Journal of Statistical Planning and Inference 174 (2016) 104–128

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Nonparametric weighted estimators for biased data

Fabienne Comte[a,*], Tabea Rebafka[b]

[a] MAP5, UMR 8145 CNRS, Sorbonne Paris Cité, Paris Descartes University, France

[b] Sorbonne Universités, Université Pierre et Marie Curie, UMR 7599 CNRS, Laboratoire de Probabilités et Modèles aléatoires, France

ARTICLE INFO

Article history:
Received 11 January 2015
Received in revised form 22 January 2016
Accepted 26 January 2016
Available online 3 February 2016

Keywords:
Adaptive density estimation
Bandwidth selection
Biased data
Fluorescence lifetimes
Model selection
Weighted estimators

ABSTRACT

Starting from a real data example in fluorescence, the problem of nonparametric estimation of a density in a biased data model is considered. Bias correction can be done in two ways: either an estimator is computed with the data and in a second time a correction (plug in estimator) is applied, or weights are directly associated with the data so that a direct estimator of the quantity of interest (weighted estimator) is obtained. In both cases, kernel and projection estimation strategies with bandwidth or model selection devices are developed. The bandwidth selection is inspired from a procedure recently proposed by Goldenshluger and Lepski (2011). Risk bounds are proved showing that the final data-driven estimators perform an automatic finite sample bias–variance tradeoff. A simulation study compares the two bias-correction methods and the different model or bandwidth selection methods. Finally real fluorescence data are studied.

1. Introduction

In various application settings, functional estimation can be difficult because the observations are not a sample from the distribution of interest: this may be due to noise, missing data, censored or truncated observations. In this paper biased data models are considered where the cumulative distribution function (cdf) of the observations, denoted by G, is the result of a (known) nonlinear distortion of the distribution of interest, say F. More precisely, the cdf G and F are related by some known link function H by

$$G(x) = H(F(x)), \quad x \in \mathbb{R}. \tag{1}$$

This paper is concerned with the estimation of the probability density function (pdf) f associated with F from a sample $Z_1, \ldots, Z_n$ with distribution G.

A special case of model (1) is the pile-up model, where a random variable Z with distribution G is defined as the minimum of a random number N of independent and identically distributed (i.i.d.) random variables $Y_1, \ldots, Y_N$ with distribution F. This model is our leading example. It is encountered for example in biostatistics when considering the time until the outbreak of a tumor originated from a clonogenic cell in the presence of a random number of competing clonogens (Tsodikov, 2001). Another example in physics is given by the arrival time of the fastest of a random number of emitted photons (O'Connor and Phillips, 1984). The latter is the setting this paper started from and it will be described in more detail below.

Various extensions and other examples may be considered pointing out the relevance of the model given by (1) from an application viewpoint. For example, the maximum of a random number of i.i.d. random variables corresponds in actuarial science to modeling the largest claim received by an insurer in a given time interval (Li and Zuo, 2004), or in transportation

[*] Corresponding author.
E-mail addresses: fabienne.comte@parisdescartes.fr (F. Comte), tabea.rebafka@upmc.fr (T. Rebafka).

http://dx.doi.org/10.1016/j.jspi.2016.01.008
0378-3758/© 2016 Elsevier B.V. All rights reserved.

theory to the modeling of the maximal accident-free distance of a shipment of, say, explosives, with a random number of defective explosives which may explode and cause an accident during transport (Shaked and Wong, 1997).

It is worth mentioning that our model can be related to other biased data contexts, which have been studied from various points of view by several authors: strategies for estimating cumulative distribution functions are proposed by Gill et al. (1988), Wu and Mao (1996), Wu (1997), Efromovich (2004b) and El Barmi and Simonoff (2000); the specific case of length-biased sampling has been studied in many papers, see Vardi (1982), Jones (1991), de Uña-Álvarez (2004), de Uña-Álvarez and Rodríguez-Casal (2006) and Asgharian et al. (2002), among others.

The interest and difficulty of the present work lies in the fact that we have three aims.

(1) The primary concern of the paper is the nonparametric estimation of the pdf f of the distribution of interest F in the model given by (1) based on an i.i.d. sample $Z_1, \ldots, Z_n$ with distribution G and known link function H.

(2) Our second question is about a methodological point of view. We want to determine which general approach of density estimation should be used. Indeed, we consider hereafter kernel estimators and projection estimators and wonder which are to be preferred. More precisely, projection estimators with model selection devices and kernel estimators with data-driven bandwidth selection are constructed for the model given by (1). Adaptive projection estimators correspond to methods originally described by Barron et al. (1999), see also Lerasle (2012) for developments more specific to density estimation. These methods have been applied to survival analysis and biased data by Efromovich (2004a,b) and Brunel et al. (2005); more recently, wavelet projection estimators have been studied by Chesneau (2010), Cutillo et al. (2014). For the bandwidth selection of the kernel estimator the recent approach of Goldenshluger and Lepski (2011) is applied to our model and considered from both a pointwise and a global point of view. Here "pointwise" refers to the estimation of the density on an interval with point-by-point bandwidth selection, in contrast to a unique global bandwidth in the "global" strategy. One may expect better results for the pointwise method when the function under consideration has inhomogeneous smoothness on the interval.

(3) Thirdly, the more model specific question is how to take into account the distortion H in the estimation procedure. Indeed, different properties of the model (1) give rise to two different strategies to correct the bias in the data. The first way is a sort of global correction of a primary estimator of g, which we call plug-in estimator, as in Navarro et al. (2015). The other way consists in using a standard density estimator of f while associating specific weights with all observations to correct the bias more locally, and we call it weighted estimator. This has been done in a different context in Rebafka et al. (2010). The same type of question arises in density estimation for censored data: the so-called Inverse Probability Correction Weights (IPCW) can be applied to the data, or a final correction can be applied to a functional estimator, see Brunel et al. (2005).

The combination of every adaptive kernel and projection estimator with each bias correction strategy finally gives rise to six different estimation procedures that are worth being compared. In this paper theoretical results on the mean-square risk of the estimators, more precisely, oracle-type risk bounds are provided. Namely the finite sample risk bounds for the adaptive kernel estimators are new. For adaptive projection estimators, part of the proofs follow the line of Brunel et al. (2005) which makes the novelty of the results less decisive. Furthermore, a simulation study is conducted to calibrate all methods and to find out how the different estimation procedures compare in specific settings. To avoid a huge number of models, simulations are carried out for the pile-up model and an application to real fluorescence data is provided. The questions in order are: Which strategy to correct the bias has a better performance? Do projection or kernel estimators provide better results? How do the pointwise and the global strategy for bandwidth selection compare in specific examples? We are not aware of any other empirical study answering to these questions.

The paper is organized as follows. Section 2 presents the leading example and the general model. In Section 3, kernel and projection estimators are defined, and risk bounds are given in order to show why a data-driven selection of bandwidth or model is required. Section 4 explains how these procedures are performed and provides theoretical results ensuring that these strategies reach their aim and deliver an adequate data-driven squared bias–variance compromise. In the simulation study (Section 5) different aspects of the estimators are compared. Section 6 summarizes our findings. Finally, Appendix A presents the proofs for the theoretical results of the paper.

2. Model and assumptions

2.1. Notations

Let $u : \mathbb{R} \mapsto \mathbb{R}$ and $v : \mathbb{R} \mapsto \mathbb{R}$ be two real functions. If u is a one-to-one map, denote u^{-1} its inverse function, that is the function verifying $u^{-1}(u(x)) = u(u^{-1}(x)) = x$ for all x. The standard convolution product is given by $u * v(x) = \int u(t)v(x-t)\mathrm{d}t$. Denote by $\| \ . \ \|_p$ the $\mathbb{L}^p$-norm given by $\|u\|_p^p = \int |u(x)|^p \mathrm{d}x$ and by $\| \ . \ \|_\infty$ the $\mathbb{L}^\infty$-norm, $\|u\|_\infty = \sup_{x\in\mathbb{R}} |u(x)|$. The inner product $\langle \cdot, \cdot \rangle$ is defined by $\langle u, v \rangle = \int u(t)v(t)\mathrm{d}t$.

2.2. Our leading example: the pile-up model in time-resolved fluorescence

Fluorescence is the phenomenon of photon emission by excited molecules. An important feature is the duration that the molecule spends in the excited state before emitting a photon, also called fluorescence lifetime. As the probability