Journal of Statistical Planning and Inference 146 (2014) 1–19

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Wavelet density estimation for weighted data

Luisa Cutillo[a,*], Italia De Feis[b], Christina Nikolaidou[c], Theofanis Sapatinas[c]

[a] Dipartimento di Statistica e Matematica per la Ricerca Economica, Università degli studi di Napoli "'Parthenope",
Via Generale Parisi n. 13, Naples, Italy
[b] Istituto per le Applicazioni del Calcolo "M. Picone" (CNR), Naples, Italy
[c] Department of Mathematics and Statistics, University of Cyprus, Cyprus

ARTICLE INFO

Article history:
Received 4 December 2012
Received in revised form
24 September 2013
Accepted 28 September 2013
Available online 9 October 2013

Keywords:
Density estimation
Fourier series
Kernel smoothing
Mean integrated squared error
Wavelets
Weighted distributions

ABSTRACT

We consider the estimation of a density function on the basis of a random sample from a weighted distribution. We propose linear and nonlinear wavelet density estimators, and provide their asymptotic formulae for mean integrated squared error. In particular, we derive an analogue of the asymptotic formula of the mean integrated square error in the context of kernel density estimators for weighted data, admitting an expansion with distinct squared bias and variance components. For nonlinear wavelet density estimators, unlike the analogous situation for kernel or linear wavelet density estimators, this asymptotic formula of the mean integrated square error is relatively unaffected by assumptions of continuity, and it is available for densities which are smooth only in a piecewise sense. We illustrate the behavior of the proposed linear and nonlinear wavelet density estimators in finite sample situations both in simulations and on a real-life dataset. Comparisons with a kernel density estimator are also given.

1. Introduction

Let X be a random variable with cumulative distribution function (c.d.f.) F and probability density function (p.d.f.) f with respect to Lebesgue measure on the line. Assuming to have available n independent direct realizations $X_1, X_2, \ldots, X_n$ of X, the optimal nonparametric estimate of f can be easily obtained (see, e.g., Silverman, 1986; Wand and Jones, 1995; Efromovich, 1999).

In practice, it may happen that drawing a direct sample from X is impossible and some kind of bias is introduced in the sampling scheme. So, we consider the problem of nonparametric estimation of f given a sample of n independent and identically distributed observations $X_1^w, X_2^w, \ldots, X_n^w$ from a weighted distribution with p.d.f. f^w given by

$$f^w(x) = \frac{w(x)f(x)}{\mu_w}, \tag{1}$$

where w is the so-called weighting (or biasing) function and $0 < \mu_w = \mathbb{E}(w(X)) < \infty$ (see, e.g., Patil et al., 1988). In what follows, it is always assumed that w is a given function satisfying $0 < c_1 \le w(x) \le c_2 < \infty$ for all x.

The concept of biased data is well known and its practical applications range from social sciences and biology to economics and quality control. These observations arise when a sampling procedure chooses an observation with probability that depends on the value of the observation. The traditional length-bias size or size-biased $(w(x) = x)$,

[*] Corresponding author. Tel.: +39 0 815474865.
E-mail address: luisa.cutillo@uniparthenope.it (L. Cutillo).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.09.015

for f supported on the positive half-line, occurs when the probability of an observation is proportional to the length of the observation itself, and arises naturally in industrial and work sampling, in sampling from stochastic processes such as queues, telephone networks and renewal processes (see, e.g., Cox, 1969; Vardi, 1982; Dewanji and Kalbfleisch, 1987; Kvam, 2008). Other examples include wild life population, line transect sampling and ecology (Eberhardt, 1978; Patil and Rao, 1978; Hanberry et al., 2012), cell cycle kinetics and early screening for disease (Zelen and Feinleib, 1969; Zelen, 1974, 2004), genome-wide linkage studies (Terwilliger et al., 1997), and epidemiologic cohort studies (Keiding, 1991; Gail and Benichou, 2000; Gordis, 2000; Sansgiry and Akman, 2000; Scheike and Keiding, 2006).

The fundamental result in the theory of weighted data is from Cox (1969), where the following estimator of F was suggested

$$\tilde{F}(x)^{-1}\hat{\mu}_w \sum_{i=1}^{n} w^{-1}(X_i^w)\mathbb{I}(X_i^w \le x), \tag{2}$$

where $\hat{\mu}_w = n\{\sum_{i=1}^{n} w^{-1}(X_i^w)\}^{-1}$ and $\mathbb{I}(A)$ denotes the indicator function of the set A. Hence, for weighted data, this estimator plays the same role as the empirical distribution function for direct data. Later, Vardi (1982) and Vardi (1985) showed that $\tilde{F}$ is the nonparametric maximum likelihood estimator of F for this situation, and that $\hat{\mu}_w$ is a $\sqrt{n}$-consistent estimator of μ_w.

Bhattacharyya et al. (1988) and Jones (1991) proposed kernel estimators of f for weighted data from model (1). The former estimator, although a little more ad-hoc, first estimates f^w by an ordinary kernel estimator on $X_1^w, X_2^w, \ldots, X_n^w$ and then uses the relationship between f and f^w to obtain an estimator of f. The latter estimator is more natural and is derived from kernel smoothing of the nonparametric maximum likelihood estimator $\tilde{F}$ of F given in (2). Jones (1991) showed that his estimator has various advantages over the estimator proposed by Bhattacharyya et al. (1988), including better asymptotic mean integrated squared error (MISE) properties. Multivariate extensions for both kernel density estimators were considered in Ahmad (1995). Whereas minimax results in the univariate case, for a Hölder class of smooth functions, were obtained in Wu (1995) and Wu and Mao (1996). A Fourier series estimator of f for weighted data from model (1) was proposed by Jones and Karunamuni (1997), while a transformation-based estimator was suggested by El Barmi and Simonof (2000). Sharp minimax results of a blockwise shrinkage estimator of f for weighted data from model (1), based on a projection estimator on trigonometric polynomial spaces and with a threshold, for a Sobolev class of smooth functions, were obtained in Efromovich (2004a), while sharp minimax results for an estimator of F, via a projection on trigonometric bases too, and of f by differentiation, for a class of analytic c.d.f.'s, were derived in Efromovich (2004b). A penalized projection estimator of f was built by Brunel et al. (2009) and the exact minimax rate of convergence under the L^2-risk over a particular Besov class was proved. This estimator generalizes the Efromovich–Pinsker adaptive estimator (Efromovich, 2004b) allowing generic orthonormal bases.

Recently, the density estimation problem for biased data using wavelets has also been addressed in several papers. Chesneau (2010) constructs an adaptive estimator of f based on an L^p-version of the BlockShrink algorithm initially developed by Cai (2002) for another statistical framework using dyadic wavelets. The author proves that the estimator attains near optimal rates of convergence. Ramirez and Vidackovic (2010) discuss the more general context of stratified size biased data proposing a nonlinear dyadic wavelet density estimator for such data. They prove consistency of their estimator in the MISE sense.

In this paper, we propose linear and nonlinear wavelet estimators of f for weighted data from model (1) using nondyadic wavelets as in Hall and Patil (1995a,b) and Hall and Penev (2001). We provide asymptotic formulae for MISE. We derive an analogue of the asymptotic formula of the MISE in the context of kernel density estimators for weighted data, admitting an expansion with distinct squared bias and variance components. For nonlinear wavelet density estimators, unlike the analogous situation for kernel or linear wavelet density estimators, this asymptotic formula of the MISE is relatively unaffected by assumptions of continuity, and it is available for densities which are smooth only in a piecewise sense. Moreover, it is shown that nonlinear wavelet density estimators possess a property which guarantees a high level of robustness against oversmoothing, not encountered in the context of kernel (see van Eeden, 1985; van Es, 2001) or linear wavelet density estimators for weighted data.

The paper is organized as follows. In Section 2, we briefly discuss basic elements on orthonormal wavelets, and describe the proposed linear and nonlinear wavelet density estimators for weighted data. In Section 3, we discuss the asymptotic MISE formulae for both wavelet density estimators in the context of smooth densities. For nonlinear wavelet density estimators it is shown that the MISE formula is unaffected by discontinuities. In Section 4, we illustrate the numerical performance of the proposed linear and nonlinear wavelet density estimators in finite sample situations by considering both a simulated and a real-life dataset reporting an example of length-bias sampling scheme. We also provide comparisons with the weighted kernel density estimator proposed by Jones (1991). In Section 5, we provide concluding remarks. Finally, in Section Appendix A (Appendix), we provide the proofs of the theoretical results stated in Section 3.

2. Wavelet density estimators for weighted data

The term wavelet is used to refer to a set of orthonormal basis functions generated by dilation and translation of a compactly supported *scaling function* (or *father wavelet*), Φ, and a *mother wavelet*, Ψ, associated with an r-regular ($r>0$) multiresolution analysis of $L^2(\mathbb{R})$, the space of squared integrable functions on the line (see, e.g., Mallat, 1999).