. Certainly, there are more adaptive methods for non-parametric density estimation (see, for example, Kooperberg and Stone (1991) and Dias (1998a)) but they are more difficult to implement.

ACKNOWLEDGMENT

I would like to thank the referee for the comments and suggestions that made this work better and clearer. Part of this work was done while the author was visiting scholar at Department of Statistics, University of California, Berkeley. This research was partially supported by FAPESP grant no. 99/00261-0 and CNPq grant no. 300644/94-9.

References

- de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, New York.
- Dias, R. (1998a). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.
- Dias, R. (1998b). Prior selection by an adaptive smoothing splines approach, *Random Operator and Stochastic Equations* **6**: 57–60.
- Dias, R. (1999). Sequential adaptive non parametric regression via h-splines, *Communications in Statistics: Computations and Simulations* **28**: 501–515.

- Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm, *J. of the Amer. Stat'l. Assn.* **88**: 495–504.
- Hall, P., Nussbaum, M. and Stern, S. E. (1997). On the estimation of a support curve of indeterminate sharpness, *J. Multivariate Anal.* **62**(2): 204–232.
- Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation, *Computational Statistics 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. (1988). Fast computation of fully automated log-density and log-hazard estimators, *SIAM J. on Scientific and Stat'l. Computing* **9**: 363–379.
- Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *Ann. of Statistics* **10**: 795–810.
- Stone, C. J. (1990). Large-sample inference for log-spline models, *Ann. of Statistics* **18**: 717–741.