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

Ronaldo Dias and Nancy L. Garcia

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 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.

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

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.

[*] 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 and nancy@ime.unicamp.br

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,

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

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 $e^{S_1}/(\int e^{S_1}) = e^S/(\int e^S)$. We are going to require, as Dias (1998), that $\int_{\mathcal{X}} S = 0$, to ensure uniqueness of the elements in $\mathcal{F}_\mu$.

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 coefficients 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, an element of $\mathcal{H}_K$ can be written as

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

where

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

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 $\Theta_0 = \{\theta \in \mathbb{R}^K : \sum_j^K \theta_j = 0\}$.

Assuming that the density f belongs to $\mathcal{F}_\mu$, we have that there exists K such that f is well approximated by functions in $\mathcal{H}_K$. Consequently, there exists vector $\theta = (\theta_1, \ldots, \theta_K)$ such that the log-likelihood of $\mathbf{X}$ is given by

$$L_K(\theta|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\langle\theta, M(X_i)\rangle_K - \log\int e^{\langle\theta, M\rangle_K}. \qquad (1.1)$$

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 1.1, Theorem 1.1, Lemma 1.2 and Proposition 1.4.

Lemma 1.1 *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 Θ and θ^* in Θ_0.

Proposition 1.1 *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

$$\hat{\theta}_n^{(K)} = \arg\max_{\theta\in\Theta_0} L_K(\theta|\mathbf{X}). \qquad (1.2)$$

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

$$h(\theta, \bar{M}(\mathbf{X})) = 0, \tag{1.3}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} M_j(X_i) = \bar{M}_j, \quad j \in \{1, \ldots, K\}. \tag{1.4}$$

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

$$\frac{\partial L_K(\theta|\mathbf{X})}{\partial\theta} := h(\theta, M^*(\mathbf{X})) = 0, \tag{1.5}$$

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,

$$h_j(\theta, \mathbf{u}) = u_j - \frac{\int \exp(\langle\theta, M(z)\rangle_K) M_j(z) dz}{\int \exp(\langle\theta, M(z)\rangle_K) dz}, \tag{1.6}$$

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 1.2 *Let θ_0 be the unique solution of*

$$h(\theta, \int f(x) M(x) d\mu(x)) = 0 \tag{1.7}$$

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 1.3 *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.*

However, the density estimate $\hat{f}_K$ strongly depends on the number of basis functions K which regularizes the optimization problem (1.1). In fact, in the context of nonparametric density estimation using basis functions approach one of the most challenging problem is how to select the number of basis functions. 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, see for example Kooperberg and Stone (1991), Gu (1993), Antoniadis (1994) De Vore, Petrova and Temlyakov (2003), Bodin, Villemoes and Wahlberg (2000), Kohn, Marron and Yau (2000). Dias (2000) and Dias and Garcia (2003) suggested to use a proxy of the Kullback-Leibler distance in order to select the number of basis functions. The goal of this work is to prove that this selection criterion provides a consistent estimator of the dimension of the approximant space.

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

$$D_n(K) = \int f \log \hat{f}_K. \tag{1.8}$$

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

$$Z_n(K) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_K(X_i), \tag{1.9}$$

it is easy to prove their equivalence.

Proposition 1.4 *For any fixed K,*

$$D_n(K) - Z_n(K) = \sum_{j=1}^{K} \hat{\theta}_{nj}^{(K)} \Big(\int f(x) M_j(x) d\mu(x) - \frac{1}{n}\sum_{i=1}^{n} M_j(X_i) \Big) \longrightarrow 0 \tag{1.10}$$

$n \longrightarrow \infty$ *almost surely.*

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

$$\hat{K} = \arg\min_{K} \mathrm{LP}(n, K), \tag{1.11}$$

where $\mathrm{LP}(n, K) = n\, Z_n(K) - c_n\, K$ and $c_n > 0$. This includes the most common information criteria for model selection such as AIC estimator ($c_n = 2$) and BIC estimator ($c_n = \log n$).

We have the following main results:

Theorem 1.1 *If* $\lim_{n\to\infty} c_n/n = 0$ *and* $\liminf_{n\to\infty} c_n/\sqrt{n} = 0$ *then* $\hat{K}$ *is a strongly consistent estimator for* K*.*

Theorem 1.2 *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 next section, 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 estimator).

2 Proof of the main results

Proof of Theorem 1.1. 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 1.3 we have that

$$Z_n(K) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_K(X_i) \to \mathbb{E}[\log f(X)] := m_K \text{ a.s.} \tag{2.1}$$

where X is a random variable with density f.

On the other hand,

$$\begin{aligned} Z_n(l) &= \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_l(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{j=1}^{l} \hat{\theta}_j^{(l)} M_j(X_i) - \log \int e^{\langle \hat{\theta}^{(l)}, M\rangle_l} \right] \qquad (2.2) \\ &\to \sum_{j=1}^{l} \theta_j^{(l)} \mathbb{E}[M_j(X)] - \log \int e^{\langle \theta^{(l)}, M\rangle_l} := m_l \text{ a.s.} \qquad (2.3) \end{aligned}$$

where $\hat{\theta}^{(l)}$ and $\theta^{(l)}$ are the solutions of (1.2) and (1.7) respectively, replacing K by l.

Therefore, by Jensen's inequality

$$\begin{aligned} m_l - m_K &= \int f \log \frac{f_l}{f} = \mathbb{E}\left[\log \frac{f_l}{f}\right] \\ &\leq \log \mathbb{E}\left[\frac{f_l}{f}\right] = \log \int f \frac{f_l}{f} = 0. \qquad (2.4) \end{aligned}$$

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 $l < K$

$$\lim_{n\to\infty} \frac{1}{n}\left(Z_n(K) - Z_n(l)\right) = m_K - m_l > 0 \text{ a.s..} \tag{2.5}$$

By (2.5) as $n \to \infty$

$$\begin{aligned} \frac{1}{n}(\mathrm{LP}(n,K) - \mathrm{LP}(n,l)) &= \frac{1}{n}(Z_n(K) - Z_n(l)) - \frac{c_n}{n}(K-l) \\ &= (m_K - m_l)(1 + o(1)) - \frac{c_n}{n}(K-l) \text{ a.s..} \qquad (2.6) \end{aligned}$$

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 (2003) 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 random variable. Therefore,

$$\lim_{n\to\infty} \frac{1}{n}(Z_n(K) - Z_n(l)) = O\left(\frac{1}{\sqrt{n}}\right) \text{ a.s..} \tag{2.7}$$

By (2.7) as $n \to \infty$

$$\mathrm{LP}(n,K) - \mathrm{LP}(n,l) \quad = \quad O\left(\sqrt{n}\right) \; + \; c_n(K-l) \text{ a.s..} \tag{2.8}$$

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 1.2 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 $n \to \infty$.

Acknowledgments This work was partially funded by FAPESP grant 02/01554-5 and CNPq grants 301054/1993-2, 300644/1994-9 and 475763/2003-3.

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).

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. (2003). A spline approach to nonparametric test of hypotheses, *Brazilian Journal of Probability and Statistics. To appear.*

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

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.

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).

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.