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 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. The number of basis functions is the smoothing parameter and is estimated by minimizing a proxy of the Kullback-Leibler distance which equivalent to get the maximum likelihood cross-validation.

Keywords: density estimation; B-splines; likelihood ratio test; partitions of unity.

1 Introduction

Let $X_1, \ldots, X_n$ be i.i.d. random variables with an unknown probability density

*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

function f on a finite domain $\mathcal{X}$. 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 (Leonard, 1978) $f = e^g/(\int e^g)$. 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 loglikelihood (Silverman, 1982; 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 parameter. This parameter needs to be estimate and the estimation procedures require computational intensive methods. Differently from other smoothing methods, the proposed procedure does not penalize the likelihood function and the general solution for this optimization problem 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 which is shown to be equivalent to maximum likelihood cross validation, the coefficients of the expansion are found by maximizing the likelihood. Section 4 presents some simulation 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,....

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 $X_1, \ldots, X_n$ i.i.d. random variables on a finite domain $\mathcal{X}$, assume that $f = e^g / \int e^g, \quad g \in \mathcal{N}$, 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 loglikelihood equation is,

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

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 $\hat{\theta} = \hat{\theta}^{(K)} = (\theta_1, \ldots, \theta_K) \in \Theta \subseteq \mathbb{R}^K$ the maximizer of $L_K(\theta)$. Our estimate will be $f_K = e^{\hat{g} - \log \int e^{\hat{g}}}$ 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 $\sum_{j=1}^{n} \theta_j = 0$, since $\int M_j = 1$ (normalized cubic B-splines) for $j = 1, \ldots, K$. 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,

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

Note that, the equality holds in (2.2) 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. Thus, $L_K(\theta)$ 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 of parameters $\hat{\theta}$ 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 $c_\theta : \Theta \subset \mathbb{R}^K \to \mathbb{R}$, $K \geq 1$,

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

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)

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

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 $t_1, \ldots, t_p$ be sequence of knots such that $-\infty, < t_1, < t_2 < \cdots < t_{p-1} < t_p < \infty$. Suppose $(-\infty, t_1]$ and $[t_p, \infty)$ have at least one observed value and four or more observations lie in the compact sets $[t_1, t_2], \ldots [t_{p-1}, t_p]$, then $\hat{\theta} = \theta^K$ exists.

3 Computing the number of basis functions

One may notice the density estimate 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 a test statistic $\frac{f(X)}{f_K(X)}$ based on the likelihood ratio test is taken. For $(\hat{\theta}, K)$ this statistic should be close to 1, and consequently $E[\log(f(X)/f_K(X))]$

roughly 0. This approach is equivalent to minimize de Kullback-Leibler distance (not a metric, for using Hellinger pseudo metric see Dias (1999)). For this, given any density f, let $g = \log f$ and consider the Kullback-Leibler measure for the difference between f and f_K

$$\begin{aligned} d_{KL}(f, f_K) &= \int (\log f - \log f_K) f \\ &= \int (g - g_K) e^g \end{aligned} \tag{3.1}$$

Of course, we cannot compute $d_{KL}(f, f_K)$ 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 $\hat{K} = \arg\min_{K \in \{1,\ldots,n\}} d_{KL}(f, f_K)$.

Note that to minimize the Kullback-Leibler distance is equivalent also to maximize the usual cross validation function. To see, denote the *leave-one-out* estimate

$$f_K^{-i}(X_i) = \frac{1}{n-1} \sum_{l \neq i}^{n} \exp(\langle \theta, M(X_l) \rangle_K - a(\theta)),$$

where $a(\theta) = \log \int \exp(\langle \theta, M(x) \rangle_K) dx$. Then

$$n^{-1} \log \prod_{i=1}^{n} f_K^{-i}(X_i) = n^{-1} \sum_{i=1}^{n} \log f_K^{-i}(X_i).$$

Define the maximum likelihood cross validation function to be

$$CV(K) = n^{-1} \sum_{i=1}^{n} \log \hat{f}_K^{-i}(X_i). \tag{3.2}$$

Under this procedure the best K would be $K_{CV} = \arg\max_K CV(K)$. Note that this procedure is highly computational intensive and might take a considerable amount of time to get a density estimate. However, it is not difficult to show that by disregarding the *leave-one-out* effect

$$E[CV(K)] \approx E[\int f \log \hat{f}_K]. \tag{3.3}$$

Comparing (3.3) with Kullback-Leibler distance, we can see

$$E[CV(K)] \approx \int f \log f - E[d_{KL}(f, f_K)].$$

Noting that the first term of the right hand does not depend on K, we can expect to approximate the optimal K by minimizing $d_{KL}(f, f_K)$. Observe that, the Kullback-Leibler distance, $d_{KL}(f, f_K) = \int f \log f - \int f \log f_K$ and

$$\int f \log f_K = E[\log f_K(X)]$$

is the only term which depends on K and it can be approximated by

$$E[\log f_K(X)] \approx \frac{1}{n} \sum_{i=1}^{n} \log f_K(X_i).$$

Thus, in fact, the optimization relies on a proxy of the Kullback-Leibler.

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 on the order statistics.

Algorithm 3.1

1. *For* $K \in \{1, \ldots, m\} \quad m < n$,
2. *get the maximizer of* $L_K(\theta)$ *and compute* f_K, *and*
3. *choose* $\hat{K}$ *that maximizes* $\frac{1}{n} \sum_{i=1}^{n} \log f_K(X_i)$.
4. *Compute the maximizer of* $L_{\hat{K}}(\theta)$ *and deliver* $f_{\hat{K}}$

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 entirely code is written is Splus and R, does not require any dynamic loading and it can be implemented in any personal computer where Splus (or R) is running.

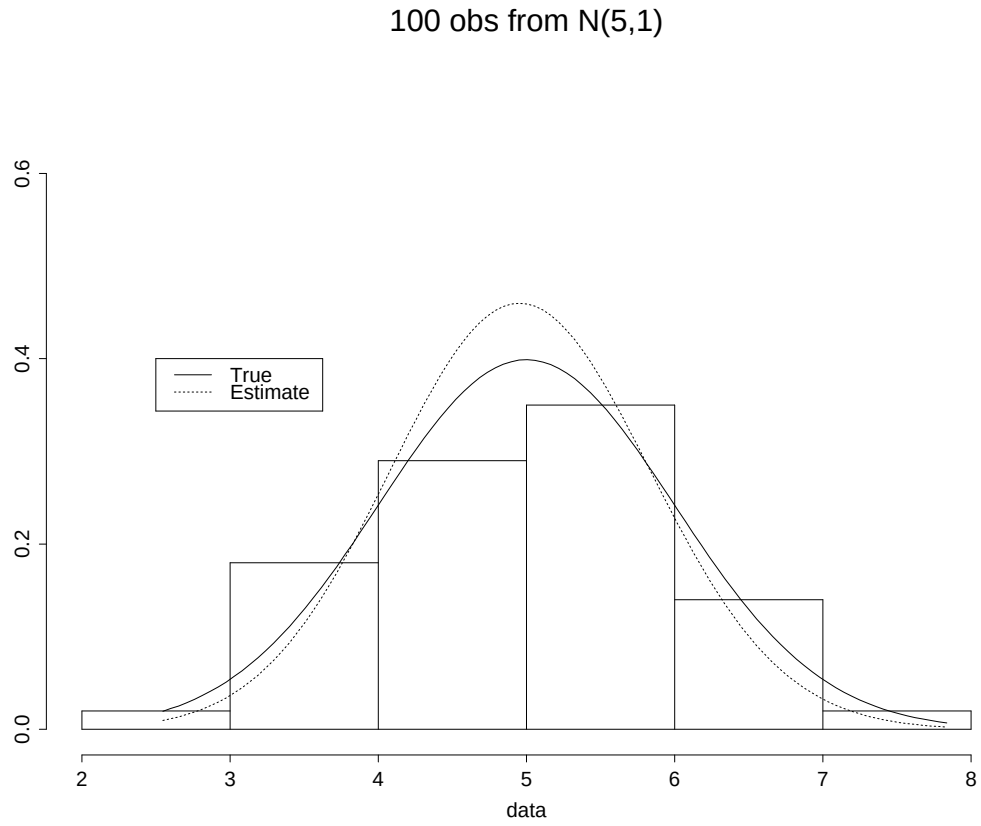

Figure 4.1: One hundred observations from N(5,1) and the optimal K is 3.

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.

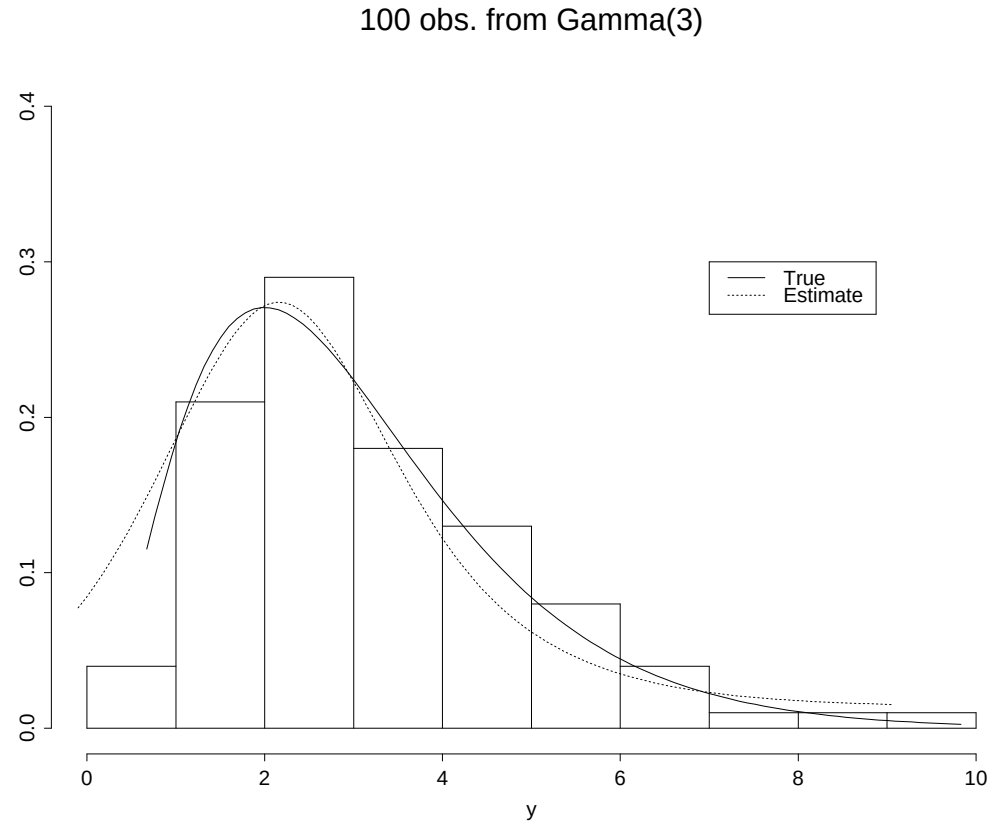

Figure 4.2: two hundred observations from Gamma(3) and the optimal K is 4.

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.

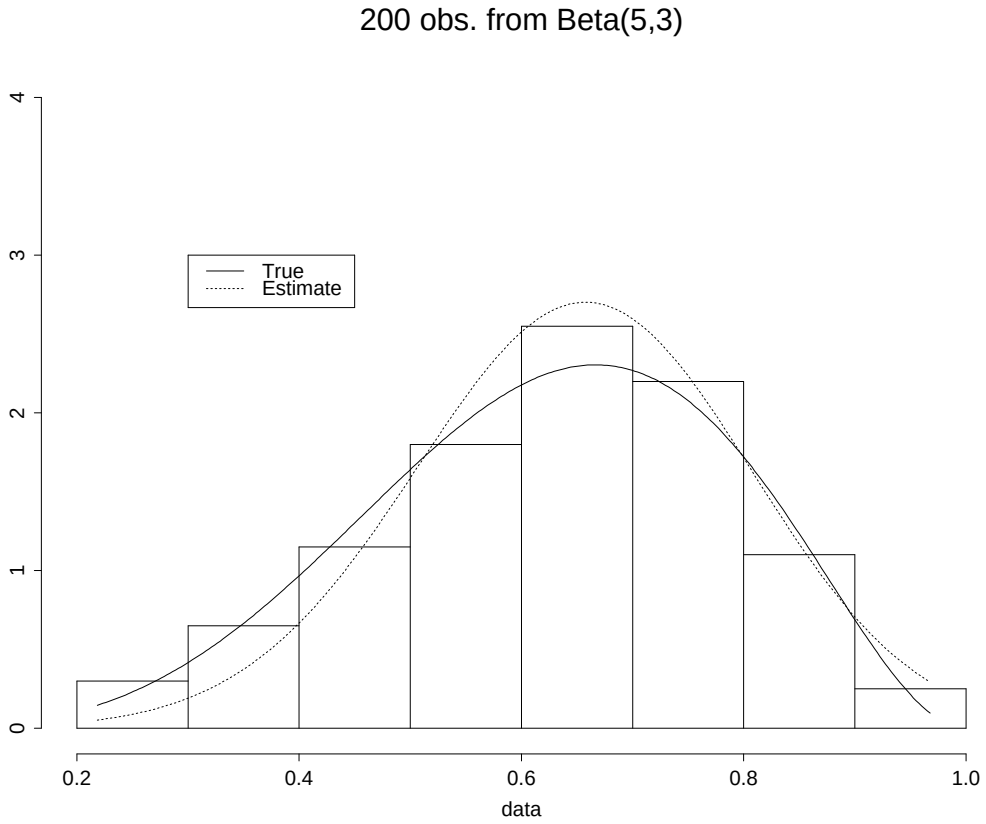

Figure 4.3: The data contain two hundred observations from Beta(5,3) and the optimal K is 4.

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.

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

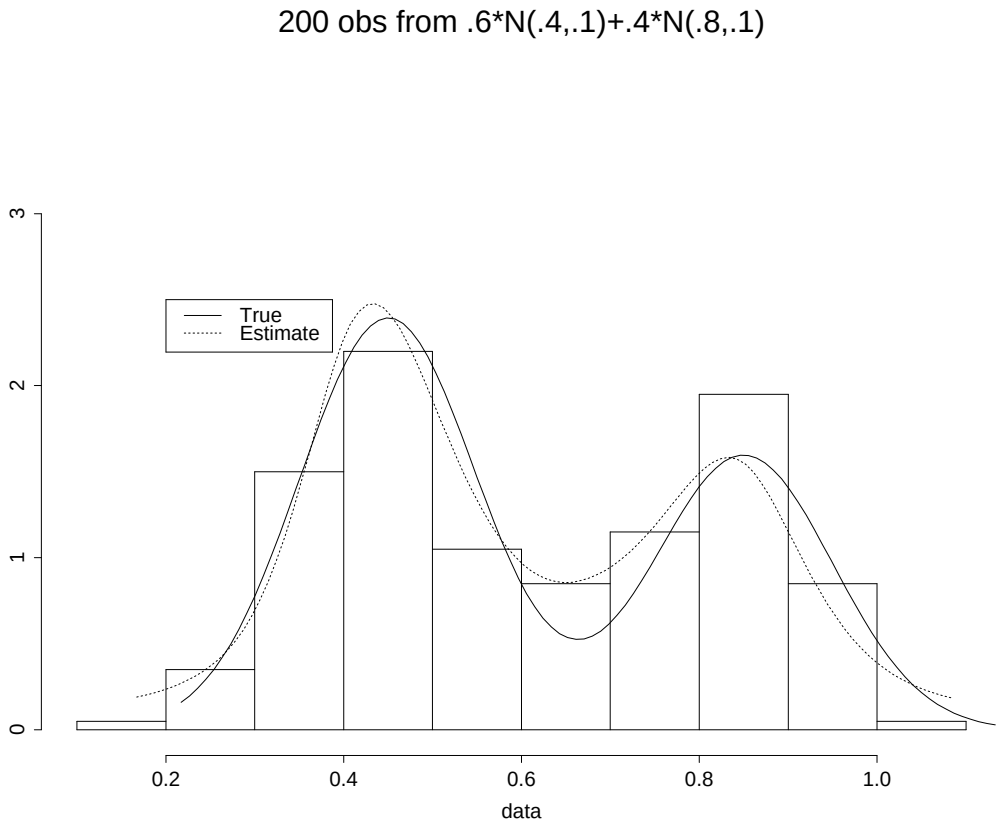

Figure 4.4: The data contain two hundred observations from mixed normal distributions and K is 7.

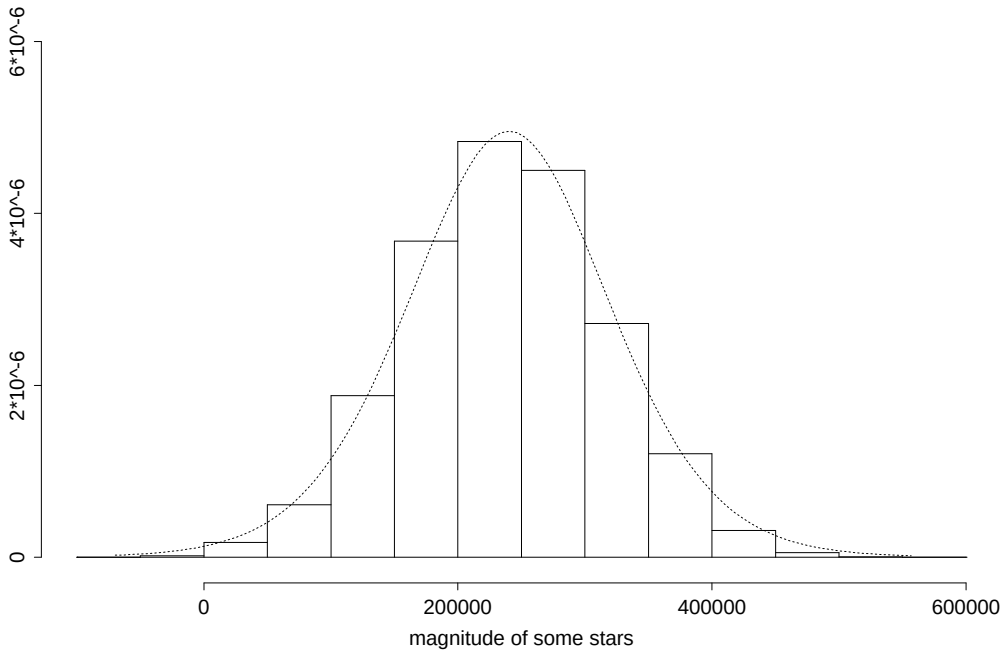

Figure 4.5: Comparison of the density estimate provided by this method and the histogram for the data Magnitute of stars.

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 nonparametric estimation can be a starting guess for a parametric model. Simulations have shown that K=3 is appropriate to estimate a single mode. Also, placement knots on the order statistics is much better when one knows a priori that the data might have outliers. It does a very good job estimating non pathological data sets. Certainly, there are more adaptive methods for nonparametric density estimation (see, for example, Kooperberg and Stone (1991) and Dias (1998a)) but they are more difficult to implement.

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**: issue 2.

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

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.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *Ann. of Statistics* **10**: 795–810.