Statistics and Probability Letters 84 (2014) 223–230

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Fourier methods for smooth distribution function estimation

José E. Chacón [a,*], Pablo Monfort [a], Carlos Tenreiro [b]

[a] Departamento de Matemáticas, Universidad de Extremadura, E-06006 Badajoz, Spain

[b] CMUC, Department of Mathematics, University of Coimbra, Apartado 3008, 3001-454 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 11 May 2013
Received in revised form 15 October 2013
Accepted 16 October 2013
Available online 22 October 2013

Keywords:
Fourier analysis
Kernel distribution estimation
Mean integrated squared error
Optimal bandwidth
Sinc kernel

ABSTRACT

The limit behavior of the optimal bandwidth sequence for the kernel distribution function estimator is analyzed, in its greatest generality, by using Fourier transform methods. We show a class of distributions for which the kernel estimator achieves a first-order improvement in efficiency over the empirical estimator.

1. Introduction

The kernel estimator of a distribution function was introduced independently by Tiago de Oliveira (1963), Nadaraya (1964) and Watson and Leadbetter (1964) as a smooth alternative to the empirical estimator. It is defined as the distribution function corresponding to the well-known kernel density estimator. Precisely, given independent real random variables $X_1, \ldots, X_n$ with common and unknown distribution function F, assumed to be absolutely continuous with density function f, the kernel estimator of $F(x)$ is

$$F_{nh}(x) = n^{-1} \sum_{j=1}^{n} K\left(h^{-1}(x - X_j)\right),$$

where $h > 0$ is the bandwidth and the function K will be referred to as the integrated kernel, since it is assumed that $K(x) = \int_{-\infty}^{x} k(y)dy$ for some integrable function k, called kernel, having unit integral over the whole real line.

Classical references on kernel distribution function estimators include Yamato (1973), which provided mild necessary and sufficient conditions for its consistency in uniform norm, Azzalini (1981), Swanepoel (1988) and Jones (1990) on asymptotic squared error analysis of the estimator, or Sarda (1993); Altman and Léger (1995) and Bowman et al. (1998), and more recently Polansky and Baker (2000) and Tenreiro (2006), on data-driven bandwidth selection. There are also other recent papers on different aspects of kernel distribution function estimation, like Tenreiro (2003), Swanepoel and Van Graan (2005), Janssen et al. (2007), and Giné and Nickl (2009), Berg and Politis (2009); Chacón and Rodríguez-Casal (2010); Mason and Swanepoel (2012) or Tenreiro (2013). See Servien (2009) for a detailed survey on distribution function estimation, not limited to kernel-type methods.

* Corresponding author.
E-mail addresses: jechacon@unex.es (J.E. Chacón), pabmonf@unex.es (P. Monfort), tenreiro@mat.uc.pt (C. Tenreiro).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.10.010

This paper is devoted to the study of the kernel distribution function estimator from the point of view of the mean integrated squared error,

$$\mathrm{MISE}(h) \equiv \mathrm{MISE}_n(h) = \mathbb{E}\int_{-\infty}^{\infty}\{F_{nh}(x) - F(x)\}^2 dx.$$

In this sense, the optimal bandwidth h_{0n} is the value of $h > 0$ minimizing $\mathrm{MISE}(h)$. The existence of such a bandwidth was proved in Theorem 1 of Tenreiro (2006) under very general assumptions, and Proposition 2 in the same paper showed that $h_{0n} \to 0$ whenever the Fourier transform of k is not identically equal to 1 on any neighborhood of the origin. This condition can be considered mild as well, since it is satisfied for any finite-order kernel; however, it does not hold for a superkernel (see Chacón et al., 2007).

The purpose of this note is to show how to use Fourier transform techniques for the analysis of kernel distribution estimators. Particularly, expressing the MISE in terms of characteristic functions allows us to obtain a result on the limit behavior of the optimal bandwidth sequence in its most general form so that it also covers the case of a superkernel, and to explore its consequences showing the peculiar properties of the use of superkernels and the sinc kernel in kernel distribution function estimation. Precisely, it is shown in Section 2 that in some situations the sequence h_{0n} does not necessarily tend to zero. Moreover, we exhibit a class of distributions for which the kernel distribution estimator presents a first-order improvement over its empirical counterpart, opposite to the usual situation, where only second-order improvements are possible (see Remark 3). Our findings are illustrated in Section 3 through two representative examples.

2. Main results

Recall from Chacón and Rodríguez-Casal (2010) that the kernel distribution function estimator admits the representation

$$F_{nh}(x) = \int F_n(x - hz)dK(z), \tag{1}$$

where F_n denotes the empirical distribution function (here and below integrals without integration limits are meant over the whole real line). Using this, and standard properties of the empirical process, it is possible to obtain a decomposition of $\mathrm{MISE}(h) = \mathrm{IV}(h) + \mathrm{ISB}(h)$, where the integrated variance $\mathrm{IV}(h) = \int \mathrm{Var}\{F_{nh}(x)\}dx$ and the integrated squared bias $\mathrm{ISB}(h) = \int\{\mathbb{E}[F_{nh}(x)] - F(x)\}^2 dx$ can be expressed in the following exact form:

$$\mathrm{IV}(h) = n^{-1}\iiint \big\{F\big(x - h(y \vee z)\big) - F(x - hy)F(x - hz)\big\}dK(y)dK(z)dx, \tag{2}$$

$$\mathrm{ISB}(h) = \iiint \{F(x - hy) - F(x)\}\{F(x - hz) - F(x)\}dK(y)dK(z)dx, \tag{3}$$

with $y \vee z$ standing for $\max\{y, z\}$.

Note that the representation (1) and the exact expressions (2) and (3) also make sense for $h = 0$, implying that the kernel distribution estimator reduces to the empirical distribution function for $h = 0$, for which the well-known MISE formula reads $\mathrm{MISE}(0) = \mathrm{IV}(0) = n^{-1}\int F(1-F)$ whenever $\psi(F) = \int F(1-F)$ is finite. Moreover, it is not hard to check that $\int |x|dF(x) < \infty$ and $\int |y\,k(y)|dy < \infty$ ensure that $\mathrm{MISE}(h)$ is finite for all $h > 0$, so those two minimal conditions will be assumed henceforth. Note that the required condition that F have a finite mean is slightly stronger than $\psi(F) < \infty$ since $\psi(F) \leq 2\int |x|dF(x)$.

2.1. Limit behavior of the optimal bandwidth sequence

Denote by φ_g the Fourier transform of a function g, defined as $\varphi_g(t) = \int e^{itx}g(x)dx$. As in Chacón et al. (2007), the key to understand the limit behavior of the optimal bandwidth sequence is to use Fourier transforms to express the MISE criterion. Abdous (1993) provided a careful account of the necessary conditions under which the MISE can be expressed in terms of Fourier transforms. The proof of his Proposition 2 implicitly derives formulas for $\mathrm{ISB}(h)$ and $\mathrm{IV}(h)$ in terms of φ_k and φ_f for $h > 0$. We reproduce this result here for completeness, and show that it can be extended to cover the case $h = 0$ as well.

Theorem 1. *If $\int |x|dF(x) < \infty$ and $\int |y\,k(y)|dy < \infty$ then, for all $h \geq 0$, the IV and ISB functions can be written as*

$$\mathrm{IV}(h) = (2\pi)^{-1}n^{-1}\int t^{-2}|\varphi_k(th)|^2\{1 - |\varphi_f(t)|^2\}dt,$$

$$\mathrm{ISB}(h) = (2\pi)^{-1}\int t^{-2}|1 - \varphi_k(th)|^2|\varphi_f(t)|^2 dt.$$

Particularly, note that for $h = 0$ the previous result yields a Parseval-like formula for distribution functions,

$$\psi(F) = \int F(1 - F) = (2\pi)^{-1}\int t^{-2}\{1 - |\varphi_f(t)|^2\}dt, \tag{4}$$