Consistent estimator for basis selection based on a proxy of the Kullback-Leibler distance.

Ronaldo Dias* and Nancy L. Garcia

Universidade Estadual de Campinas

13/02/2006

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 basis functions approach. The number of basis functions acts as the smoothing parameter and it is estimated by minimizing a penalized proxy of the Kullback-Leibler distance which includes as particular cases AIC and BIC criteria. We prove that this estimator is consistent.

JEL classification: C14

Keywords: non-parametric density estimation; B-splines; Wavelets; information criteria.

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

1 Introduction

Suppose we have a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a cumulative distribution F which is absolutely continuous with respect to a dominant Lebesgue measure μ. Moreover, assume that the density of F, $f = \frac{dF}{d\mu}$, has compact support $\mathcal{X}$. Define $\mathcal{F}_\mu$ be the class of density functions such that,

$$\mathcal{F}_\mu = \{h : \mathbb{R} \to [0, \infty) : h(x) = \frac{e^{S(x)}}{\int_{\mathcal{X}} e^{S(x)} d\mu(x)} \text{ and } \int_{\mathcal{X}} e^{S(x)} d\mu(x) < \infty\},$$

where the function S is of the class $C^2(\mathbb{R})$. It is easy to see that the elements in $\mathcal{F}_\mu$ are not identifiable since for any function S_1 such that $S_1 = S + c$, we have $e^{S_1}/(\int e^{S_1}) = e^S/(\int e^S)$. To ensure uniqueness of the elements in $\mathcal{F}_\mu$ a side condition, such as $\int_{\mathcal{X}} S = 0$ (or $S(x_0) = 0$ for some fixed x_0), is required, see for example, Gu (1993) and Dias (1998).

Consider the problem of finding the maximum likelihood estimator of f. It is well known (see for example, Silverman (1986); Pagan and Ullah (1999) and Dias (1994)) that such optimization problem is unbounded over the class of all smooth functions. In fact, the optimizer is a sum of delta functions. To avoid the *Dirac's disaster* one might want to apply penalized likelihood procedure or one may assume that f can be well approximated by a function belonging to a finite dimensional space $\mathcal{H}_K$ which is spanned by K (fixed) basis functions, such as Fourier expansion, wavelets, B-splines, natural splines. See, for example, Silverman (1986), Kooperberg and Stone (1991), Vidakovic (1999), Dias (1998) and Dias (2000). Although this fact might lead one to think that the nonparametric problem becomes a parametric problem, one notices that the number of parameters can be as large as the number of observations, and there may be difficulties in estimating the density. Moreover, if the number of observations is large, the system of equations for exact solution is too expensive to solve. This is an inheritance from the approximation theory of functions.

In fact, one of the most challenging problem is how to select the dimension of the

approximant space. A similar problem is encountered in the field of image processing where the level of resolution needs to be determined appropriately. Several authors suggested algorithms in order to provide a good choice of the dimension of the approximant space as a function of the sample size, see for example Gu (1993), Antoniadis (1994) De Vore, Petrova and Temlyakov (2003), Bodin, Villemoes and Wahlberg (2000), Kohn, Marron and Yau (2000). In some cases, it is possible to obtain convergence rates for the distance between the true density function and the estimated one (see for example, Stone (1980), Stone (1990), Fenton and Gallant (1996a), Fenton and Gallant (1996b)). However, all of these procedures including adaptive ones (Kooperberg and Stone (1991), Luo and Wahba (1997) and Dias (1998)) deal with a non-random choice of K.

Differently, assuming that there exists a finite but unknown K such that f is an element of $\mathcal{H}_K$, Dias (2000) and Dias and Garcia (2004) suggested to use a proxy of the Kullback-Leibler distance in order to determine $\hat{K}$, an estimate of the true dimension. The Kullback-Leibler distance is attractive not only because of the linearization of the logistics transformation, but also it is asymptotically equivalent to maximize the likelihood function and to minimize the Hellinger distance. The goal of this work is to prove that depending on the selection criterion one obtains consistent estimator of the dimension of the space, that is, as the sample size increases $\hat{K}$ converges to K (almost surely or in probability).

A technical assumption of the proofs is that the true model belongs to the set of candidate models. Notice that this requirement rules out the case of infinite dimension. Consistency results without this assumption cannot be obtained by a modification of our proofs, it requires totally different techniques and it is not going to be addressed in this work. This difficulty arises in other contexts, for example, Finesso (1992) proved the consistency for the BIC estimator of the order of a finite memory Markov chain using the Law of Iterated Logarithm, while Csiszár and Shields (2000) had to use a completely different approach to eliminate the finiteness restriction.

This paper is organized as follows. Section 2 formalizes the problem and review the estimation procedure for the density function when the true dimension K is known. Section 3 describes the penalized likelihood problem to obtain an estimate of K and presents the conditions for the penalization term c_n under which strong and weak consistency are achieved. Moreover, bounds on the rates of convergence are shown in Section 4. The proofs of the main theorems are presented separately in the Appendix. Simulation results comparing the rates of convergence of some criteria are presented in Section 5, including the well-known Akaike, Bayesian and Hannan and Quinn Information Criteria.

2 Previous results

From now on, we assume that f is an element of $\mathcal{H}_K$ and it can be written as

$$f = \frac{e^{S_f}}{\int e^{S_f}}$$

where

$$S_f = \sum_{j=1}^{K} \theta_j M_j \quad \text{with } \int e^{S_f} < \infty$$

and $M_1, \ldots, M_K$ are normalized basis functions that span $\mathcal{H}_K$ such that $\int M_j = 1$. As pointed before, in order to enforce one-to-one correspondence we need the restriction $\int S_f = 0$ and then $\sum_{j=1}^{K} \theta_j = 0$, since $\int M_j = 1$. For any $K > 0$, let $\Theta_0 = \{\theta \in \mathbb{R}^K : \sum_j^K \theta_j = 0\}$.

Consequently, there exists vector $\theta = (\theta_1, \ldots, \theta_K)$ such that the log-likelihood of $\mathbf{X}$ is given by

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

The vector of coefficients θ are unknown and need to be determined. One of the most common standard statistical procedure in nonparametric estimation, is to determine θ using maximum likelihood method. For fixed K, the asymptotics of the density

estimator were studied by Dias (2000) and are presented in Lemma 2.1, Theorem 2.1, Lemma 2.2 and Proposition 2.4.

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

It is not difficult to show that $L_K(\theta|\mathbf{X})$ 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 $\mathbb{R}^K$ and θ^* in Θ_0.

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

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

$$\hat{\theta}_n^{(K)} = \arg\max_{\theta\in\Theta_0} L_K(\theta|\mathbf{X}). \tag{2.2}$$

Notice that, in fact,

$$L_K(\theta|\mathbf{X}) = \langle\theta, \bar{M}\rangle_K - \log\int e^{\langle\theta,M\rangle_K},$$

then $\hat{\theta}_n^{(K)}$ is the unique solution of the equation

$$h(\theta, \bar{M}(\mathbf{X})) = 0, \tag{2.3}$$

where $\bar{M}(\mathbf{X})$ is a K-dimensional vector with j-th components given by

$$\frac{1}{n}\sum_{i=1}^{n} M_j(X_i) = \bar{M}_j, \quad j \in \{1,\ldots,K\}. \tag{2.4}$$

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

$$\frac{\partial L_K(\theta|\mathbf{X})}{\partial\theta} := h(\theta, M^*(\mathbf{X})) = 0, \tag{2.5}$$

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

$$h_j(\theta, \mathbf{u}) = u_j - \frac{\int \exp(\langle \theta, M(z)\rangle_K) M_j(z) dz}{\int \exp(\langle \theta, M(z)\rangle_K) dz}, \tag{2.6}$$

for $j \in \{1, \ldots, K\}$. Therefore, $\hat{\theta}_n^{(K)}$ is an M-estimator and since $\theta \mapsto h_\theta$ is continuous we have the following result.

Proposition 2.2 *Let θ_0 be the unique solution of*

$$h(\theta, \int f(x) M(x) d\mu(x)) = 0 \tag{2.7}$$

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

Thus, the density estimate is, for fixed K

$$\hat{f}_K = e^{\hat{S} - \log \int e^{\hat{S}}},$$

where $\hat{S} = \langle \hat{\theta}, M\rangle_K$ with $\hat{\theta} = \hat{\theta}_n^{(K)}$.

Proposition 2.3 *For fixed K, the density estimates $\hat{f}_K(\cdot) = f_K(\cdot|\hat{\theta}_n)$ converge pointwise almost surely (a.s.) to f as n goes to infinity.*

In order to provide an appropriate K, one might want to choose K that minimizes the Kullback-Leibler distance between the true f and the random function $\hat{f}_K$, $d(f, \hat{f}_K) = \int (\log f - \log \hat{f}_K) f$ or equivalently

$$D_n(K) = \int f \log \hat{f}_K. \tag{2.8}$$

Of course, we cannot compute $D_n(K)$ from the data, since it requires the knowledge of f. Defining a proxy of this distance by

$$Z_n(K) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_K(X_i), \tag{2.9}$$

it is easy to prove their equivalence.

Proposition 2.4 *For any fixed K,*

$$D_n(K) - Z_n(K) = \sum_{j=1}^{K} \hat{\theta}_{nj}^{(K)} \Big(\int f(x) M_j(x) d\mu(x) - \frac{1}{n} \sum_{i=1}^{n} M_j(X_i) \Big) \longrightarrow 0 \tag{2.10}$$

$n \longrightarrow \infty$ almost surely.

3 Main results

Since $Z_n(K)$ is strongly related to the likelihood, it increases as K increases. Notice that K acts as the control parameter (smoothing parameter) between adaptiveness (large values of K) and smoothness (small values of K). Therefore, a reasonable way of defining the best K is to penalize $Z_n(K)$ for large values of K. In fact, define $\hat{K} = \hat{K}_n$ as

$$\hat{K} = \arg\max_{\ell} \mathrm{LP}(n, \ell), \tag{3.1}$$

where $\mathrm{LP}(n, \ell) = n\, Z_n(\ell) - c_n\, \ell$ and $c_n > 0$. This includes the most common information criteria for model selection such as AIC estimator ($c_n = 2$), BIC estimator ($c_n = \log n$) and Hannan and Quinn (HQIC),($c_n = c \log\log n$, where c is a positive constant.

Under suitable conditions on c_n, we show that in the set up of density estimation, these criteria consistently estimate the true dimension of the model.

Theorem 3.1 *If* $\lim_{n\to\infty} c_n/n = 0$ *and* $\lim_{n\to\infty} c_n/\sqrt{n} = \infty$ *then* $\hat{K}$ *is a strongly consistent estimator for* K*.*

Theorem 3.2 *If* $\lim_{n\to\infty} c_n/n = 0$ *and* $\lim_{n\to\infty} c_n = \infty$ *then* $\hat{K}$ *is a weakly consistent estimator for* K*.*

Remark: From the proofs of the theorems presented in the Appendix, we can see that the conditions above are sharp, that is, if $c_n \not\to \infty$, then the estimator $\hat{K}$ is not consistent (e.g. AIC estimator). On the other hand, if $c_n/\sqrt{n} \to 0$ and $c_n \to \infty$ the estimator is weakly but not strongly consistent (e.g. BIC and HQIC estimators). Intuitively, since we are working with penalized likelihood estimation for finite samples, the larger the sample size the heavier the penalization needed to avoid the *Dirac's disaster*.

4 Rates of convergence

The fact that we are considering a fixed model K lead us to the problem of computing rates of convergence for consistent estimators of K. We are denoting an estimator $\hat{K}$ to be *weakly consistent* if

$$\lim_{n\to\infty} \mathbb{P}(\hat{K} = K) = 1.$$

In general, even if an estimator is not consistent, it will converge in distribution, that is, there exists a sequence γ_j, $j = 0, 1, \ldots$ such that

$$\lim_{n\to\infty} \mathbb{P}(\hat{K} = j) \;=\; \gamma_j, \quad j = 0, 1, \ldots.$$

Consistency corresponds to $\gamma_K = 1$ and $\gamma_j = 0$ for all $j \neq K$. However, an useful way of comparing among consistent estimators would be to give the rate of convergence of such estimators. For the estimators considered in this paper we were able to determine their rates of convergence.

Theorem 4.1 *If* $\lim_{n\to\infty} c_n/n = 0$ *and* $\lim_{n\to\infty} c_n = \infty$, *then*

(i) For $l < K$, *we have*

$$\mathbb{P}\left(\hat{K} = l\right) \leq e^{-\frac{n}{\sigma^2}(m_K - m_l)^2}.$$

(ii) For $l > K$, *we have*

$$\mathbb{P}\left(\hat{K} = l\right) \leq e^{-c_n(l-K)t}\left(\frac{1}{1-2t}\right)^{l-K/2}, \qquad \textit{for all } t < 1/2.$$

(iii) If, furthermore, $\lim_{n\to\infty} c_n/\sqrt{n} = \infty$, *then*

$$\mathbb{P}\left(\hat{K} > K\right) \leq e^{-c_n^2/n\sigma^2},$$

where $m_l := \mathbb{E}[\log f_l(X)]$ *is defined by (5.3) and* σ^2 *is the asymptotic variance of* $\sqrt{n}[Z_n(l) - Z_n(K)]$.

Again, we can see the major role played by the penalization term. For strongly consistent estimators (c_n increasing faster than $\sqrt{n}$) we obtain faster convergence rates. Notice that, for moderate sample sizes one can choose the penalization term in order to bound the probability of overestimating.

5 Simulated and real data

In this section, we present a real data application and some simulation results comparing several estimation criteria, including AIC, BIC among others. Recall that we estimate $\hat{K} = \hat{K}_n$ as

$$\hat{K} \;=\; \arg\max_{\ell} \mathrm{LP}(n, \ell), \tag{5.1}$$

where $\mathrm{LP}(n, \ell) = n\, Z_n(\ell) - c_n\, \ell$ and $c_n > 0$.

The real dataset consists of 11,130 individual observations of Average Hourly Earnings obtained from Current Population Survey in the United States during the years 1992–1998. It can be downloaded from R statistical package using Ecdata. Figure 5.1 shows the final fit obtained using 6 basis functions through the penalty $c_n = n^{2/7}$ which provides a strongly consistent estimator.

In order to compare the performance of the several criteria we simulated several samples from the distribution spanned by seven basis functions presented in Figure 5.2 which were generated as

$$f_7(x) = \sum_{j=1}^{7} \theta_j M_j(x) \tag{5.2}$$

and $M_1, \ldots, M_7$ are normalized logspline basis functions.

We performed 100 replication for each criteria and computed the number of times the true dimension was estimated as well as the mean and standard deviation. Table 5 shows that for very small samples using AIC ($c_n = 2$) is better than the others although it does provide an estimator which is not even consistent. On the other hand, the strongly consistent estimators ($c_n = n^{.52}$ and $c_n = n^{2/7}$) heavily underestimate

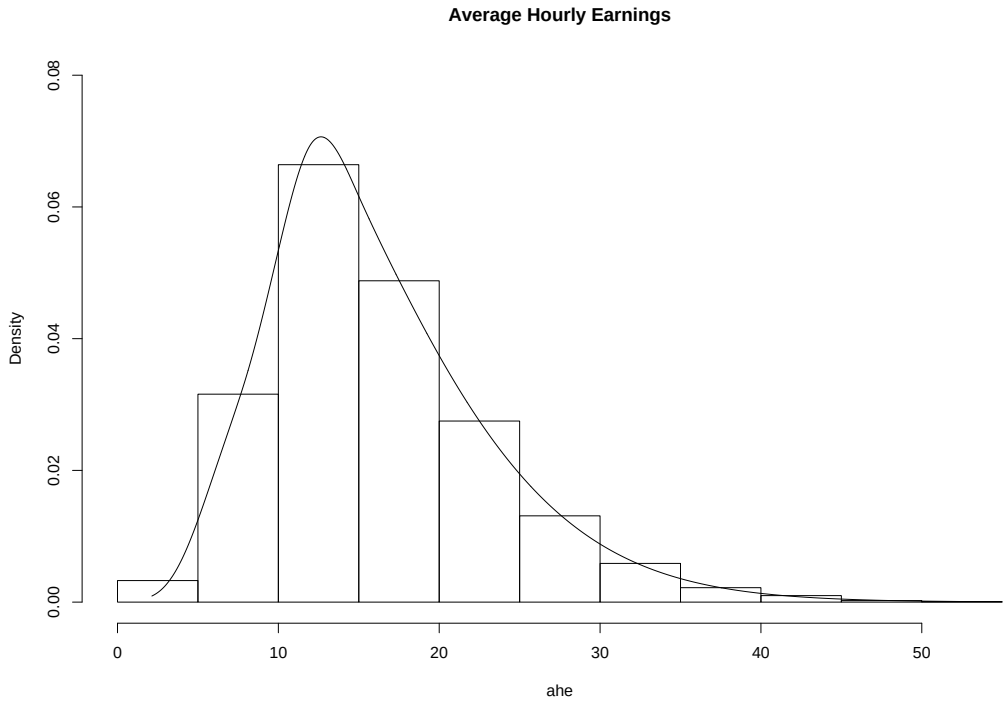

Figure 5.1: Density estimation based of 6 knots obtained through $c_n = n^{2/7}$.

the dimension of the space for small sample sizes. However, we can see that even for moderate sample sizes (around 500 or 1,000), they begin to approach a 100% of hits. Notice that $c_n = n^{.52}$ is barely above the threshold between strongly and weakly consistent estimators. Overall, the weakly consistent estimators (included here are $c_n = \log\log n$ and BIC corresponding to $c_n = \log n$) have good performances although as expected they do not always hit the true value for large samples.

	$n = 100$			$n = 250$		
c_n	% $[\hat{K} = 7]$	Average	S.D.	% $[\hat{K} = 7]$	Average	S.D.
2	55	6.56	1.01	73	7.78	1.43
$\log\log n$	51	7.95	1.99	73	7.74	1.46
$\log n$	25	5.89	0.77	92	6.95	0.48
$n^{.52}$	11	5.22	0.63	6	5.12	0.47
$n^{2/7}$	46	6.18	1.06	69	7.09	0.79
	$n = 500$			$n = 1000$		
c_n	% $[\hat{K} = 7]$	Average	S.D.	% $[\hat{K} = 7]$	Average	S.D.
2	85	7.27	0.89	85	7.28	0.91
$\log\log n$	76	7.46	1.07	71	7.36	0.64
$\log n$	92	7.17	0.63	100	7.00	0.00
$n^{.52}$	14	5.52	0.73	98	6.97	0.22
$n^{2/7}$	87	7.35	0.92	99	7.01	0.10
	$n = 5000$			$n = 10000$		
c_n	% $[\hat{K} = 7]$	Average	S.D.	% $[\hat{K} = 7]$	Average	S.D.
2	7	8.04	0.60	1	8.32	0.70
$\log\log n$	12	7.91	0.40	0	8.25	0.74
$\log n$	79	7.21	0.40	49	7.51	0.50
$n^{.52}$	100	7.00	0.00	100	7.00	0.00
$n^{2/7}$	100	7.00	0.00	100	7.00	0.00

Table 5.1: Simulation results for 100 replication for a density with true dimension equals to 7.

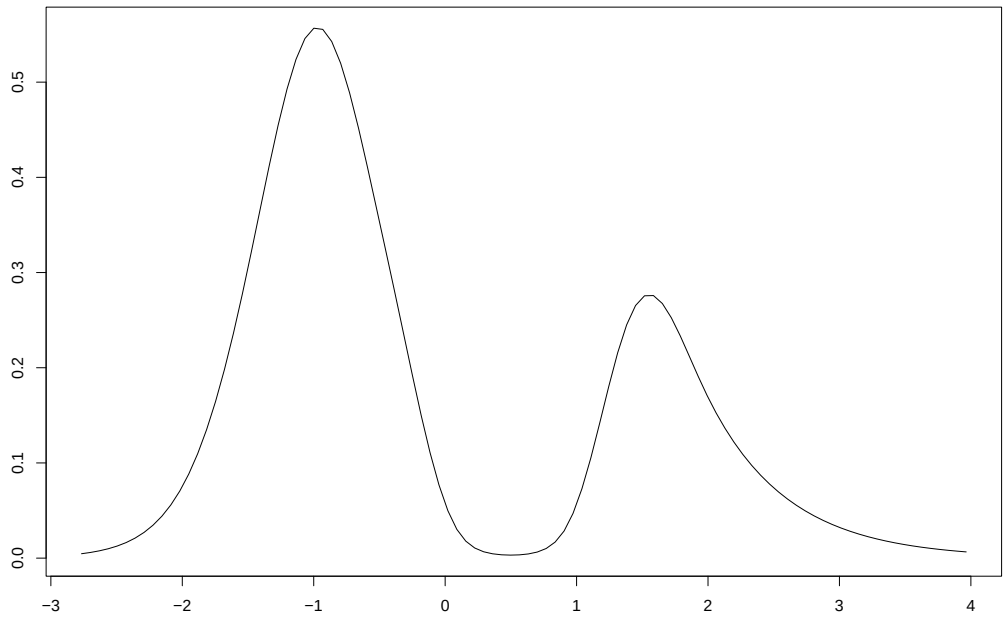

Figure 5.2: True density spanned by 7 basis functions

Acknowledgments We would like to thank Prof. Klaus Vasconcellos and the anonymous referees for valuable suggestions that greatly improved this paper. This work was partially funded by FAPESP grant 02/01554-5 and CNPq grants 301054/1993-2, 300644/1994-9 and 475763/2003-3.

Appendix: Proofs

Proof of Theorem 3.1. Suppose that K is the true dimension of the approximant space to be determined, that is $f \in \mathcal{H}_K$.

(a) Assume that $l < K$, we are going to prove that, with probability 1, for large values of n

$$\mathrm{LP}(n, l) < \mathrm{LP}(n, K).$$

By Proposition 2.3 we have that

$$Z_n(K) \;=\; \frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_K(X_i) \;\to\; \mathbb{E}[\log f(X)] \;:=\; m_K \text{ a.s.} \tag{5.3}$$

where X is a random variable with density f.

On the other hand,

$$\begin{aligned} Z_n(l) &= \frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_l(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{l}\hat{\theta}_j^{(l)}M_j(X_i) - \log\int e^{\langle\hat{\theta}^{(l)},M\rangle_l}\right] \qquad (5.4)\\ &\to \sum_{j=1}^{l}\theta_j^{(l)}\mathbb{E}[M_j(X)] - \log\int e^{\langle\theta^{(l)},M\rangle_l} \;:=\; m_l \text{ a.s.} \qquad (5.5)\end{aligned}$$

where $\hat{\theta}^{(l)}$ and $\theta^{(l)}$ are the solutions of (2.2) and (2.7) respectively, replacing K by l.

Therefore, by Jensen's inequality

$$\begin{aligned} m_l - m_K &= \int f\log\frac{f_l}{f} = \mathbb{E}\left[\log\frac{f_l}{f}\right] \\ &\le \log\mathbb{E}\left[\frac{f_l}{f}\right] \;=\; \log\int f\frac{f_l}{f} = 0. \qquad (5.6)\end{aligned}$$

where $\log f_l = \sum_{j=1}^{l}\theta_j^{(l)}M_j - \log\int e^{\langle\theta^{(l)},M\rangle_l}$. Moreover, equality holds only if $f_l = f$ and this mean that l is the true dimension of the approximant space contradicting our hypothesis. It follows that for $l < K$

$$\lim_{n\to\infty}\frac{1}{n}\left(Z_n(K) - Z_n(l)\right) = m_K - m_l > 0 \text{ a.s..} \tag{5.7}$$

By (5.7) as $n \to \infty$

$$\begin{aligned}\frac{1}{n}(\mathrm{LP}(n,K) - \mathrm{LP}(n,l)) &= \frac{1}{n}(Z_n(K) - Z_n(l)) - \frac{c_n}{n}(K-l) \\ &= (m_K - m_l)(1+o(1)) - \frac{c_n}{n}(K-l) \text{ a.s..} \qquad (5.8)\end{aligned}$$

Since $c_n/n \to 0$ as $n \to \infty$, we have with probability 1, for large values of n

$$\mathrm{LP}(n,l) < \mathrm{LP}(n,K).$$

(b) Assume that $l > K$, we are going to prove that, with probability 1, for large values of n

$$\mathrm{LP}(n,l) < \mathrm{LP}(n,K).$$

Similarly to the arguments used by Dias and Garcia (2004) we can prove that, $Z_n(l) - Z_n(K) \to 0$ a.s. and $\sqrt{n}[Z_n(l) - Z_n(K)]$ converges in distribution to a normally distributed zero mean random variable. Therefore,

$$\lim_{n\to\infty} \frac{1}{n}(Z_n(K) - Z_n(l)) = O\left(\frac{1}{\sqrt{n}}\right) \text{ a.s..} \tag{5.9}$$

By (5.9) as $n \to \infty$

$$\mathrm{LP}(n,K) - \mathrm{LP}(n,l) \;=\; O\left(\sqrt{n}\right) + c_n(K-l) \text{ a.s..} \tag{5.10}$$

Since $c_n/\sqrt{n} \to \infty$ as $n \to \infty$, we have with probability 1, for large values of n

$$\mathrm{LP}(n,l) < \mathrm{LP}(n,K).$$

Proof of Theorem 3.2 For $l > K$, we can interpret $2n(Z_n(l) - Z_n(K))$ as the likelihood ratio test statistic. It is well known (see for example, Ferguson (1996)) that $2n(Z_n(l) - Z_n(K))$ has a limiting chi-square distribution. Therefore,

$$2n(Z_n(l) - Z_n(K)) = O_P(1)$$

where O_P means bounded in probability. Since, $c_n \to \infty$, we have

$$\mathrm{LP}(l) - \mathrm{LP}(K) = -n(Z_n(l) - Z_n(K)) + c_n(l-K) = O_P(1) + c_n(l-K) \xrightarrow{P} \infty$$

as $n \to \infty$ and we conclude that $\hat{K} \xrightarrow{P} K$ as $n \to \infty$.

Proof of Theorem 4.1

(i) For $l < K$ there exists a positive constant ξ^2 such that

$$\sqrt{n}(Z_n(K) - m_K - Z_n(l) + m_l) \to N(0, \xi^2)$$

in distribution and $n \to \infty$ where m_l is given by (5.3). In this case,

$$\begin{aligned}
\mathbb{P}(\hat{K} = l) &\leq \mathbb{P}\left(LP(l) - LP(K) > 0\right) \\
&= \mathbb{P}\left(n(Z_n(l) - Z_n(K)) > c_n(l-K)\right) \\
&= \mathbb{P}\left(\sqrt{n}(Z_n(l) - Z_n(K) - m_l + m_K) > \frac{c_n}{\sqrt{n}}(l-K) + \sqrt{n}(m_K - m_l)\right) \\
&\leq \left(\exp\left\{\frac{-c_n}{\sqrt{n}\xi}(l-K) - \frac{\sqrt{n}}{\xi}(m_K - m_l)\right)^2\right\} \\
&= O\left(\exp\left\{\frac{-n}{\xi^2}(m_K - m_l)^2\right\}\right) \qquad (5.11)
\end{aligned}$$

where the last inequality was obtained by using the Chernoff bounds $\mathbb{P}(Z \leq a) \leq e^{-at}\mathbb{E}(e^{Zt})$ optimized for normal random variables.

(ii) On the other hand, for $l > K$ we have

$$2n(Z_n(K) - Z_n(l)) \to \chi^2_{(l-K)}$$

in distribution as $n \to \infty$. Therefore,

$$\begin{aligned}
\mathbb{P}(\hat{K} = l) &= \mathbb{P}\left(LP(l) - LP(K) > 0\right) \\
&= \mathbb{P}\left(2n(Z_n(l) - Z_n(K)) > 2c_n(l-K)\right) \\
&\leq \exp\left\{-c_n(l-K)t\left(\frac{1/2}{1/2 - t}\right)^{(l-K)/2}\right\} \qquad (5.12)
\end{aligned}$$

for all $t < 1/2$ where the last inequality was obtained by using the Chernoff bounds $\mathbb{P}(Z \leq a) \leq e^{-at}\mathbb{E}(e^{Zt})$ for chi-square random variables.

(iii) If, furthermore, $\lim_{n\to\infty} c_n/\sqrt{n} = \infty$, then for $l > K$, there exists a positive constant σ^2 such that

$$\sqrt{n}[Z_n(l) - Z_n(K)] \to N(0, \sigma^2)$$

weakly. In this case,

$$\begin{aligned}
\mathbb{P}(\hat{K} > K) &\leq \sum_{l=K+1}^{\infty} \mathbb{P}\left(LP(l) - LP(K) > 0\right) \\
&= \sum_{l=K+1}^{\infty} \mathbb{P}\left(n(Z_n(l) - Z_n(K)) > c_n(l-K)\right) \\
&= \sum_{l=K+1}^{\infty} \mathbb{P}\left(\frac{\sqrt{n}(Z_n(l) - Z_n(K))}{\sigma} > \frac{c_n}{\sqrt{n}}\frac{(l-K)}{\sigma}\right) \\
&\approx \sum_{l=K+1}^{\infty} 1 - \Phi\left(\frac{c_n}{\sqrt{n}}\frac{(l-K)}{\sigma}\right) \\
&\leq \sum_{l=K+1}^{\infty} \exp\left\{\frac{-c_n^2}{n}\frac{(l-K)^2}{\sigma^2}\right\} \\
&= \sum_{l=1}^{\infty} \exp\left\{\frac{-c_n^2}{n}\frac{l^2}{\sigma^2}\right\} \\
&= O\left(\exp\left\{\frac{-c_n^2}{n\sigma^2}\right\}\right) \qquad (5.13)
\end{aligned}$$

where $a_n \approx b_n$ means $\lim_n a_n/b_n = 1$.

References

Antoniadis, A. (1994). Wavelet methods for smoothing noisy data, *Wavelets, images, and surface fitting (Chamonix-Mont-Blanc, 1993)*, A K Peters, Wellesley, MA, pp. 21–28.

Bodin, P., Villemoes, L. F. and Wahlberg, B. (2000). Selection of best orthonormal rational basis, *SIAM J. Control Optim.* **38**(4): 995–1032 (electronic).

Csiszár, I. and Shields, P. C. (2000). The consistency of the BIC Markov order estimator, *Ann. Statist.* **28**(6): 1601–1619.

De Vore, R., Petrova, G. and Temlyakov, V. (2003). Best basis selection for approximation in L_p, *Found. Comput. Math.* **3**(2): 161–185.

Dias, R. (1994). Density estimation via h-splines, *University of Wisconsin-Madison.* Ph.D. dissertation.

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dias, R. (2000). A note on density estimation using a proxy of the Kullback-Leibler distance, *Brazilian Journal of Probability and Statistics* **13**(2): 181–192.

Dias, R. and Garcia, N. L. (2004). A spline approach to nonparametric test of hypotheses, *Brazilian Journal of Probability and Statistics.*

Fenton, V. and Gallant, R. (1996a). Convergence rates of SNP density estimators, *Econometrica* **64**(3): 719–727.

Fenton, V. and Gallant, R. (1996b). Qualitative and asymptotic performance of snp density estimators, *Journal of Econometrics* **74**: 77–118.

Ferguson, T. S. (1996). *A course in large sample theory*, Texts in Statistical Science Series, Chapman & Hall, London.

Finesso, L. (1992). Estimation of the order of a finite Markov chain, *Recent advances in mathematical theory of systems, control, networks and signal processing, I (Kobe, 1991)*, Mita, Tokyo, pp. 643–645.

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

Kohn, R., Marron, J. S. and Yau, P. (2000). Wavelet estimation using Bayesian basis selection and basis averaging, *Statist. Sinica* **10**(1): 109–128.

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

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines, *Journal of the American Statistical Association* **92**: 107–116.

Pagan, A. and Ullah, A. (1999). *Nonparametric econometrics*, Cambridge University Press, Cambridge.

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

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators, *Ann. Statist.* **8**(6): 1348–1360.

Stone, C. J. (1990). Large-sample inference for log-spline models, *Ann. of Statistics* **18**: 717–741.

Vidakovic, B. (1999). *Statistical modeling by wavelets*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.